PHW 250B Week 15 Reader

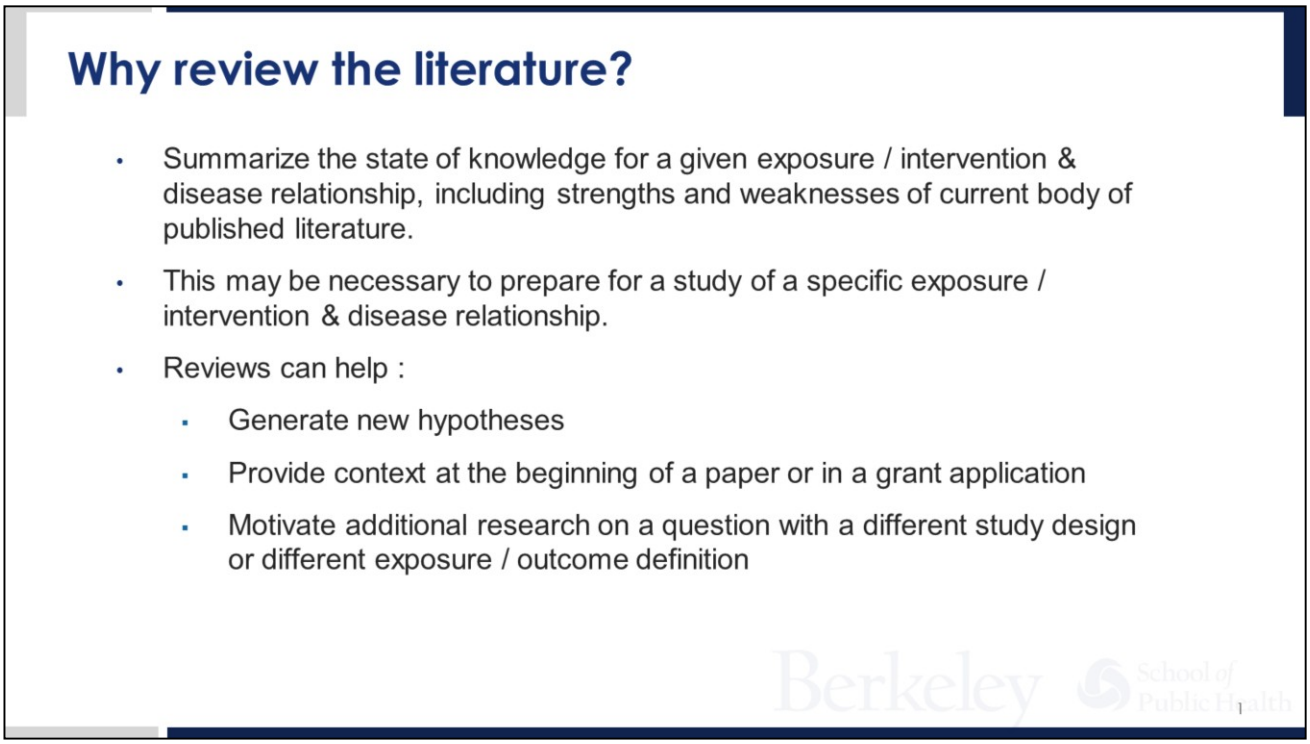
Topic 1: Systematic Reviews & Meta-Analyses

Topic 2: Population Intervention Effects

Topic 3: Spillover Effects

Topic 4: Transparency and Reproducibility

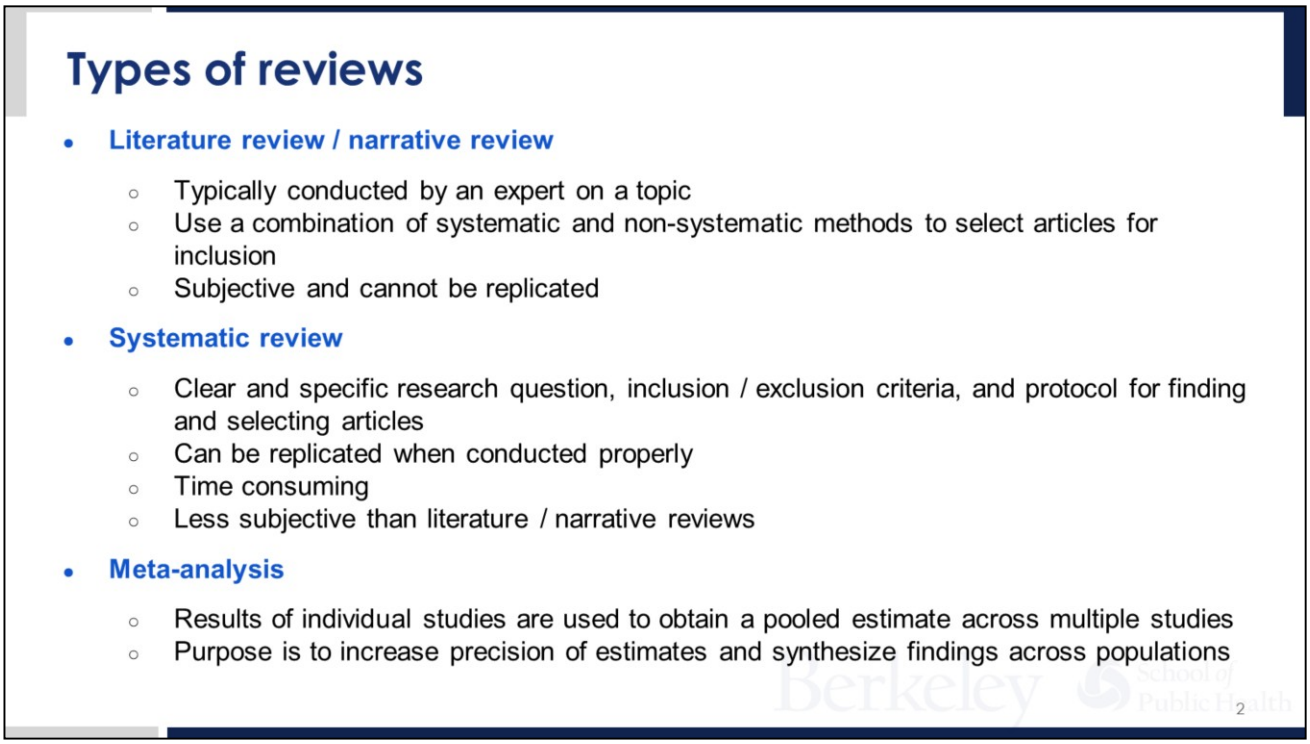
PRESENTER: Let's turn next to a popular and important tool in epidemiology called a systematic review.



It's probably obvious to you why we want to review the epidemiology public health and medical literature. Frequently, we want to summarize the state of knowledge for some specific exposure and intervention relationship, or a relationship between exposure and disease. And we want to look at the strengths and weaknesses of a current body of published literature.

You might be a policymaker wanting to understand how to make policy based on evidence. You might be a health planner wanting to understand what the prior work suggests you should do in your county, state, or local health department. Or you might be preparing to conduct a new study looking at the risks from some specific exposure or intervention, and their relationship to disease.

Systematic reviews and reviews in general, because there are some reviews that are non systematic, can all generate new hypotheses, things that you might wish to study, relationships you might wish to study. They can provide important context at the beginning of a paper or in the grant application process. A systematic review can be presented as the justification for the need to do the work that's suggested in the grant. Reviews might also motivate additional research on a question with a different study design than has been used in the past or with a different exposure and outcome definition.



It's helpful to think about the broad area of reviews in general, breaking things down into some different categories. And three categories we like to use are literature reviews. These are sometimes called narrative reviews, systematic reviews. And then a separate category that we'll talk about in a subsequent module is meta analysis, which is often done to supplement a systematic review. So in a literature review or narrative review, this is generally written after research by an expert who knows a topic really well and knows the literature.

It uses a combination of systematic and non-systematic methods for finding the articles that are included and very heavily relies on the experts knowledge of the field. But unfortunately, these are generally subjective and they can't be replicated by other people. In contrast, a systematic review is set up, if done properly, with a clear and very specific research question that includes criteria for inclusion and exclusion of studies and a protocol for finding and selecting articles.

The idea behind this is that the systematic review could be replicated if it's done properly and reported properly by some other investigator. Systematic reviews are time consuming. They very much take up a large portion of work, energy, and effort on the part of usually a team that's doing them. And their real strength is that they're less subjective than literature or narrative reviews.

Types of reviews

- **Literature review / narrative review**
  - Typically conducted by an expert on a topic
  - Use a combination of systematic and non-systematic methods to select articles for inclusion
  - Subjective and cannot be replicated
- **Systematic review**
  - Clear and specific research question, inclusion / exclusion criteria, and protocol for finding and selecting articles
  - Can be replicated when conducted properly
  - Time consuming
  - Less subjective than literature / narrative reviews
- **Meta-analysis**
  - Results of individual studies are used to obtain a pooled estimate across multiple studies
  - Purpose is to increase precision of estimates and synthesize findings across populations

Now, often after a systematic review is conducted, a meta analysis is then performed. And in a meta analysis, the results of individual studies are used to obtain a pooled or averaged estimate across multiple studies. The purpose of a meta analysis is to increase the precision of estimates and to synthesize findings across different populations. So the meta analysis is what tries to provide us with a summary answer of a specific question after reviewing all the literature. And you have to judge the meta analysis quality and how well it was done and the statistical methods as we'll learn about in a little bit to know whether or not you can put much faith in the estimate that it provides.

Example narrative review

OPEN ACCESS Freely available online

PLoS MEDICINE

Policy Forum

Hygiene, Sanitation, and Water: Forgotten Foundations of Health

Jamie Bartram[1], Sandy Cairncross[2]*

1 Water Institute, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, North Carolina, United States of America, 2 London School of Hygiene & Tropical Medicine, London, United Kingom

This is the introductory article in a four-part *PLoS Medicine* series on water and sanitation.

Globally, around 2.4 million deaths (4.2% of all deaths) [1] could be prevented annually if everyone practised appropriate hygiene and had good, reliable sanitation and drinking water. These deaths are mostly of children in developing countries from diarrhoea and subsequent malnutrition, and from other diseases attributable to malnutrition.

How is an opportunity to prevent so many deaths (and 6.6% of the global burden of disease in terms of disability-burden. Even using the most conservative scenarios, the long-term sequelae due to diarrhoea in early childhood contribute more DALYs than do the deaths [3].

Regrettably, it is no surprise that much ill health is attributable to a lack of HSW. Globally, nearly one in five people (1.1 billion individuals) habitually defecates in the open. Conversely, 61% of the world's population (4.1 billion people) has some form of improved sanitation at home—a basic hygienic latrine or a flush toilet. Between these two extremes, many households rely on dirty, unsafe latrines or shared toilet facilities [4]. Not only can it

quality standards [5]. Reliable safe water at home prevents not only diarrhoea but guinea worm, waterborne arsenicosis, and waterborne outbreaks of diseases such as typhoid, cholera, and cryptosporidiosis.

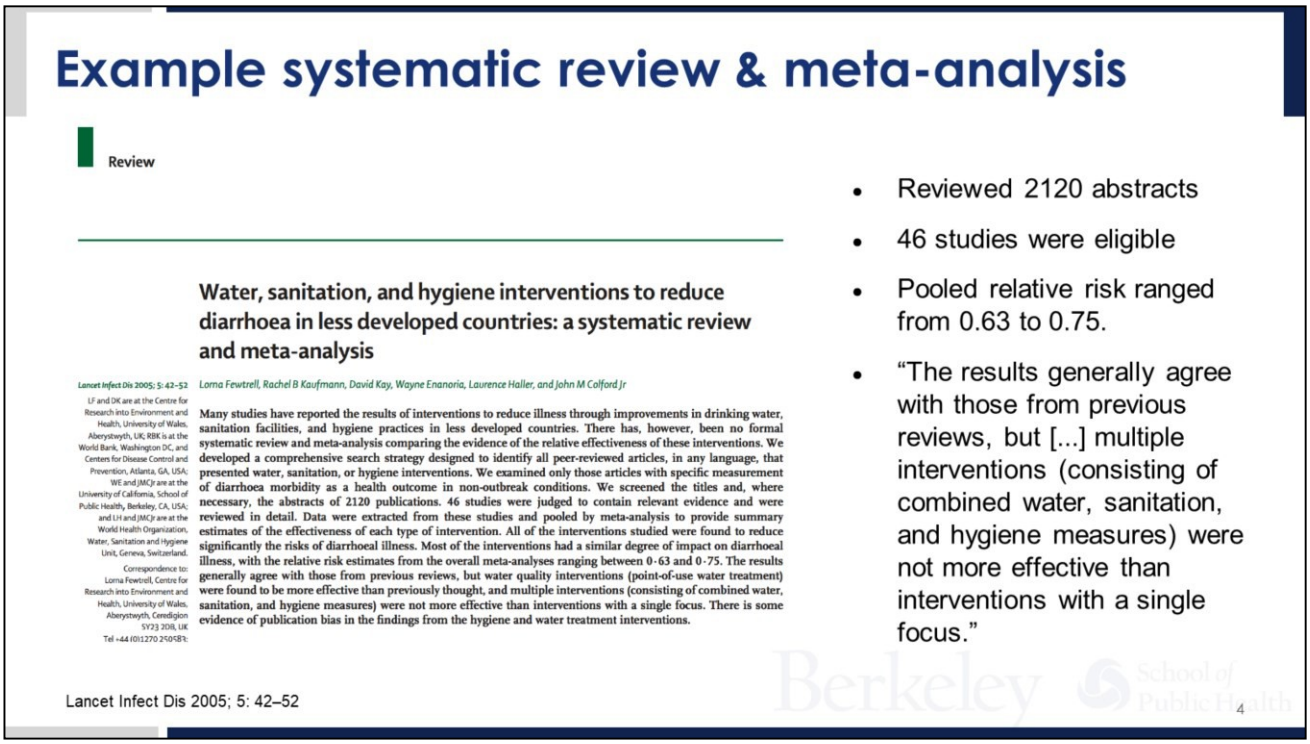
Much of the impact of water supply on health is mediated through increased use of water in hygiene. For example, hand washing with soap reduces the risk of endemic diarrhoea, and of respiratory and skin infections, while face washing prevents trachoma and other eye infections. A recent systematic review of the literature [6] confirmed that hygiene, particularly hand washing at delivery and postpartum,

Bartram J, Cairncross S (2010) Hygiene, Sanitation, and Water: Forgotten Foundations of Health. PLoS Med 7(11): e1000367. doi:10.1371/journal.pmed.1000367

- Example of a very general review on a topic rather than a narrow research question.
- "A massive disease burden is associated with deficient hygiene, sanitation, and water supply and is largely preventable with proven, cost-effective interventions."

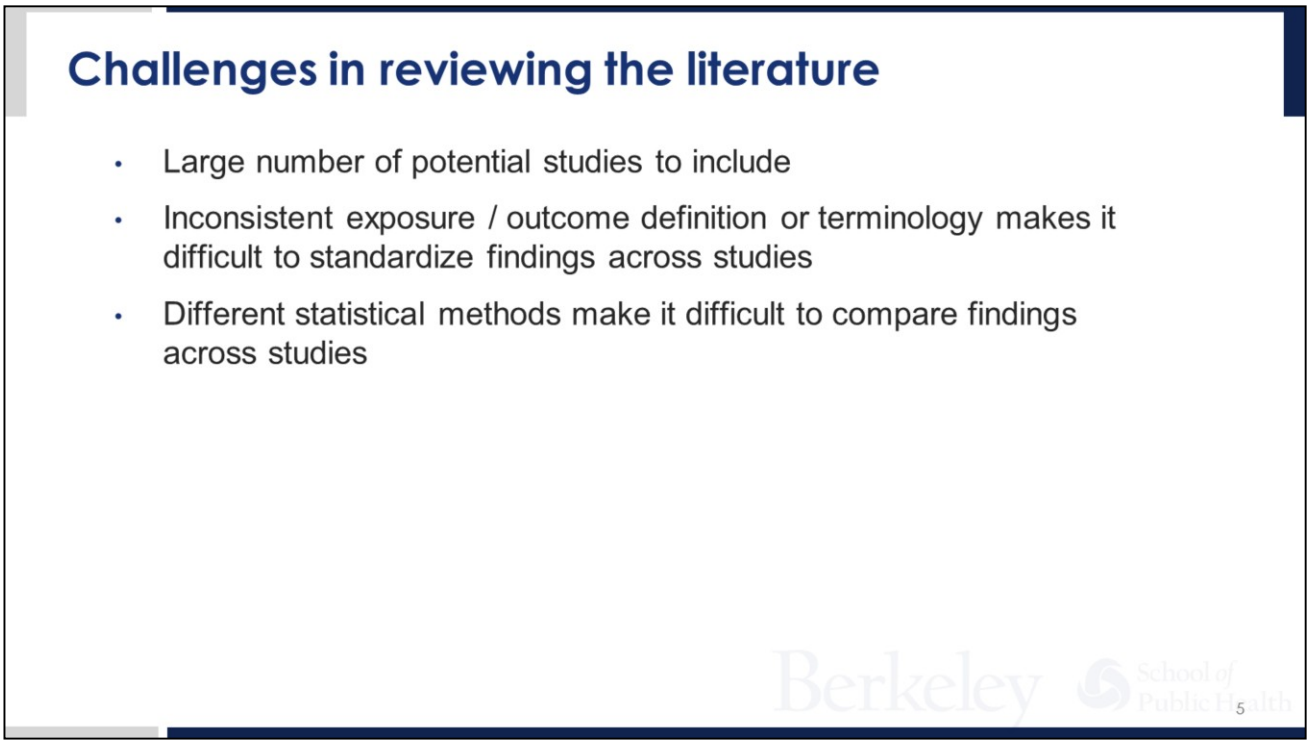
Let's discuss some specific examples of different types of reviews. And these are reviews that were done in advance or in the midst of the water, sanitation, and hygiene issues that led to WASH benefits that we've discussed previously in the course. So these reviews refer to a body of literature that supported the need to do the WASH benefits trials.

So this example narrative review by Jamie Bartram and Sandy Cairncross called Hygiene, Sanitation, and Water, the Forgotten Foundations of Health, rely on these two world renowned experts to talk about the WASH field. But it's an example of a very general review on a topic rather than a narrow research question. This review doesn't set out to answer a research question, but rather to cover the field kind of at a 30,000 foot level. They write in the review, a massive disease burden is associated with deficient hygiene, sanitation, and water supply, and is largely preventable with proven cost effective interventions. So a statement like that is then used by subsequent investigators to justify more specific work to answer a specific question, such as what was done in the WASH benefits trials.



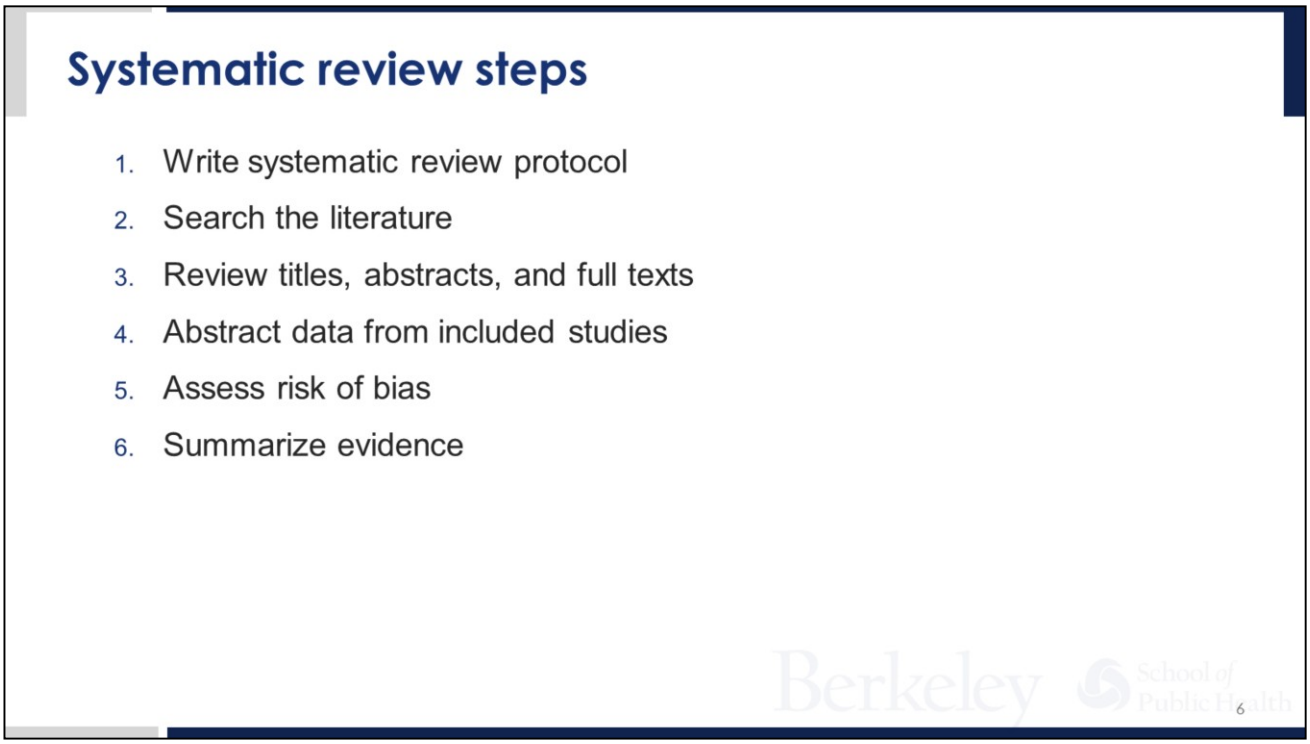
This next article that I was involved with is a formal systematic review and meta-analysis. And in this review, we reviewed 2,120 articles by looking at the titles and abstracts to see if they related to the question we were trying to get at, which was whether these interventions could reduce diarrhea in less developed countries. And after that process was completed, we felt that 46 studies were eligible for inclusion. Then we analyzed the results from all the studies in a way that allowed us to average them and we saw that the pooled relative risk ranged from 0.63 to 0.75.

We wrote that the results generally agree with those from previous reviews, but multiple interventions consisting of combined water, sanitation, and hygiene measures were not more effective than interventions with a single focus. And this in particular was starting to set the stage for WASH benefits, because as you recall, WASH benefits tested the idea of combining these interventions in one program.



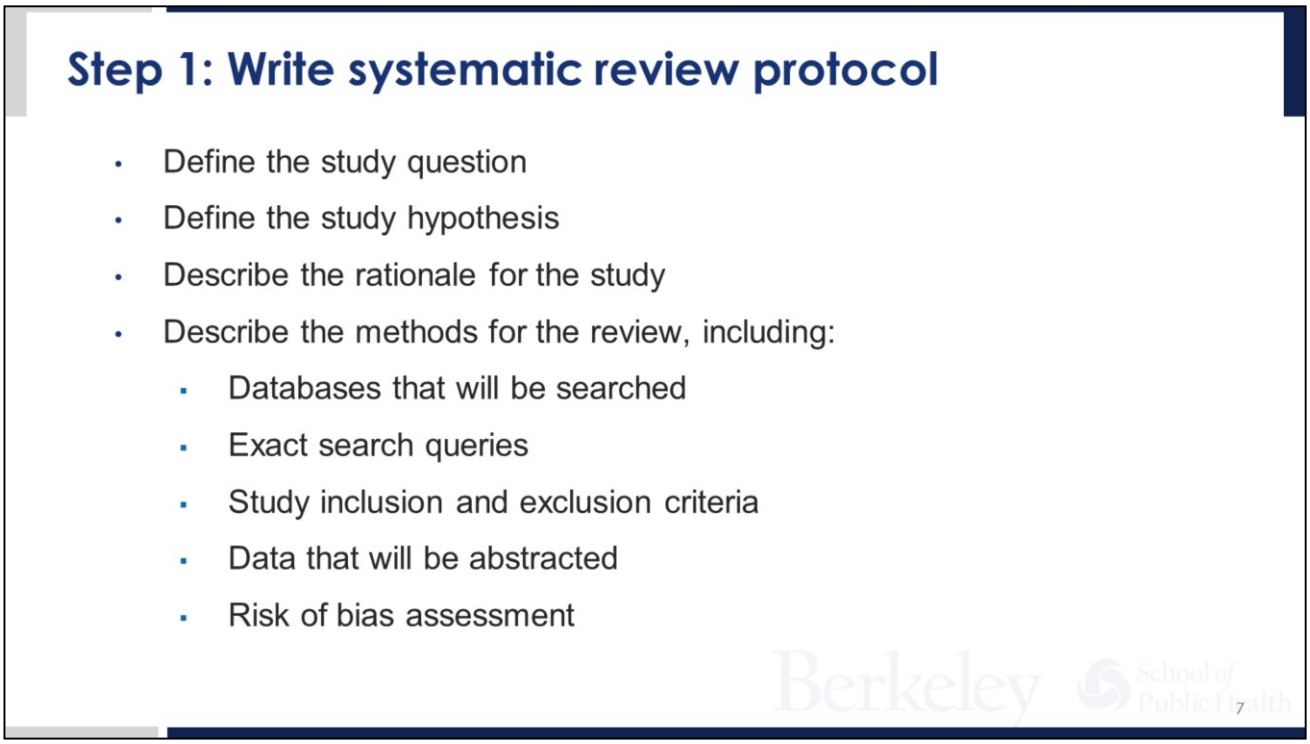
There are many challenges in reviewing the literature when conducting a systematic review. There can be a large and therefore time consuming number of potential studies that might need to be reviewed and included.

Unfortunately, there are very often inconsistent exposure and outcome definitions or terminology that can make it difficult to standardize findings across studies. Different authors might measure different things in different ways. You know for example, when studying diarrhea, different scales are used, different frequency measures are used, and so forth. So it can be hard to compare apples to apples and oranges to oranges when doing a systematic review. And then because different statistical methods are used in the individual studies, it can be difficult to compare the findings across different studies.

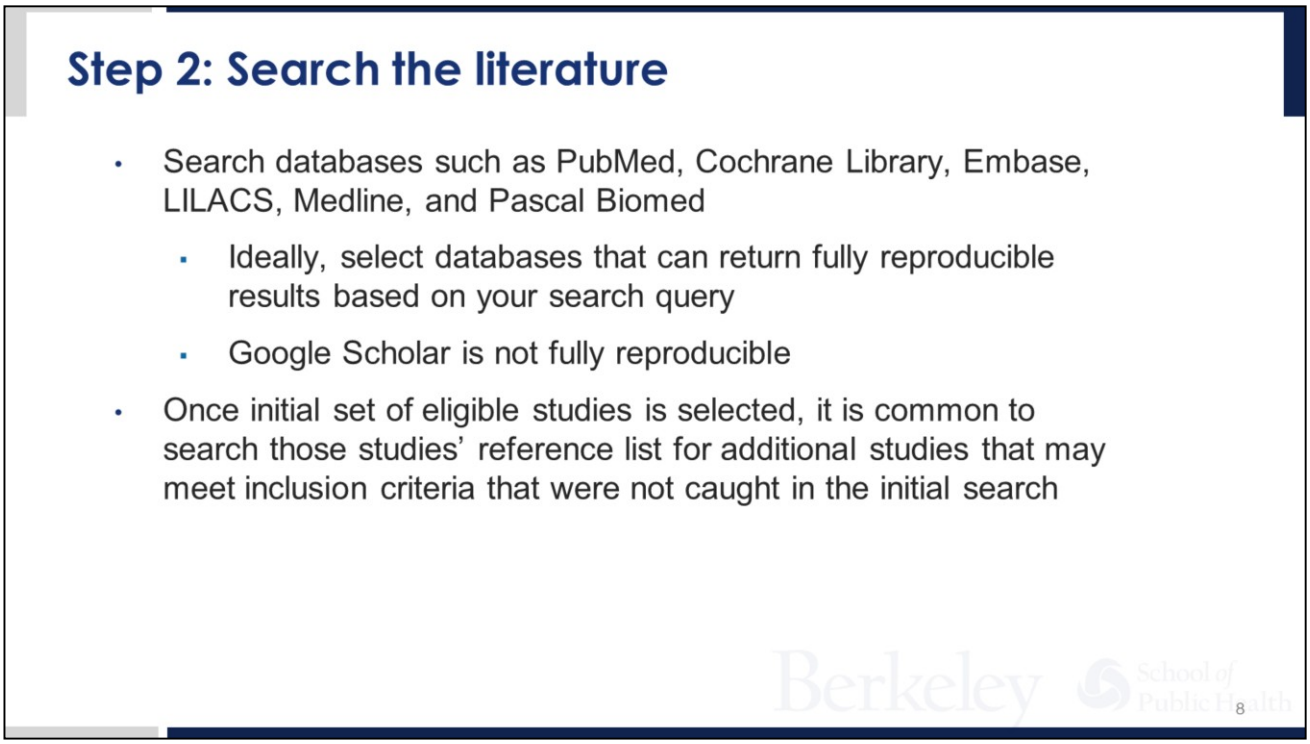


So to take a high level view of the steps in a systematic review, like any research project, it should start off with a review protocol. And this helps distinguish it from a narrative review. A narrative review generally never has a systematic review protocol prepared for it. Then the literature is searched, the authors conducting the systematic review will review the titles and the abstracts. And then from that group, go through and review the full text of the articles that seem most relevant.

The conductors of the systematic review will then abstract data from all the studies that they've decided meet the inclusion criteria and aren't excluded by other criteria. They'll next assess whether there was bias in each study that might have led to erroneous conclusions. And then finally, they'll summarize the evidence in the systematic review. Now, notice that what is not here is there's no summary estimate of an effect. That's a specific step that's coming later in a meta-analysis.



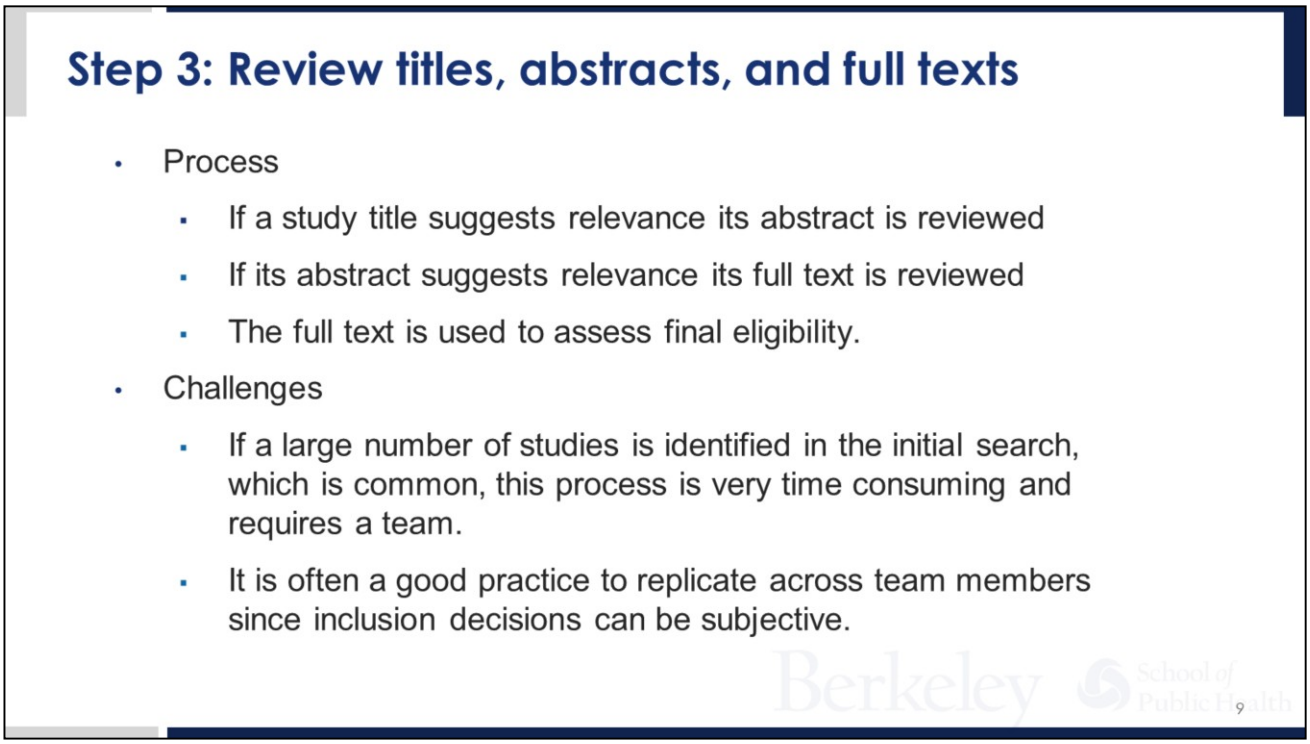
So the first step in a systematic review is to write the protocol. And a good protocol will very clearly define the study question and study hypothesis and describe the rationale for the study, such as in the WASH benefit study, what were the prior studies that led us to think that WASH benefits should be conducted? Then the methods for the review will be presented that will include what databases will be searched? What were the exact search terms and queries that were used? What were the study inclusion and exclusion criteria? What data will be abstracted from the individual articles? And how will the risk of bias in each of the individual studies be assessed? What tools will be used to assess bias?



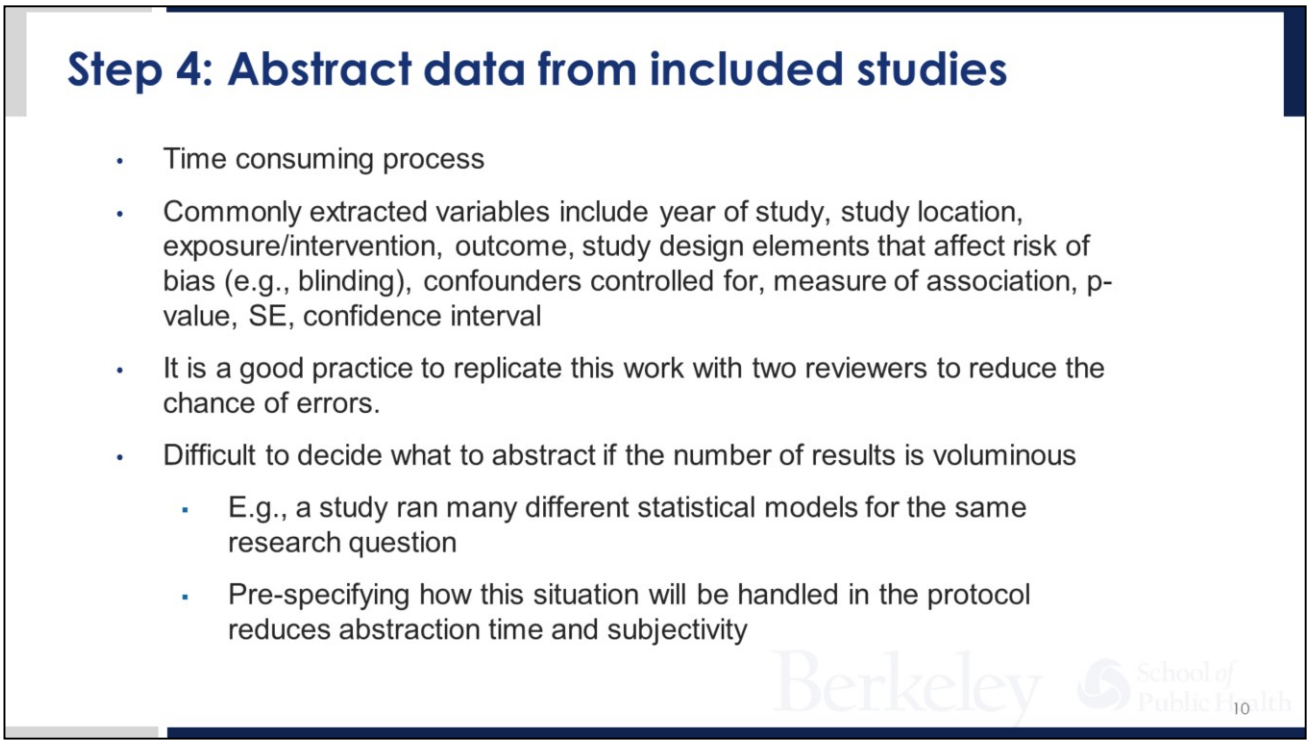
The next step is to search the literature after the protocol is written. And there is a wealth of databases that should be searched in most systematic reviews depending on the specific topic. But these might include the PubMed database, the Cochrane Library of prior systematic reviews, MBase, which is the European database for public health and medical articles, LILACS, which includes Latin American articles, Medline, Pascal Biomed, and there are many others. Ideally, one would select databases that can return fully reproducible results based on your search query.

In other words, if you hand your search query to somebody else, they repeat it, they would get the same results. Although if they do it out at a later date, they might retrieve more articles if additional articles have been published. Note that Google Scholar is not considered to be fully reproducible. Although it can be a useful tool, very often after a little bit of time elapses, its results can change because it's updated so often.

So once an initial set of eligible studies is selected, it's common to search those study reference lists themselves for additional studies that may meet inclusion criteria that were not caught in the initial search. This is called snowball searching. So you found an article that seems relevant, you then go to its reference list and search for additional articles in that reference list.



The third step is to review titles, abstracts, and the full text of the articles that have been reviewed. So if a study title suggests that the articles relevant, it's abstract is reviewed. If the abstract suggests it's relevant, it's full text is reviewed. And then the full text is used in assessing whether at the end the studies should be included or not in the systematic review. This can be a very challenging procedure, because if there's a really large number of studies identified in the initial search, which is commonly the case, the process is very time consuming and generally requires a team. And it's often good practice to replicate across team members since inclusion decisions can be subjective. So in other words, both team members are more than two team members, each take the selection criteria for the articles and see if they retrieve the same articles, and review which ones they differ on.



The fourth step is, again, time consuming. And that's too abstract data from the studies that have been included. So commonly extracted variables that usually are taken out of all the studies that are included would include the year of the study, the author of the study, location, the exposure, intervention, and outcome that were measured. What were the study design elements that might affect the risk of bias?

For example, was blinding used? Was randomization used? How were confounders controlled for? What measures of association were used? Were p values used? Were standard errors reported? Were confidence intervals reported?

And for all of these elements, they would be abstracted from each of the articles. And then once again, it's also good practice here to replicate this work with two or more reviewers to reduce the chance of errors. It can be difficult to decide what to abstract if the number of results is really voluminous.

For example, if a study ran many different statistical models for the same research question, one should have a procedure for deciding which model to abstract and report. And ideally, a team would pre-specify how this would be handled in such a situation in the protocol for the systematic review, because this would reduce abstraction time and subjectivity. For example let me give you a concrete situation, say that an article reports different models with different numbers of confounders or covariance, which is very common. Often what systematic reviewers will do will be to decide that we're going to report or collect either the most highly adjusted model that is the most covariance or the least adjusted model, the fewest covariance. So these sorts of decisions are best pre-specified in the protocol for the systematic review.

Example of extracted data

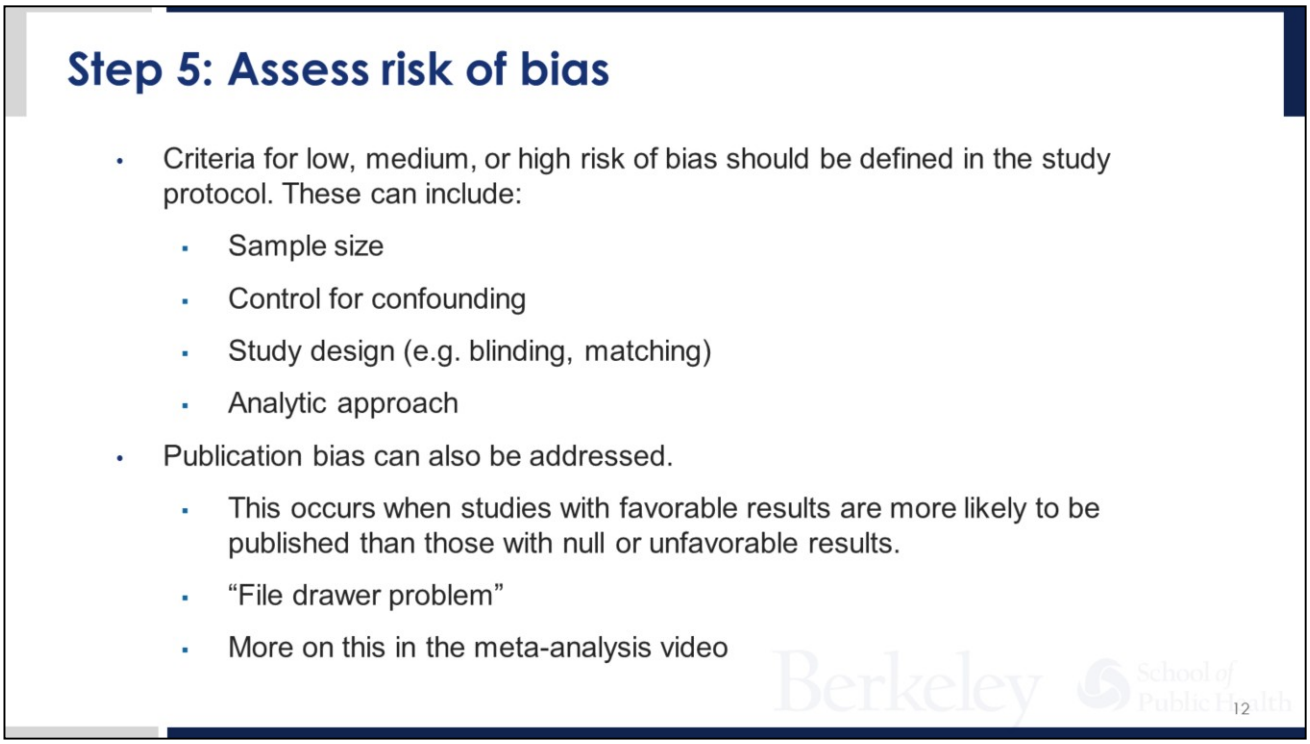
Reference	Intervention	Country (location)	Study quality*	Health outcome	Age group	Measure	Estimate (95% CI)
Khan,[11] 1982	Handwashing with soap	Bangladesh (unstated)	Good	Diarrhoea	All	RR†	0·62 (0·35–1·12)‡
Torún,[12] 1982	Hygiene education	Guatemala (rural)	Poor	Diarrhoea	0–72 months	RR†	0·81 (0·75–0·87)‡
Sircar et al,[13] 1987	Handwashing with soap	India (urban)	Good	Watery diarrhoea	0–60 months	RR†	1·13 (0·79–1·62)
					>5 years	RR†	1·08 (0·86–1·37)
				Dysentery	0–60 months	RR†	0·67 (0·42–1·09)
					>5 years	RR†	0·59 (0·37–0·93)
				Combined outcome	Combined ages	RR†	0·97 (0·82–1·16)‡
Stanton et al,[14] 1988; Stanton and Clemens,[15] 1987	Hygiene education	Bangladesh (urban)	Good	Diarrhoea	0–72 months	IDR	0·78 (0·74–0·83)‡
Alam et al,[16] 1989	Hygiene education (and increased water supply)	Bangladesh (rural)	Good	Diarrhoea	6–23 months	OR	0·27 (0·11–0·66)‡

Lancet Infect Dis 2005; 5: 42–52

11

Here's an example of what extracted data would look like. So in this neutral article we've been talking about earlier looking at WASH interventions, we see in each row individual articles that were published. And so this table lays out clearly for the reader what interventions were attempted in each study, where it was done, what the quality of the study was. And there's a separate note in the paper that describes how quality is defined for each study. What the health outcome was that was measured. And note that for the Sircar article, because three different outcomes were measured, those are each reported separately.

What ages were studied? How was the risk estimated? Was it with a relative risk measure or an incidence density ratio measure or an odds ratio? And then what was that estimate of effect?

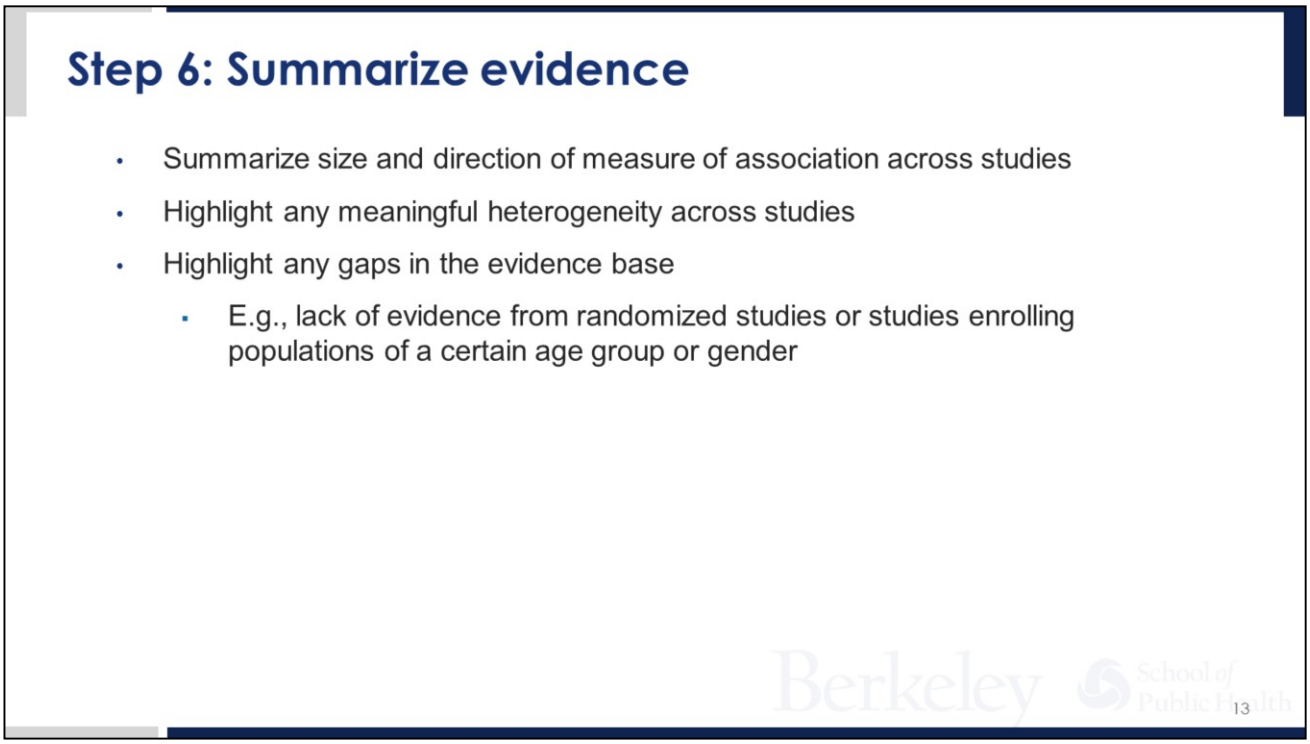


In the next step, the risk for bias is assessed in the study. So prior to conducting the study, the systematic review authors should indicate what their criteria are for low, medium, or high risk of bias. And this would be defined clearly in their study protocol. And these could include what was the sample size.

Smaller sample sizes, of course, are more at risk for bias. How was confounding controlled? Was it controlled? What study design was used? Was there blinding or matching involved? Was there randomization involved? And then how were the analysis done?

So all of these can be part of the decision about whether a study is going to be graded as low, medium, or high risk of bias. And that's a more advanced topic than we can cover right here, but there's approaches to this. And then also the important issue of publication bias can be addressed.

So this occurs when studies with favorable results are more likely to be published than those with null or unfavorable results. This is sometimes referred to as the "file drawer problem" because investigators who have negative results, even though they did a well conducted study, might file away their article in a file drawer and never publish it. So that's called the file drawer problem. We'll talk about that a bit more is the next video on meta-analysis.



And in the next step, the authors will summarize the size and direction of the measure of association across the different studies and highlight any meaningful heterogeneity across the studies. And in these first two steps about summarizing the size and direction and highlighting the heterogeneity, these are much more rigorously done in a meta analysis that's built on top of a systematic review. So this summary of the size and direction of association might just be reporting it and not averaging it. And then the heterogeneity might just be pointed out without addressing it or adjusting for it. But we'll talk later in the meta-analysis video about those concepts.

And then if there are gaps in the evidence base, those should be highlighted by the authors. For example, if there are no randomized studies contributing to the body of literature, or if the studies enrolled only populations of a certain age group or gender so that the results aren't generalizable to other age groups or genders, that's important to note.

Reporting systematic reviews

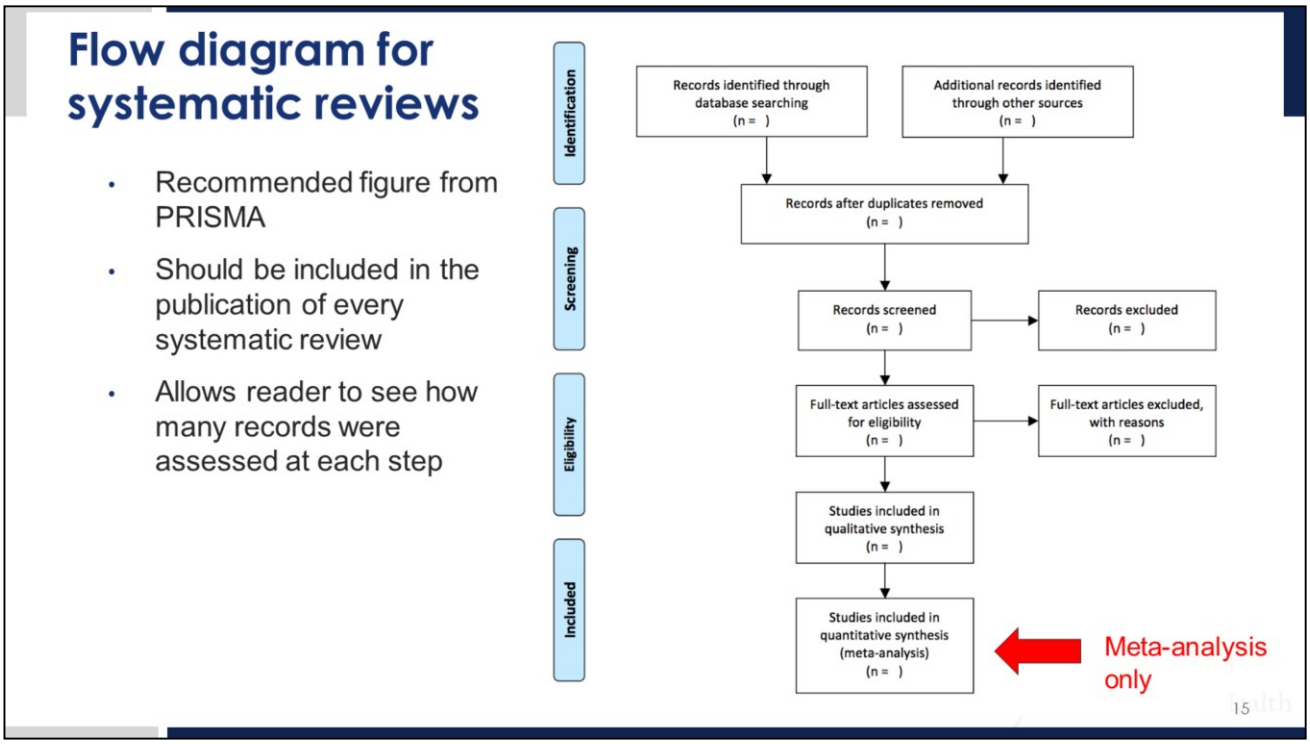
- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses http://www.prisma-statement.org/
- PRISMA checklist for reporting includes specific guidelines on what to report in a publication of a systematic review

PRISMA 2009 Checklist

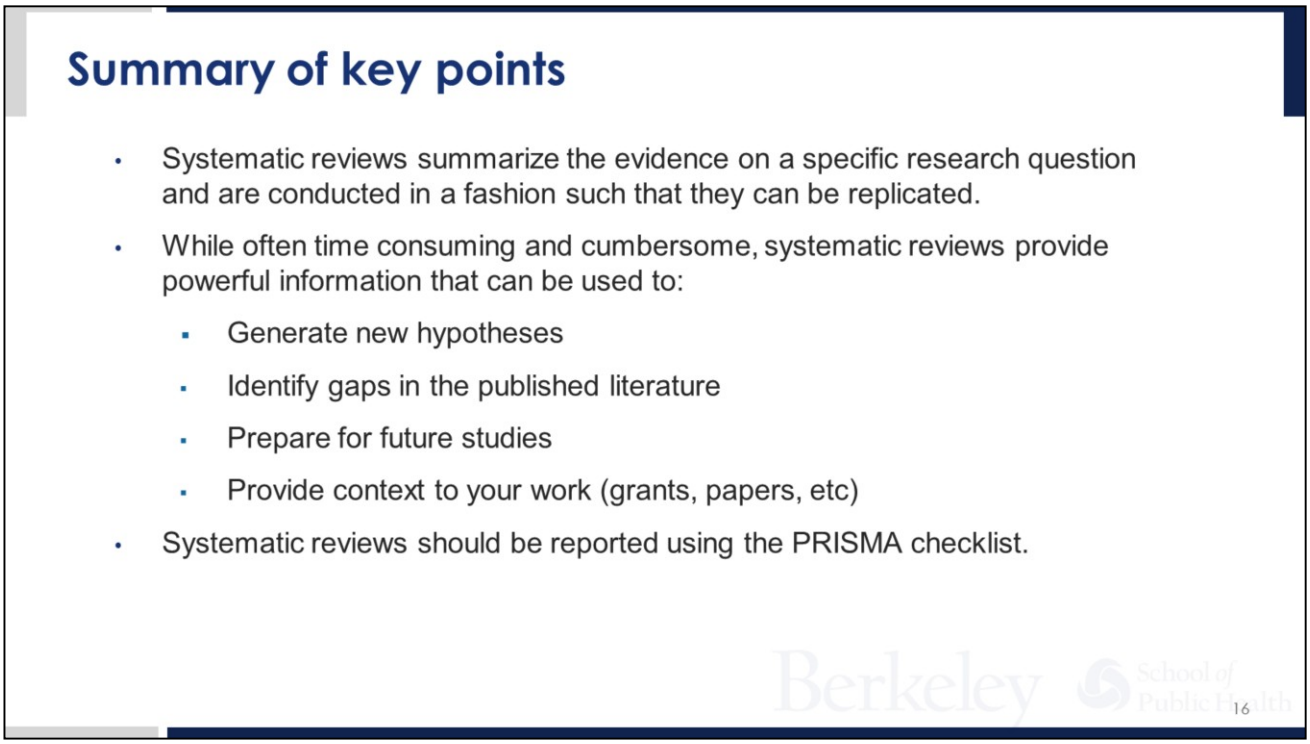
Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			

14

There is a very widely used approach to reporting systematic reviews called the PRISMA checklist or guidelines. PRISMA stands for the preferred reporting items for systematic reviews and meta-analysis. And this includes very specific guidelines on what to report in a publication of a systematic review. And you see a small portion of the checklist here.

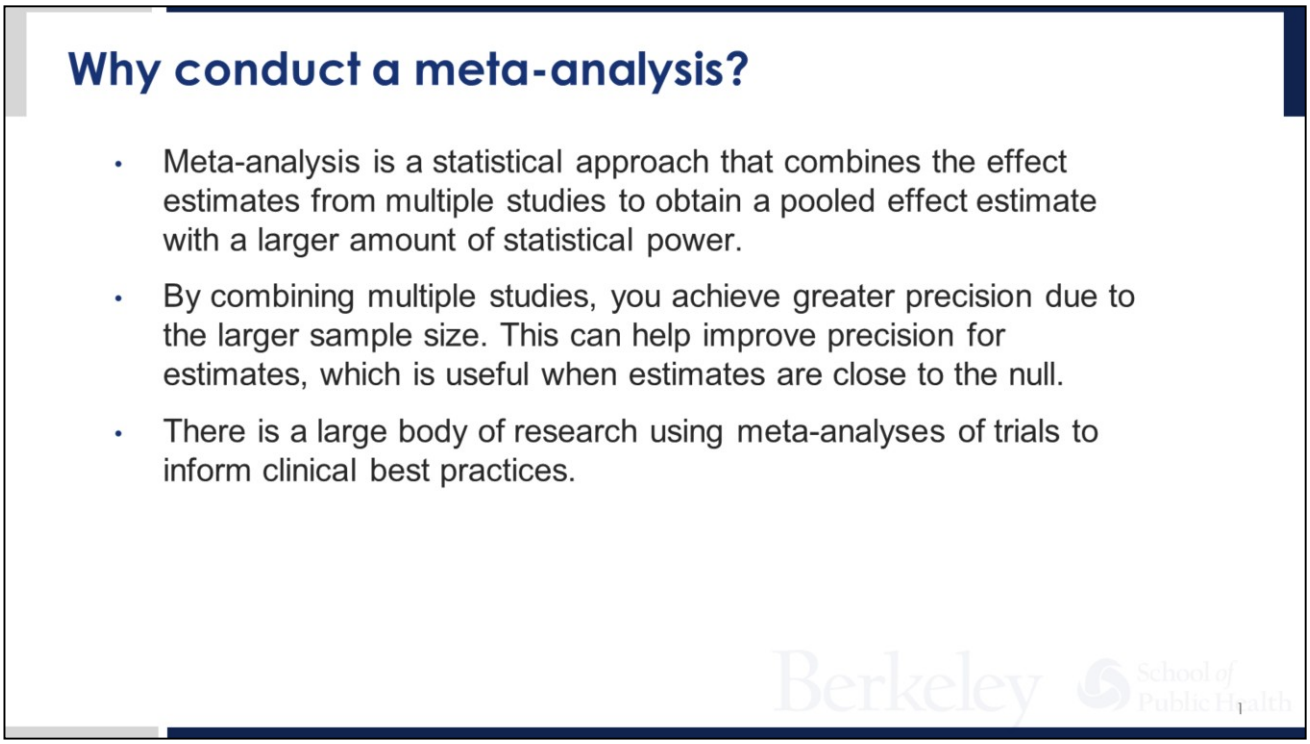


This slide shows the flow diagram for a systematic review. This allows the reader to track what the authors have done in their identification of the literature. And this comes right out of the PRISMA guidelines and really belongs in every systematic review that's published. And so at each step, which includes the identification of articles, the screening of articles, the eligibility of articles, and the final inclusion of articles, at each step, the number of articles kept and removed is shown. And usually the specific articles that were kept or removed are itemized in a separate appendix. And note that in the last step, at the very bottom, when we say studies are included in a quantitative synthesis, that step is done only in a meta-analysis, which we'll come to in the next video.



So in summary, systematic reviews summarize the evidence on a specific research question and are conducted in a fashion such that they can be replicated. While they're often time consuming and cumbersome, systematic reviews provide powerful information that allow us to generate new hypotheses, identify gaps in the published literature, prepare for future studies, and provide context for our work, such as in grants or papers. And in general, systematic reviews should be reported using the PRISMA checklist.

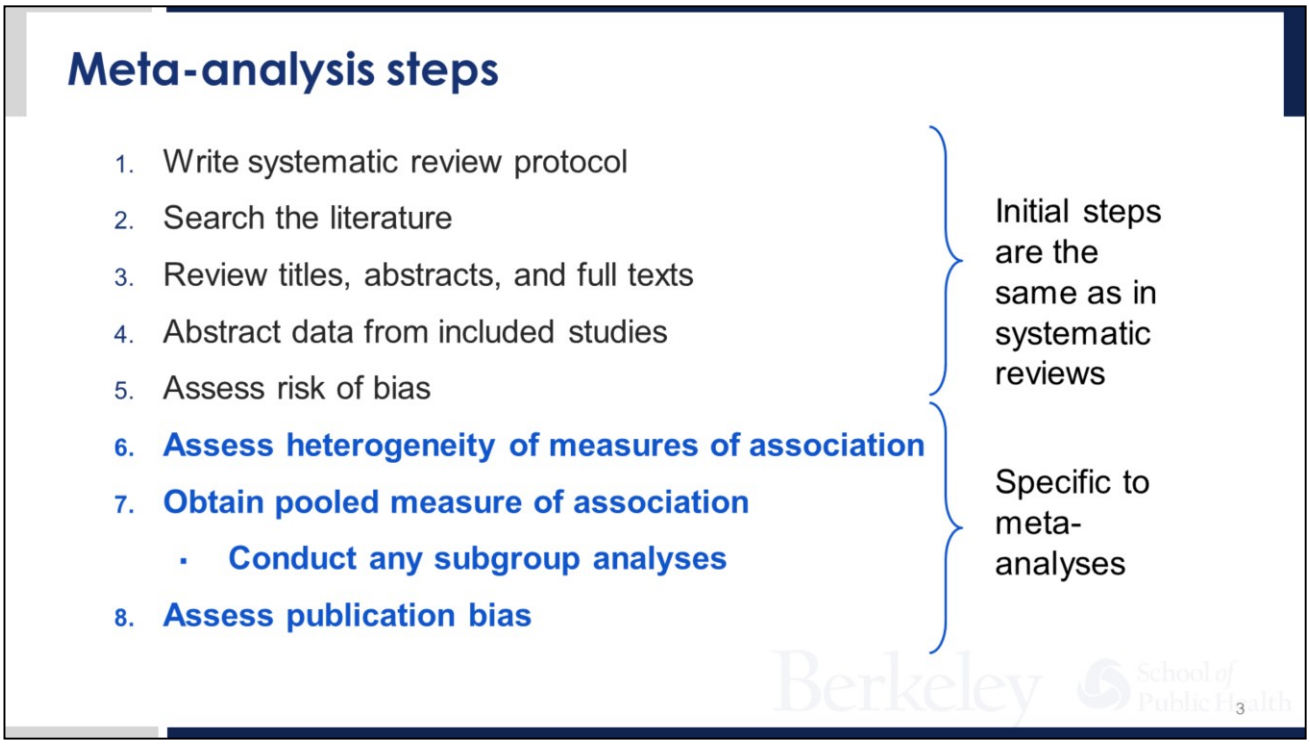
JACK COLFORD: Let's continue our discussion by talking about a technique called meta-analysis. Meta-analyses are really important tools that epidemiologists use to summarize and estimate effects across a body of literature. So we're moving now from working on just one study to taking multiple studies together and coming up with a summary estimate. And there are different techniques for doing this.



So in this approach, we're going to use a statistical technique that will combine the effects from multiple individual different studies, and thereby get a pooled effect estimate. And this gives us a larger amount of statistical power, because we're summarizing across multiple prior results. And by increased statistical power, what I'm specifically referring here to is the fact that we'll have a greater precision. Because of a larger sample size, we'll have a tighter confidence interval, generally around the point estimate that we get. And that is greater precision.

And this is particularly helpful if we are working in a situation where even the summary estimates are close to the null value of no effect. Having a tighter confidence interval will allow us to see whether or not perhaps the body of literature does not cross the null in its confidence interval. There's really a huge body of research that uses meta-analyses, particularly meta-analyses of trials in order to inform best clinical practices. You know, multiple smaller studies may have been done on a particular drug treatment or a particular medical device, or some other sort of intervention. These can then be pooled together to have increased statistical power to come up with a summary result across all of them.

One of the best tools for finding previously conducted systematic reviews and meta-analyses is the Cochrane Database of Systematic Reviews. So you can find this online. We're showing you the link here. Just a really powerful way on any topic to go and see if someone has prescribed and laid out very carefully a prior meta analysis and systematic review.



In the earlier lecture, we talked about the steps first in a systematic review. And in many situations, it's possible to continue beyond a systematic review to step 6, 7, and 8 and conduct a meta-analysis. And the key steps in a meta-analysis are to assess whether there is heterogeneity between the different studies that are used.

So let's say I'm studying 10 different prior studies in my meta-analysis. I want to see how variable are the results from those 10 different studies. That's heterogeneity. Then I went to obtain a pooled measure of association. That is, how do I average it? Let's say I have 10 cumulative incidence ratios, one from each study. How do I average those 10 cumulative influence ratios, or whatever measure of effect I've used? How do I average them properly?

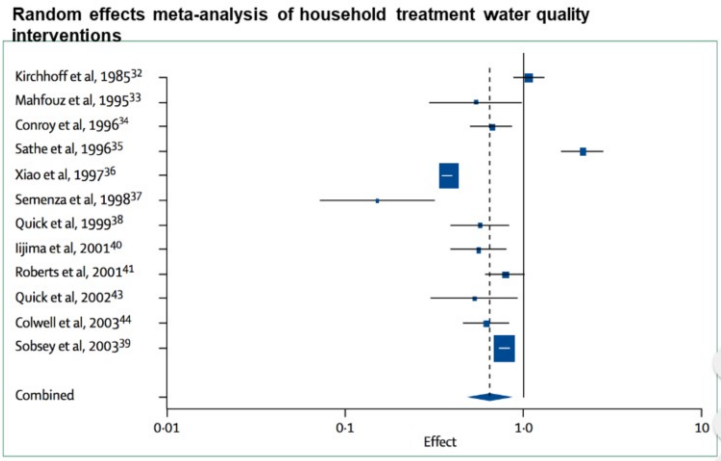
And I might choose to look at subgroup analyses. So for example, let's say I've done a meta-analysis of 10 prior studies, but I'm most interested in the effects of my-- in this case, let's make it an example of a drug. I'm most interested in the effect of a particular drug on an outcome. I might choose to go in and look at only the effect of the drug on the young patients in the study versus the old patients in the study.

So I would extract the data, if it were provided, on the younger patients in the 10 studies, and get a summary estimate for them, and then do the same for the older patients. So I could have those as subgroups. I could do that for other covariates as well, if the studies had published separate results themselves for those subgroups.

And then finally, we want to assess whether we have evidence of publication bias. This is the concept that perhaps some studies are missing from the literature and weren't published, either because of their quality, or their results weren't attractive or appealing to the investigators or the journals. So they never saw the light of day.

Step 6: Assess heterogeneity of measures of association

- Examine **forest plot** for heterogeneity
- Each study is one row in the plot
- This plot shows the combined effect estimate, but this is estimated after this initial assessment.
- The dashed line corresponds to the pooled estimate.
- The solid line corresponds to the null.
- Most effect estimates are to the left of the null and have similar value, suggesting little heterogeneity.



Fewtrell et al. Lancet Infect Dis 2005; 5: 42–52

So in the next step, we want to assess the heterogeneity of the different measures of association that are part of our systematic review. We're using here the systematic review by Lorna Fewtrell and a few others to show the effects of household water quality treatment interventions on diarrhea as the outcome.

So each row in this forest plot is an individual study. And what the plot is doing, then, is showing a combined effect estimate after gathering all the individual level studies. So all the rows are showing effect estimates for each individual study, but the diamond at the bottom is showing the statistically proper average across all of these studies.

And in this plot, the dashed line is showing us where the pooled estimate is-- that diamond estimate at the bottom. The middle of that diamond estimate is the single point estimate for all of the studies above that are averaged. And the solid line is drawn to show where the null value would be. The null value here would be no effect-- would be 1.0.

So since the diamond does not cross 1.0, it suggests that over this whole body of literature, there's evidence to suggest there's an effect-- that household water quality treatment interventions do lower the risk of diarrhea. And we also can just note that most of the individual estimates in this example are to the left of the null and have a similar value.

That would suggest little heterogeneity. However, there is obviously some heterogeneity. Look at the Semenza article in the middle. That has a much lower result, but also a very wide confidence interval. And look at the Sathe article-- the fourth study down. It goes in the opposite direction. But most of the studies are either showing a reduction in diarrhea or no effect, or the one study that shows a higher value.

Sources of heterogeneity

- Individual studies' measures of association may differ due to:
  - Population characteristics (age, sex, race)
  - Year of study
  - Location of study
  - Study participants' symptoms / clinical characteristics
  - Variables treated as confounders
  - Study design
  - Sample size
  - Statistical analyses

Many factors can lead to heterogeneity between different studies, because the studies might have been conducted on different age, sex, or races, the year of the study might have been different, the location of the study might have been different, the symptoms that were measured in the patients in the studies could be different, as were the clinical characteristics.

What each author chose to treat as confounders could differ between the studies. The design of the study could be different. The sample size of the study could be different. Or the way in which the authors analyzed the data statistically could be different. And that could impact how the relative measure of effect or the difference measure of effect is presented for each study.

Step 6: Assess heterogeneity of measures of association

Test for heterogeneity using a Q-statistic (a type of chi-square test)

Null hypothesis: no heterogeneity among original studies

MoA: measure of association

i: indicator for each study

MoA_s: summary measure of association

w_i: weight of ith study

$$Q = \sum_{i=1}^{k} w_i(\mathrm{MoA}_i - \mathrm{MoA}_s)^2$$

$$\mathrm{MoA}_s = \frac{\sum_i w_i \times \mathrm{MoA}_i}{\sum_i w_i}$$

- There are different approaches to weighting
- Larger studies usually have larger weights
- Under the null, the Q statistic follows a chi square distribution with k-1 degrees of freedom, where k is the number of studies
- P-value<0.2: reject the null hypothesis, conclude that heterogeneity is statistically significant

We need a way statistically to assess this heterogeneity across the studies. And we'll teach you now about two different approaches, one called the Q-statistic and one called the I squared statistic. The null hypothesis that we are starting with is that there is no heterogeneity among the group of studies that comprise our meta-analysis.

So the formula on the right is going to be the statistical approach to deciding whether or not we believe heterogeneity is present. So the formula is composed of several elements. The MoA is the Measure of Association. And it can either be at the level of the individual study-- that's MoA sub i. Or it can be a summary estimate-- MoA sub s, or Summary Measure of Association. So you see in the second formula there how the Summary Measure of Association is calculated.

It's the sum of the weight for each individual study times the measure of association for that study divided by the sum of the weights of all the studies. Now, there are different approaches to calculating w sub i. So if I have 10 studies in a meta-analysis, there'll be 10 w sub i's. And those values are calculated by a separate formula that just gives each study a certain weight.

And usually, larger studies have a larger weight. And after we've gotten our Summary Measure of Association-- MoA sub s-- we then put it into the formula at the top over here, which is the Q-statistic. So you see the expression for the Q-statistic. We take the difference between each individual study's Measure of Association-- the MoA sub i. We then subtract from it the Summary Measure of Association or the Standardized Measure of Association.

Step 6: Assess heterogeneity of measures of association

Test for heterogeneity using a Q-statistic (a type of chi-square test)

Null hypothesis: no heterogeneity among original studies

MoA: measure of association

i: indicator for each study

MoA_s: summary measure of association

w_i: weight of ith study

$$Q = \sum_{i=1}^{k} w_i(\mathrm{MoA}_i - \mathrm{MoA}_s)^2$$

$$\mathrm{MoA}_s = \frac{\sum_i w_i \times \mathrm{MoA}_i}{\sum_i w_i}$$

- There are different approaches to weighting
- Larger studies usually have larger weights
- Under the null, the Q statistic follows a chi square distribution with k-1 degrees of freedom, where k is the number of studies
- P-value<0.2: reject the null hypothesis, conclude that heterogeneity is statistically significant

We square that value, that difference, and multiply it by the w sub i for that particular study. So we would do this 10 times. We would add up those values. And the sum of those values would be the Q-statistic. Now, how do we tell whether the Q-statistic is statistically significant? Well, we go and look up in a chi-square table a value for a test statistic that has k minus 1 degrees of freedom.

And in this situation, k is the number of studies in the meta-analysis. So if we have 10 studies, it would be 10 minus 1, or 9 degrees of freedom. So we look up a value for Q with 9 degrees of freedom. And then we see the Q that we got from our process of completing these formula to the right here, we compare that against what our Q-statistic would be for 9 degrees of freedom. And if our Q-value is larger than that, we reject the null. If we reject the null, we reject the idea that there's no heterogeneity, and then we conclude that there is heterogeneity.

Usually we look up a Q-value that has a statistical value-- a P-value of 0.2. So in these tests of heterogeneity, we generally use 0.2 as our hypothesis against which we're testing it. So again, if we have a Q-statistic larger than a Q that has a P-value of 0.2 from our 9 degrees of freedom in this example, then we would conclude that heterogeneity is statistically significant in our meta-analysis.

Step 6: Assess heterogeneity of measures of association

Test for heterogeneity using a I^2 statistic

- Assesses the percentage of variability due to heterogeneity rather than chance.

$$I^2 = \big((Q - \mathrm{df})/Q\big) \times 100$$

- Q: defined as in previous slide
- df: degrees of freedom (k-1)
- If I^2 > 50% heterogeneity is substantial

Another approach to testing for heterogeneity is to use a different statistic called the I squared statistic. And this approach assesses the percentage of variability that's due to heterogeneity other than chance. And the formula here for this is I squared is given by the Q-value from the prior slide-- same Q-value-- minus the degrees of freedom divided by the Q-value times 100, to put it in percentage terms. And once again, the degrees of freedom is k minus 1. And with this result, if I squared is greater than 50%, it is decided or usually concluded that heterogeneity is substantial.

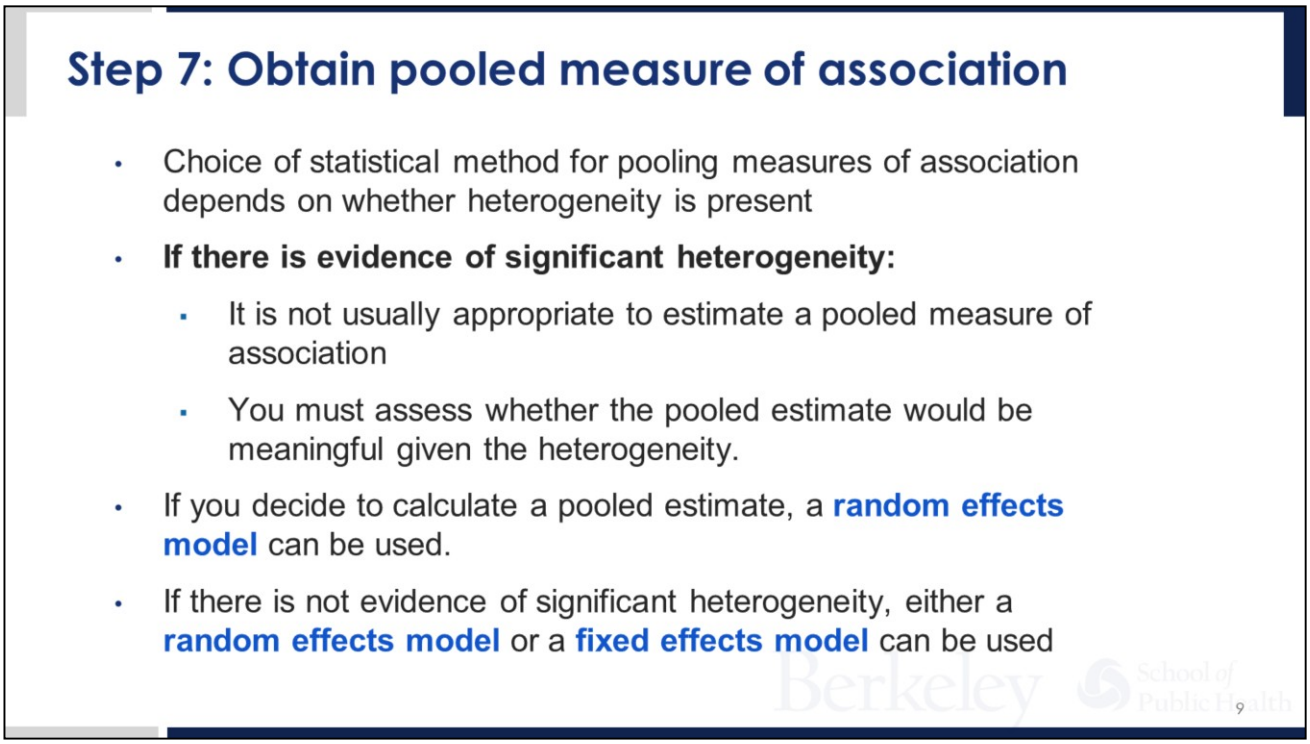
Step 7: Obtain pooled measure of association

- MoA_s: summary or pooled measure of association
- w_i: weight of ith study
- Weights are defined differently depending on the statistical method:
  - **Fixed effects**
    - Inverse variance
    - Mantel-Haenszel
    - Peto
  - **Random effects**
    - DerSimonian Laird

$$\mathrm{MoA}_s = \frac{\sum_i w_i \times \mathrm{MoA}_i}{\sum_i w_i}$$

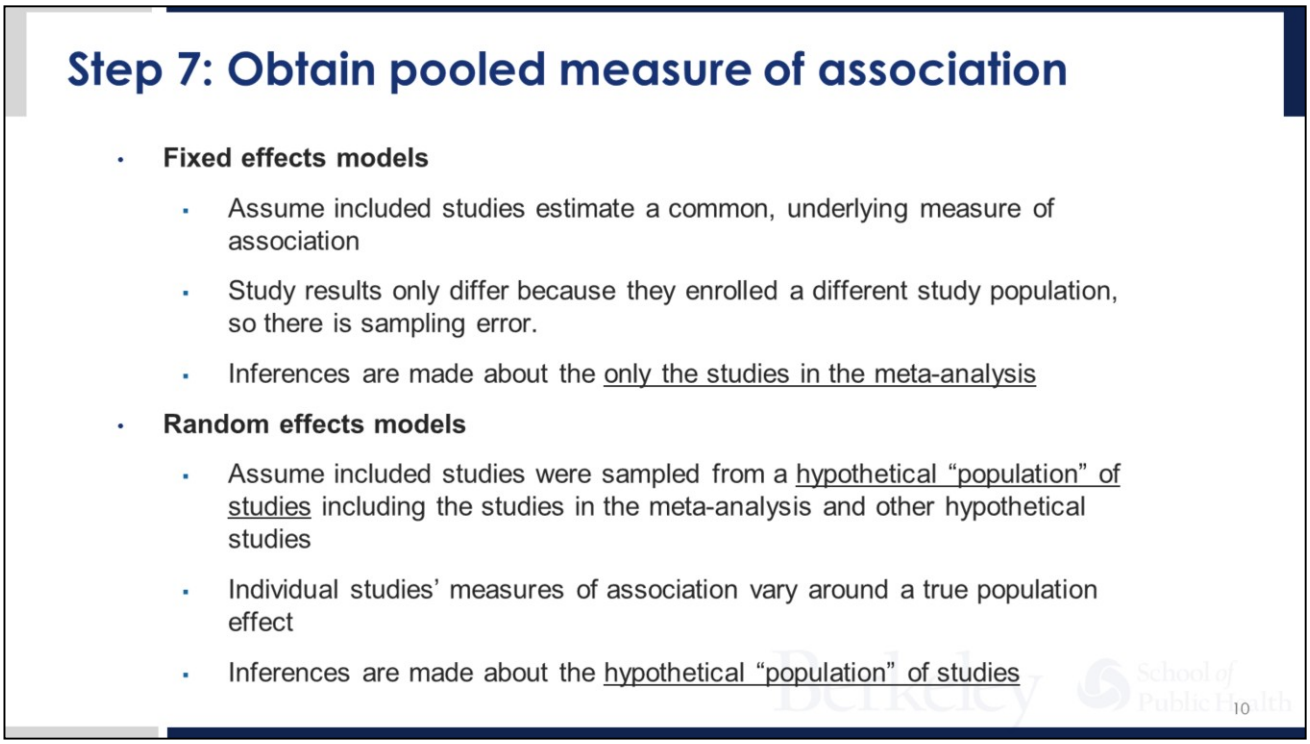
Now, to do these calculations, we need a pooled measure of association. So this MoA sub s is a summary or pooled measure of association. And the wi, once again, is the weight of the study. So the weights are defined differently, depending on various different statistical methods. And there are several. And we won't go into details for these.

But there are some models of family of models-- I'll talk about these in a moment-- called fixed effects models, which include an inverse variance approach, a Mantel-Haenszel approach, or a Peto approach. And then there's another model called the DerSimonian Laird random effects model. And all of these make use of the Summary Measure of Association that you see here. It's the wi that can differ, depending on which of these model choices is made.



So our next step is to obtain a pooled measure of association. Now, our choice of a specific statistical method for finding a pooled measure of association depends on whether heterogeneity is present. So if in our prior testing, either using Q-statistic or I squared that we developed in the last few slides, if we have evidence of significant heterogeneity, then it's usually not appropriate to estimate a pooled measure of association.

So we have to assess whether the pooled estimate would be meaningful given the heterogeneity. If there is heterogeneity present, most authors would argue not to present a pooled estimate, but rather to just present the separate estimates. Now, if we decide to calculate a pooled estimate, we could use a random effects model, even when heterogeneity is present, because a random effects model, it turns out, will have a larger confidence interval. But if there's not evidence of significant heterogeneity, we can use either a random effects model or a fixed effects model. If there's not much heterogeneity, the random effects model and the fixed effects model will give similar results.

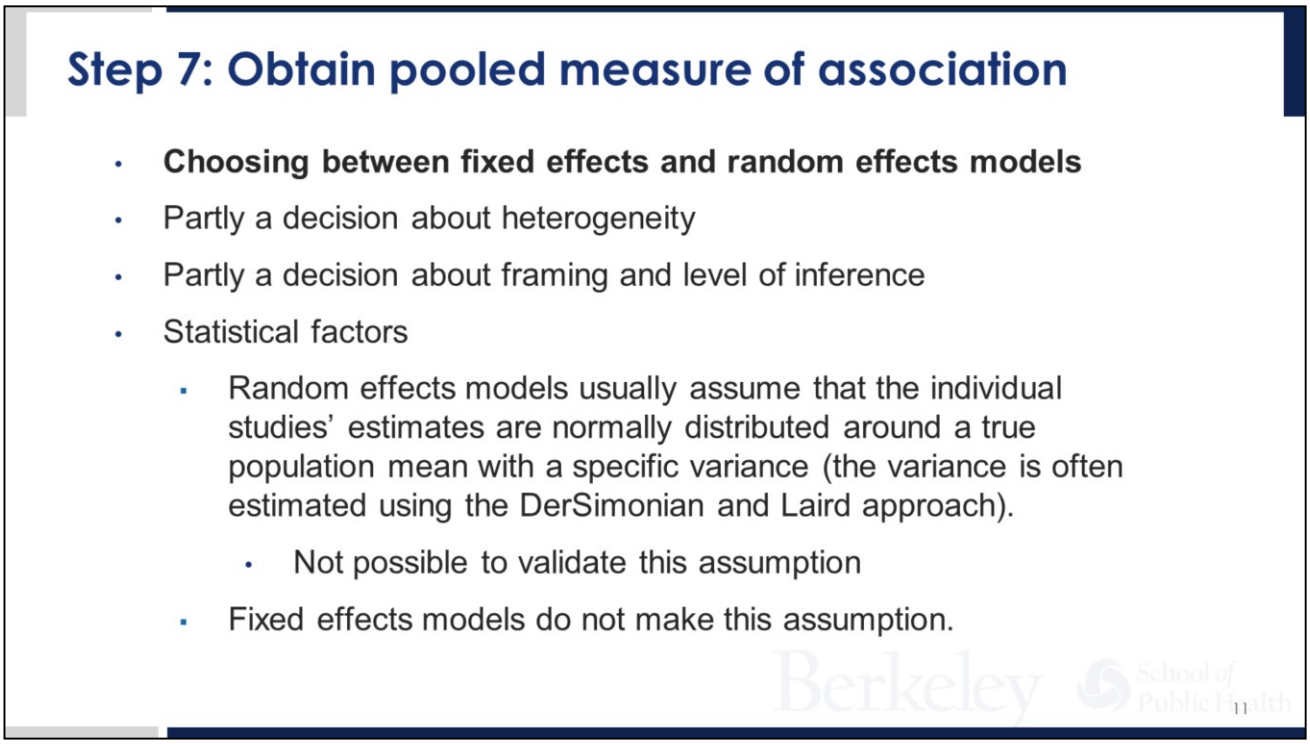


In fixed effects models, there's an assumption that the studies we've included are estimating a common underlying measure of association, and that the study results that differ between the individual studies in our meta-analysis differ only because they enrolled a different study population. So there's just only sampling error.

We're making inference here only about the studies in our meta-analysis. Now, this is different than the assumption that goes into a random effect model. In a random effects model, we're assuming that the studies that we included in our meta-analysis were themselves as a group just a sample from a larger hypothetical population of studies that include the studies in our meta-analysis, but other hypothetical studies that we missed-- that we weren't able to get.

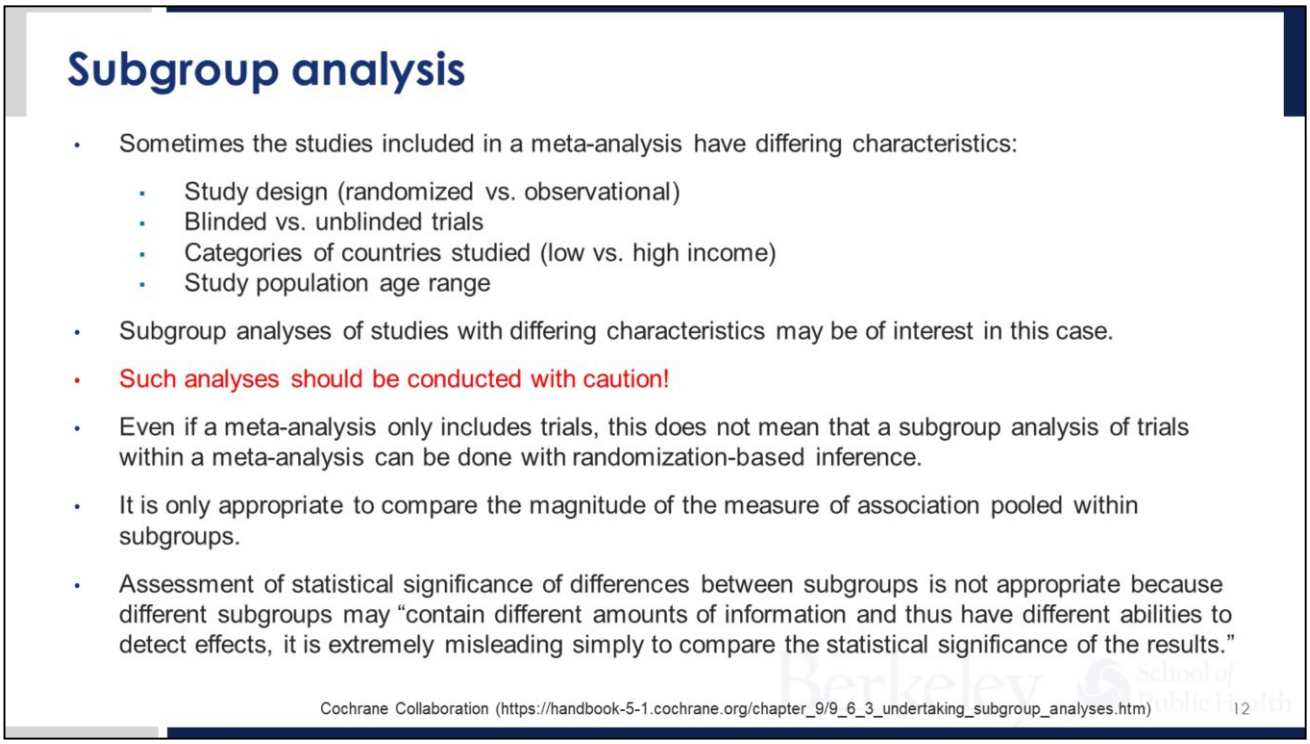
So our grouping of studies in a meta-analysis is just a sampling of the broader world of studies that are out there on this topic. And the assumption in a random effects model is that the individual studies measures of association are varying around this kind of outside true population effect. And so inferences we make in the random effects model are made about the hypothetical larger population of studies, not just the studies that we sampled from. So I hope you're understanding the difference between the fixed effects and random effects model.

The fixed effect is focused on just the studies that we have found and identified. That's our universe of studies. But in a random effects model, the idea is that we believe we've missed some studies, and we're trying to generalize or draw inference out to those other studies that we might have missed, in terms of what the population mean was.



So in order to choose between fixed effects and random effects, investigators usually use heterogeneity as part of that decision. If there's a lot of heterogeneity-- if the P-value for when you tested the Q-statistic. Or if your I square is over 50, many or most investigators would choose to use a random effects model.

And partly it's a decision about how you're framing your results in the meta-analysis. Are you trying to talk just about the studies that you located, or are you actually trying to talk more broadly about the whole likely universe of studies that's out there? The random effects models usually assume that the individual study estimates are normally distributed around a true population mean with some specific variance. And this variance is estimated by this DerSimonian Laird approach that I mentioned. But it's not possible to validate this assumption. And the fixed effects models don't make this assumption.



In many meta-analyses, investigators will decide to do subgroup analysis. And this is because the studies within a meta-analysis might have different characteristics that are worth exploring. For example, some of the studies might be randomized trials. Others might be observational studies. The investigators might then choose to conduct subgroup analyses separately for each type of study design. Or some of the studies might be blinded and others might be unblinded.

After pooling all the studies into one overall estimate, the investigator might choose to look at the blinded studies separately from the unblinded studies to see if there's consistency in the estimates between the two. There might be different categories of countries that were studied, such as low versus high income countries that might be interesting to explore in subgroup analyses. Or the study populations may have different age ranges that could be explored.

And one of the powers of meta-analysis is the ability to do this, because these different characteristics wouldn't be present with enough data in any one study. So by pooling across more studies, these characteristics of interest can be studied. But these have to be conducted with caution, because even though a meta-analysis includes trials, it doesn't mean that a subgroup analysis of trials within a meta-analysis can be done with randomization-based inference. And it's only appropriate to compare the magnitude of the measure of association pooled within subgroups.

An assessment of statistical significance of difference between subgroups is usually not appropriate, because different subgroups might have different amounts of information, and thus have different abilities to detect effects. So it's misleading to statistically compare between different subgroup analyses. So in other words, if I've done a meta-analysis overall, and then I do a subgroup analysis for the old versus the young, it's not encouraged to do a statistical test to compare the results of the old to the results of the young because of all these reasons I've just mentioned.

Example of subgroup analysis in meta-analysis

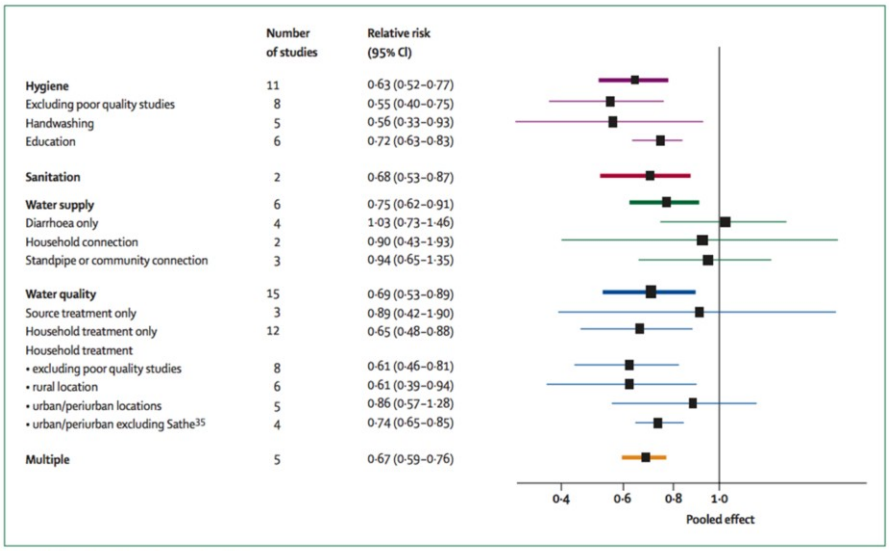


Figure 3: Summary of meta-analysis results

Fewtrell et al. Lancet Infect Dis 2005; 5: 42–52

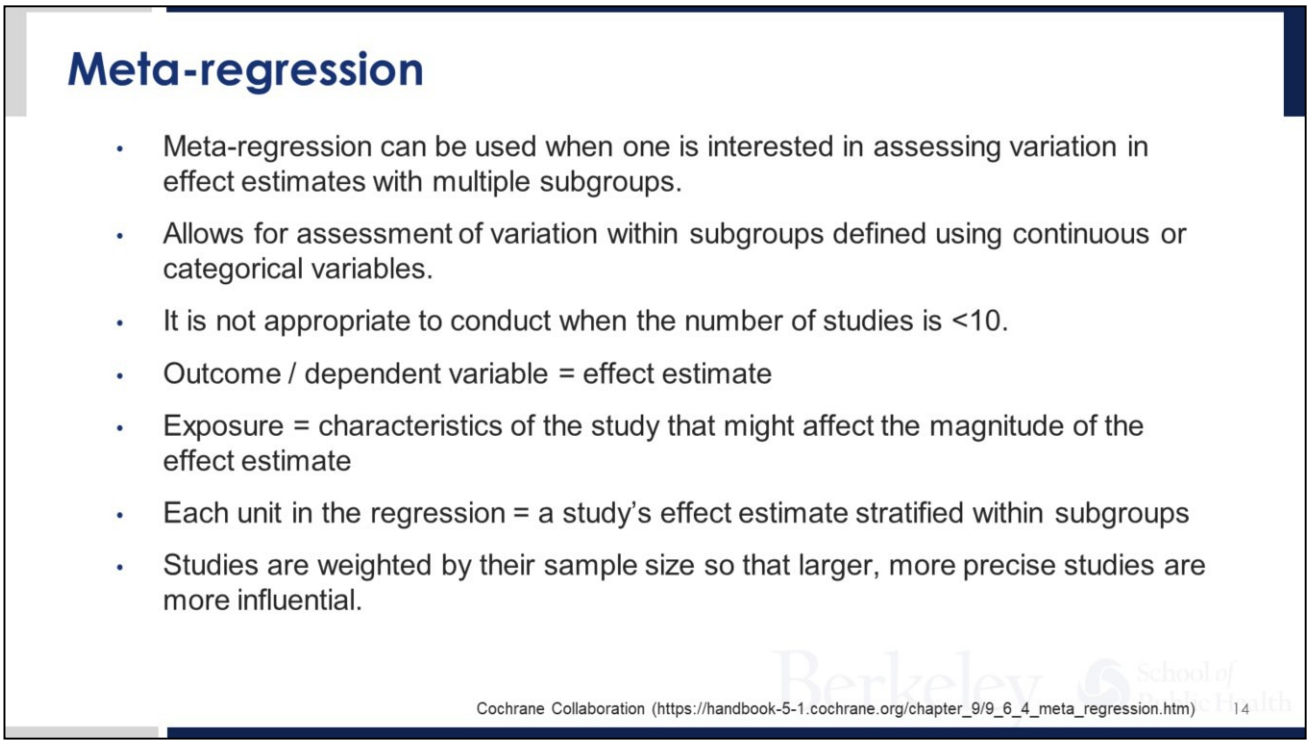
13

Let's look at an example of a subgroup analysis from the Fewtrell meta-analysis. So this, again, was a study looking at whether different water and hygiene and sanitation interventions reduced diarrhea. So the way the forest plot is set up here is the null effect at 1 would imply in any given study that there was no difference between the intervention group and the control group. And results to the left would imply a reduction in diarrhea. Results to the right would be an increase in diarrhea.

So the overall meta-analysis was done. We saw that earlier. But then subgroup analyses were done. And we see those here, where the subgroups are defined by the different types of interventions. So the hygiene interventions are separate from the sanitation interventions, which are separate from the water supply interventions, and separate from the water quality interventions, and then separate from interventions that use multiple interventions.

So for instance, in the hygiene group, there were 11 different studies that focused on hygiene interventions. And you see that it had a relative risk summarized in the meta-analysis of 0.63. So suggesting a reduction in diarrhea among the hygiene grouping of interventions when a meta-analysis is done of just that subgroup.

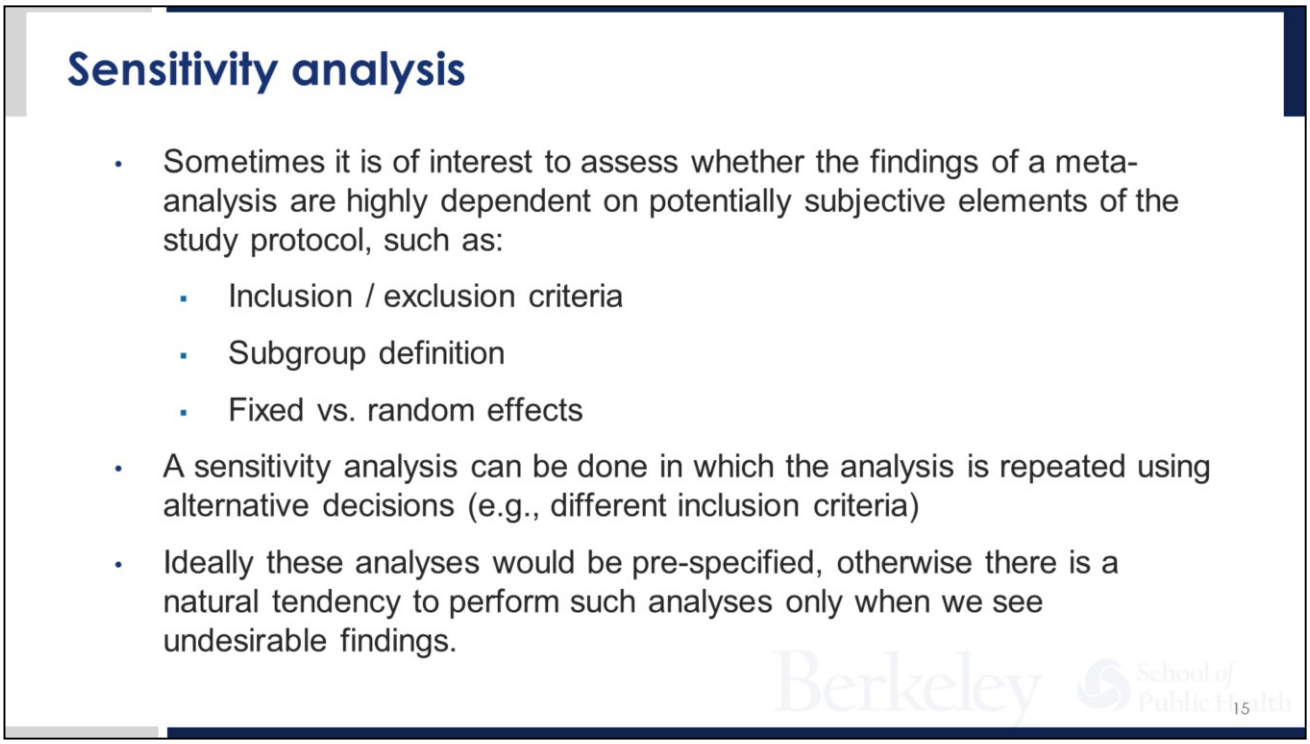
Similarly for sanitation, it was 0.68, for water supply 0.75, for water quality 0.69, and for multiple interventions, 0.67. So you see kind of at a higher level here that there is a consistency between the different interventions when the groups are broken down into these specific areas. And then furthermore, in some of the subgroups here, we also separated into specific interventions within that subgroup. So take a look at the water quality subgroup, for example. We separated the three studies that looked at only source treatment, the 12 studies that looked at only household treatment, and so forth. So this is essentially a subgroup within a subgroup.



Another technique used in meta-analysis is called meta-regression. And meta-regression is a technique that can be used when one is interested in assessing variation in effect estimates across the multiple studies in the meta-analysis. This technique allows for an assessment of variation within studies that are defined using continuous or categorical variables.

So let's say, for example, I've got 10 studies in my meta-analysis. And the mean age in the 10 different studies is different. I could then regress the outcome of the 10 studies-- in other words, the 10 effect estimates-- against the mean age of the 10 studies to see whether there's a relationship between age and the effect estimate of the studies.

Now, usually this isn't done when the number of studies is fewer than 10. And as I mentioned, the outcome or dependent variable in this regression is the effect of each study. And the exposure here is the characteristic of the study that might affect the magnitude of the effect estimate. And each unit in the regression is a study's effect estimate stratified within subgroups. So studies are weighted by their sample size so that larger or more precise studies are more influential.



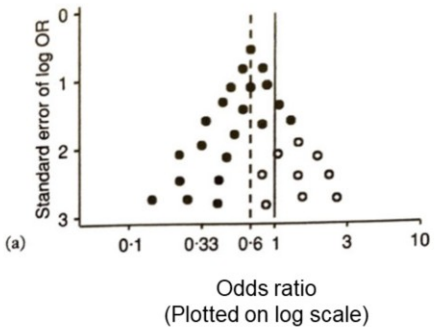
Another useful analysis in the meta-analysis can be a sensitivity analysis. So this idea here is to see whether the findings of the meta-analysis are highly dependent on subjective elements of the study protocol, such as inclusion criteria or exclusion criteria, the definitions of the subgroups, whether fixed or random effects were used.

A sensitivity analysis can be done in which the analysis is repeated using alternative decisions. For example, changing the inclusion criteria for studies, and seeing whether that impacts the meta-analysis results. So for example, I could do a meta analysis that includes blinded and unblinded trials, then repeat that analysis excluding the unblinded trials, and seeing whether or not the result I get is different when the unblinded trials were excluded.

So usually, these types of analysis would be pre-specified so you aren't on a hunting expedition. Because otherwise, there'd be a natural tendency to perform the analysis only when we see undesirable findings. And then we would try to do sensitivity analyses to find a situation where we got desirable findings.

Use funnel plots to assess publication bias

- Measure of association on one axis
- Sample size or measure of precision (e.g., standard error) plotted on the other axis
- If no publication bias occurred, we would expect the graph to resemble a funnel.
  - Larger studies with higher precision will be closer to the overall pooled estimate
  - Smaller studies with lower precision will be distributed symmetrically on either side of the average.
- The plot to the right depicts an example with no publication bias.



Egger et al. *Systematic Reviews in Health Care*. BMJ Publishing Group, 2001.

One tool to assess whether there's publication bias in a body of literature that we've identified in a meta-analysis is through a funnel plot. And in a funnel plot, the sample size or the measure of precision or standard error is plotted on one axis, and the measure of effect is plotted on the other axis-- the measure of association.

So on this plot example here, we see on the x-axis the odds ratio plotted on the log scale, and the standard error of the log the odds ratio plotted on the y-axis. Now, if there had been no publication bias, we would expect the graph that we see in this plot such as this to represent a funnel. It would be fully filled out.

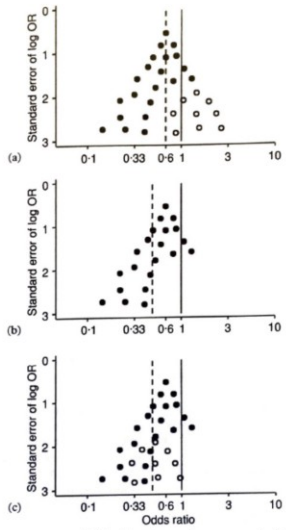
And note that larger studies with higher precision will be closer to the overall pooled estimate, generally, and smaller studies with lower precision would be distributed symmetrically on either side of the average. So they'd be farther away from the average effect. Now, the plot to the right is depicting an example where there's believed to be no publication bias, because the funnel looks complete. And the open circles here are referring to studies with a smaller sample size, and the closed circles are referring to studies with larger sample size.

Use funnel plots to assess publication bias

a) Symmetrical plot with **no publication bias**
   - **Open circles:** smaller studies with no statistically significant effects

a) Asymmetrical plot **with publication bias.** Smaller studies with no statistically significant effects are missing from the plot.

a) Asymmetrical plot with **bias due to poor quality of smaller studies**
   - **Open circles:** smaller studies with poor quality



(Plotted on log scale)

Dashed line: pooled odds ratio

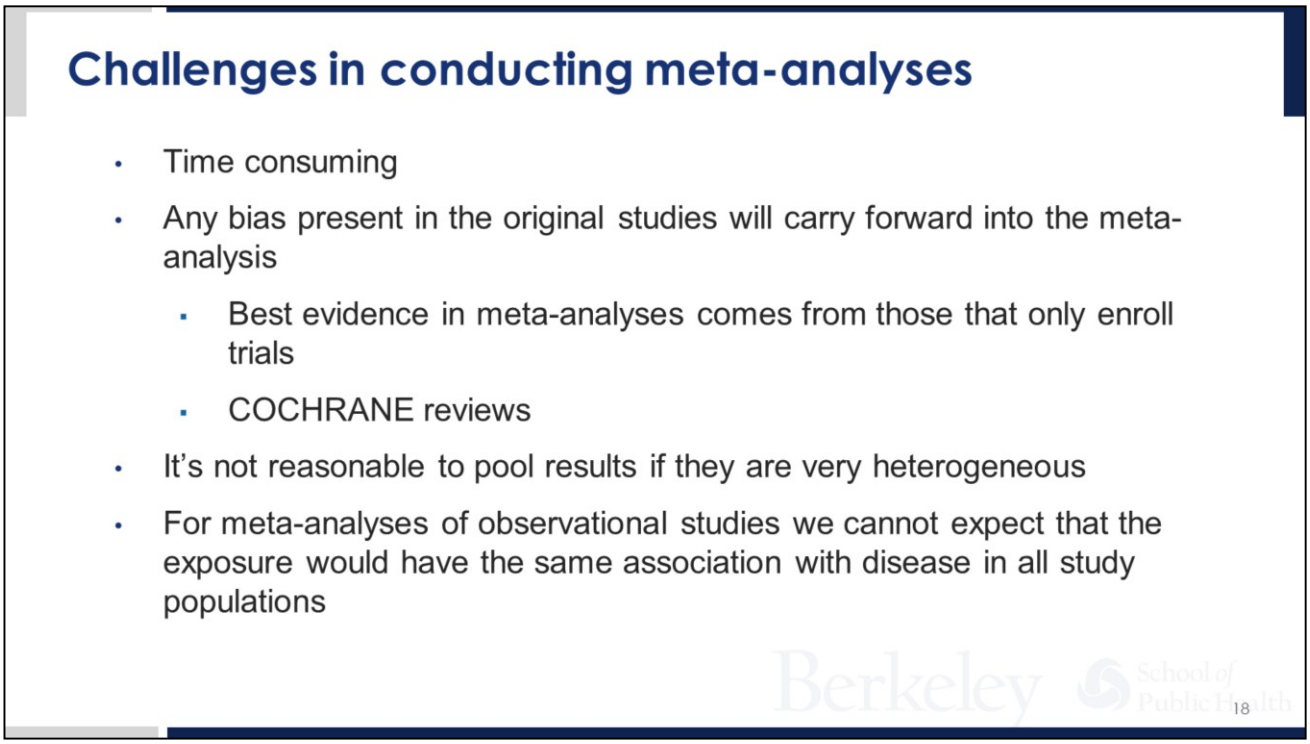
Solid line: null (OR = 1)

Egger et al. *Systematic Reviews in Health Care*. BMJ Publishing Group, 2001.

17

And that's shown again here in these several different examples of funnel plots. In the top one, the idea is that there is no publication bias. And as I said, the open circles are smaller studies, and they show no statistically significant effects. The second plot-- the asymmetrical plot here, you see how part of the funnel is missing. This has publication bias present. And the studies that are missing are the smaller studies with no statistically significant effects. They are apparently missing from the plot.

And then the third example is showing an asymmetrical plot with bias due to poor quality of smaller studies. Here the open circles are smaller studies with poor quality. And so these quality measures are made during your review of the literature. And so that's an interpretation you need to bring when you look at the funnel plot, because you've separately quantified whether the studies are of high or low quality.



Meta-analyses can be very challenging. They're time consuming. And any bias present in the original studies will carry forward into the meta-analysis. The best evidence in meta-analysis generally comes from those meta-analyses that only enrolled trials. Cochrane reviews are a great way to see prior meta-analyses and systematic reviews that have been done.

Now, if the results suggest a lot of heterogeneity between the studies, then one should stop at the stage of just presenting the individual results, and not present the summary meta-analysis result. And for meta-analyses of observational studies-- which are commonly done, even though it's considered more powerful to do an analysis of trials-- we can't expect that the exposure would have the same association with disease in all the study populations. So there's usually going to be even more variability in what we see in a meta-analysis of observational studies.

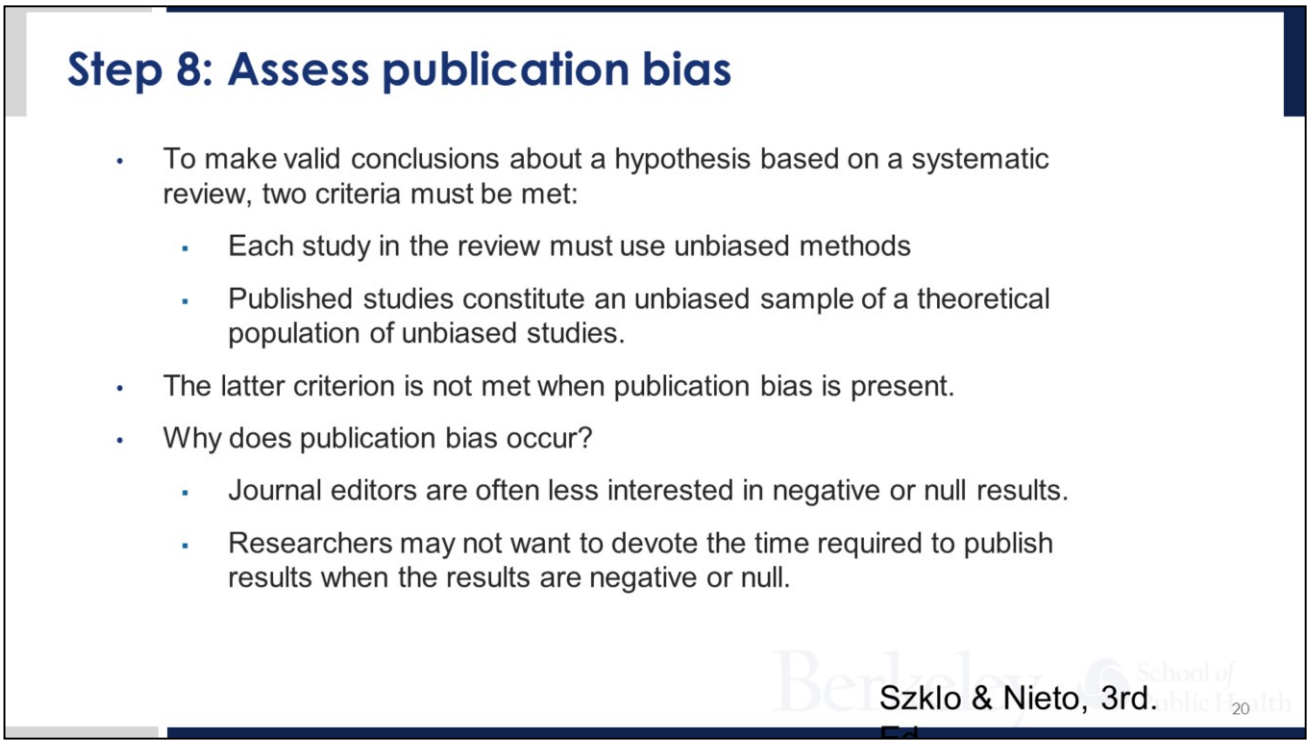
Reporting meta-analyses

- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses http://www.prisma-statement.org/
- PRISMA checklist for reporting includes specific guidelines on what to report in a publication of a systematic review

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
METHODS			

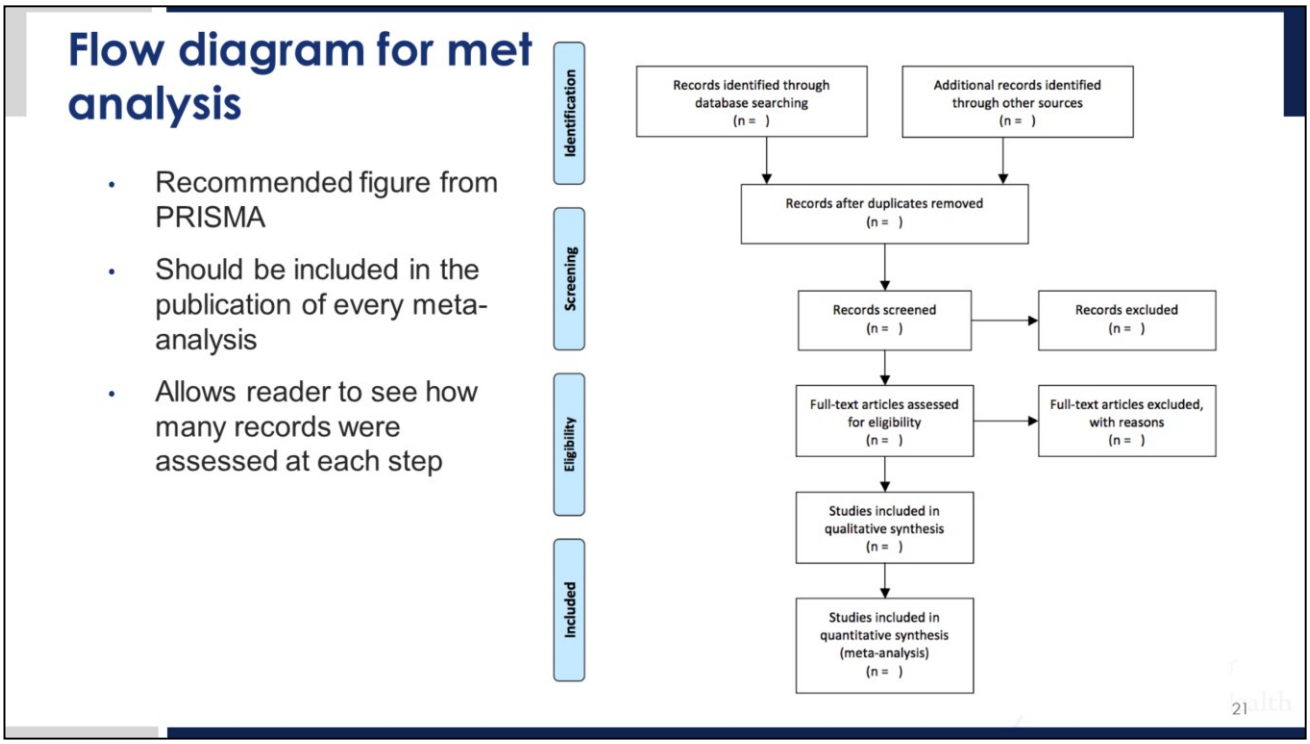
We've discussed before the PRISMA checklist for how to report systematic reviews. That also has the guidelines for reporting a meta-analysis.



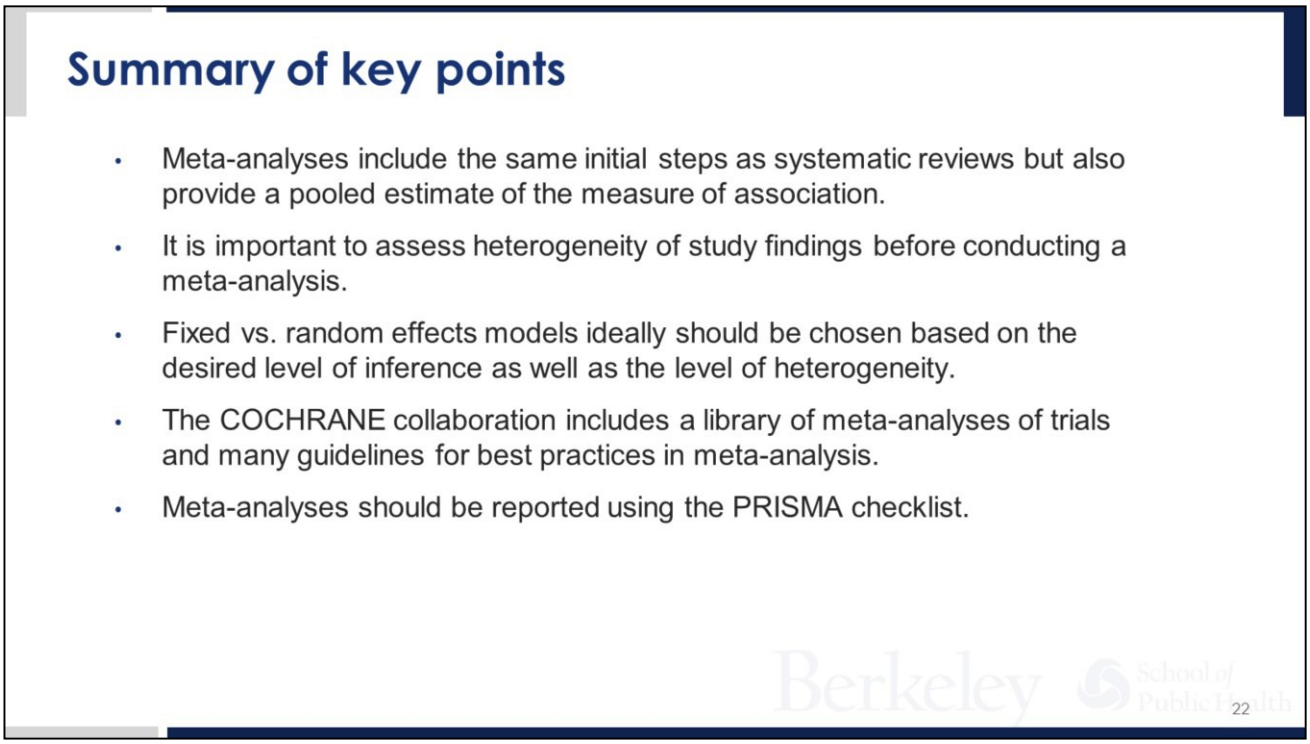
And our final step after we've done our funnel plots is to assess publication bias. And in order to do that, we want to see whether each study in the review used unbiased methods. So we go through each study and review what methods the authors used.

And published studies constitute an unbiased sample of a theoretical population of unbiased studies. This is a criterion that we believe has to be true for our meta-analysis to arrive at an accurate result. When this second criterion is not met, then we believe publication bias is present. So why does publication bias occur?

Well, it's well known that journal editors are often less interested in publishing negative or no results, because it's not as exciting for the journal to published studies with results like that. And also, researchers may not want to devote the time needed to published studies when the results of their studies are negative or null.



This flow diagram for meta-analysis is similar to what we saw for systematic reviews and through the different stages of the analysis, including identification of studies, the screening of studies, the eligibility of studies-- that was part of our systematic review. And then the studies that are included in our final quantitative synthesis or meta-analysis is that last box at the bottom.



So to summarize, meta-analysis include the same initial steps as systematic reviews, but they also provide a pooled estimate of the measure of association-- pooled or averaged across the entire number of studies in the meta-analysis. It's important to check for heterogeneity of study findings before conducting a meta-analysis, because if heterogeneity is significantly present, we don't want to do a summary result.

Fixed versus random effects models should be chosen based on the desired level of inference, as well as the level of heterogeneity. This was the idea that a fixed effects model are essentially or universe of studies is just the studies we've identified, whereas in a random effects model, the universe of studies is we believe we've missed some studies, and our sample is just a sampling of that universe of studies. We've mentioned several times how important the Cochrane collaboration is. And it has a library of meta-analyses and systematic reviews, and also has a lot of tools for best practices in meta-analysis. And meta-analyses should be reported using the PRISMA checklist.

 International Journal of Epidemiology 2002;31: 1-5

EDITORIAL

Where now for meta-analysis?

Matthias Egger, Shah Ebrahim and George Davey Smith

In the short time since its introduction, meta-analysis, the statistical pooling of the results from independent but 'combinable' studies, has established itself as an influential branch of clinical epidemiology and health services research, with hundreds of meta-analyses published in the medical literature each year.[1] This issue of the *International Journal of Epidemiology* contains several papers[2-9] that address methodological issues in meta-analytic research, a review article on where we stand with systematic reviews in observational epidemiology[10] and three meta-analyses of observational studies.[11-13] Publication of a themed issue on meta-analysis by an epidemiological journal begs several questions: Where does meta-analysis come from? Does it deserve the attention it is currently getting? And where should it be going next?

The statistical basis of meta-analysis reaches back to the 17th century when, in astronomy, intuition and experience suggested that combinations of data might be better than attempts to select amongst them.[14] In the 20th century the distinguished statistician Karl Pearson (Figure 1), was, in 1904, probably the first medical researcher using formal techniques to combine data from different studies when examining the preventive effect of serum inoculations against enteric fever.[15] However, such techniques were not widely used in mec[jcine for many years to come. In contrast to medicine, the social sciences and in particular psychology and educational research, demonstrated early interest in the synthesis of research findings. In 1976 the psychologist Gene Glass coined the term 'meta-analysis' in a paper entitled 'Primary, Secondary and Meta-analysis of Research'.[16] Three years later the British physician and epidemiologist Archie Cochrane drew attention to the fact that people who want to make informed decisions about health care do not have ready access to reliable reviews of the available evidence.[17] In the 1980s meta-analysis became increasingly popular in medicine, particularly in the clinical trial fields of cardiovascular disease, oncology, and perinatal care. In the 1990s the foundation of The Cochrane Collaboration,[18] an international network of health care professionals who prepare and regularly update systematic reviews ('Cochrane Reviews') facilitated the conduct of meta-analyses in all areas of health care.

The achievements of meta-analysis in the realm of clinical trial research are impressive. First, meta-analysis helped to overcome the problem first identified by Pearson, that 'any of the groups ... are far too small to allow of any definite opinion being formed at all, having regard to the size of the probable error involved'. Although the size of trials published in general health care journals has been increasing since 1948 (see the paper by McDonald *et al.*[7] in this issue), many trials fail to detect, or exclude with certainty, a modest but clinically relevant difference in the effects of two therapies. This means that the conclusions from several small trials will often be contradictory and confuse those seeking guidance. The meta-analytic approach may overcome this problem by combining trials evaluating the same intervention in a number of smaller, but comparable, trials. Early examples include meta-analyses of trials of beta-blockers in secondary prevention after myocardial infarction, [19] of a short course of corticosteroids given to women about to give birth prematurely[20] and of adjuvant tamoxifen in early breast cancer.[21] A welcome effect of the surge of systematic reviews and meta-analysis, and of evidence-based medicine in general, is the dismantling of the magnificence and splendour of the full professor, who was used to argue casually based on status and opinion, but is now confronted by well-informed junior members of staff and consumers of health services.

Second, meta-analysis may highlight areas where there is a lack of adequate evidence and thus identify where further

Figure 1 Distinguished statistician Karl Pearson is seen as the first medical researcher to use formal techniques to combine data from different studies

Editorial Office, International Journal of Epidemiology, Department of Social Medicine, University of Bristol, UK.

Correspondence: Matthias Egger, Department of Social Medicine, Canynge Hall, Whiteladies Road, Bristol BSS 2PR, UK. E-mail: m.egger@bristol.ac.uk

studies are needed. For example, a period of starvation is common practice after gastrointestinal surgery, but a recent meta-analysis22 of randomized controlled trials concluded that keeping patients 'nil by mouth' may do more harm than good, and that a large trial is required to clarify this issue. About half of Cochrane reviews and a fifth of meta-analyses published in medical journals conclude that the evidence is inappropriate and that a large trial is needed.[23] Indeed, as Iain Chalmers pointed out, systematic reviews of existing trials and registers of ongoing trials should be seen as prerequisites for scientific and ethical trial design.[24]

Third, meta-analyses offer a sounder basis for subgroup analyses, particularly if they are based on individual participant data.[25] For example, the meta-analysis of individual patient data from 55 trials of tamoxifen in operable breast cancer showed that the benefit of tamoxifen was much smaller and non-significant in women reported to have oestrogen receptor negative disease.[26] Based on these findings, oestrogen receptor status is now used to inform treatment decisions.

Finally, the realization that the results from meta-analysis are not always trustworthy[27,28] led to research into the numerous ways in which bias may be introduced, and the development of methods to detect the presence of such bias. For example, several studies have examined the influence of unpublished trials, trials published in languages other than English, and of trial quality on the results of meta-analyses of randomized controlled trials. The authors used a 'meta-epidemiological' approach29 and considered collections of meta-analyses in which component trials had been classified according to characteristics such as publication status or study quality, thus ensuring that the treatment effects are compared only between studies in the same meta-analysis.[30] Figure 2 shows a 'meta-meta-analysis' of these studies which includes the study by Jiini *et al.* published in this issue, on language bias.[31] Combined results indicate that, on average, unpublished trials will underestimate treatment effects by about 10%, trials published in languages other than English will overestimate effects by the same amount and trials not indexed in MEDLINE will overestimate effects by about 5%. Trials with inadequate or unclear concealment of allocation and trials that are not double blind overestimate treatment effects by about 30% and 15%, respectively. The quality of trials thus appears to be a more important source of bias than the reporting and dissemination of trials. However, as pointed out by Clarke in his commentary,[32] the influence of language bias and other reporting biases may still be large in meta-analyses based on few trials. Also, the size of effects will differ across individual meta-analyses, perhaps depending on specialty, type of active and control intervention and trial design.

Considerable progress has also been made in our understanding of how best to detect bias, and deal with bias in meta-analysis, using graphical and statistical methods.[33] In this issue Higgins and Spiegelhalter[4] revisit the clinical trials of the effect of magnesium infusion in myocardial infarction, a well-known example where bias may explain the discrepancy between meta-analyses of small trials which showed a clear treatment effect[34] and the subsequent large Fourth International Study of Infarct Survival (ISIS-4)[35] which showed no effect. Using Bayesian methods the authors show how scepticism can be formally incorporated into the analysis and that such an approach would have led to appropriate caution before the results of the mega-trial became available. In his commentary,[36] Woods argues that the degree of scepticism required would have been extreme, considering the laboratory studies which made a beneficial effect of magnesium biologically plausible. Woods offers an alternative explanation for the discrepant findings between ISIS-4 and the smaller trials: myocardial protection by magnesium is abolished when treatment is delayed until after reperfusion has occurred, as was the case in ISIS-4, but not the smaller trials.

This debate must continue, but the magnesium example, and other meta-analyses that were later contradicted by single large

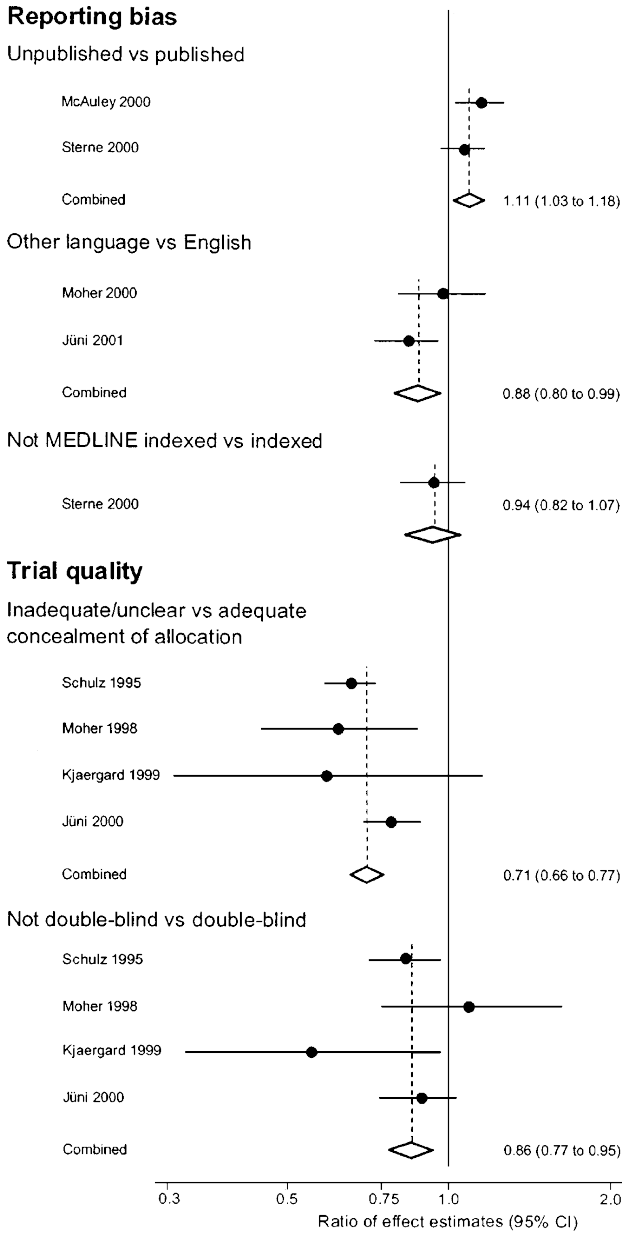


Figure 2 Meta-analysis of empirical studies of reporting bias and trial quality. All studies compared estimates of treatment effects within a large number of meta-analyses and calculated ratios of effect estimates for this purpose. A ratio of estimates below I indicates that trials with the characteristic (for example published in a language other than English) showed a more beneficial treatment effect

trials, [37] has certainly demonstrated that the pooling of trials in meta-analysis may not always be appropriate. It is therefore important to distinguish between systematic reviews and meta-analysis: it is always appropriate and desirable to systematically review a body of data, but it may sometimes be inappropriate, or even misleading, to statistically pool results from separate studies.[38] Indeed, it is our impression that reviewers often find it hard to resist the temptation of combining studies when such meta-analysis is questionable or clearly inappropriate. This point is particularly pertinent to systematic reviews of observational studies. A clear distinction should be made between meta-analysis of randomized controlled trials and meta-analysis of epidemiological studies: consider a set of trials of high methodological quality that examined the same intervention in comparable patient populations: each trial will provide an unbiased estimate of the same underlying treatment effect. The variability that is observed between the trials can confidently be attributed to random variation and meta-analysis should provide an equally unbiased estimate of the treatment effect, with an increase in the precision of this estimate. A fundamentally different situation arises in the case of epidemiological studies, for example case-control studies, cross-sectional studies or cohort studies. Due to the effects of confounding and bias, such observational studies may produce estimates of associations that deviate from the true causal effects beyond what can be attributed to chance. Combining a set of epidemiological studies will thus often provide spuriously precise, but biased, estimates of associations.[39] The thorough consideration of heterogeneity between observational study results, in particular of possible sources of confounding and bias, will generally provide more insights than the mechanistic calculation of an overall measure of effect. This is illustrated by the systematic review of epidemiological studies of homocysteine and the risk of coronary heart disease published in this issue.[11] The association was weak for cohort studies (combined odds ratio [OR]= 1.06, 95% CI: 0.99-1.13), stronger for nested case-control studies (OR= 1.23, 95% CI: 1.07-1.41) and strongest for standard case-control studies (OR = 1.70, 95% CI: 1.50-1.93), as shown in the Figure in Clarke's commentary.36 The strength of the association thus varies inversely with the strength of the study design, which surely must be taken into account when interpreting these findings.

The importance of different sources of bias will vary across different areas of epidemiological enquiry. For example, confounding and differential measurement error is a serious problem in studies of exposures that are closely linked to lifestyle, for example dietary intake of beta-carotene, but may be of considerably less relevance in genetic epidemiology.[40] Publication bias, conversely, may be a particular problem in studies of genetic factors. For example, several meta-analyses of small case-control studies found substantial associations between the angiotensin converting enzyme (ACE) insertion/deletion polymorphism and the risk of myocardial infarction.[41]A[2] When plotting the odds ratios from the 19 studies included in Agerholm-Larsen's[43] analysis against their standard error in a 'funnel plot', it is clear that the effect is large in small case-control studies but only modest in larger studies (Figure 3).[43] The name 'funnel plot' is based on the fact that effect estimates from small studies will scatter more widely at the bottom of the graph, with the spread narrowing among larger studies. In the absence of bias the plot will thus resemble a symmetrical inverted funnel. The degree of

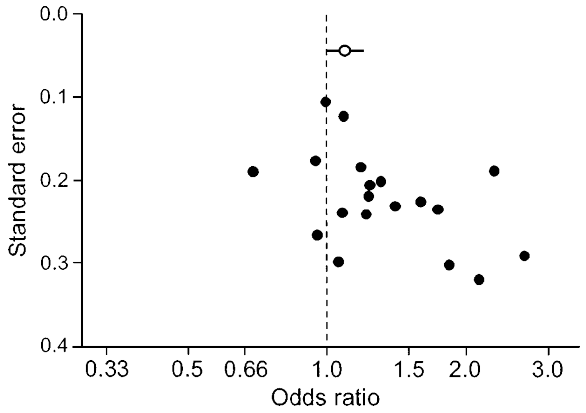


Figure 3 Funnel plot of 19 case-control studies of angiotensin converting enzyme (ACE) gene polymorphism in coronary heart disease (from Agerholm-Larsen *et* a/.[43]). Odds ratios comparing the ACE *DD* genotype with the *ID! II* genotype are plotted against their standard error. The open circle and horizontal line shows the point estimate and 95% CI of the ISIS genetic study (from Keavney *et* a!.[44])

asymmetry observed for the ACE gene polymorphism studies, which includes the studies in Whites published up to 1998, is unlikely to be due to chance (P = 0.033 by regression test[37]). Based on these findings, the results of the large ISIS genetic study,[44] which was based on 4629 myocardial infarction cases and 5934 controls and published in 2000, are hardly surprising: the estimated risk ratio was 1.10, with confidence intervals (1.00-1.21) that exclude the effects seen in the earlier meta-analyses.41,42

Where now for meta-analysis? In her review article,[10] Dickersin takes observational epidemiology to task for being far behind, and argues that methodological research is urgently required in this area. We agree, and will continue to be interested in publishing such research. The *International Journal of Epidemiology* will also participate actively in the development of reporting guidelines for epidemiological studies, similar to the Consolidated Standard for Reporting Trials (CONSORT).[45] Such guidelines are required to facilitate the assessment of the quality of epidemiological studies and will be helpful not only to systematic reviewers and meta-analysts, but also to editors of, and referees for, epidemiological journals. Dickersin's analysis of the instructions for authors on the preparation of systematic reviews made us realize that our instructions urgently need updating, and this process has now started. Finally, Dickersin is critical of journal editors who do not treat systematic reviews and meta-analyses as original research, thus depriving those who specialize in this area from a form of academic reward. We stress that at the *IJE* we do consider well-conducted systematic reviews and meta-analyses as original research and publish them as such. However, we believe that there continues to be a place for reviews that express an informed opinion, as Dickersin does in her review,[10] and (we hope) we do in this editorial.

Acknowledgements

We thank Ken Schulz and Lise Kjaergard for kindly providing unpublished data. We are grateful to the MRC Health Services Research Collaboration for funding a workshop in November

2000, which helped identify topical issues in meta-analysis. Bristol is the lead centre of the MRC Health Services Research Collaboration.

References

1 Egger M, Davey Smith G, O'Rourke K. Rationale, potentials and promise of systematic reviews. In: Egger M, Davey Smith G, Altman DG (eds). *Systematic Reviews in Health Care: Meta-Analysis in Context.* London: BMJ Books, 2001, pp.23-42.

2 Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. *Int J Epidemiol* 2002;31:72-76.

3 Song F, Khan KS, Dinnes J, Sutton AJ. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int J Epidemiol* 2002;31:88-95.

4 Higgins JPT, Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int J Epidemio/* 2002;31:96-104.

5 Vale CL, Tierney JF, Stewart LA. Effects of adjusting for censoring on meta-analyses of time-to-event outcomes. *Int J Epidemio/* 2002;31: 107-11.

6 Clark OAC, Castro AA. Searching the LILACS database improves systematic reviews. *Int J Epidemio/* 2002;31:112-14.

7 McDonald S, Westby M, Clarke M, Lefebvre C and the Cochrane Centres' Working Group on 50 Years of Randomized Trials. Number and size of randomized trials reported in general health care journals from 1948 to 1997. *Int J Epidemio/* 2002;31:125-27.

8 Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol* 2002;31:150-53.

9 Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HY, Vail A. Meta-analyses involving cross-over trials: methodological issues. *Int J Epidemiol* 2002;31:140-49.

10 Dickersin K. Systematic reviews in epidemiology: why are we so far behind? *Int J Epidemiol* 2002;31:6-12.

11 Ford ES, Smith SJ, Stroup DF, Steinberg KK, Mueller PW, Thacker SB. Homocyst(e)ine and cardiovascular disease: a systematic review of the evidence with special emphasis on case-control studies and nested case-control studies. *Int J Epidemiol* 2002;31:59-70.

12 Missmer SA, Smith-Warner SA, Spiegelman D *et al.* Meat and dairy food consumption and breast cancer: a pooled analysis of cohort studies. *Int J Epidemiol* 2002;31:78-85.

13 Fischbach LA, Goodman KJ, Feldman M, Aragaki C. Sources of variation of *Helicobacter pylori* treatment success in adults worldwide: a meta-analysis. *Int J Epidemio/* 2002;31:128-39.

14 Plackett RL. Studies in the history of probability and statistics: VII. The principle of the arithmetic mean. *Biometrika* 1958;45:130-35.

15 Pearson K. Report on certain enteric fever inoculation statistics. *Br Med J* 1904;3:1243-46.

16 Glass GV. Primary, secondary and meta-analysis of research. *Educ Res* 1976;5:3-8.

17 Cochrane AL. 1931-1971: a critical review, with particular reference to the medical profession. In: *Medicines for the Year 2000.* London: Office of Health Economics, 1979.

18 Bero L, Rennie D. The Cochrane Collaboration. Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995;274:1935-38.

19 Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis* 1985;17:335-71.

20 Crowley P. Corticosteroids prior to preterm delivery. In: Enkin MW, Keirse MJNC, Renfrew MJ, Neilson JP (eds). *Pregnancy and Childbirth Module.* Oxford: Update Software, 1994, p.2955.

21 Early Breast Cancer Trialists' Collaborative Group. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28 896 women. *N Engl J Med* 1988;319:1681-92.

22 Lewis SJ, Egger M, Sylvester PA, Thomas S. Early enteral feeding versus 'nil by mouth' after gastrointestinal surgery: systematic review and meta-analysis of controlled trials. *Br Med J* 2001;323:773-76.

23 Schwarzer G, Antes G, Tallon D, Egger M. *Review Publication Bias? Matched Comparative Study of Cochrane and Journal Meta-analyses.* BMC Meeting Abstracts: 9th International Cochrane Colloquium 2001, 1: pc142.

24 Chalmers I. Using systematic reviews and registers of ongoing trials for scientific and ethical trial design, monitoring, and reporting. In: Egger M, Davey Smith G, Altman DG (eds). *Systematic Reviews in Health Care: Meta-analysis in Context.* London: BMJ Books, 2001, pp.429-43.

25 Clarke MJ, Stewart LA. Systematic reviews of evaluations of prognostic variables. In: Egger M, Davey Smith G, Altman DG (eds). *Systematic Reviews in Health Care: Meta-Analysis in Context.* London: BMJ Books, 2001, pp.109-21.

26 Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet* 1998;35 **I:** 1451-67.

27 Thompson SG, Pocock SJ. Can meta-analysis be trusted? *Lancet* 1991; **338:**1127-30.

28 Egger M, Davey Smith G. Misleading meta-analysis. Lessons from 'an effective, safe, simple' intervention that wasn't. *Br Med J* 1995;310: 752-54.

29 Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. *Br Med J* 1997;315:617-19.

30 Sterne JAC, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002; (In press).

31 Jiini P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *Int J Epidemiol* 2002;31:115-23.

32 Clarke M. Searching for trials for systematic reviews: what difference does it make? *Int J Epidemiol* 2002;31:123-24.

33 Sterne JA, Egger M, Davey Smith G. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. *Br Med J* 2001;323:101-05.

34 Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *Br Med J* 1991;303:1499-503.

35 Collaborative Group. ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58 050 patients with suspected acute myocardial infarction. *Lancet* 1995;345:669-87.

36 Clarke R. An updated review of the published studies of homocysteine and cardiovascular disease. *Int J Epidemiol* 2002;31:70-71.

37 Egger M, Davey Smith G, Schneider M, Minder CE. Bias in meta-analysis detected by a simple, graphical test. *Br Med J* **1997;315:** 629-34.

38 O'Rourke K, Detsky AS. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. *J Clin Epidemiol* 1989;42:1021-24.

39 Egger M, Schneider M, Davey Smith G. Spurious precision? Meta-analysis of observational studies. *Br Med J* 1998;316:140-45.

40 Clayton D, McKeigue PM. Epidemiologic methods for studying genes and environmental factors in complex diseases. *Lancet* 2001;358: 1356-60.

41 Samani NJ, Thompson JR, O'Toole L, Channer K, Woods KL. A meta-analysis of the association of the deletion allele of the angiotensin-converting enzyme gene with myocardial infarction. *Circulation* 1996; 94:708-12.

42 Staessen JA, Wang JG, Ginocchio G *et al.* The deletion/insertion polymorphism of the angiotensin converting enzyme gene and cardiovascular-renal risk. *J Hyper/ens* 1997;15:1579-92.

43 Agerholm-Larsen B, Nordestgaard BG, Tybjaerg-Hansen A. ACE gene polymorphism in cardiovascular disease: meta-analyses of small and large studies in whites. *Arterioscler Thromb Vase Biol* 2000;20:484-92.

44 Keavney B, McKenzie C, Parish S *et al.* Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls. International Studies of Infarct Survival (ISIS) Collaborators. *Lancet* 2000;355:434-42.

45 Altman DG, Schulz KF, Moher D *et al.* The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94.

Psycoological Methods
199 . Vol. 3, :,lo 4, 486-504

1082-989X/98/$3.00

Fixed- and Random-Effects Models in Meta-Analysis

Larry V. Hedges
University of Chicago

Jack L. Vevea
University of North Carolina at Chapel Hill

There are 2 families of statistical procedures in meta-analysis: fixed- and random-effects procedures. They were developed for somewhat different inference goals: making inferences about the effect parameters in the studies that have been observed versus making inferences about the distribution of effect parameters in a population of studies from a random sample of studies. The authors evaluate the performance of confidence intervals and hypothesis tests when each type of statistical procedure is used for each type of inference and confirm that each procedure is best for making the kind of inference for which it was designed. Conditionally random-effects procedures (a hybrid type) are shown to have properties in between those of fixed- and random-effects procedures.

The use of quantitative methods to summarize the result<; of several empirical research studies, or *meta-analysis,* is now widely used in psychology, medicine, and the social sciences. Meta-analysis usually involves describing the results of each study by means of a numerical index (an estimate of effect size, such as a correlation coefficient, a standardized mean difference, or an odds ratio) and then combining these estimates across studies to obtain a summary. Two somewhat different statistical models have been developed for inference about average effect size from a collection of studies, called the *fixed-effects* and *random-effects* models. (A third alternative, the *mixed-effects* model, arises in conjunction with analyses involving study-level covariates or moderator variables, which we do not consider in this article; see Hedges, 1992.)

Fixed-effects models treat the effect-size parameters as fixed but unknown constants to be estimated and usually (but not necessarily) are used in conjunction with assumptions about the homogeneity of effect parameters (see, e.g., Hedges, 1982; Rosenthal & Rubin, 1982). Random-effects models treat the effect-size parameters as if they were a random sample from a population of effect parameters and estimate hyperparameters (usually just the mean and variance) describing this population of effect parameters (see, e.g., DerSimonian & Laird, 1986; Hedges, 1983; Schmidt & Hunter, 1977). Although inference procedures based on these models have been available for well over a decade, there is still considerable confusion about the differences between them.

The purpose of this article is to clarify the conceptual distinctions between the models themselves and provide analytic information about the properties of statistical procedures (significance tests and confidence intervals) based on them. First, we discuss fixed- and random-effects models for meta-analysis in conjunction with fixed- and random-effects models in the more familiar context of analysis of variance (ANOVA), emphasizing that choice of model depends on the inferences the analyst wishes to make. Next, we develop the sampling properties of fixed- and random-effects procedures in detail, followed by those of a hybrid, which we call *conditionally random-effects* procedures. Then we compare the rejection rates of fixed-, random-, and conditionally random-effects tests analytically and address the adequacy of confidence intervals developed by means of these three procedures. Finally, we offer suggestions about applications of these methods in meta-analysis.

Fixed-Effects Versus Random-Effects Models in Meta-Analysis

There has been a great deal of confusion about the difference between fixed- and random-effects models

Larry V. Hedges, Departments of Education. Psychology, and Sociology, University of Chicago; Jack L. Vevea, Department of Psychology, University of North Carolina at Chapel Hill.

Correspondence concerning this article should be addressed to Larry V. Hedges, Department of Education, University of Chicago, Chicago, Illinois 60637. Electronic mail may be sent to hedge@cicero.spc.uchicago.edu.

in meta-analysis. Because this is part of a larger issue of conditional versus unconditional analyses in statistics, differences of opinion about the appropriateness of these analyses are likely to persist for some time (see Camilli, 1990, for a discussion of the conditionality issue in another context). However, we attempt here to clarify some of the issues.

The choice between fixed- and random-effects procedures has sometimes been framed as entirely a question of homogeneity of the effect-size parameters. That is, if all of the studies estimate a common effect-size parameter, then fixed-effects analyses are appropriate. However, if there is evidence of heterogeneity among the population effects estimated by the various studies, then random-effects procedures should be used. Although it is true that fixed- and random-effects analyses give similar answers when there is in fact a common population effect size, the inference models remain distinct. Moreover, there may be situations in which the fixed-effects analysis is appropriate even when there is substantial heterogeneity of results (e.g., when the question is specifically about a particular set of studies that have already been conducted).

We argue that the most important issue in determining statistical procedure should be the nature of the inference desired. If the analyst wishes to make inferences only about the effect-size parameters in the set of studies that are observed (or to a set of studies identical to the observed studies except for uncertainty associated with the sampling of participants into those studies), this is what we call a *conditional* inference. One might say that conditional inferences about the observed effect sizes are intended to be robust to the consequences of sampling error associated with sampling of participants (from the same populations) into studies. Strictly speaking, conditional inferences apply to *this* collection of studies and say nothing about other studies that may be done later, could have been done earlier, or may have already been done but are not included among the observed studies. Fixed-effects analysis procedures are appropriate for making conditional inferences.

In contrast, the analyst may wish to make a different kind of inference, one that embodies an explicit generalization beyond the observed studies. In this case, the studies that are observed are not the only studies that might be of interest. Indeed, one might say that the studies observed are of interest only because they reveal something about a putative population of studies that are the real object of inference. If the analyst wishes to make inferences about the parameters of a population of studies that is larger than the set of observed studies and that may not be strictly identical to them, we call this an *unconditional* inference. Random-effects analysis procedures are designed to facilitate unconditional inferences.

An Analogy to the ANOVA

To understand the fixed- and random-effects models in meta-analysis, it is helpful to place the problem in a context that is more familiar to many researchers: the **ANOV**A Consider meta-analyses for which the data from different studies are directly comparable, so that the raw data from all the studies can be analyzed together. That is, assume that each study compares a treatment group with a control group and that all studies measure the outcome on the same scale and there is homogeneity of within-group variance across studies. This is not a case that arises in practice very often (because studies tend to use different outcome measures and have different sampling plans that lead to different variances), but it illustrates the ideas in a form that is easy to understand. We lose no generality in assuming that the common within-group variance is I, which means that the raw mean difference in each study is an estimate of the standardized mean difference (Glass's effect size). It can be shown that the application of meta-analysis and the ANOVA yield the same results when applied to such a situation (Olkin & Sampson, 1998).

In this case, the data layout is a 2 (treatments) x k (studies) design, and we can apply ordinary two-factor ANOVA to analyze the data. The treatment factor is fixed, and the main effect of treatment corresponds to the average effect size (it is the weighted average of the study-specific treatment contrasts across studies). The Treatment x Studies interaction describes how much variation there is across studies in the study-specific treatment effects. The test of this interaction is a test that the study-specific treatment effect parameters are identical across studies.

Different tests for the fixed effect of treatment are appropriate depending on whether the studies factor is considered to be fixed or random. If the studies factor is treated as fixed, then the design is a fixed-effects design, and there is only one source of uncertainty, the within-group sampling error£, and only one true variance component, the error variance cr_E^2. The appropriate test uses only within-group variation (the mean square within groups) as the error term for testing the main effect of treatment. This corresponds to the in-

tuition that the only source of uncertainty in inferences about the means of the observed studies is the sampling of participants into the studies. In this case, the inference is about the treatment-effect parameters in the particular studies included in the design.

If the studies factor is considered to be random, there are three sources of uncertainty: the within-group sampling error e, the (random) effect of study . and the Treatment x Study interaction a . Each of these sources of uncertainty has a corresponding variance component: Cl/, Cl_{1}/, and CiaJ/. The fixed effects test for the main effect of treatment is incorrect; the correct error term is the Treatment x Studies interaction. This error tenn includes two components of variance: one due to within-study (within-group) sampling error c_{I}.^{2} and one due to the Study x Treatment interaction $\hat{c}_{I}$Of course, the reason that this is the appropriate error tenn is that variance of the treatment contrast (and its a<;sociated mean square) also includes a component of variance due to the Treatment x Studies interaction eral/. In this case, the inference about treatment-effect parameters is to the mean effect in a population of studies (a population of potential levels of the studies factor) from which the observed studies are a random sample.

The test for the Treatment x Studies interaction is the same in the two models, albeit with a slightly different interpretation. In the studies-fixed model, the test is that all the treatment-effect parameters in the observed studies are equal. In the studies-random model, the test is that the Treatment x Studies interaction variance component $cr'''_{1}{}^{2}$ is zero. This variance component describes the variance across all studies (in the putative population of studies) of the study-specific treatment-effect contrasts.

Inference to Other Studies

Fixed-effects models. In conditional (fixed-effects) models, inferences are, in the strictest sense, limited to the factor levels represented in the sample. Of course, conditional models are widely used in primary research, and the generalizations made from them by researchers are typically not constrained precisely to factor levels in the study. For example, generalizations about treatment effects in fixed-effects ANOV A are usually not constrained to apply only to the precise levels of treatment found in the experiment but are typically viewed as applying to similar treatments as well, even if they were not explicitly part of the experiment.

How are such inferences justified? Typically, they are justified on the basis of an a priori (extraempirical) decision that other levels (other treatments) are sufficiently like those in the sample that their behavior will be identical. The key point is that generalization to levels not present in the sample requires an assumption that the levels are similar to those in the sample-one not justified by a fonnal sampling argument.

Inference to studies not identical to those in the sample can be justified in meta-analysis by the same intellectual devices used to justify the corresponding inferences in primary research. Specifically, inferences may be justified if the studies are judged a priori to be sufficiently similar to those in the study sample. Note, however, that the inference process has two distinct parts. One part is the generalization from the study sample to a universe of identical studies, which is supported by a sampling theory rationale. The second part is the generalization from the universe of studies that are identical to the sample to a universe of sufficiently similar (but nonidentical) studies. This second part of the generalization is supported not by a sampling argument, but by an extrastatistical one.

Random-effects models. In random-effects models, inferences are not limited to studies represented in the sample. Instead, the inferences, for example, about the mean or variance of effect-size parameters, apply to the universe of studies from which the study sample was obtained. In effect, the warrant for generalization to other studies is through a classical sampling argument. Because the universe contains studies that differ in their characteristics and those differences find their way into the study sample by the process of random sampling, generalizations to the universe pertain to studies that are not identical to those in the study sample.

By using a sampling model of generalization, the random-effects model seems to avoid subjective difficulties that plague the fixed-effects model in generalizations to studies not identical to the study sample. That is, one does not have to ask, "How similar is similar enough?" Instead another question, "Is this new study part of the universe from which the study sample was obtained?" can be substituted. If study samples were obtained from well-defined sampling frames through overtly specified sampling schemes, this might be an easy question to answer. That, however, is virtually never the case **in** meta-analysis (and is unusual in other applications of random-effects models). The universe is usually rather ambiguously specified, and consequently, the ambiguity in generalization based on random-effects models is that it is

difficult to know precisely what the universe is. In contrast, the universe is clear in fixed-effects models, but the ambiguity arises in deciding if a new study might be similar enough to the studies already contained in the study sample.

The random-effects model does provide the technical means to address an important problem that is not handled in the fixed-effects model, namely, the additional uncertainty introduced by the inference to studies that are not identical (except for the sample of people involved) to those in the study sample. Inference to (nonsampled) studies in the fixed-effects model occurs outside of the technical sampling theory framework, and hence, any uncertainty it contributes cannot be evaluated by technical means within the model. In contrast, the random-effects model does incorporate between-study variation into the sampling uncertainty used to compute tests and estimates.

Although the random-effects model has the advantage of incorporating inferences to a universe of studies exhibiting variation in their characteristics, the definition of the universe may be ambiguous. A tautological universe definition could be derived by using the sample of studies to define the universe as "a universe from which the study sample is representative." Such a population definition remains ambiguous; moreover, it may not be the universe definition desired for the use of the information produced by the synthesis. For example, if the study sample includes many studies of short-duration, high-intensity treatments but the likely practical applications usually involve low-intensity, long-duration treatments, the universe defined implicitly by the study sample may not be the universe most relevant to applications.

One potential solution to this problem might be to explicitly define a structured universe in terms of study characteristics and to consider the study sample as a stratified sample from this universe. Estimates of parameters describing this universe could be obtained by weighting each stratum appropriately. For example, if one half of the studies in the universe to which one wishes to generalize are long-duration studies but only one third of the study sample has this characteristic, the results of each long-duration study must be weighted twice as much as the short-duration studies.

Statistical Inference in Meta-Analysis

In this article, we assume that there are effect-size estimates from k independent studies. Denote the population effect size (effect-size parameter) in the ith study by θ_i and its estimate (the sample effect-size estimate) by T_i.

Study	Effect-size parameter	Effect-size estimate	Conditional variance of T given θ
1	θ_1	T_1	v_1
2	θ_2	T_2	v_2
k			

We assume that the T_i is normally distributed about the corresponding θ; with known variance v_i. That is, we assume that

$$T_i \sim N(\theta_i, v_i) \quad i = 1, \ldots, k. \tag{1}$$

This assumption is very nearly exactly true for effect sizes such as the Fisher z-transformed correlation coefficient and standardized mean differences transformed by the Hedges-Olkin variance-stabilizing transformation (Hedges & Olkin, 1983). However, for effect sizes such as the untransfonned standardized mean difference, correlation coefficient, or the log-odds ratio, the results are not exact but remain true as large-sample approximations.

Statistical Inference in Fixed-Effects Meta-Analysis

If a series of k studies can reasonably be expected to share a common effect size θ, or if we are interested in the mean of the effect sizes in the series of studies, it is natural to estimate θ by pooling estimates from each of the studies. If the sample sizes of the studies differ, then the estimates from the larger studies will usua11y be more precise than the estimates from the sma11er studies. In this case, it is reasonable to give more weight to the more precise estimates when pooling. This leads to weighted estimators, and the weights that minimize the variance give weight inversely proportional to the variance in each study. This is intuitively clear in that smaller variance (i.e., more precision) should lead to a larger weight. The optimal weights are given by

$$w_i = 1/v_i. \tag{2}$$

Thus, the weighted mean that minimizes the variance can be written as

$$T_{\cdot} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} \quad (3)$$

Note that f. is also the maximum-likelihood estimator of 0 under this model.

The sampling variance *v. of* $T_{\cdot}$ is simply the reciprocal of the sum of the weights, namely,

$$v_{\cdot} = \frac{1}{\sum_{i=1}^{k} w_i}, \quad (4)$$

and the standard error SE(T.) off. isjust the square root of *v.,* that is, SE(T.) = $\sqrt{v_{\cdot}}$ Because $T_1, \ldots, T_k$ is normally distributed, it follows that f. also is normally distributed.

Tests and confidence intervals for the mean. If $T_1, \ldots, T_k$ estimates the s e underlying effect size $\theta_1 = \ldots = \theta_k = \theta$, then $T_{\cdot}$ estimates 0 and a 100(1 - ex)% confidence interval for e is given by

$$L = T_{\cdot} - z_{\alpha} \sqrt{v_{\cdot}} \quad \text{and} \quad U = T_{\cdot} + z_{\alpha} \sqrt{v_{\cdot}} \quad (5)$$

where z_{002} is the two-tailed critical value of the standard normal distribution and *v.* is the sampling variance of f. given by Equation 4.

A 100a% significance test of the nu11 hypothesis that $\theta = \theta_0$ could be obtained by means of the statistic

$$Z = (T_{\cdot} - \theta_0)/\sqrt{v_{\cdot}}, \quad (6)$$

which has the standard normal distribution when $\theta = \theta_0$. The one-sided test rejects the null hypothesis whenever $Z > z'''$.

Testing for heterogeneity of effect-size parameters. Before pooling the estimates of effect size from a series of *k* studies, it is important to determine whether the studies can reasonably be described as sharing a common effect size. A statistical test for the homogeneity of population effect sizes is formally a test of the hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ versus the alternative that at least one of the effect sizes 0; differs from the remainder.

An exact small-sample test of $\mathbf{H_0}$ {which is also the likelihood ratio test of this hypothesis) is based on the statistic

$$Q = \sum_{i=1}^{k} w_i (T_i - \overline{T}_{\cdot})^2, \quad (7)$$

where f. is the weighted estimator of effect size given in Equation 3. The test statistic *Q* is the sum of squares of the *Ti* about the weighted mean f. where the ith square is weighted by the reciprocal of the variance of *T;.* Because $w_i = 1/v_i$ and $(T_i - T_{\cdot})^2$ can be seen as a (crude) estimate of between-study variation, each term of *Q* can also be interpreted as a ratio of between-study to within-study variances, meaning that *Q* can be interpreted as a comparison of between- to within-study variance.

If all *k* studies have the same population effect size (i.e., if H_0 is true), then the test statistic *Q* has a chi-square distribution with *k* - 1 degrees of freedom. Therefore, if the obtained value of *Q* exceeds the 100(1 - *a)%* critical value of the chi-square distribution with *k* - 1 degrees of freedom, we reject the hypothesis that the ei are equal.

Example. The results of 14 studies of gender differences in field articulation ability were reported by Hyde (1981), who reported the effect-size estimate (as a standardized mean difference) and total sample size for each study. The data are listed in Table 1, where, for each study, column 2 gives the unbiased estimate of effect size corresponding to Hyde's standardized mean difference. To carry out the fixed-effects analysis, first compute the sampling variance *v;* for each study. Hyde reported the total sample size for each study but not the sample sizes of each group. Consequently, we compute the variance here from the formula, on the basis of the assumption that the sample sizes of the two groups within a study are (approximately) equal (see Hedges, 1981), namely, $v_i = 4(1 + d/18)/N_i$, where N_i is the total sample size in the *ith* study.

The fixed-effects analysis depends on the sums of three variables: the weights $(w_i = 1/v)$, the weights x the effect sizes (w;d;), and the weights x the effect sizes squared (w;d/). The analysis can be carried out by using a packaged computer program (e.g., **SAS** 6.2, 1996, or **SPSS,** 8.0, 1998) or a spreadsheet to compute these three variables and their sums. Column 3 of Table 1 gives the value of v_i for each study, and columns 4, 5, and 6 give the values of w_i, $w_i d_i$, and $w_i d_i^2$, along with their sums at the bottom of the columns.

The fixed-effects weighted mean effect size given in Equation 3 is just f. == (l 18.050)/(216.700) = 0.545, and its variance, given in Equation 4, is just *v.* = 1/216.700 == .004615, which yields a standard error of $SE(T_{\cdot}) = \sqrt{.004615} = .068$. The 95% confidence interval for 0, given by Equation 5 is

$$0.412 = 0.545 - 1.96(.068) \quad 0 \quad 0.545 + 1.96(.068) = 0.678.$$

Table 1
Effect-Size Data From 14 Studies of Gender Difference in Field Articulation

Study	*N*	*d*	*v*	*w*	*wd*	*wtP-*	w2	*w**	*w*d*
1	60	0.76	.071	13.990	10.632	8.081	195.718	7.783	5.915
2	140	1.15	.033	30.035	34.540	39.721	902.093	11.075	12.736
3	30	0.48	.137	7.290	3.499	1.680	53.145	5.150	2.472
4	30	0.29	.135	7.422	2.152	0.624	55.086	5.216	1.513
5	30	0.65	.140	7.124	4.630	3.010	50.748	5.066	3.293
6	46	0.84	.095	10.568	8.877	7.457	111.681	6.595	5.540
7	40	0.70	.106	9.423	6.596	4.617	88.790	6.130	4.291
8	34	0.50	.121	8.242	4.121	2.061	67.938	5.608	2.804
9	76	0.18	.053	18.923	3.406	0.613	358.094	9.104	1.639
IO	163	0.17	.025	40.603	6.903	1.173	1,648.630	12.251	2.083
11	97	0.77	.044	22.577	17.384	13.386	509.711	9.872	7.602
12	44	0.27	.092	10.901	2.943	0.795	118.825	6.723	1.815
13	78	0.40	.052	19.118	7.647	3.059	365.484	9.148	3.659
14	43	0.45	.095	10.485	4.718	2.123	109.927	6.563	2.953
Total				216.700	118.050	88.399	4,635.868	106.285	58.315

Note. These data are from Hyde (1981). *v* = conditional variance; *w* = weight, fixed effects model; *w** = weight, random-effects model.

The test of homogeneity of effect sizes is computed from Equation 7 as

$$Q = 88.399 - (118.050)21(216.700) = 24.090,$$

and because this value exceeds 22.36, the 95% critical value of the chi-square distribution with (14 - 1) = 13 degrees of freedom, we reject the hypothesis that the effect-size parameters are the same in all of the studies.

Statistical Inference in Random-Effects Meta-Analysis

In this section, we describe procedures for estimating the mean μ of the effect-size distribution underlying the results of a series of studies using an analysis based on the random-effects model. There are obvious similarities between estimating a common underlying effect size by talcing the mean of the estimates and estimating the mean of the effect-size distribution. In both procedures, the pooled estimate is usually computed by taking the weighted mean across studies of the sample effect-size estimates, and it is not unusual for either of these estimates to be called the *average* effect size. However, note that the quantity to be estimated (the mean of the effect-size distribution) in random-effects models does not have exactly the same interpretation as the one to be estimated (the single or average underlying effect size) in fixed-effects models. In the case of random-effects models, for example, some individual effect-size parameters may be negative even though μ is positive. That corresponds to the substantive idea that some realizations of the treatment may actually be harmful even if the average effect of the treatment μ, is beneficial.

The variance of estimates of effect size. In the fixed-effects model, the effect sizes 0; are fixed, but unknown, constants. Under this assumption, the variance of *Ti* is simply *vi.* In the random-effects model, the 0; are not fixed but are themselves treated as random and have a distribution of their own. Therefore, it is necessary to distinguish between the variance of *T;* assuming a fixed 0i and the variance of *T;* incorporating the variance of 0 as well. The former is the *conditional* sampling variance of *T;,* and the latter is the *unconditional* sampling variance of *Ti.*

It is convenient to decompose the observed effect-size estimate into fixed and random components

$$Ti = 6; + E.; = \mu + \{;; + Ei, \qquad (8)$$

where *E.;* is a sampling error of *T;* as an estimate of 0;, and 0i can itself be decomposed into the meanμ of the population from which the 0s are sampled and the error t of 0i as an estimate ofμ. In this decomposition, only μ is fixed, and one can assume both t;and the e,.s are random, with expected value zero. The variance of E; is *v;,* the conditional sampling variance of *T;,* which is known. The variance of the population from which $t_{1, \ldots, k}$ are sampled is T^2. Equivalently, one might say that $,-^2$ is the variance of the population from which the study-specific effect parameters $0_{1, \ldots, 6k}$ are sampled. Frequently $,-^2$ is called the between-study variance component.

Because the effect size 6; is a value obtained from a distribution of potential 0; values, the unconditional sampling variance of T_i involves σ^2. A direct argument shows that this sampling variance is

$$v_i^* = v_i + \sigma^2. \quad (9)$$

Methods of estimation for random-effects models have been suggested in different meta-analytic contexts by DerSimonian and Laird (1986), Hedges (1983), and Schmidt and Hunter (1977). They use the method of moments to estimate the between-study variance component and are analogous to the methods often used to estimate variance components in ANOVAs of balanced designs.

Estimating the between-studies variance component. Estimation of the between-studies variance component τ^2 uses the same principles as estimation of the variance components in the ANOVA. One estimate of σ^2 is

$$\sigma^2 = \begin{cases} \dfrac{Q-(k-1)}{c} & \text{if } Q \geq k-1 \\ 0 & \text{if } Q < k-1 \end{cases} \quad (10)$$

where c is given by

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \quad (11)$$

and w_i are the weights given in Equation 2 used in the fixed-effects analysis. Estimates of τ^2 are set to 0 when $Q - (k - 1)$ yields a negative value, because τ^2, by definition, cannot be negative.

If the within-study sampling-error variances $v_1, \ldots, v_k$ used to construct the weights w_i are known exactly and the estimate is not truncated at 0, then thew; are constants, and the estimate is unbiased, a result that does not depend on assumptions about the distribution of the random effects (or the conditional distribution of the effect sizes themselves). Inaccuracies in the estimation of the v_i (and hence thew;) may lead to biases, although they are usually not substantial. The truncation of the estimate at zero is a more serious source of bias, although it improves the accuracy (reduces its mean square error about the true τ^2) of estimates of σ^2. This bias can be substantial when k is small but decreases rapidly when k becomes larger. Table 2 gives the bias of $\hat{\sigma}^2$ when $v_1 = \ldots = v_k = v$ for values of k and τ^2. The table shows that the absolute bias of σ^2 can be as much as .2 to .3 for $k = 3$ and $\sigma^2 = .33v$, leading to relative biases that are well over 50%. This result underscores the fact that estimates of $\hat{\sigma}^2$ computed from only a few studies should be treated with caution. For $k > 20$, the biases are much smaller, and relative biases are only a few percent.

Note that other methods of estimation of variance components are available. For example, maximum-likelihood estimation under either a restricted model (first estimating the mean and then estimating the variance component conditional on the estimate of the mean) or unrestricted model (estimating the mean and the between-studies variance component simultaneously) is used in other problems and can be used in meta-analysis (see, e.g., Raudenbush & Bryk, 1985). ln the case of equal v_i, which we examine, the method of moments is identical to restricted maximum-likelihood estimation. These methods are iterative and therefore more complex to implement in general, usually requiring specialized computer programs.

Testing the significance of the effect-size variance component. The test that $\tau^2 = 0$ in the random-effects model is the same as the test of homogeneity in the fixed-effects model, using the Q statistic. The reason is that if $\sigma^2 = 0$, then $\theta_1 = \theta_2 = \ldots = \theta_k = \mu$; thus, the effect-size parameters are fixed, but unknown, constants. This is analogous to the situation with the F tests in the one-way random- and fixed-effects ANOVAs. In the ANOVA, the null distributions of the test statistics are identical, but the nonnull distributions of the F ratios differ. Similarly, although the null distributions of the Q statistics are identical in fixed- and random-effect-size models, the nonnull distributions of Q differ under the two models.

Estimating the mean effect size. The logic of using weighting is the same in random-effects procedures as it is in fixed-effects procedures, but the choice of weights differs somewhat because random-effects models include in their definition of *variance* a component of variance σ^2 associated with between-study differences in effect parameters, which fixed-effects models do not. That is, the total variance v_i^* for the ith effect-size estimate T_i is defined by $v_i^* = \tau^2 + v_i$. Because the additional component of variance is the same for all studies, it both increases the total variance of each effect size estimate and tends to make the total variances of the studies (the v_i^*) more equal than the sampling-error variances (the v_i).

Because the true value of τ^2 is rarely known, we usually substitute an estimate of this variance compo-

Table 2
Bias E(f2) - ,-2 and Relative Bias [E(f²) - -r]h² in the Estimator ,f of i-2 *Based on k Studies*

	Bias			Relative Bias			
k	$\tau^2 = 0$	$\tau^2 = .33$	$\tau^2 = .67$	$\tau^2 = 1.00$	$\tau^2 = .33$	$\tau^2 = .67$	$\tau^2 = 1.00$
2	0.484	0.429	0.389	0.358	1.286	0.583	0.358
3	0.368	0.296	0.248	0.213	0.889	0.372	0.213
4	0.308	0.229	0.179	0.144	0.688	0.268	0.144
5	0.271	0.187	0.137	0.104	0.562	0.205	0.104
6	0.244	0.158	0.108	0.078	0.474	0.163	0.078
7	0.224	0.137	0.088	0.060	0.410	0.132	0.060
8	0.208	0.120	0.073	0.047	0.360	0.110	0.047
9	0.195	0.106	0.061	0.038	0.319	0.092	0.038
10	0.185	0.095	0.052	0.030	0.286	0.078	0.030
20	0.128	0.043	0.014	0.005	0.128	0.022	0.005
30	0.104	0.024	0.005	0.001	0.072	0.008	0.001
40	0.090	0.015	0.002	0.000	0.044	0.003	0.000
50	0.080	0.010	0.001	0.000	0.029	0.001	0.000
60	0.073	0.007	0.000	0.000	0.020	0.001	0.000
70	0.068	0.005	0.000	0.000	0.014	0.000	0.000
80	0.063	0.003	0.000	0.000	0.010	0.000	0.000
90	0.060	0.002	0.000	0.000	0.007	0.000	0.000
100	0.057	0.002	0.000	0.000	0.005	0.000	0.000

nent such as that given in Equation IO into Equation 9 in place of τ^2 (see, e.g., DerSimonian & Laird, 1986; Hedges, 1983; Hedges & Olkin, 1985). This yields

$$T_.^* = \frac{\sum_{i=1}^{k} w_i^* T_i^*}{\sum_{i,c,J}^{k} w_i^*} \qquad (12)$$

where the weight w_i^* is an estimated optimal weight that is th reciprocal of an estimate of the total variance of T_i^* given by

$$w_i^* = 1/(v_i^*) = 1/(v_i + f^2). \qquad (13)$$

Here, we use the * to distinguish the weights, means, and variances in the random-effects procedure from the corresponding quantities in the fixed-effects procedure.

The sampling variance $v_.^*$ of the random-effects estimate (of the mean of the effect-size distribution) $T_.^*$ is given by the reciprocal of the sum of the ran- dom-effects weights, that is,

$$v_.^* = \frac{1}{\sum_{r-1} w_i^*}. \qquad (14)$$

The standard error $SE(T_.^*)$ of the mean effect estimate $T_.^*$ is just the square root of its sampling variance, that is, $SE(T_.^*) = v_.^*$. Note that whenever the between-studies variance component (estimate) $f^2 > 0$, the standard error $v_.^*$ of the mean estimated using the random-effects procedure will be larger than $v_.$, the standard error of the mean estimated using the fixed-effects procedure. If $f^2 = 0$, the standard errors (and the mean estimates) of the random- and fixed-effects procedures will be identical.

Note that some writers who advocate random-effects procedures (e.g., Hunter & Schmidt, 1990, p. 147) advocate the use of suboptimal weights that correspond to the fixed-effects weights, presumably because they assume that τ^2 is small. This will not make a great deal of difference if τ^2 is indeed small, but assuming that $\tau^2 = 0$ when it is not will lead to an underestimate of the variance $v_.^*$.

Although it is possible to obtain very precise estimates of 0; if the sample sizes are reasonably large in each study, the precision of the estimate of τ^2 depends primarily on the number k of studies. Therefore, if the number of studies is small, the estimates of the weights may be fairly imprecise even though the sample size in each study is quite large.

Tests and confidence intervals for the mean. 1f the random effects are approximately normally distributed, the weighted mean $T_.^*$ is approximately normally distributed about the mean effect-size param-

eter μ that it estimates. As in the fixed-effects case, the fact that this mean is normally distributed with the variance given in Equation 14 leads to straightforward procedures for constructing tests and confidence intervals. An approximate 100(1 - ex)% confidence interval for the mean effect μ is given by

$$L^* = \bar{T}_{,}^* - z0/_{.12}\ v,^*,;;\ \mu.,;\ \bar{T}\text{-}^* + za1_2{}^{\mathrm{f}},v.^* = U^*, \quad (15)$$

where *za12* is the two-tailed critical value of the standard normal distribution (e.g., zo.12 = 1.96 for u = .05 and 95% confidence intervals) and $v.^*$ is the variance of $T.^*$ given in Equation 14.

Significance tests corresponding to the confidence intervals also can be constructed. An approximate test of whether the mean effect μ differs from a predefined constant l.Lo (e.g., to test if μ - μ.0 = 0) by testing the null hypothesis H0: μ = μ0, uses the statistic

$$Z^* = \frac{T^* - \mu_Q}{(v.^*)11.2}$$

The one-sided test consists of rejecting H_0 at level a: (that is, decide that the effect parameter differs from μ0) if the value of Z* exceeds the *100a%* critical value of the standard normal distribution. That is, reject H_0 if $V' > Za\_$• For example, for a one-sided test that μ = 0 at a: = .05 level of significance, reject the null hypothesis if the value of Z* exceeds 1.645.

Example. Return to the studies of gender differences in field articulation ability discussed above and whose results are reported in Table l. The random-effects analysis depends on the sums of six variables: the fixed-effects weights (w; = *1/v;),* the fixed-effects weights x the effect sizes (wid;), the fixed-effects weights x the effect sizes squared *(wid/),* the fixed-effects weights squared (w/), the random-effects weights (w;*), and the random-effects weights x the effect sizes *(w;*dJ.* The sums of the first four of these variables are needed to compute the variance component estimate, which in tum is used to compute the random-effects weights in the last two variables. The analysis can be carried out by using a packaged computer program (e.g., SAS 6.12, 1996, or **SPSS** 8.0, 1998) or a spreadsheet, to compute these six variables and their sums. In Table 1, columns 4-9 give the values of *w;, w;d;,* w;d/, w/, *w;*,* and w;*d;, for each study, along with their sums at the bottom of the columns.

The first step in the random-effects analysis is to compute the between-studies variance component estimate. The homogeneity test statistic *Q* was computed in the fixed-effects analysis as Q = 24.090. The constant *c* given in Equation l l is computed from the sums of *w;* and *W/* as

$$c = 216.700 - (4{,}635.868)/(216.700) = 195.307,$$

and the variance component estimate itself is computed from Equation 10 as

$$,.^2 = [24.090 - (14 - 1)]/195.307 = 0.057.$$

Given that $f^2 = 0.057$, the random-effects weights are computed, using Equation 13, as $w;^* = 1/(v; + 0.057)$, Note that in a spreadsheet, this computation can easily be automated, but in a computer program package such as SAS 6.12, 1996, or SPSS 8.0, 1998, the random-effects analysis will require two passes through the data: one to obtain the sums necessary to compute f^2 and a second pass to compute the sum of the *w,** and *w;*d,.*

The random-effects weighted mean effect size given in Equation 12 is just $T.^* = (58.315)/(106.285) = 0.549$, and its variance, given in Equation 14, is just $v.^* = 1/106.285 = 0.009408$, which yields a standard error of $SE(T.^*) = \sqrt{(0.009408)} = .097$. The 95% confidence interval for μ, given by Equation 15, is

$$0.359 = 0.549 - 1.96(.097)\ ,;;\ \mu\ ,;;.\ 0.549 + 1.96(.097) = 0.739.$$

Comparing the results of the random-effects analysis with the fixed-effects analysis of the same data given in the previous example reveals that the weighted means computed in the two analyses are almost identical ($T.^* = 0.549$ vs. f. = .545), but the standard error computed in the random-effects analysis is substantially larger than that in the fixed-effects analysis, $SE(T.^*) = 0.097$ versus $SE(T.) = 0.068$. The reason for the difference in standard errors is the substantial between-study heterogeneity in the effect sizes. The between-study variance component estimate (0.057) is about two thirds as large as the average (0.086) of the sampling error variances. Consequently, when this substantial additional component of variance is included as part of the sampling uncertainty of the mean in the random-effects model, the standard error of the mean sharply increases.

Conditional Choice of Random-Effects Procedures

We have treated the choice between fixed- and random-effects procedures as an a priori one based primarily on conceptual issues, not on the outcomes of the analysis. However, the emphasis on homogeneity

as the criterion for choosing between fixed- and random-effects procedures has led some to use the statistical test of heterogeneity as the sole criterion for choice among statistical procedures. That is, a preliminary test is conducted to determine whether $\tau^2 > 0$; if the null hypothesis that $,-^2 = 0$ is rejected, a random-effects procedure is used, and if the hypothesis $\tau^2 = 0$ is not rejected, then a fixed-effects procedure is used. The properties of such a conditional procedure were first analyzed by Chang (1992), who investigated them by means of simulation. We call this procedure the *conditionally random-effects* procedure, to emphasize that the choice of random (vs. fixed) effects is conditional on the outcome of the test that $,-2 > 0$. This procedure has appeal because it is simple and offers an alternative to either the fixed- or random-effects procedures. It is also equivalent to the practice of the many meta-analysts who use the test for homogeneity of effect size, to determine whether effects differ across studies, and then use random-effects analysis procedures if homogeneity is rejected, and fixed-effects analysis procedures if homogeneity is not rejected.

Comparing Inferenee Procedures

To compare the performance of inference procedures, it is essential to clarify the inferences being drawn by them. Specifically, we must clarify the difference between *conditional* inferences drawn about the mean of the specific set of effect parameters in the set of studies analyzed and *unconditional* inferences about the mean of the population of effect parameters from which the observed study parameters are a random sample.

It is important to distinguish between the underlying statistical *model,* which is determined by the inferences desired, and the statistical *procedure* used, which is a choice made by the analyst to accomplish the inferential purpose. If the analyst chooses to make conditional inferences (by conditioning on the studies in the data set), the statistical model has been determined because the effect parameters are treated as fixed for inference. If the analyst chooses to make unconditional inferences, the statistical model treats the effect sizes as a sample (even if no real sampling has been done), and thus, they are treated as random effects.

We analyze the performance of fixed, random, and conditionally random statistical procedures under both conditional and unconditional inference models. First, we examine the rejection rates of tests on means and the probability content of confidence intervals when conditional inferences are desired. Then we consider the rejection rates of tests on means and the probability content of confidence intervals when unconditional inferences are desired.

Analytic comparisons among the three inference methods are not easy in general, but substantial insight can be obtained by examining the case in which the effect-size parameters are normally distributed and all the conditional variances are equal, that is, $v_1 = \ldots = vk = v$ (which usuaJly is equivalent to the condition that the within-study sample sizes are equal). In this case, the weights in all three inference procedures are equal across studies. Consequently, the mean effect sizes under each procedure are simply the unweighted means of the T;, The variances computed using the fixed- and random-effects procedures wi11 differ, however, in that the fixed-effects estimate of the variance will be $v. = vlk$ and the random-effects estimate of the variance is $v.* = (v + f^2)/k$.

The random effects and conditionally random-effects confidence intervals and tests on the mean effect size involve the use of the variance component estimate f^2 in place of τ^2 construct the weights used in computing the mean $T.*$ and its variance v.*. If the number of studies is small, the estimate of the variance component will not be very accurate. It is reasonable to ask therefore how accurate these inferences might be when the number of studies is small.

Conditional Inferences

When conditional inferences are desired, the analyst conditions on the study characteristics, so that the effect parameters are fixed but unknown constants for inference. Statistical inferences are about the mean of the effect parameters in the experiments analyzed. Specifically, when conditional inferences are desired, the test on the mean effect size is a test of the null hypothesis $H_0: 6 = 0_0$, where 0_0 is some predefined constant (often 0) and the relevant confidence intervals are confidence intervals for 0, the mean of the k effect-size parameters $0_1, .\text{-}\text{-} , 6k$ in the k studies being analyzed. Conditionally, the sampling distribution of the mean effect-size estimate is normal with a variance v/k.

In the fixed-effects analysis, the effect parameters are treated as fixed, and the variance of the mean is exactly correct. Therefore, the rejection rates and probability content of the confidence intervals based on the fixed-effects procedures can be calculated di-

reedy and correspond ex!_Ctly to the nominal values. Because the variance of $T.*$ computed using the random- and conditionally random-effects procedures includes a contribution from the variance component estimate T^2, it will be too large. Therefore, the test on the mean effect size will have a rejection rate that is less than the nominal significance level when H_0 is true, and the probability content of confidence intervals will be larger than the nominal values.

The variance of the mean effect-size estimate using random- and conditionally random-effects procedures depends on the variance component estimate $-r^2$. Therefore, both the probability content of the confidence intervals and the rejection rates of the test statistic can be evaluated by conditioning on f^2 and averaging over values of $\cdot r2$ (weighting according to their probability density). The details are given in the Appendix. Table 3 contains the probability content of (nominally) 95% confidence intervals for 0, based on the three procedures, for sample sizes ranging from 2 studies to I00 studies and degrees of heterogeneity ranging from 0% to 100% of the sampling-error variance. These values of heterogeneity seem plausible, as $\tau^2 = 0$ corresponds to complete homogeneity, whereas $\tau^2 = .33v$ corresponds to the situation in which 75% of the total variance among effect sizes is sampling error, and $\tau^2 = v$ corresponds to the situation in which 50% of the total variance is sampling error, which are consistent with the range of values found in the survey conducted by Schmidt (1992).

As expected, the random- and conditionally random-effects procedures lead to confidence intervals that have larger than nominal probability content. In other words, they produce confidence intervals that are too wide. The extent of the exaggeration in width depends on the number of studies and the amount of between-study heterogeneity, but it can be profound if both are large.

Table 4 contains the rejection rate of the (nominal $a = .05$) one-tailed test of the null hypothesis that $0 = 0$ for each of the inference procedures. The amounts of heterogeneity (variance between effect-size parameters) are the same as in Table 3. It is evident that the rejection rates are exactly nominal for the fixed-effects procedures but lower than nominal for the nmdom- and conditionally random-effects procedures. The extent to which rejection rates are too low for random- and conditionally nmdom-effects procedures depends on heterogeneity and the number of studies, but the departure from nominal can be profound when both are large. Table 5 contains the rejection rates for the same test when the null hypothesis is false and $0 = .25\ v$, which gives some indication of power. In general, the rejection rate is highest for the fixed-effects test, followed by that of the conditionally random-effects test and then the random-effects test.

Table 3
Probability Content of 95% Confidence Intervals Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Conditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$-r^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.950	.952	.962	.950	.956	.969	.950	.961	.975	.950	.965	.980
3	.950	.952	.963	.950	.957	.970	.950	.961	.977	.950	.966	.982
4	.950	.952	.963	.950	.957	.971	*.950*	.962	.978	.950	.968	.983
5	.950	.952	.962	.950	.957	.971	.950	.963	.979	.950	.970	.985
6	.950	.952	.962	.950	.958	.971	.950	.964	.979	.950	.971	.985
7	.950	.952	.962	.950	.958	.972	.950	.965	.980	.950	.973	.986
8	.950	.952	.962	.950	.958	.972	.950	.966	.980	.950	.974	.987
9	.950	.952	.961	.950	.959	.972	.950	.967	.981	.950	.976	.987
10	.950	.952	.961	.950	.959	.972	.950	.968	.981	.950	.977	.988
20	.950	.952	.959	.950	.961	.973	.950	.974	.984	.950	.986	.991
30	.950	.952	.958	.950	.963	.973	.950	.979	.985	.950	.990	.992
40	.950	.952	.957	.950	.965	.973	.950	.982	.986	.950	.992	.993
50	.950	.952	.957	.950	.966	.974	.950	.984	.986	.950	.993	.993
60	.950	.952	.956	*.950*	.967	.974	.950	.985	.987	.950	.993	.993
70	.950	.951	.956	.950	.968	.974	.950	.986	.987	.950	.993	.993
80	.950	.951	.956	.950	.969	.974	.950	.986	.987	.950	.994	.994
90	.950	.951	*.955*	.950	.970	.975	.950	.987	.987	.950	.994	.994
100	.950	.951	.955	.950	.970	.975	.950	.987	.987	.950	.994	.994

Table 4

Rejection Rates for Nominal α = .05, One-Tailed Tests of the Hypothesis Thar δ = 0 Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Conditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	**RE**	FE	CR	**RE**	FE	**CR**	**RE**	FE	**CR**	RE
2	.050	.048	.039	.050	.044	.032	.050	.039	.026	.050	.035	.022
3	.050	.048	.039	.050	.043	.031	.050	.039	.025	.050	.034	.020
4	.050	.048	.039	.050	.043	.031	.050	.038	.024	.050	.033	.019
5	.050	.048	.039	.050	.043	.031	.050	.037	.024	.050	.031	.018
6	.050	.048	.040	.050	.043	.031	.050	.036	.023	.050	.029	.017
7	.050	.048	.040	.050	.042	.031	.050	.035	.023	.050	.028	.017
8	.050	.048	.040	.050	.042	.031	.050	.035	.023	.050	.027	.016
9	.050	.048	.040	.050	.042	.031	.050	.034	.022	.050	.026	.016
10	.050	.048	.041	.050	.042	.031	.050	.033	.022	.050	.025	.015
20	.050	.048	.042	.050	.040	.031	.050	.028	.020	.050	.017	.013
30	.050	.048	.043	.050	.039	.031	.050	.025	.020	.050	.014	.012
40	.050	.049	.044	.050	.038	.031	.050	.022	.019	.050	.012	.012
50	.050	.049	.045	.050	.037	.030	.050	.021	.019	.050	.012	.OIi
60	.050	.049	.045	.050	.036	.030	.050	.020	.018	.050	.011	.011
70	.050	.049	.045	.050	.035	.030	.050	.019	.018	.050	.011	.011
80	.050	.049	.046	.050	.034	.030	.050	.018	.018	.050	.011	.011
90	.050	.049	.046	.050	.034	.030	.050	.018	.018	.050	.011	.011
100	.050	.049	.046	.050	.033	.030	.050	.018	.018	.050	.011	.Oll

Table 5

Rejection Rates for Nominal α = 0.05, One-Tailed Tests of the Hypothesis that δ = 0 When the True δ = .25 v, Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Conditional Inferences

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.098	.097	.078	.098	.086	.064	.098	.078	.053	.098	.070	.044
3	.113	.111	.089	.113	.098	.073	.113	.088	.059	.113	.077	.047
4	.126	.123	.IOI	.126	.109	.082	.126	.096	.065	.126	.083	.051
5	.139	.135	.112	.139	.120	.091	.139	.104	.072	.139	.087	.055
6	.151	.147	.124	.151	.130	.100	.151	.111	.Q78	.151	.092	.060
7	.163	.159	.135	.163	.140	.108	.163	.118	.084	.163	.095	.064
8	.174	.170	.146	.174	.149	.117	.174	.125	.090	.174	.099	.068
9	.185	.18I	.157	.185	.159	.126	.185	.131	.097	.185	.102	.072
10	.196	.192	.167	.196	.168	.134	.196	.138	.103	.196	.106	.076
20	.299	.293	.269	.299	.254	.218	.299	.196	.166	.299	.141	.123
30	.391	.385	.363	.391	.333	.299	.391	.254	.232	.391	.185	.176
40	.475	.468	.448	.475	.407	.375	.475	.313	.298	.475	.236	.233
50	.549	.543	.524	.549	.475	.447	.549	.373	.364	.549	.293	.291
60	.615	.609	.592	.615	.538	.514	.615	.433	.427	.615	.350	.350
70	.672	.668	.653	.672	.595	.575	.672	.491	.488	.672	.408	.408
80	.723	.718	.705	.723	.647	.631	.723	.547	.544	.723	.464	.464
90	.766	.763	.751	.766	.693	.680	.766	.598	.597	.766	.517	.517
100	.804	.800	.791	.804	.735	.725	.804	.646	.645	.804	.568	.568

Unconditional Inferences

When unconditional inferences are desired, the analyst treats the effect parameters as a sample from a population and estimates the mean and variance of that population. Statistical inferences are about the mean (and possibly the variance) of the population from which the study effect sizes were sampled. Specifically, when unconditional inferences are desired, the test on the mean effect size is a test of the null hypothesis H_0: $\mu = \mu_{,0,}$ where μ_0 is some predefined constant (often 0) and the relevant confidence intervals are confidence intervals for μ, the mean of the population from which the k effect-size parameters $0_{1,}$ -.- , ek in the k studies being analyzed were sampled. Unconditionally, the 5'.ampling distribution of the mean is normal with a variance $(v + \tau^2)$/k.

In the fixed-effects analysis, the effect parameters are treated as fixed, and the variance of the mean effect-size estimate is underestimated a'i $v!k$. The rejection rates and probability content of the confidence intervals based on the fixed-effects procedures can be calculated directly but do not correspond to the nominal values. The test on the mean effect size will have a rejection rate that is larger than the nominal significance level when H_0 is true, and the probability content of confidence intervals will be less than the nominal values.

The random- and conditionally random-effects procedures should produce more accurate estimates of the variance of the mean effect. However, they will not be exactly correct because both involve using the variance component estimate f^2 in place of the true value of the variance component τ^2. Both the probability content of the confidence intervals and the rejection rates of the test can be evaluated by conditioning on T^2 and averaging over values of $,.^2$ (weighting according to their probability density). The details are given in the Appendix.

Table 6 contains the probability content of (nominally) *95%* confidence intervals for μ, based on the three procedures, for sample sizes ranging from 2 studies to 100 studies and degrees of heterogeneity ranging from 0% to 100% of the sampling error variance. As expected, the fixed-effects procedure yields confidence intervals that have exactly the nominal probability content when $\tau^2 = 0$, but when $\tau^2 > 0$, they have lower than nominal probability content (are too narrow). Random- and conditionally random-effects procedures lead to confidence intervals that have slightly larger than nominal probability content when $\tau^2 = 0$ (because they overestimate τ^2 in this case) and less than nominal content (are too narrow) when $\tau^2 > 0$. The extent of the underestimation in width is greater when the number of studies is small and the

Table 6
Probability Content of 95% Confidence Intervals Based on Fixed-Effects (FE), Random-Effects ***(RE),*** *and Conditionally Random-Effects* ***(CR)*** *Procedures: Unconditional Inferences*

	$\tau^2 = 0$			$\tau^2 = .33$			,-2 = .67			$\tau^2 = 1.00$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.950	.951	.962	.911	.913	.936	.871	.879	.913	.834	.848	.893
3	.950	.951	.963	.911	.915	.940	.871	.883	.921	.834	.857	.905
4	.950	.951	.963	.911	.916	.941	.871	.887	.924	.834	.865	.912
5	.950	.951	.962	.911	.917	.942	.871	.890	.927	.834	.871	.916
6	.950	.951	.962	.911	.918	.943	.871	.893	.929	.834	.877	.919
7	.950	.951	.962	.911	.919	.943	.871	.896	.930	.834	.882	.922
8	.950	.951	.962	.911	.920	.943	.871	.898	.931	.834	.886	.924
9	.950	.951	.961	.911	.920	.943	.871	.900	.932	.834	.891	.926
IO	.950	.951	.961	.911	.921	.944	.871	.902	.933	.834	.894	.927
20	.950	*.951*	.959	.911	.925	.945	.871	.917	.939	.834	.919	.937
30	.950	.951	.958	.911	.929	.945	.871	.927	.942	.834	.932	.941
40	.950	.951	.957	.911	.931	.946	.871	.933	.943	.834	.939	.943
50	.950	.951	.957	.911	.933	.946	.871	.938	.945	.834	.942	.944
60	.950	.951	.956	.911	.935	.947	.871	.941	.945	.834	.944	.945
70	.950	.951	.956	.911	.937	.947	.871	.943	.946	.834	.945	.946
80	.950	.951	.956	.911	.938	.947	.871	.944	.946	.834	.946	.946
90	.950	.951	*.955*	.911	.939	.947	.871	.945	.947	.834	.947	.947
100	.950	.951	.955	.911	.941	.948	.871	.946	.947	.834	.947	.947

heterogeneity is large, and it can be substantial in extreme cases. However, the fixed-effects procedures are always further from nominal than either random- or conditionally random-effects procedures.

Table 7 contains the rejection rate of the (nominal ex = .05) one-tailed test of the null hypothesis that μ = 0 for each of the inference procedures. The amounts of heterogeneity (variance between) effect-size parameters are the same as in Table 6. It is evident that the rejection rates for the fixed-effects procedures are exactly nominal when $-r^2 = 0$, but those of the random- and conditionally random-effects procedures are lower than nominal. The rejection rates for all procedures are higher than nominal when $-r^2 > 0$, but those of the random- and conditionally random-effects procedures are closer to nominal than those of the fixed-effects procedures. When $\tau^2 > 0$, the rejection rate converges to the nominal as the number of studies increases and is reasonably close to nominal if $k > 20$. Table 8 contains the rejection rates for the same test when the null hypothesis is false and μ = .25 v, which gives some indication of power. In general, the rejection rate is highest for the fixed-effects test, followed by that of the conditionally random-effects test and then the random-effects test. However, in comparing the power of these tests, it is important to **remember that** the null (μ = 0) rejection rate of the fixed-effects procedure is also higher than nominal when $-r^2 > O$.

Effects of Unequal Variances

The calculations in this article were based on equal within-study conditional variances ($v_1 = \ldots = vk$). These calculations provide an indication of the magnitude of the differences in performance among procedures that might be expected. It is reasonable to ask how these results would generalize to more realistic cases of unequal conditional variances. Fortunately, qualitative arguments suggest that the differences in performance among procedures will remain similar even when the conditional variances are unequal.

With unequal variances, the rejection rates under the null hypothesis and probability content of confidence intervals for fixed-effects analyses would still be exactly correct in the case of conditional inferences. Random- and (to a lesser extent) conditionally random-effects procedures overestimate the variance of the mean effect size under the conditional inference model, although the exact amount will depend on the configuration of conditional variances. Thus, these procedures will have smaller than nominal rejection rates under the null hypothesis and confidence intervals that have higher than nominal probability content.

Table 7
Rejection Rates for Nominal a "' *.05, One-Tailed Tests of the Hypothesis That* μ = *0, Based on Fixed-Effects (FE),Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Unconditional Inferences*

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.00$		
k	FE	**CR**	**RE**	FE	**CR**	**RE**	FE	**CR**	**RE**	FE	**CR**	RE
2	.050	.049	.039	.077	.075	.057	.102	.095	.071	.122	.112	.082
3	.050	.049	.039	.077	.073	.054	.102	.092	.066	.122	**.106**	.075
4	.050	.049	.039	.077	.072	.054	.102	.089	.064	.122	.101	.071
5	.050	.049	.039	.077	.071	.053	.102	.087	.063	.122	.097	.069
6	.050	.049	.040	.077	.071	.053	.102	.085	.062	.122	.093	.067
7	.050	.049	.040	.077	.070	.053	.102	.083	.061	.122	.090	.066
8	.050	.049	.040	.077	.070	.053	.102	.082	.060	.122	.087	.064
9	.050	.049	.040	.077	.069	.053	.102	.080	.060	.122	.085	.063
10	.050	.049	.041	.077	.069	.053	.102	.079	.059	.122	.082	.063
20	.050	.049	.042	.077	.066	.053	.102	.070	.056	.122	.068	.057
30	.050	.049	.043	.077	.064	.052	.102	.064	.055	.122	.060	.055
40	.050	.049	.044	.077	.062	.052	.102	.060	.054	.122	.056	.054
50	.050	.049	.045	.077	.061	.052	.102	.057	.053	.122	.054	.053
60	.050	.049	.045	.077	.060	.052	.102	.056	.053	.122	.053	.053
70	.050	.049	.045	.077	.058	.052	.102	.054	.052	.122	.053	.052
80	.050	.049	.046	.077	.058	.052	.102	.053	.052	.122	.052	.052
90	.050	.049	.046	.077	.057	.051	.102	.053	.052	.122	.052	.052
100	.050	.049	.046	.077	.056	.051	.102	.052	.052	.122	.052	.052

Table 8
Rejection Rates for Nominal a= *.05, One-Tailed Tests of the Hypothesis Thatμ= 0 When the Trueμ= .25 v, Based on Fixed-Effects (FE), Random-Effects (RE), and Conditionally Random-Effects (CR) Procedures: Unconditional Inferences*

	$\tau^2 = 0$			$\tau^2 = .33$			$\tau^2 = .67$			$\tau^2 = 1.0$		
k	FE	CR	RE	FE	CR	RE	FE	CR	RE	FE	CR	RE
2	.098	.093	.076	.132	.120	.096	.159	.138	.111	.181	.151	.121
3	.113	.111	.089	.147	.139	.106	.174	.158	.117	.196	.169	.124
4	.126	.123	.101	.160	.150	.116	.188	.166	.124	.209	.173	.124
5	.139	.135	.112	.173	.161	.125	.200	.173	.131	.221	.177	.128
6	.151	.147	.124	.185	.171	.134	.212	.180	.138	.233	.180	.133
7	.163	.159	.135	.197	.180	.144	.223	.186	.145	.243	.184	.138
8	.174	.170	.146	.208	.190	.152	.234	.193	.151	.254	.187	.143
9	.185	.181	.157	.219	.199	.161	.244	.199	.158	.263	.191	.153
10	.196	.192	.167	.229	.208	.170	.254	.205	.165	.273	.194	.158
20	.299	.293	.269	.324	.288	.251	.342	.260	.228	.355	.230	.209
30	.391	.385	.363	.406	.359	.325	.416	.310	.287	.423	.269	.258
40	.475	.468	.448	.478	.424	.393	.480	.359	.343	.482	.311	.306
50	.549	.543	.524	.542	.483	.456	.538	.407	.396	.535	.354	.351
60	.615	.609	.592	.600	.537	.514	.589	.453	.446	.582	.396	.395
70	.672	.668	.653	.651	.586	.567	.635	.497	.493	.624	.438	.437
80	.723	.718	.705	.696	.631	.615	.676	.539	.536	.662	.477	.477
90	.766	.763	.751	.736	.672	.659	.713	.579	.577	.696	.515	.515
100	.804	.800	.791	.771	.709	.698	.746	.617	.615	.727	.550	.550

In the case of unconditional inferences, the rejection rate of fixed-effects tests under the null hypothesis would be larger than nominal and the probability content of confidence intervals smaller than nominal whenever $\tau^2 > 0$. This is because fixed-effects procedures would underestimate (to an extent that depends on the configuration of conditional variances) the actual unconditional variance of the mean effect-size estimate. Random- and (and to a lesser extent) conditionally random-effects procedures would provide rejection rates and confidence intervals whose probability content were closer to nominal but would not be exactly nominal because they involved substituting an estimate of $\hat{\tau}^2$ for the exact value of τ^2 in computing weights and the variance of the mean effect size. However, as the number of studies increased, the performance of both conditionally random and random-effects procedures would converge to nominal.

Conclusion

The selection of a statistical procedure in meta-analysis (and elsewhere) should be determined by the inferences one wishes to make. Once the inference goals are clear, a statistical procedure appropriate for those goals should be chosen. If conditional inferences (i.e., inferences about the parameters in the collection of studies observed) are desired, then fixed-effects procedures should be used. Heterogeneity of effects should not necessarily be a reason to abandon fixed-effects analyses. However, substantial heterogeneity may suggest that explanatory analyses are desirable to gain an understanding of that heterogeneity. Using random-effects or conditionally random-effects analyses to make conditional inferences will result in less powerlul tests of significance (tests that have a higher actual significance level than the nominal) and confidence intervals that are too wide.

If unconditional inferences (i.e., inferences about the population from which the observed studies are sampled) are desired, then either conditionally random- or random-effects procedures should be used. Unless there is almost perfect homogeneity of effects ($\tau^2 = 0$ or nearly so), fixed-effects tests will reject more often than expected (have a higher actual significance level than nominal). Random-effects tests and confidence intervals will not give results that are exactly nominal either, particularly if the number of studies is small, but their perlormance will be closer to nominal than fixed-effects tests. If the number of studies is very small (say less than five), random-effects tests should be regarded as only approximate.

Conditionally random-effects procedures (which correspond to fixed-effects procedures when the be-

tween-studies variance component is not statistically significant and random-effects procedures when it is significant) generally have performance in between that of fixed- and random-effects procedures. If the inference one wishes to make is not entirely clear, conditionally random-effects procedures might be a reasonable compromise between fixed- and random-effects procedures. However, we believe that clarifying the inference desired (followed by the choice of either fixed- or random-effects procedures) is generally more desirable.

References

Camilli, G. (1990). The test of homogeneity for 2 x 2 contingency tables: A review of and some personal opinions on the controversy. *Psychological Bulletin, 108,* 135-145.

Chang, L. (1992). *A power analysis of the test of homogeneity in effect size meta-analysis.* Unpublished doctoral dissertation, Michigan State University, East Lansing.

DerSimonian, **R.,** & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials,* 7, 177-188.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 106--128.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388-395.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17,* 279-296.

Hedges, **L. V.,** & **Olkin,** I. (1983). Clustering estimates of effect magnitude from independent studies. *Psychological Bulletin, 93,* 563-573.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hunter, J.E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bia. in research findings.* Newbury Park, CA: Sage.

Hyde, J. S. (1981). How large are cognitive gender differences: A meta-analysis using omega and d. *American Psychologist, 36,* 892-901.

Olkin, I., & Sampson, A. (1998). Comparison of meta-analysis versus analysis of variance of individual patient data. *Biometrics, 54,* 272-277.

Patnaik, P. B. (1949). The noncentral X^2- and F-distributions and their applications. *Biometrika, 36,* 202-232.

Raudenbush, ***S.* W.,** & **Bryk, A.** *S.* (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, JO,* 75-98.

Rosenthal, **R.,** & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin, 92,* 500--504.

SAS 6.12 [Computer software]. (1996). Cary, NC: The SAS Institute.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173-1181.

Schmidt, F. L., & Hunter, J. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 65,* 643--661.

SPSS 8.0 for Windows [Computer softwarel. (1998). Chicago: SPSS Inc.

(Appendix follows)

Appendix

Derivations

Bias in the Estimator 1'2 of ,2

When the effect-size parameters are normally distributed and all the conditional variances are equal, that is, $v_1 = \ldots = v_k = v$, then the T_i's are independently and identically distributed as $N(\mu, v + \tau^2)$ and the (estimated) weights $w_i = w^* = 1/(v + \tau^2)$. Consequently, the homogeneity statistic Q is distributed as $[(v + \tau)/v]$ multiplied by a central chi-square variate with $(k - 1)$ degrees of freedom (Patnaik, 1949). To obtain the bias in the estimator τ^2 given in Equation 10, note that $\hat{\tau}^2$ is a function of Q given by

$$\hat{\tau}^2(Q) = \begin{cases} \frac{v}{c}(Q - (k-1)) & \text{if } Q \geq k-1. \\ 0 & \text{if } Q < k-1. \end{cases}$$

Using the fact that the probability density of Q is $[v/(v + \tau^2)]g[v/(v + \tau^2)]$, where $g(x)$ is the probability density function of a chi-square with $k - 1$ degrees of freedom, the expected value of $\hat{\tau}^2$ is

$$E(\hat{\tau}^2) = \frac{v}{c}\int_{k-1}^{\infty}\left(\frac{v}{c}(q-(k-1))\right)g\left(\frac{v}{v+\tau^2}q\right)dq,$$

where $g(x)$ is the probability density function of a chi-square with $k - 1$ degrees of freedom. The values in Table 2 were obtained by numerically integrating the expression above.

Confidence Intervals and Rejection Rates in the Conditional

Fixed-Effects Procedures

When all the conditional variances are equal, that is, $v_1 = \ldots = v_k = v$, then the T_i's are independently distributed as $N(0, v)$. and mean Tis distributed as $N(0, v/k)$. Consequently, the probability p that θ lies between the confidence limits L and U given in Equation 5 is simply

$$p = P(L < \theta < U) = P(\theta < U) - P(\theta < L),$$

which after calculating the probabilities gives

$$p = \Phi(z_u - \theta\sqrt{k/v}) - \Phi(z_l - \theta\sqrt{k/v}).$$

The rejection rate of the test of the hypothesis that $\theta = 0$ is given by $p = P[|\bar{T}/\sqrt{v/k}| > z_{\alpha/2}] = 1 - \Phi[z_{\alpha/2} - \theta\sqrt{k/v}]$.

Random-Effects Procedures

When all the conditional variances are equal, that is $v_1 = \ldots = v_k = v$, the (estimat) weight are $w_i^* = w^* = 1/(v + \tau^2)$, and consequently $T.^*$ is simply the unweighted mean Tand $v.^* = (v + \tau^2)/k$. Using Equation 10 for $\hat{\tau}^2$, the expression for $v.^*$ conditional on Q becomes

$$v.^*(Q) = \begin{cases} \frac{v}{k}(Q-(k-1)) & \text{if } Q \geq k-1. \\ v/k & \text{if } Q < k-1. \end{cases}$$

The sampling distribution of the homogeneity statistic Q is a noncentral chi-square variate with $(k - 1)$ degrees of freedom, and noncentrality parameter

$$\lambda = \sum_{i=1}^{k}(\theta_i - \bar{\theta})^2/v = k\tau^2/v,$$

and is independent of $\bar{T}$.

The $100(1-\alpha)\%$ nominal confidence interval for θ is given by

$$L^* = \bar{T} - z_{1-\alpha/2}\sqrt{v/k} < \theta < \bar{T} + z_{1-\alpha/2}\sqrt{v/k} = U^*.$$

Evaluating the probability content conditional on Q gives

$$P(L^* < \theta < U^*|Q) = P(\theta < U^*|Q) - P(\theta < L^*|Q).$$

Noting that $\bar{T} \sim N(\theta, v/k)$ and evaluating these conditional probabilities gives

$$p(Q) = \Phi\{z_{1-\alpha/2}\sqrt{v/k} - [v.^*(Q)/v]\} - \Phi\{-z_{1-\alpha/2}\sqrt{v/k} - [v.^*(Q)/v]\}$$

Integrating over the distribution of $p(Q)$ and taking account of the fonctional dependence of $v.^*(Q)$ on Q given above yields

$$p = \int_0^{k-1}\left[\Phi\left(z_{1-\alpha/2}\sqrt{\frac{v}{k}}\right) - \Phi\left(-z_{1-\alpha/2}\sqrt{\frac{v}{k}}\right)\right]f(q|\lambda)\,dq + \int_{k-1}^{\infty}\left[\Phi\left(z_{1-\alpha/2}\sqrt{\frac{v}{k}}\right) - \Phi\left(-z_{1-\alpha/2}\sqrt{\frac{v}{k}}\right)\right]f(q|\lambda)\,dq,$$

where $f(q|\lambda)$ is the probability density fonction of Q, which is a noncentral chi-square with $(k - 1)$ degrees of freedom and noncentrality parameter $\lambda = k\tau^2/v$. The results in Table 3 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $\theta = 0$ can also be evaluated conditionally on Q. The conditional one~tailed rejection rate is just

$$P\{Z^* > z_{1-\alpha}|Q\} = P\{\sqrt{k}(\bar{T} - 0)/\sqrt{v} > \sqrt{k}[z_{1-\alpha}\sqrt{v} - v.^*(Q) - 0]/\sqrt{v}|Q\}.$$

Evaluating these conditional probabilities using the fact that $\bar{T} \sim N(\theta, v/k)$ yields

$$\left[1 - \Phi\left(C_\alpha - \frac{\bar{\theta}\sqrt{k}}{\sqrt{v}}\right)\right]\int_0^{k-1} f(q|\lambda)\,dq + \int_{k-1}^{\infty}\left[1 - \Phi\left(C_\alpha\sqrt{\frac{q}{(k-1)}} - \frac{\bar{\theta}\sqrt{k}}{\sqrt{v}}\right)\right]f(q|\lambda)\,dq,$$

where f (qi>..) is the probability density function of Q. The values in Tables 4 and 5 were obtained by numerically integrating the expression above.

Conditionally Random-Effects Procedures

In the conditionally random effects procedures, the random-effects procedures are used if Q is statistically significant, and **the** fixed-effects procedures, are used otherwise. Thus, the expression for $v.c$ conditional on Q becomes

$$v.(Q) = \begin{cases} vQ!k\,(k-1) & \text{if } Q \geq Q_{.95} \\ vlk & \text{if } Q < Q_{.95}, \end{cases}$$

where $Q_{.95}$ is the 95th percentile point of the chi-square distribution with k - 1 degrees of freedom. The sampling distribution of the homogeneity statistic Q is a noncentral chi-square variate, with $(k$ - 1) degrees of freedom and noncentrality parameter

$$>.. = \sum_{i=1}^{k} (9; - 0)2/v = k\text{-}r21v,$$

and is independent of T.

The 100(1-o:)% nominal confidence interval for ii is given by

$$LC = T\text{-}z_{,,,\_12VV,c} < 0 < T + Z_{,,,\_12}|v.c = if.$$

Evaluating the probability content conditional on Q gives

$$P\{Lc < 0 < UC'|Q\} == P\{6 < U(Q)|Q\} - P\{0 < L(Q)|Q\}.$$

Noting that T - N (0, v/k) and evaluating these conditional probabilities gives

$$p(Q) == 2\langle l\rangle\{Za12vlkv,c\ (Q)/v)]\} - 1.$$

Integrating over the distribution of $p(Q)$ and taking account the functional dependence of $v.*\ (Q)$ on Q, given above, yields

$$p = [2\Phi(C_{\alpha/2}) - 1]\int_0^{Q.95} f(q|\lambda)\,dq + \int_{Q.95}^{\infty}\left[2\Phi\left(C_{\alpha/2}\sqrt{\frac{q}{(k-1)}}\right) - 1\right] f(q|\lambda)\,dq,$$

where f (qlll.) is the probability density function of Q (which is just a noncentral chi-square with (k - 1) degrees of freedom and noncentrality parameter>.. = k:r^2/v). The results in Table 3 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $0 = 0$ can also be evaluated conditionally on Q. The conditional one-tailed rejection rate is just

$$P\{zC > Za\,|Q\} = P\{\sqrt{k}(T - 0)/yv.c\ (Q) > y'k[za\,y'v.c\ (Q) - 0J!\backslash lv\,|Q\}.$$

Evaluating these conditional probabilities using the fact that T - N[ii, vlk] yields

$$\left[1 - \Phi\left(C_\alpha - \frac{\bar{\theta}\sqrt{k}}{\sqrt{v}}\right)\right]\int_0^{Q.95} f(q|\lambda)\,dq + \int_{Q.95}^{\infty}\left[1 - \Phi\left(C_\alpha\sqrt{\frac{q}{(k-1)}} - \frac{\bar{\theta}\sqrt{k}}{\sqrt{v}}\right)\right] f(q|\lambda)\,dq,$$

where f (qlll.) is the probability density function of Q. The values in Tables 4 and 5 were obtained by numerically integrating the expression above.

Confidence Intervals and Rejection Rates in the Unconditional Model

Fixed-Effects Procedures

When the effect-size parameters are normally distributed and all the conditional variances are equal, that is $v, = \ldots = vk = v$, then the $T;s$ are independently distributed as N(μ, v + -r^2), and mean Tis distributed as N[μ,, $(v$ + -r^2)/k]. Consequently, the probability p that μ, lies between the confidence limits L and U given in Equation 5 is simply

$$p = P(L < \mu < U) = P(\mu, < U) - P(\mu, < L),$$

which after evaluating the probabilities gives

$$p = 2\Phi\{z_{\alpha/2}\sqrt{[v/(v + \tau^2)]}\} - 1.$$

The rejection rate of the test of the hypothesis that μ = 0 is given by

$$p = P\{T!v'[vlk] > z_{\cdot} = 1 - \langle l\rangle\{za\,V[v/(v + -r^2)] - \mu,\backslash,l[k!(v + -r^2)]\}.$$

Random-Effects Procedures

When all the conditional variances are equal, that is $v_I = \ldots = vk = v$, the (estimated) weights are $w/ = w^* = 1/(v + t^2)$, and consequently $T.^*$ is simply the unweighted mean $T.$, and v.* = (v + t^2)/k. Using Equation 10 for ;-2, the expression for $v.*$ conditional on Q becomes

$$v.^*(Q) = \begin{cases} -\{vQlk(k-1) & \text{if } Q \geq k\text{-}\ I \\ vlk & \text{if } Q < k\text{-}\ 1. \end{cases}$$

The sampling distribution of the homogeneity statistic Q is $(v + \tau^2)/v$ multiplied by a central chi-square vari , with $(k$ - I) degrees of freedom, and is independent of T.

The 100(1-o:)% nominal confidence interval foT μ, is given by

$$L^* = T\text{-}\ z_{012}y'v,^* < T + z_{,.12}v'v,^* = U^*.$$

Evaluating the probability content conditional on Q gives

$$P[L^* < \mu < U^*|Q] = P[\mu < U(Q)|Q] - P[\mu < L(Q)|Q].$$

Noting that $T \sim N(\mu, (v + \tau^2)/k)$ and evaluating these conditional probabilities give

$$p(Q) = 2\Phi\{z_{\alpha/2}\sqrt{[kv.^C(Q)]/\sqrt{(v+\tau^2)}}\} - 1.$$

Integrating over the distribution of $p(Q)$ and taking account of the functional dependence of $v.^*(Q)$ on Q given above yield

$$P = \int_0^{q_{.95}} \left[1 - \Phi\left(C_\alpha \sqrt{\frac{v}{v+\tau^2}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq + \int_{k-1}^{\infty} \left[1 - \Phi\left(C_\alpha \sqrt{\frac{vq}{(v+\tau^2)(k-1)}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq,$$

where $j(q)$ is the probability density function of Q. The results in Table 6 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $\mu = 0$ can also be evaluated conditionally on Q. The conditional one-tailed rejection rate is just

$$\begin{aligned} &P\{Z^* > Z_\alpha|Q\} \\ &\quad = P\{v'k(T - \mu)/\sqrt{(v+\tau^2)} \\ &\quad > v'k[Z_\alpha\sqrt{v.^*(Q)} - \mu]/\sqrt{(v+\tau^2)}\,|Q\}. \end{aligned}$$

valuating these conditional probabilities using the fact that $T \sim N[6, (v + \tau^2)/k]$ yields

$$\left[1 - \Phi\left(C_\alpha \sqrt{\frac{v}{v+\tau^2}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] \int_0^{k-1} f(q)\,dq + \int_{k-1}^{\infty} \left[1 - \Phi\left(C_\alpha \sqrt{\frac{vq}{(v+\tau^2)(k-1)}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq,$$

where $f(q)$ is the probability density function of Q. The values in Tables 7 and 8 were obtained by numerically integrating the expression above.

Conditionally Random-Effects Procedures

In the conditionally random effects procedure, the random-effect,; procedures are used if Q is statistically significant, and the fixed-effects procedures are used otherwise. Thus, the expression $v.C$ for the variance of T conditional on Q becomes

$$v.(Q) = \begin{cases} v.(Q)lk & \text{if } Q > q_{.95} \\ v.lk & \text{if } Q < q_{.95}, \end{cases}$$

where $q_{.95}$ is the 100a% point of the chi-square distribution with $(k - 1)$ degrees of freedom. The sampling distribution of the homogeneity statistic Q is $(v + \tau^2)/v$ x a central chi-square variate, with $(k - 1)$ degrees of freedom, and is independent of T.

The I 00(1-rx)% nominal confidence interval for μ is given by

$$Le = T - z_{\alpha/2}\sqrt{v.c} < \mu < T + z_{\alpha/2}\sqrt{v.c} = if.$$

Evaluating the probability content conditional on Q gives

$$P(l,c < \mu < if|Q] = P[\mu < U(Q)|Q] - P[\mu < L(Q)|Q].$$

Noting that $T \sim N(\mu., (v + \tau^2)/k)$ and evaluating these conditiomd probabilities give

$$p(Q) = 2\Phi\{z_{\alpha/2}\sqrt{[kv.^C(Q)]/\sqrt{(v+\tau^2)}}\} - 1.$$

Integrating over the distribution of $p(Q)$ and taking account of the functional dependence of $v.c(Q)$ on Q given above yield

$$P = \int_0^{q_{.95}} \left[1 - \Phi\left(C_\alpha \sqrt{\frac{v}{v+\tau^2}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq + \int_{k-1}^{\infty} \left[1 - \Phi\left(C_\alpha \sqrt{\frac{vq}{(v+\tau^2)(k-1)}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq,$$

wheref(q) is the probability density function of Q and Qa is defined above. The results in Table 6 are obtained by numerically integrating the expression above.

The rejection rate of the test of the hypothesis that $\mu = 0$ can also be evaluated conditionally on Q. The conditional one-tailed rejection rate is just

$$\begin{aligned} &P\{Z^* > z_\alpha|Q\} \\ &\quad = P\{v'k(T - \mu)/\sqrt{(v+\tau^2)} \\ &\quad > v'k[z_\alpha\sqrt{v.^*(Q)} - \mu]/\sqrt{(v+\tau^2)}\,|Q\}. \end{aligned}$$

Evaluating these conditional probabilities using the fact that $T \sim N[6, (v + \tau^2)/k]$ yields

$$\left[1 - \Phi\left(C_\alpha \sqrt{\frac{v}{v+\tau^2}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] \int_0^{q_{.95}} f(q)\,dq + \int_{q_{.95}}^{\infty} \left[1 - \Phi\left(C_\alpha \sqrt{\frac{vq}{(v+\tau^2)(k-1)}} - \frac{\mu\sqrt{k}}{\sqrt{v+\tau^2}}\right)\right] f(q)\,dq,$$

where $f(q)$ is the probability density function of Q. The values in Tables 7 and 8 were obtained by numerically integrating the expression above.

Received December 19, 1997
Revision received March 16. 1998
Accepted July 5, 1998 ■

Berkeley School of Public Health

Population intervention effects

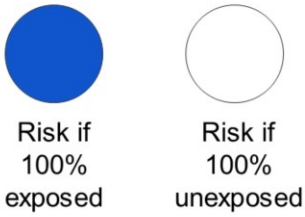
PHW250 B – Andrew Mertens

ANDREW MERTENS: This video will introduce population intervention effects. These are essentially another type of measure of association. So you've learned about relative risks, risk differences, and population attributable fractions. Population intervention effects are a class of parameters, which is just another way of saying a quantity we intend to estimate that can measure in a factor and association. And they're not defined quite as narrowly as, for example, a population attributable fraction. But they have the potential to provide much more realistic and policy-relevant information.

Motivation

- Studies often contrast two scenarios:
  - 1) if everyone was treated or exposed
  - 2) of no one was treated or exposed

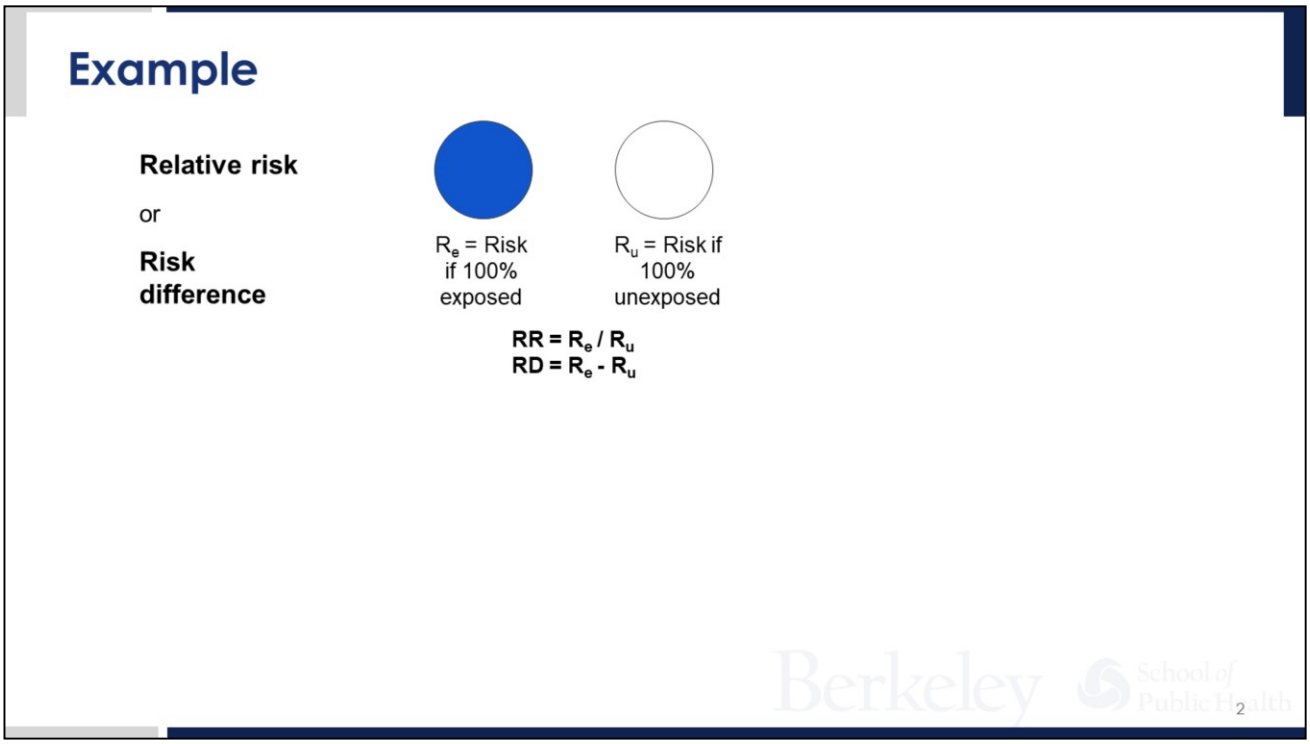
Relative risk:



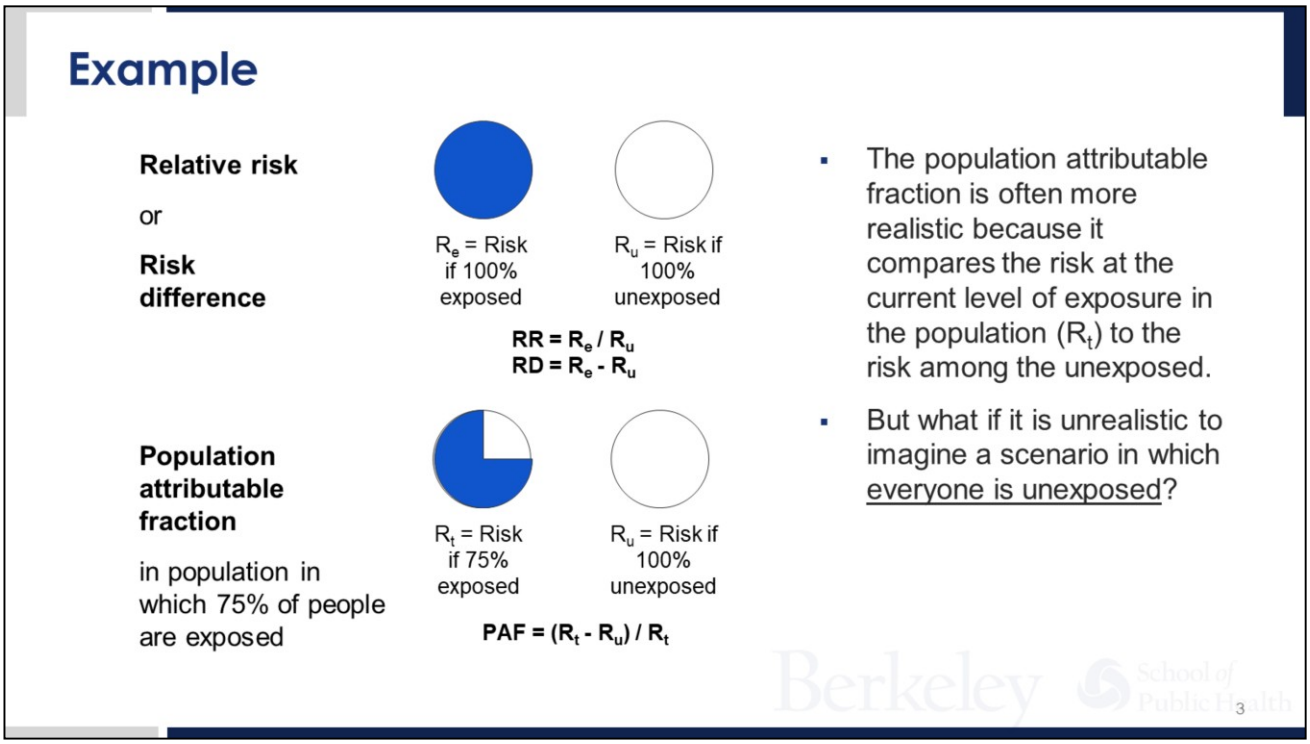
- However, real-world effects are likely to differ if it is unlikely that an entire population will receive or not receive the intervention or exposure.

In epidemiology, our studies often contrasted two scenarios, one in which everyone was treated or exposed and one in which no one was treated or exposed. So for example, if we estimate a relative risk, and for our example here, let's say it's in a trial, the blue circle is indicating that in the treatment group, everyone is receiving the treatment, or 100% of people are treated or exposed. And the white circle is indicating a scenario in which no one is receiving treatment, or 100% of people are unexposed.

This can be a very useful contrast when we're trying to understand the etiology of a particular disease. But these are not going to reflect real-world effects. And that's because it's unlikely that an entire population would receive or not receive an intervention or exposure.



So let's draw this out a little further. And below these two, I've added the formula, the relative risk, and the risk difference. So it's just RE, the risk if 100% is exposed, and RU, the risk if 100% is unexposed. And we can divide or subtract those to get the relative risk or the risk difference.



And let's contrast this with the population attributable fraction.

And our circle on the bottom left here is showing that in the total population, 75% of people are exposed, which is indicated by this blue shading in the pie here. And so RT corresponds to 75% exposed or 75% treated. And RU corresponds to no one exposed or treated. And then the population attributable fraction takes the difference in RT and RU and divides by RT.

So we often learn in epidemiology that the population attributable fraction is more realistic because it's comparing the observed scenario with the observed level of exposure to one in which we counterfactually remove the exposure from the population. And sometimes this is a useful counterfactual to imagine. But what if it's unrealistic to imagine a scenario in which everyone is exposed? We need more flexible parameters that we can tailor to our particular research questions. And this is what population intervention effects can provide to us.

Example: SHEWA-B

- SHEWA-B: The Sanitation Hygiene Education and Water Supply in Bangladesh Program
- Implemented by UNICEF and the Government of Bangladesh from 2007-2012
- Targeted 20 million people
- Interim evaluation in 2009 using matched control areas
- Fell short of targets for child illness and health behavior
- Poor results could reflect, poor design, poor implementation, or both

Huda et al., 2012. https://doi.org/10.1016/j.socscimed.2011.10.042
Benjamin-Chung et al., 2017. doi: 10.2105/AJPH.2017.303686

I'll use an example of a program evaluation. The program was called SHEWA-B, the Sanitation, Hygiene, Education, and Water Supply in Bangladesh program. And this program was implemented by UNICEF and the government of Bangladesh in Bangladesh from 2007 to 2012. And it was a very large-scale program targeting around 20 million people. And when I say a target, what I mean is they intended to deliver the program to 20 million people to have potentially a sweeping impact on household-level water and sanitation.

In 2009, an interim evaluation was conducted using matched control areas. So they identified villages very similar to those that should have received the program to improve sanitation and hygiene. But the control villages did not receive that program. And this 2009 interim evaluation found that SHEWA-B had fallen short of their targets for improving child illness and health behaviors.

And so at this point, the program still had a few more years to go. And the program officers had the question of whether these interim results reflected a poor design, a poor implementation, or both. Poor design would mean that the interventions themselves were not sufficient to improve health behavior and reduce illness. Poor implementation would suggest that the program design was adequate but that it wasn't delivered to a sufficient number of, people, or it wasn't delivered properly. And it's possible that both of these factors could have been at play.

Example: SHEWA-B

- SHEWA-B: The Sanitation Hygiene Education and Water Supply in Bangladesh Program
- Implemented by UNICEF and the Government of Bangladesh from 2007-2012
- Targeted 20 million people
- Interim evaluation in 2009 using matched control areas
- Fell short of targets for child illness and health behavior
- Poor results could reflect, poor design, poor implementation, or both

Ideal scenario:

R_i = Risk when 80% receive program

Observed scenario:

R_o = Risk if 40% receive program

Risk difference for ideal vs. observed scenario: $R_i - R_o$

This is a population intervention effect.

Huda et al., 2012. https://doi.org/10.1016/j.socscimed.2011.10.042
Benjamin-Chung et al., 2017. doi: 10.2105/AJPH.2017.303686

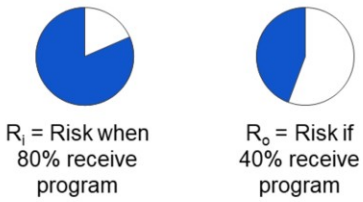
5

They observed, in 2009, that 40% of people who were supposed to receive the SHEWA-B program actually received it. That's our observed scenario, shown in this pie on the right hand side. And UNICEF's ideal scenario in our example is that 80% of the population received the program. This is quite a big gap, between 40% and 80%. So we can call the 80% scenario their ideal scenario and the 40% the observed scenario.

And these are denoted by R sub I and R sub O. And we could take the risk difference for the ideal versus observed scenario. And that's RI minus RO. This is an example of a population intervention effect.

Population intervention effects

- **Population intervention effects** are "causal effects tied to contrasts between the observed population and exposure distributions under realistic interventions"



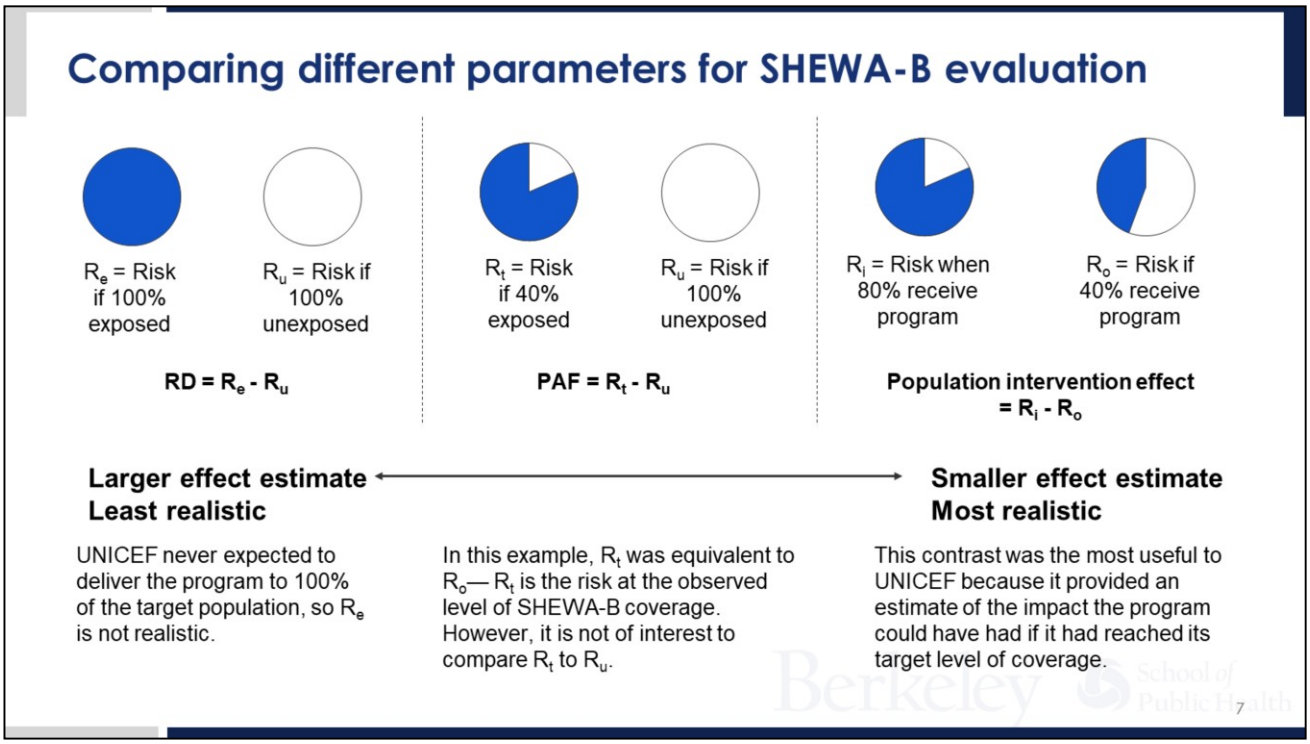
R_i = Risk when 80% receive program

R_o = Risk if 40% receive program

Population intervention effect
= R_i - R_o

Westreich et al., 2016 doi: 10.2105/AJPH.2016.303226

I'll delve into this in a little more detail, but here's the definition. Population intervention effects are causal effects tied to contrast between the observed population and exposure distributions under realistic interventions. So again, we're comparing RI, which is our ideal counterfactual scenario, to RO, which is our realistic observed scenario.

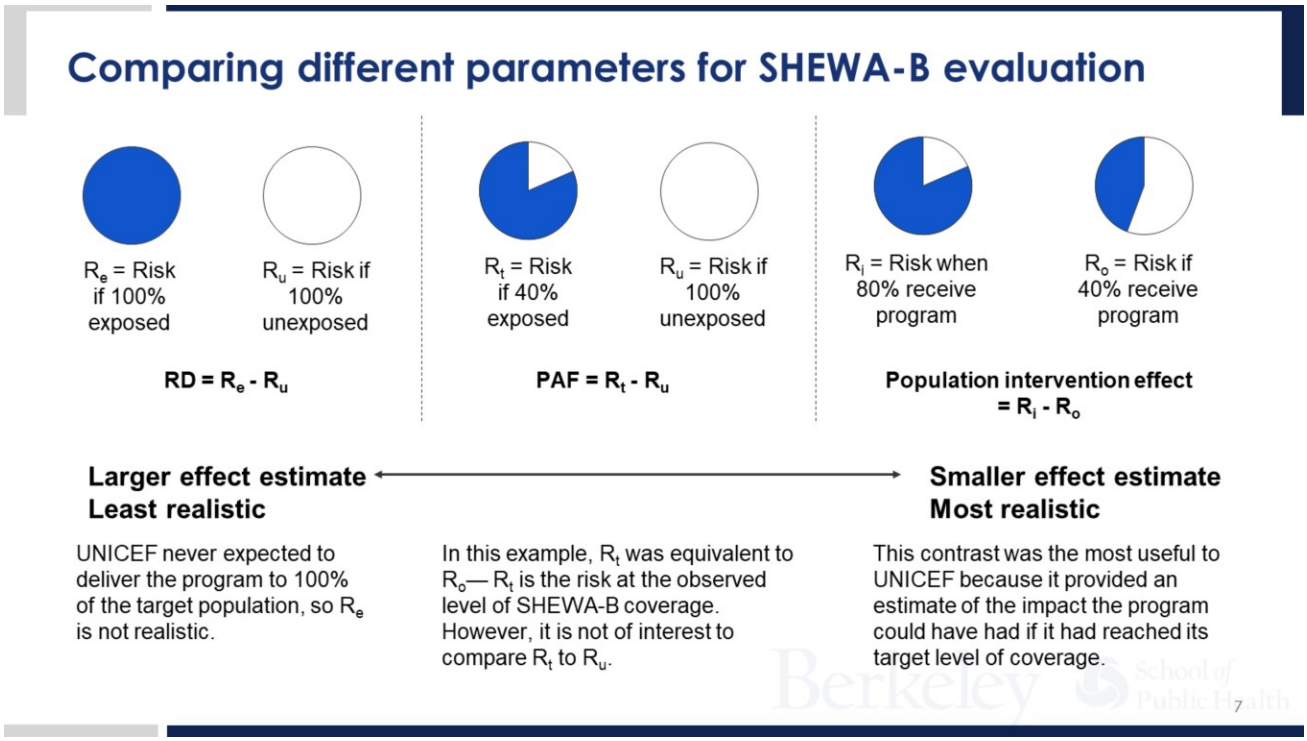


Let's compare different possible parameters that could have been used in this SHEWA-B evaluation. On the left hand side, we have the traditional risk difference. In this case, we would compare RE minus RU, the risk of illness if 100% of people were exposed or 100% got the program, and RU, if 100% were unexposed or no one got the program.

In the middle, we have the population attributable fraction. And I've written it a little bit differently, just for continuity here. Normally, the PAF is RT minus RU divided by RT. But for our purposes, let's just say it's RT minus RU. This would compare the risk if 40% were exposed, which is consistent with the observed total population level of exposure, minus RU, if no one was exposed.

And then we have the population intervention effect, which is comparing RI, the risk when 80% receive the population, which is the ideal for SHEWA-B, to RO, the risk if 40% received the program, which is consistent with what was observed. And as we move from left to right, the effect estimate will get smaller.

Now how do I know this? Well, assuming that there is a linear pattern, an effect size that's correlated with exposure with the percentage of people receiving the program, this would mean that because our contrasts are closer to each other-- in other words, 80% coverage is closer to 40% than 100% is to 0% exposed-- this means that we would expect the population intervention effect to be smaller in most cases than the risk difference. And population attributable fraction would be in the middle.



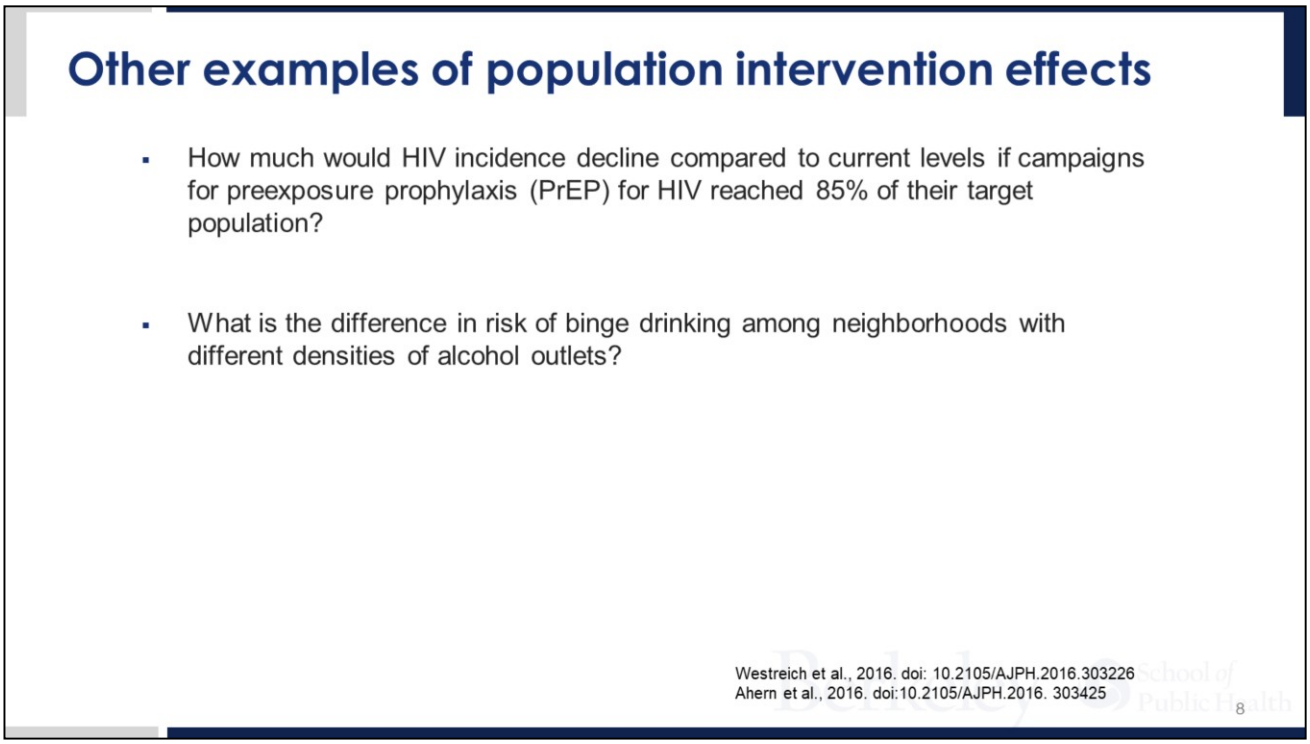
We can also think about how realistic these different parameters are. So comparing a scenario in which everyone received SHEWA-B to a scenario in which no one received SHEWA-B is the least realistic scenario we can come up with. And the population intervention effect scenarios are the most realistic. And the population attributable fraction scenarios fall somewhere in between.

So moving from left to right, for the risk difference, it's not realistic because UNICEF never expected to deliver the program to 100% of the target population. And so RE is not realistic. And similarly, R sub U is also not realistic, because we're really not interested in a situation where nobody is receiving the program.

Remember, their specific question after the interim evaluation was why the program wasn't working at that point. And so they're already at 40% coverage. They want to know how impactful would the program have been if it had been delivered to 80% instead.

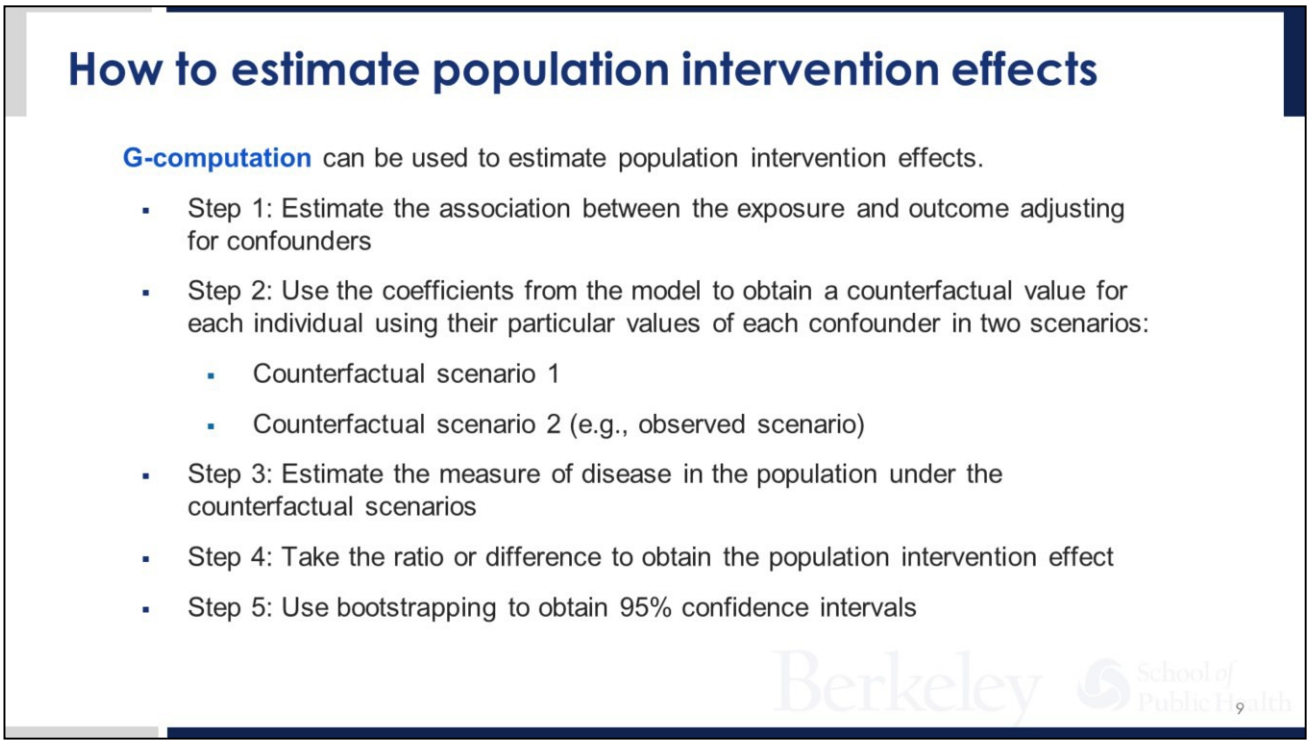
Looking at the center, at the population attributable fraction, the RT was equivalent to RO. And so this parameter is comparing the current level of exposure to a scenario where no one's exposed. And that's, again, not really of interest based on the question that they asked in the interim evaluation.

The last parameter, population intervention effect, is most useful because it provides an estimate of the impact the program could have had if it reached its target level of coverage compared to the level of coverage it actually achieved. So in other words, the population intervention effect parameter is telling us how much of a difference in health illness we would expect to see if the program had had ideal coverage levels compared to observed coverage levels.



Here's two other examples of population intervention effects. One is, how much would HIV incidence decline compared to current levels if campaigns for pre-exposure prophylaxis for HIV reached 85% of their target population? So pre-exposure prophylaxis involves taking chemotherapy preventively to reduce the risk of contracting HIV. And there's numerous campaigns in place to promote this type of prophylaxis among people at high risk of contracting HIV. And so this parameter is assessing how much greater the decline in incidence would be if the campaign reached 85% of their target population compared to the current percentage that they reach.

Another question, what's the difference in risk of binge drinking among neighborhoods with different densities of alcohol outlets? So you can define different counterfactual scenarios based on different densities of alcohol outlets-- very low and very high densities-- and assess how the risk of binge drinking varies, depending on these densities. And when we say densities, what we mean is in a local area, how close together and how many alcohol outlets are in place? These are places that sell alcohol.



We can use G-computation to estimate population intervention effects. In the first step, we estimate the association between our exposure and outcome, adjusting for confounders. In the second step, we use the coefficients from the model to obtain a counterfactual value for each individual, using their particular values of each confounder under two scenarios. And this is where it's a bit different from what we discussed in the previous G-computation video.

Here, we just defined two different counterfactual scenarios. And often, the second one is an observed scenario. So it's our observed level of exposure. And in the first scenario, it might be an ideal or realistic level of exposure.

Then in the third step, we estimate the measure of disease in the population under each scenario. In the fourth step, we take the ratio or difference to obtain our population intervention effect. And in the fifth step, we use bootstrapping to obtain our confidence intervals.

Summary of key points

- Studies often contrast two exposure definitions: 1) if everyone was treated and 2) of no one was treated
- However, real-world effects are likely to differ if it is unlikely that an entire population will receive or not receive the intervention or exposure.
- Population intervention effects are parameters that compare two counterfactual scenarios
- These parameters can be used to estimate effects with more realistic and policy-relevant counterfactual scenarios.

To summarize, in epidemiology the default often is to compare two exposure definitions, one where everyone is treated and one where no one's treated. This is the default, even in the causal inference methods you've learned with G-computation and IPTW. However, real-world effects are very often different because it's unlikely the entire population will receive or not receive an intervention or exposure. And so these population intervention effects are really nice, because they allow us to define customized counterfactual scenarios that are tailored to our research question, and thus, I tend to be able to provide us with more realistic and policy-relevant evidence than traditional parameters for certain kinds of research question.

HHS Public Access

Author manuscript

Am J Public Health. Author manuscript; available in PMC 2017 June 01.

Published in final edited form as:

Am J Public Health. 2016 June ; 106(6): 1011–1012. doi:10.2105/AJPH.2016.303226.

Causal impact: epidemiologic approaches for a public health of consequence

Daniel Westreich[1], **Jessie K. Edwards**[1], **Elizabeth T. Rogawski**[2], **Michael G. Hudgens**[3], **Elizabeth A. Stuart**[4], and **Stephen R. Cole**[1]

[1]Department of Epidemiology, Gillings School of Global Public Health, UNC-Chapel Hill, Chapel Hill, NC

[2]Division of Infectious Diseases and International Health, University of Virginia, Charlottesville, VA

[3]Department of Biostatistics, Gillings School of Global Public Health, UNC-Chapel Hill, Chapel Hill, NC

[4]Departments of Biostatistics, Health Policy and Management, and Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

The *causal impact framework* is a conceptual framework encompassing internal validity, external validity, and population intervention effects, which we argue can help us produce evidence of greater utility to public health decision-making.

To improve the health of a population, public health research should consider factors that can be changed, particularly exposures that are potential targets of intervention. Thus, useful public health research will focus on identifying causes of health and disease, rather than simply associated risk factors. In a trial setting, we are typically interested in contrasting two exposure distributions: (i) what if everyone were assigned to the experimental treatment, and (ii) what if everyone were assigned to the comparison treatment? Such causal public health (and clinical) research designs generally focus on ensuring internal validity(1): generating an accurate estimate of a causal effect for the people in the study, such as obtained from a double-masked randomized controlled trial with no loss to follow-up.

This focus on internal validity informs the conduct and reporting of randomized trials and the framing of observational data analysis. Yet, if our goal is to understand the potential impact of a specific public health policy in a real population, establishing internal validity is merely a first step; we must also consider external validity and population intervention impact, which together we describe as the causal impact framework.

A motivating example

Here we focus on generating evidence for public health action beginning with an individually randomized trial (or observational study conducted in place of a trial). In particular consider a randomized trial of antiretroviral agents for pre-exposure prophylaxis

Correspondence to: Daniel Westreich, CB 7435 McGavran-Greenberg Hall, Chapel Hill, NC 27599.

for HIV prevention (PrEP), in which an active arm is compared to a placebo in a group of HIV-negative volunteers who are followed up through regular clinic visits to identify new HIV infections. With perfect follow-up, a comparison of unadjusted survival curves between the active and placebo trial arms will yield a valid estimate of the causal effect of assignment to PrEP on risk of HIV acquisition in the study sample(2).

The difference in outcomes between participants *assigned* to each treatment arm, allowing for differential adherence, is referred to as the effectiveness of the treatment. We often prioritize effectiveness of an intervention strategy over its efficacy (effect under perfect adherence or in laboratory settings), because in the real world we cannot force people to adhere to an intervention. Accordingly, effectiveness is sometimes referred to as the public health effect of the treatment. However, further work is needed to translate effectiveness to an estimate of the impact of a real intervention in a real population.

External validity

An essential ingredient of the causal impact framework is the ability to generalize or transport to populations beyond that under study: thus we must consider possible differences in characteristics between our study sample and our *target population*, the population in which we want to implement the intervention(3). If the study sample is a random sample of the target population, external validity is guaranteed in expectation. However, study samples and target populations nearly always differ systematically due to factors such as inclusion/exclusion criteria in a trial and self-selection into studies through the informed consent process. These differences may modify the impact of the intervention in the target population, leading to external validity bias. In the trial described above, if PrEP is more effective at preventing HIV in women than in men, and if women are more likely to participate in the trial than men, then PrEP may appear more effective in the trial than it would be in the target population.

In such a simple case, the effectiveness of PrEP in the target population can be estimated by standardization. But if the effectiveness of the intervention varied due to a complex combination of several variables (e.g., sex, age, race) the joint distribution of which differed between the study sample and the target population, then model-based strategies are recommended(3). Of course, if one or more of those variables were unmeasured, no quantitative approach would be guaranteed to provide an accurate estimate of the true population effect (a close parallel to the problem of uncontrolled confounding in an observational study). There is a growing literature on quantitative methods to “generalize” results from a study sample to the population from which the sample was selected(3), and to “transport” results from a study sample to a different population entirely(4). However, even if we successfully estimate the effectiveness of a proposed intervention in a target population, in the causal impact framework we must make efforts to understand effect of the intervention under real-world conditions.

Population intervention effects

As noted above, in our trial setting, we are likely contrasting two exposure distributions: (i) what if everyone were assigned to PrEP, and (ii) what if everyone were assigned to placebo? But the real world effects of a population-level PrEP intervention (for example, a country-wide policy to promote PrEP to all people who meet certain criteria) will likely differ from what is estimated in a specific trial. On one hand, not everyone will be targeted by PrEP campaigns, nor will everyone targeted choose to take the treatment, and adherence may be better in a trial due to Hawthorne effects. On the other hand, preventing a single HIV infection with PrEP may prevent subsequent transmission events, a dependency not likely captured in a small study sample. Finally, implementation challenges may lead to adaptation of the intervention when scaling up from a study to a population.

Non-experimental settings raise additional challenges: observational public health research often focuses on effects of harmful exposures, rather than on interventions to limit such exposures. Using the results of a study of the effect of a harmful exposure (e.g. smoking) to estimate the potential effect of a population intervention to reduce prevalence of that exposure (e.g. mass campaign for smoking prevention) requires articulation of assumptions about the intervention in question and its side-effects(5, 6) and careful estimation (possibly using the g-methods of Robins as in(6)). Population intervention effects(5, 7), which can be thought of informally as causal effects tied to contrasts between the observed population and exposure distributions under realistic interventions, are of key importance to the causal impact framework goal of translating scientific results into policy-relevant findings, and may serve as more natural inputs into cost-effectiveness and decision-theoretic models than typically reported study results.

Remarks

There are numerous approaches for estimation of policy-relevant public health effects, including large-scale representative and pragmatic randomized trials and comparative interrupted time series(1). Despite calls for wider adoption of these methods(1), traditional randomized trials and non-experimental studies remain central to the production of evidence for public health practice. Such studies are valuable, but typically focus centrally on questions of internal validity, ignoring external validity and population intervention impact. Considering all three, as in the causal impact framework, may help us produce research more relevant to policy-making, and thus helping to produce a Public Health of Consequence.

Acknowledgments

Funding from: DP2-HD-08-4070

References

1. Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J Public Health. 2004; 94(3):400–5. [PubMed: 14998803]
2. Murnane PM, Brown ER, Donnell D, Coley RY, Mugo N, Mujugira A, et al. Estimating efficacy in a randomized trial with product nonadherence: application of multiple methods to a trial of

preexposure prophylaxis for HIV prevention. Am J Epidemiol. 2015; 182(10):848–56. [PubMed: 26487343]

3. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. Am J Epidemiol. 2010; 172(1):107–15. [PubMed: 20547574]
4. Hernán MA, Vanderweele TJ. Compound treatments and transportability of causal inference. Epidemiology. 2011; 22(3):368–77. [PubMed: 21399502]
5. Fleischer NL, Fernald LC, Hubbard AE. Estimating the potential impacts of intervention from observational data: methods for estimating causal attributable risk in a cross-sectional analysis of depressive symptoms in Latin America. J Epidemiol Community Health. 2010; 64(1):16–21. [PubMed: 19643766]
6. Westreich D. From exposures to population interventions: pregnancy and response to HIV therapy. Am J Epidemiol. 2014; 179(7):797–806. [PubMed: 24573538]
7. Hubbard AE, Laan MJ. Population intervention models in causal inference. Biometrika. 2008; 95(1):35–47. [PubMed: 18629347]

Berkeley School of Public Health

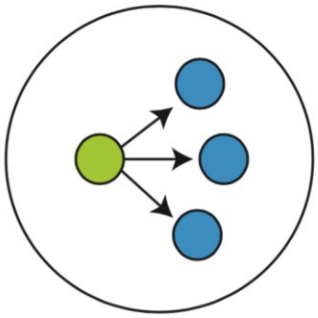
Spillover effects

PHW250 B – Andrew Mertens

Andrew Mertens: This video will talk about spillover effects.

Definition of spillover effects

A spillover is the effect of an intervention on people not targeted by an intervention but who were connected to intervention recipients socially, geographically, or by some other means.



Berkeley School of Public Health

1

Let's start with a definition of spillover effects. A spillover is the effect that an intervention has on people who are not targeted by it, but who are connected to intervention recipients. And that connection might be a social connection, a connection through geographic proximity, or through some other means. If each of the small circles in this figure represents a person, and the green person receives an intervention, and the blue people did not, a spillover effect would look at the health impacts on the blue circles. On the people who did not receive the intervention, but who are in some way connected to the person who did receive the intervention.

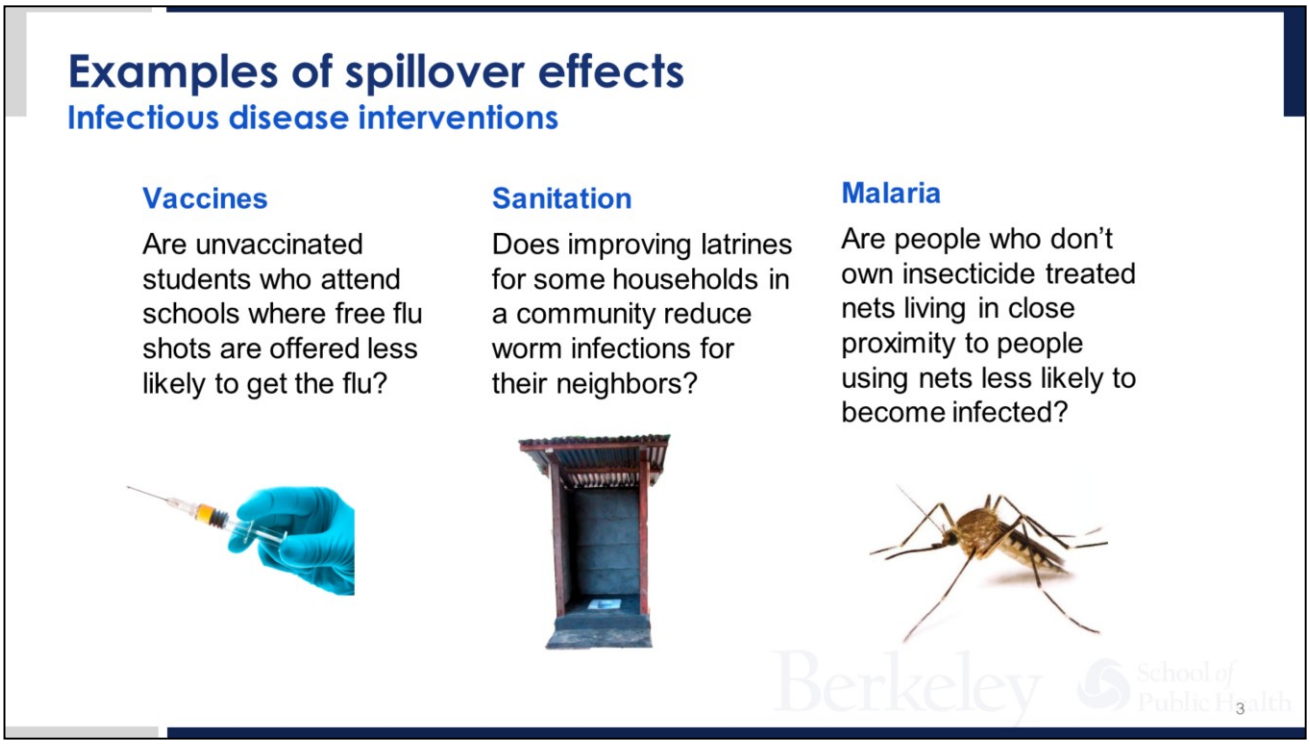
Motivation

- In most epidemiologic studies that are not focused on infectious diseases, we assume that individuals' outcomes are not correlated.
- In infectious disease studies we cannot assume that individuals are independent because of the transmissible nature of disease.
  - The presence of a spillover effect reflects this transmission — an individual who did not receive an intervention was impacted by it.
- In some non-infectious disease studies, it might not be valid to assume individuals' outcomes are not correlated:
  - What if the behaviors of some people affect those of others?
  - What if attitudes of some people affect those of others?

I'd like to provide a little more motivation before we go on. Generally in epidemiologic studies that are not focused on infectious diseases, we make a key assumption that the outcomes if individuals are not correlated with each other. So this means that if I'm conducting a study in my neighborhood of alcohol consumption practices, I'm assuming that my likelihood of alcohol consumption is not related to the likelihood of alcohol consumption of my neighbors or other people in my neighborhood. And in infectious disease studies, it's well known that we can't make this kind of assumption because of the transmissible nature of disease.

So, if a spillover effect is present, it's reflecting some form of transmission. Whether that be a pathogen that's spread between people or a behavior that's passed on from one person to another. Spillovers are present when individuals who did not receive an intervention are impacted by that intervention.

So, in some cases in non infectious disease studies, it might not be valid to make this typical assumption that we make. What if the behavior of some people affect those of others? What if the attitudes of some people affect those of others? An intervention can change the behaviors and attitudes of intervention recipients and in turn can affect the attitudes and behaviors of non recipients who come into contact with intervention recipients.



So, here are some examples of spillover effects for infectious disease intervention. The most classic one is for vaccines. You may have heard of herd immunity. Well, herd immunity is directly related to the concept of spillover effects. Here's an example of a vaccine-related research question. Are unvaccinated students who attend schools where free flu shots are offered less likely to get the flu? And that would occur if having the free program at a school greatly increases vaccination coverage and thus, if the vaccine is effective, reduces the amount of flu or influenza that's being transmitted at the school to such a small level that even students who are unvaccinated can indirectly benefit from the program.

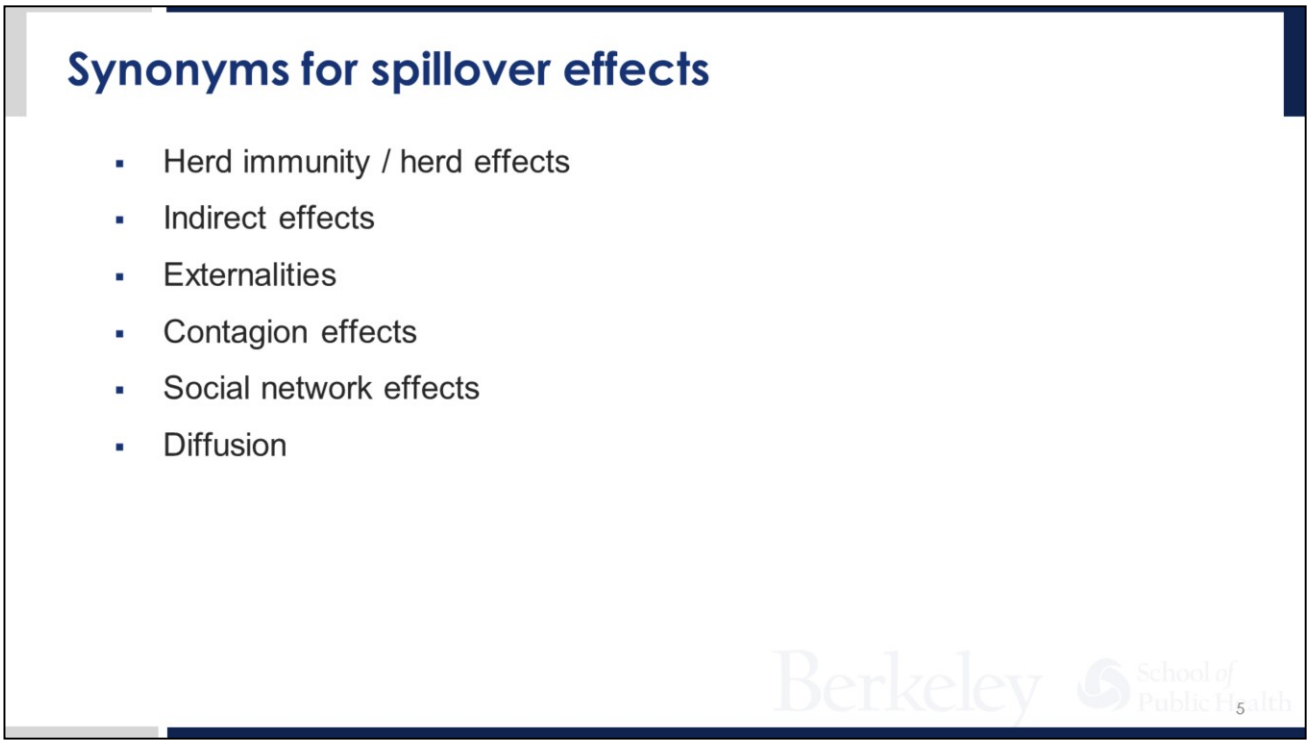
An example from sanitation. Does improving latrines for some but not all households in a community reduce the incidence of worm infections for the neighbors of people who got improved latrines? So, this is because environmental contamination affects the risk of contracting a worm infection, and a sanitation intervention might improve the environment, not only for household members who receive the improved latrine, but also in the direct vicinity around that latrine. And thus, there may be impacts on neighbors.

And then for malaria, we could ask, are people who don't own insecticide treated nets but live in close proximity to people using nets less likely to become infected? Now, if people are using insecticide treated nets, that means they're less likely to contract malaria, which is transmitted by mosquitoes. And if the amount of malaria circulating in a population goes down because some people are using nets, that means that people who don't use nets may also indirectly benefit.

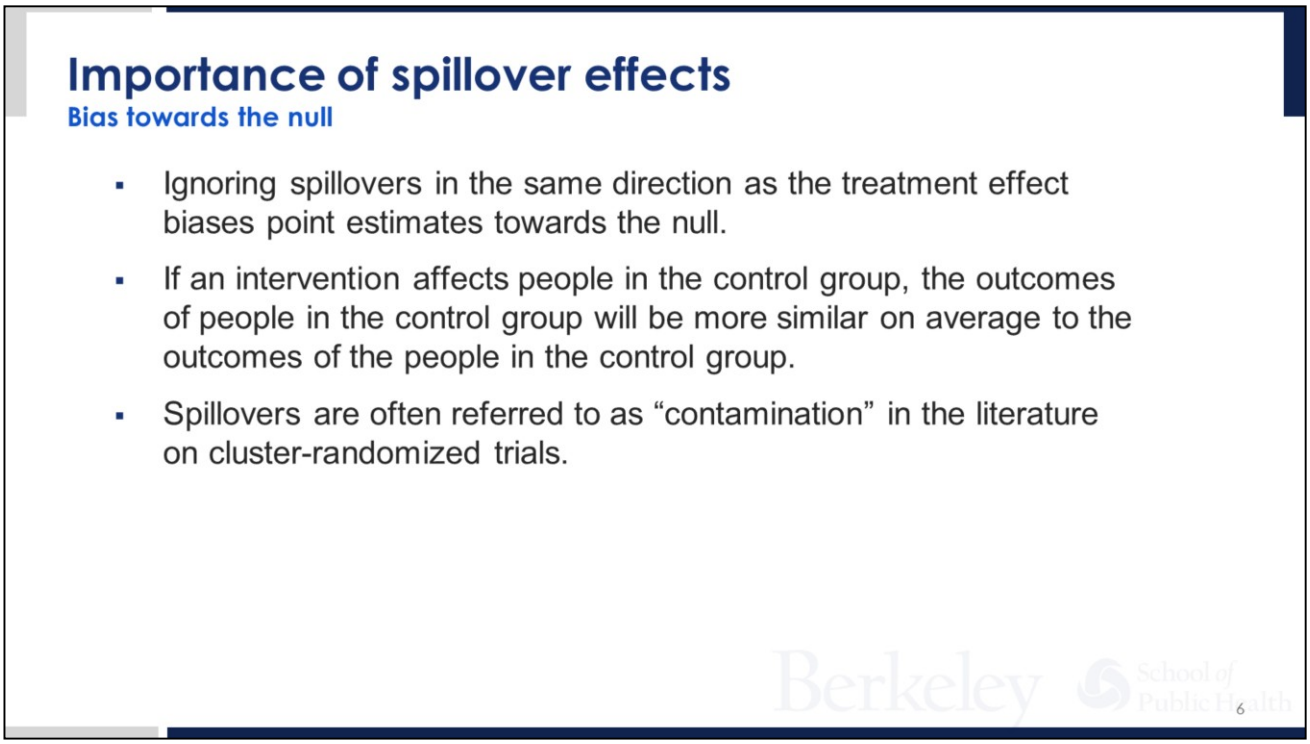
Now let's go through some examples for other types of interventions that don't focus on infectious diseases. We could look at a smoking cessation program. Are friends of participants in a smoking cessation program more likely to quit smoking than people who don't know anyone and a smoking secession program? And it's possible that by participating in this program, a person might learn a lot about smoking and learn some strategies for quitting and share that information with their friends, and so their friends may be more likely to quit as well.

The second example focuses on obesity. And this isn't quite the same as a spillover effect, but I'm including it here because it gets at this concept of diffusion. And there was a very well-known paper that came out that looked at whether if the majority of people were obese, are you more likely to be obese? And it found that that tended to be true in the United States. So this relates to this concept called diffusion, which is a form of a spillover effect. There isn't a particular intervention here, per se. But it's an example of how the health status of certain individuals affects that of others, even if it's not an infectious disease.

And then finally, for women's peer groups. Are women whose peers participate in peer support groups during pregnancy less likely to experience adverse events during their delivery? So if the friends or peers if people who are in these groups but did not participate in the groups are less likely to have adverse events, then that suggests that perhaps there was some knowledge that was shared or change in practices that occurred through a social connection with program participants.

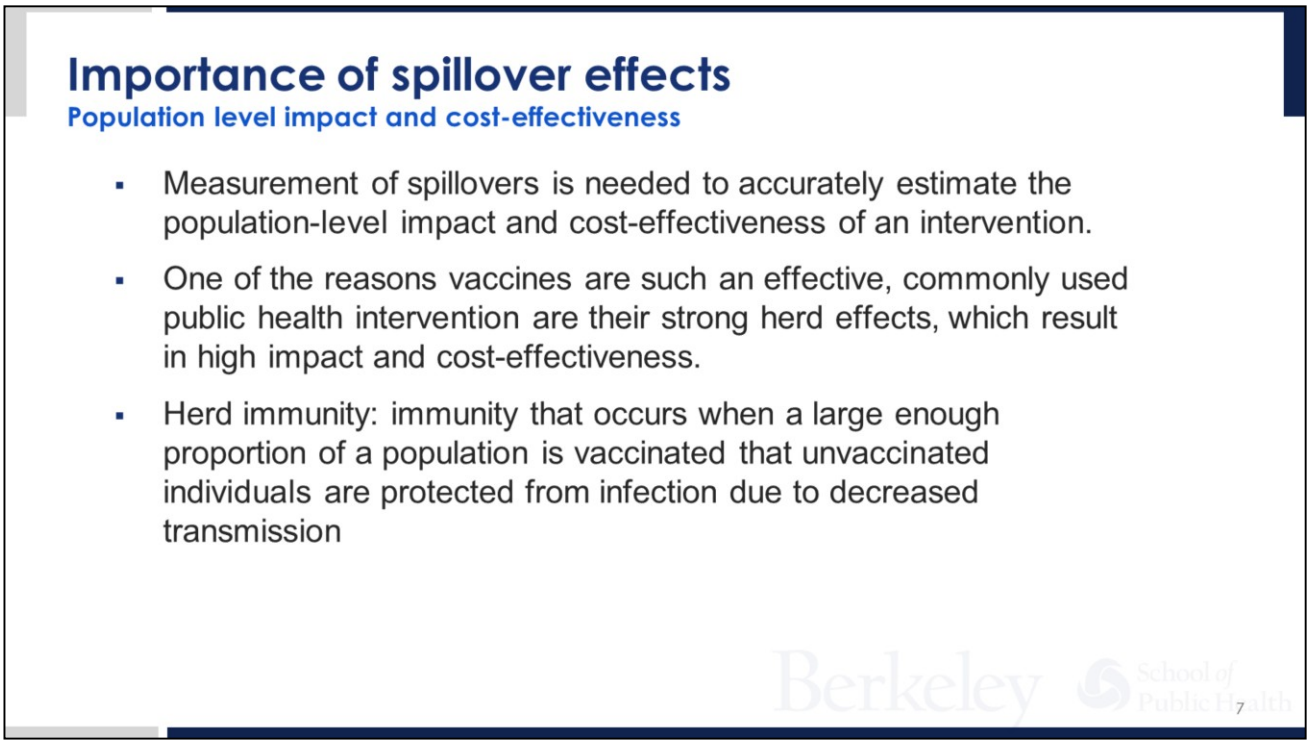


There are a lot of different terms that have been used for spillover effects. These include herd immunity, herd effects, indirect effects, externalities, contagion effects, social network effects, and diffusion. Externalities is commonly used in the economics literature. The vaccine literature tends to use the first two names. We're going to use spillover effects in this course.

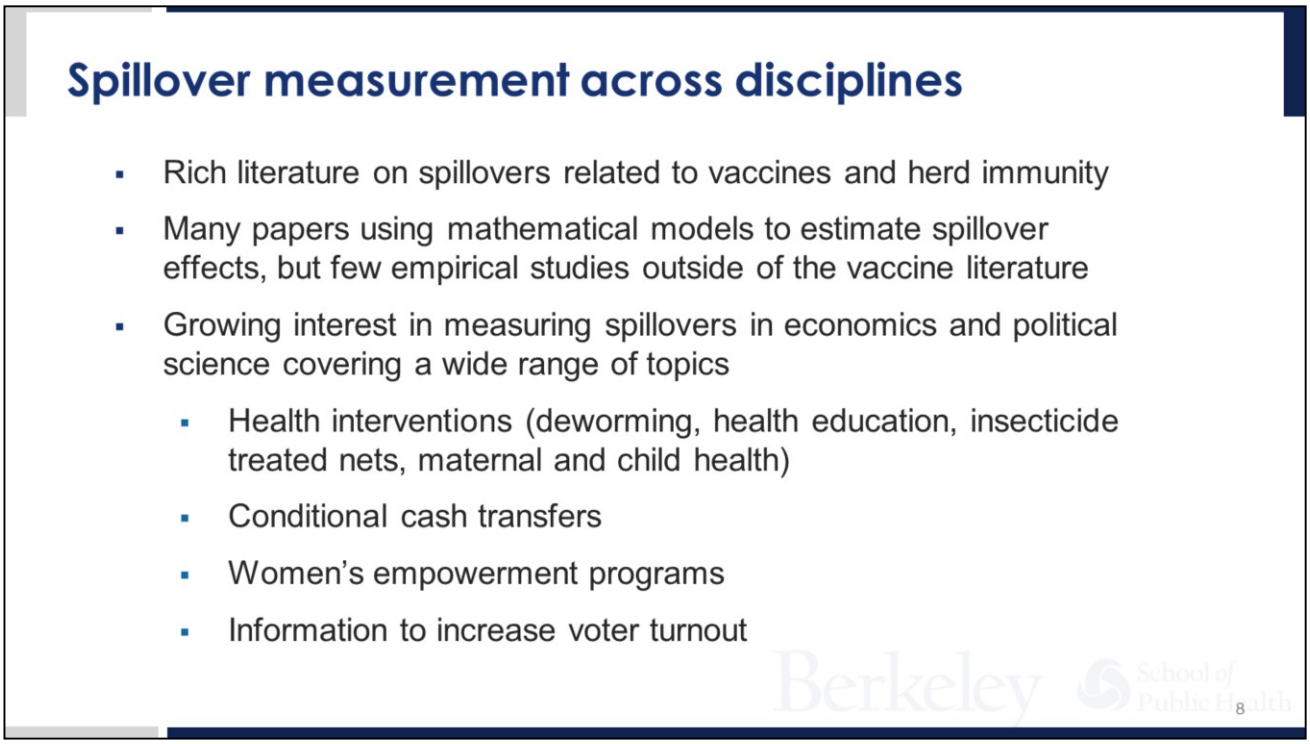


Why are spillover effects important? The first reason is that if we ignore them, it will bias our results towards the null, or bias our interpretation of our findings towards the null. So if an intervention affects people in the control group, for example, who were not receiving the intervention, that means that the outcomes of people in the control group are more similar, on average, to the outcomes of people in the intervention group.

When you learned about cluster-randomized trials, you learned about this concept of contamination. And in that context, this is something we're trying to minimize. So if the control group is experiencing a spillover effect, that means that it can no longer serve as a valid counterfactual for the intervention group because it was also impacted by intervention. And we say that contamination has occurred and that case. And when contamination occurs, the effect that we estimate is closer to the null than the one we would have estimated if we had had a pure control. So this is the first reason, is this bias towards the null.



Another reason spillover effects are important is related to their population level impact and how that affects cost effectiveness estimates for an intervention. For example, if we can vaccinate 80% of people but benefit 100% of people through spillover effects that break transmission and prevent unvaccinated people from becoming ill, we can save the cost that we would have spent to deploy the vaccine to the additional 20%, which means we have a more cost effective intervention. And we're able to achieve a population impact on 100 percent of people by only delivering an intervention to 80% of people. This is particularly important for expensive interventions such as sanitation. There's been a lot of interest in whether household sanitation interventions can produce spillover effects because if that's true, it really helps justify expanding improvements to household sanitation, which is very costly to implement.

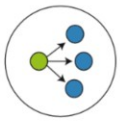


So, spillover effects have been measured in different disciplines and for different topics. The richest literature on them is related to vaccines and herd immunity. And there have been many papers that have used models, mathematical models, to estimate spillover effects but fewer empirical studies of spillover effects outside of the vaccine literature. And when we say empirical studies, we mean studies where we go and collect data in the real world. We're not just relying on our theory and using a model to estimate effects, we're using real data.

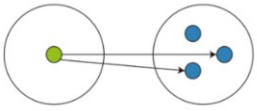
But there's been growing interest in measuring spillover effects in a variety of different disciplines, including economics, political science, as well as public health. And these have looked at health interventions such as deworming, health education, insecticide treated nets, and maternal and child health programs, conditional cash transfers, women's empowerment programs, and providing information to try to increase voter turnout. So there have been a really wide range of studies looking at spillover effects of different types of interventions.

Types of spillover effect parameters

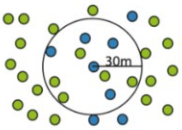
Within-cluster spillover effect



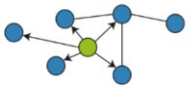
Spillover effect conditional on distance



Spillover effect conditional on treatment density



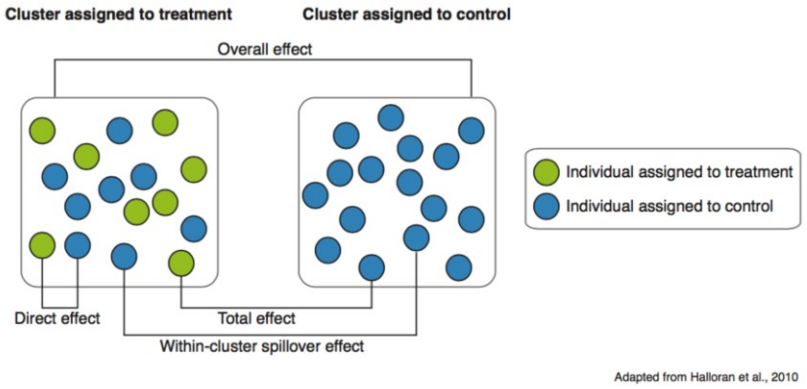
Network spillover effect



In this video, I'll introduce you to four different kinds of spillover effects parameters. There are more types than this, but these are some of the most common ones. First is within-cluster spillover effect. Then there's a spillover effect that conditions on distance, spillover effect conditional on treatment density, and a network spillover effect.

In this video, I'll be using y to indicate the outcome or disease, and X to indicate a cluster treatment assignment, and t to indicate individual treatment assignment unless otherwise specified.

Within-cluster spillover effects



Adapted from Halloran et al., 2010

- Appropriate when spillover effects are expected to only occur within clusters.
- **Example:** spillover effect of a sanitation intervention delivered to some but not all households in a village
- Critical assumption: no spillover effects between clusters

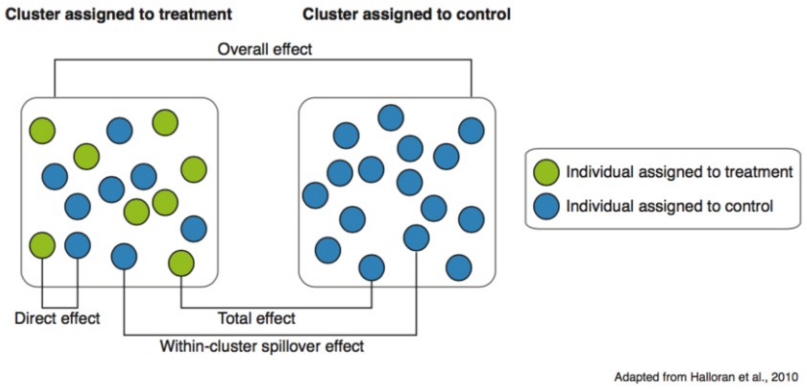
11

Let's start with within cluster spillover effects. The large circles indicate cluster treatment. So this large circle on the left is a treated cluster, and this large circle on the right is a controlled cluster. And then within each of these clusters, the small circles indicate people. So this is a cluster randomized design. The green circles are individuals who are assigned to treatment or who received treatment. And the blue circles are individuals who were assigned to control or who did not receive the intervention.

Now let's look at the different types of effects that we can estimate in a cluster randomized trial. The overall effect is an effect that is most commonly estimated, and it's the one we've emphasized so far in this course. It compares the risk or prevalence among all individuals in the cluster assigned to treatment to that of all individuals in the cluster assigned to control. Regardless of the individual level treatment assignment.

The direct effect compares treated and untreated individuals only within the treatment cluster. The total effect compares the effect on treated individuals in the treated cluster to individuals in the control cluster. And if we want to measure spillover effects within clusters, we compare untreated individuals in the treatment cluster to those in the control cluster. These blue circles here are individuals who are in the cluster assigned to treatment but who did not get treatment. And if the treatment affects others inside the cluster, we would want to measure the outcomes of these individuals and compare them to those in control.

Within-cluster spillover effects



Adapted from Halloran et al., 2010

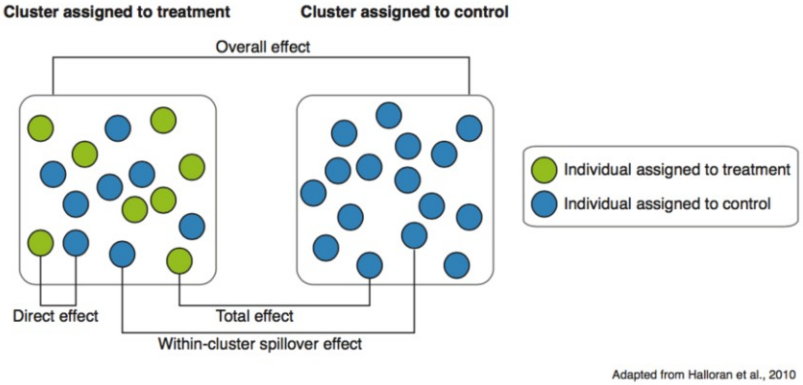
- Appropriate when spillover effects are expected to only occur within clusters.
- **Example:** spillover effect of a sanitation intervention delivered to some but not all households in a village
- Critical assumption: no spillover effects between clusters

11

So this is an appropriate type of spillover effect to estimate when it's expected that spillover effects will only occur within clusters not between clusters. An example of this would be the spillover effect of a sanitation intervention that was delivered to some but not all households in a village. We could look at the incidence of diarrhea, for example, among those who didn't get the sanitation intervention but live in a cluster where others did, to the incidence among those in a control cluster.

A critical assumption in order to estimate this type of spillover effect is that the spillover effect is confined within the cluster. That there is no contamination between clusters. In other words, the intervention cannot affect people in the control clusters. If it does, then the control cluster no longer serves as a valid comparison group or a valid counterfactual for the treatment cluster.

Within-cluster spillover effects

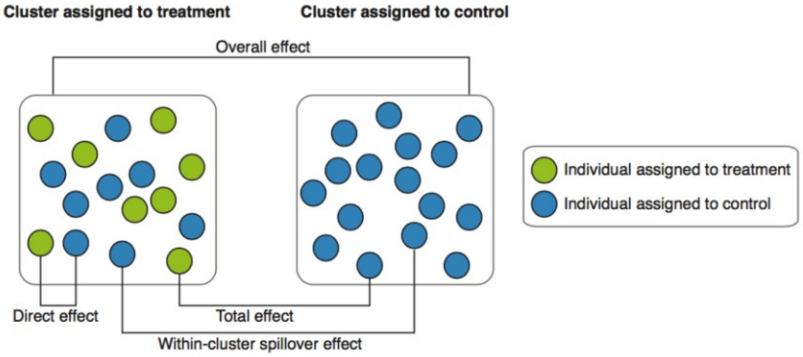


Adapted from Halloran et al., 2010

Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

I'm going to just quickly talk you through the notation for how we can estimate these different types of effects. So again, y is the outcome an x is the cluster treatment assignment. So the overall effect compares the mean outcome among those in a cluster assigned to treatment compared to the mean outcome among those in a cluster assigned to control. I'm showing this to you as a difference in means. It could also be a difference in risk. It could also be a relative risk. The key part in this equation is really what we're comparing, not whether it's on the additive or relative scale and not the measure of disease.

Within-cluster spillover effects



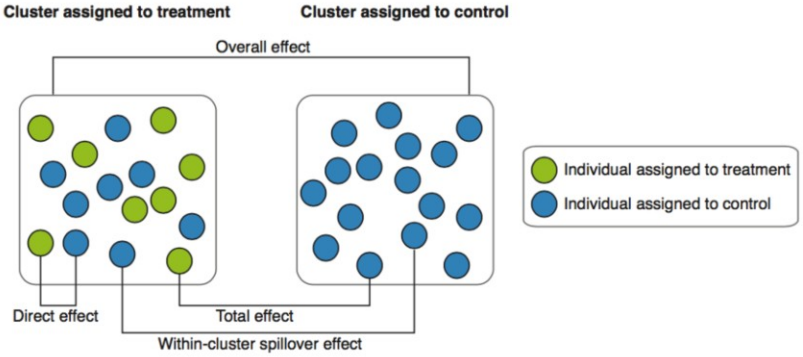
Adapted from Halloran et al., 2010

Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

Total effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 0, T = 0]$

The total effect compares the mean outcome, y, among those in the cluster assigned to treatment, x equals 1, who received the treatment themselves, t equals 1, compared to the mean outcome, y, among those in the control cluster, x equals zero, who were not treated, t equals zero.

Within-cluster spillover effects



Adapted from Halloran et al., 2010

Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

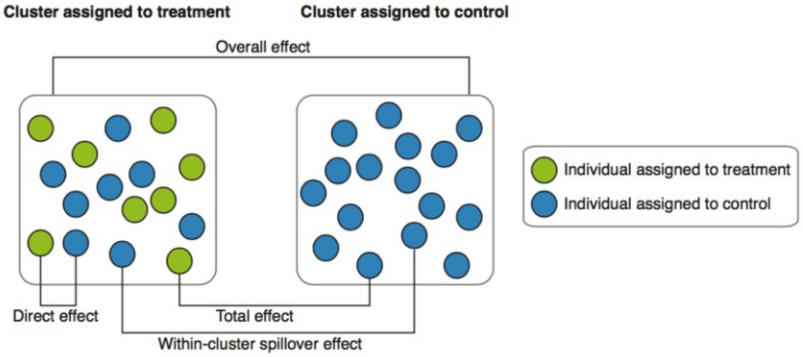
Total effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 0, T = 0]$

Direct effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 1, T = 0]$

14

The direct effect compares the mean outcome among those in the cluster assigned to treatment, x equals 1, who were treated, t equals 1, to the mean outcome among those in the treatment cluster indicated by x equals 1, were not treated, t equals zero.

Within-cluster spillover effects



Adapted from Halloran et al., 2010

Overall effect: $E[Y \mid X = 1] - E[Y \mid X = 0]$

Total effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 0, T = 0]$

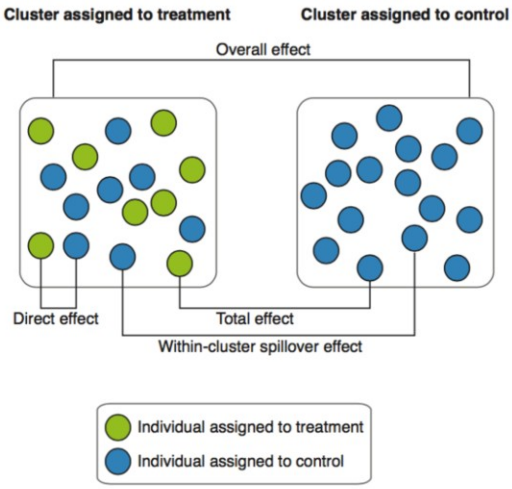
Direct effect: $E[Y \mid X = 1, T = 1] - E[Y \mid X = 1, T = 0]$

Within-cluster spillover effect: $E[Y \mid X = 1, T = 0] - E[Y \mid X = 0, T = 0]$

15

And the within cluster spillover effect compares the mean outcome, y, among those in the treatment cluster, x equals 1, who were not treated, t equals 0, to the mean outcome among those in the control cluster who are not treated, x and t equals zero.

Double-randomized trials



- Double-randomized trials are the most rigorous design for estimating within-cluster spillover effects.
  - Randomize clusters to treatment vs. control
  - Randomize individuals in the treatment arm to treatment vs. control
- Measured and unmeasured confounders are balanced:
  - between study arms
  - between individuals in the treatment arm

16

The ideal design to estimate within cluster spillover effect is a double randomized controlled trial. In this type of trial, we first randomize clusters to treatment or control. And then within the treatment arm, we randomize individuals in treatment cluster to treatment versus control. So sometimes this design is called a two stage randomized design because the randomization is in two stages.

The beauty of this design is that it can obtain balance in both measured and unmeasured confounders, not only between study arms, but also between individuals in the treatment arm. And that means that we can use randomization based inference to obtain valid direct effects and valid total and within cluster spillover effects. When I say valid, what I mean is unbiased. Assuming that we did proper randomization and that no other forms of bias occurred. So this is a really powerful design.

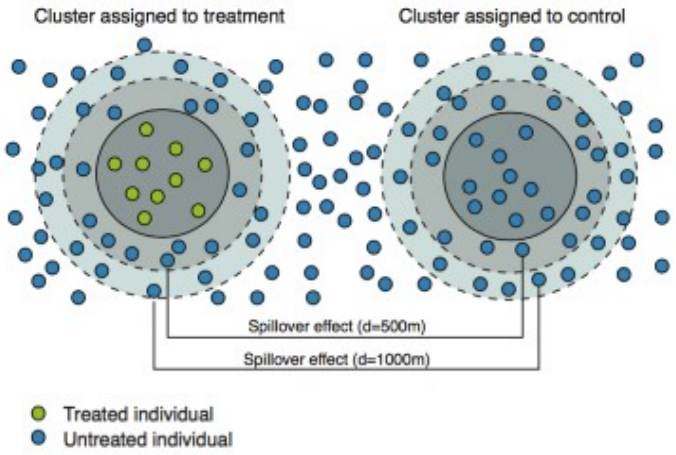
Typically we don't do the second step of randomization of individuals in the treatment arm, and we just average over the outcomes of individuals in that arm. And so this is an improvement upon that design, because there may be systematic differences between people who choose to use an intervention or sign up to be in the trial, and those who do not. This design prevents us from having those systematic differences between treated and untreated in the treatment arm.

Spillover effects and contamination

- When conducting cluster-randomized trials, the goal is to include enough distance between clusters to prevent “contamination”.
- It is important to minimize this type of contamination when estimating within-cluster spillover effects.
- But what if this contamination is of interest?
  - We can estimate spillover effects conditional on distance.

So as I mentioned earlier when we learned about cluster randomized trials, a key goal in this type of design is to include enough distance between clusters that we prevent contamination. And as I mentioned, it's important to minimize contamination when estimating within cluster spillover effects. But what if we're interested in this type of contamination? In other words, if there is an effect beyond the border of a cluster, we might be interested in that. And we can use spillover effects conditional on distance to assess this type of effect.

Spillover effect conditional on distance



Spillover effect conditional on distance (*D*):

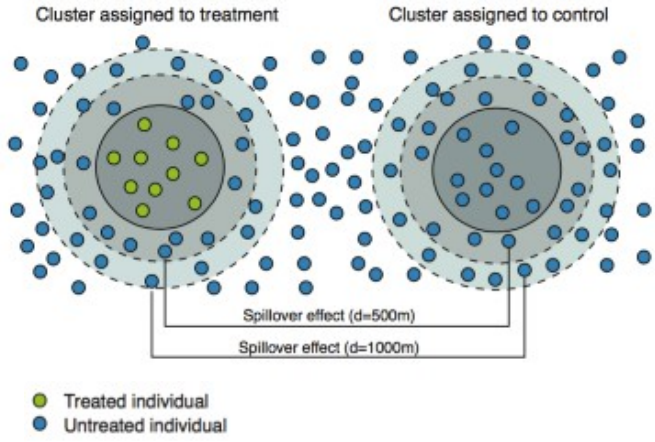
$E[Y \mid X = 1, T = 0, D = d] - E[Y \mid X = 0, T = 0, D = d]$

- Can compare magnitude of spillover effects over a distance gradient
- Appropriatewhen spillover effects are expected to extend beyond study clusters
- Important to maintain enough distance between treatment and control clusters that individuals outside of control clusters can serve as a valid counterfactual
- **Example:** Distance from clusters where everyone in the cluster received an insecticide treated bed net to preventmalaria

In this figure, the inner circle on the left is a treated cluster, and the inner circle on the right is a control cluster, and then the dashed lines around these circles indicate radii in which we conduct measurements of untreated individuals who live around the periphery of these study clusters. So within the first radius, the individuals live within 500 meters of the cluster boundary, and within the second radius, they live within 1,000 meters of the cluster boundary. And the spillover effect conditions on distance which is denoted by d. And then the parameter is defined as the mean outcome, y, among those near the treatment cluster, x equals 1, who are not treated, t equals zero, within a certain distance of that treatment cluster, D equals d, to the mean outcome among those nearer control cluster, x equals zero, who are not treated, t equals zero, within that same distance radius, D equals d.

So, you can see how we could actually look at this parameter over a gradient of distance. We could compare the effective living within 500 meters of a treatment cluster to that of living within 1,000 or 2000 meters of a treatment cluster to see if there's a gradient of spillover effects over distance. So this is an appropriate type of parameter to estimate when we expect spillover effects to extend beyond study clusters. But it's still important that we maintain enough distance between the treatment and control clusters that individuals outside control clusters can serve as a valid counterfactual for individuals outside treatment clusters.

Spillover effect conditional on distance



Spillover effect conditional on distance (*D*):

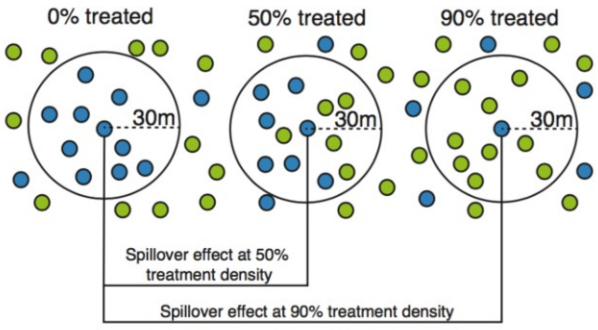
$E[Y \mid X = 1, T = 0, D = d] - E[Y \mid X = 0, T = 0, D = d]$

- Can compare magnitude of spillover effects over a distance gradient
- Appropriatewhen spillover effects are expected to extend beyond study clusters
- Important to maintain enough distance between treatment and control clusters that individuals outside of control clusters can serve as a valid counterfactual
- **Example:** Distance from clusters where everyone in the cluster received an insecticide treated bed net to preventmalaria

22

So in this design, the clusters need to be even further apart if we expect the spillover effects to occur over greater distances than compared to a within cluster design. An example of a spillover effect conditional on distance is looking at whether the distance from clusters where everyone in the cluster received an insecticide treated bed net to prevent malaria can reduce the incidence of malaria among those who live on the periphery of that cluster. A high level of net coverage could reduce the local incidence of malaria and benefit people who live around that cluster.

Spillover effect conditional on treatment density



Treated individual
Untreated individual

Spillover effect conditional on treatment density (*P*):

$E[Y \mid T = 0, P = p + \delta] - E[Y \mid T = 0, P = p]$

- This example is shown for an individually randomized trial.
- Can also be used with a cluster randomized design if the proportion
- **Example:** Assess spillover effect on malaria associated with percentage of nearby households with an insecticide treated net
- Can compare magnitude of spillover effects over a gradient of treatment density

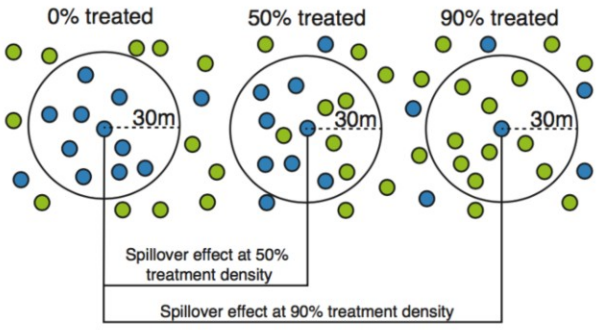
19

The next parameter we're going to talk about is the spillover effect conditional on treatment density. In this figure, we're assuming that this was a individually randomized trial. And so the spatial configuration of treatment and control, as indicated by green and blue circles, should be randomized. And then we can look at the local density of treatment within a certain distance radius.

So in this example, it's within 30 meters of an untreated individual. And on the far left, we see that no one within 30 meters was treated. So 0% were treated. And in the center, 50% of individuals within 30 meters were treated. And on the right 90% were treated. And we can compare the outcome among individuals who were untreated at 50% treatment density to 0%, and we can compare the density at 90% to those at 0%.

And so the parameter is written here in the bottom. It's the mean outcome, y, among those who were not treated, t equals zero. And then the treatment density is denoted by p. And so what we're doing is we're comparing those at different treatment density thresholds. So we have on the right-hand side of this equation, P equals p could be 0%. And then if our difference that we're interested in is 50%, delta would be equal to 50%. So that means we're comparing untreated individuals at a treatment density of 0% to those at 50% treatment density.

Spillover effect conditional on treatment density



Treated individual
Untreated individual

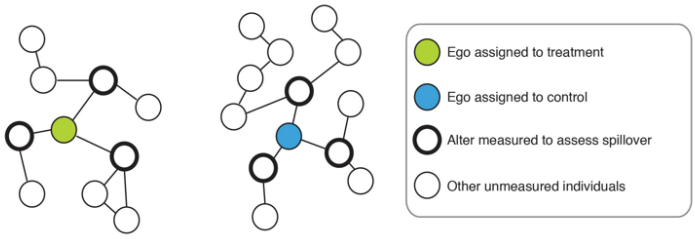
Spillover effect conditional on treatment density (*P*):

$E[Y \mid T = 0, P = p + \delta] - E[Y \mid T = 0, P = p]$

- This example is shown for an individually randomized trial.
- Can also be used with a cluster randomized design if the proportion
- **Example:** Assess spillover effect on malaria associated with percentage of nearby households with an insecticide treated net
- Can compare magnitude of spillover effects over a gradient of treatment density

And so we could also use this to assess the spillover effect on malaria associated with the percentage of nearby households with an insecticide treated net. As with the distance based spillover effect, we can also compare the magnitude of spillover effects over a gradient of treatment density. And we can also use this in a cluster randomized design, even though the example here has been for an individually randomized design. If we want to use this with a cluster randomized design, the theoretical best design would be to randomize clusters with different proportions of treatment inside the cluster, and then compare clusters with different proportions of treatment to those with no treatment control clusters. This is great in theory, but very difficult to do in practice.

Network spillover effect



- This example is shown for an individually randomized trial.
- **Example:** Spillover effect of a smoking cessation program among friends of program participants
- Requires information about social or other connectedness, which can be cumbersome to collect.

Network spillover effect:

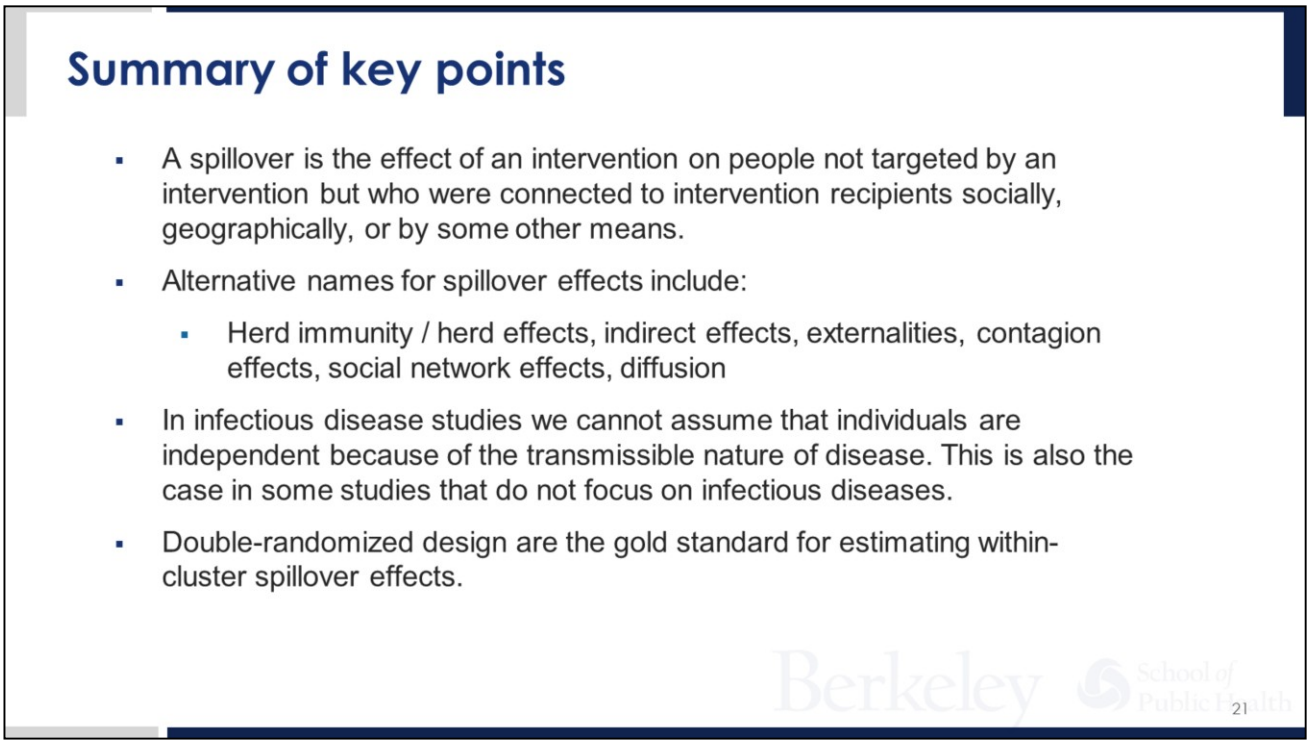
C = indicator for an alter being connected to an ego

$E[Y \mid T = 1, C = 1] - E[Y \mid T = 0, C = 1]$

Finally, network spillover effects. This is showing an example from an individually randomized trial. The green circle is someone called an ego, who's assigned to treatment. And these terms, ego and alter, are from the social network literature. The blue circle is an ego assigned to control. The reason they're called egos is that they are the initial focus of our investigation. We could say these are the individuals in the initial randomized study. Alters are people who are socially connected or connected in some other fashion to egos. And so these are denoted with thicker black circle border.

So in this case, the t means something slightly different. It means that you are connected to a person who is treated or connected to a person who is control. And c equals 1 indicates that you're a person connected. So you're an alter as opposed to an ego. So we can compare the mean outcome among an alter connected to a ego assigned to treatment to the mean outcome among alters connected to egos assigned to control.

And an example of this would be spillover effects of friends. Among friends of those who participate in this smoking cessation program. If individuals are randomly assigned to participate in this program, we could enroll their friends and then look at whether outcomes differ between friends of the intervention group and friends of the control group. This kind of parameter requires us to collect information about social connectedness or other forms of connectedness. And this can be very cumbersome to collect, which is a disadvantage. But these are also of increasing interest at this moment.



To summarize, a spillover effect is the effect of an intervention on people not targeted by the intervention but who were connected to recipients. Whether that be through social geographic or other means. And there are many different names for spillover effects, and it's important to know these because they are used inconsistently in the published literature. And infectious disease studies can't assume that individuals are independent, but this also may be the case in studies that don't focus on infectious diseases, and this is kind of why spillover effects have become of increasing interest in recent years because many investigators want to understand the extent to which non infectious disease interventions can diffuse through population. And double-randomized designs are the gold standard for estimating within cluster spillover effects, which are the most common type of spillover effects that are estimated because they're relatively convenient to estimate within cluster randomized trials.

International Journal of Epidemiology, 2018, 332–347
doi: 10.1093/ije/dyx201
Advance Access Publication Date: 2 November 2017
Education Corner

Education Corner

Spillover effects in epidemiology: parameters, study designs and methodological considerations

Andrew Mertens,[1]* Benjamin F Arnold,[1,2] David Berger,[3] Stephen P Luby,[4] Edward Miguel,[3] John M Colford Jr[1] and Alan E Hubbard[2]

[1]Division of Epidemiology, UC Berkeley School of Public Health, 101 Haviland Hall, Berkeley, CA 94720-7358, USA, [2]Division of Biostatistics, UC Berkeley School of Public Health, 101 Haviland Hall, Berkeley, CA 94720-7358, USA, [3]Department of Economics, University of California, Berkeley, CA 94720-7358, USA and [4]Division of Medicine, Stanford University, Stanford, CA 94305, USA

*Corresponding author. Division of Epidemiology, UC Berkeley School of Public Health, 101 Haviland Hall, Berkeley, CA 94720-7358, USA. E-mail: jadebc@berkeley.edu

Editorial decision 22 August 2017; Accepted 25 August 2017

Abstract

Many public health interventions provide benefits that extend beyond their direct recipients and impact people in close physical or social proximity who did not directly receive the intervention themselves. A classic example of this phenomenon is the herd protection provided by many vaccines. If these 'spillover effects' (i.e. 'herd effects') are present in the same direction as the effects on the intended recipients, studies that only estimate direct effects on recipients will likely underestimate the full public health benefits of the intervention. Causal inference assumptions for spillover parameters have been articulated in the vaccine literature, but many studies measuring spillovers of other types of public health interventions have not drawn upon that literature. In conjunction with a systematic review we conducted of spillovers of public health interventions delivered in low- and middle-income countries, we classified the most widely used spillover parameters reported in the empirical literature into a standard notation. General classes of spillover parameters include: cluster-level spillovers; spillovers conditional on treatment or outcome density, distance or the number of treated social network links; and vaccine efficacy parameters related to spillovers. We draw on high quality empirical examples to illustrate each of these parameters. We describe study designs to estimate spillovers and assumptions required to make causal inferences about spillovers. We aim to advance and encourage methods for spillover estimation and reporting by standardizing spillover parameter nomenclature and articulating the causal inference assumptions required to estimate spillovers.

Key words: Spillover effects, indirect effects, herd effects, herd immunity, diffusion, externalities, interference

Key Messages

- Spillovers are the effects of interventions on people in close physical or social proximity to intervention recipients but who do not themselves receive the intervention.
- Accurate estimation of spillover effects improves our understanding of the population-level impact and cost-effectiveness of public health interventions.
- Spillover evidence is strongest when based on individual-level outcome measurements rather than group-level outcome measurements.
- To make causal inferences about spillovers under typical assumptions: (i) factors associated with untreated individuals' proximity to treatment and their outcomes must be balanced across treatment groups, either by design or using statistical approaches; and (ii) the potential outcomes of untreated individuals in proximity to treated individuals must be independent of the treatment assignment of certain individuals in the population, who serve as a counterfactual controls.

Introduction

Public health interventions may benefit those in close physical or social proximity to intervention recipients who do not receive the intervention themselves. When such 'spillovers' are present in the same direction as direct effects on recipients, direct effects alone do not capture the full health impact and cost-effectiveness of an intervention. Most epidemiological studies measuring spillover effects have evaluated herd effects of vaccines,[1,2] but spillovers are theoretically possible for numerous other interventions that could alter disease transmission or change health behaviours. Spillovers have increasingly been measured for other interventions, particularly in economics, since evidence of spillovers can support the case for scaling up or subsidizing an intervention.[3,4] However, to date, discussion of methods for estimating spillover effects and identifying them within a causal inference framework has largely remained confined to the vaccine literature,[5–14] with few articles extending methods to studies of other interventions.[15–18]

Here we define spillover parameters using standardized notation and discuss causal inference assumptions for spillovers using non-technical language to provide an accessible introduction for epidemiologists. In conjunction with a systematic review we conducted on health spillovers of interventions in low- and middle-income countries,[19] we classified types of spillovers to make their similarities and differences more transparent. By standardizing spillover parameter definitions and articulating causal inference assumptions for spillovers, we aim to advance methods for spillover study design, estimation and reporting.

Types of spillover parameters

We describe six classes of spillover parameters that were most common in our systematic review.[19] We present individual-level counterfactual definitions of spillover effects in Box 1; the Supplement (available as Supplementary data at *IJE* online) contains average spillover effects as well as identification assumptions for each parameter. For simplicity we discuss spillovers among untreated individuals; however, spillover effects may also occur among the treated (Box 1). For example, HIV vaccine candidates provide imperfect protection to vaccinated individuals, but protection has been shown to increase as immunization coverage increases, due to reductions in transmission resulting from herd effects.[20] We define spillover parameters in the context of 'intention-to-treat' analyses, which estimate the true causal effect of interventions with high adherence. Inferences about spillovers are more complicated when there is imperfect adherence (e.g. in 'per-protocol' analyses),[16] and a formal discussion of that setting is beyond the scope of this paper.

To make causal inferences, investigators typically invoke the Stable Unit Treatment Value Assumption (SUTVA),[22] which states that an individual's potential outcome is not affected by the treatment assignment of other individuals in the population. This is also known as the assumption of no interference.[23] SUTVA does not hold when spillovers are present because an individual's potential outcome depends on their own treatment assignment and the treatment assignment of other individuals connected to them. The most common and theoretically tractable method to make causal inferences about spillover effects is by making the 'partial interference' assumption,[24] which states that there are no spillovers between clusters of individuals but allows for spillovers among individuals within the same cluster.[24] This assumption underpins the validity of many cluster-randomized trials. Studies can minimize spillovers between clusters by including buffer zones between clusters, as is common in cluster-randomized trials.[25–27]

Studies that assume no spillovers (i.e. that SUTVA holds) typically index counterfactuals only by an

Box 1. Spillover parameter definitions

Here, we provide definitions of individual-level average spillover effects using counterfactual notation. The Supplement (available at *IJE* online) includes additional details and identification assumptions. Figures cited below provide visual representations of each parameter.

We present the first three types of parameters in a two-stage randomized trial in which the treatment regimen is defined such that at least one individual in a treated cluster receives treatment (a_1), and in control clusters, all individuals are allocated to control (a_0). Let $Y_{ij}(a)$ be the potential outcome for individual j in cluster i, where a_i denotes a vector of treatment assignments for individuals in cluster i. The individual potential outcome averaging over different configurations of a_i is $\bar{Y}_{ij}(a, a)$ and is a function of the treatment regimen (a) and the individual's treatment assignment (a).

Cluster-level spillover effects (Figure 1)

Cluster-level spillover effect: $SE_{ij}(a_1; a_0) = \bar{Y}_{ij}(a_1; 0) - \bar{Y}_{ij}(a_0; 0)$

This parameter compares the mean potential outcome of an individual assigned to control in a treatment cluster with treatment regimen a_1 with their mean potential outcome if they were assigned to control in a cluster assigned to control (a_0).

Direct effect: $DE_{ij}(a_1) = \bar{Y}_{ij}(a_1; 1) - \bar{Y}_{ij}(a_1; 0)$

This parameter compares the mean potential outcome of an individual assigned to treatment in a treatment cluster with treatment regimen a_1 with their mean potential outcome if they were assigned to control in a cluster with treatment a_1.

Total effect: $TE_{ij}(a_1; a_0) = \bar{Y}_{ij}(a_1; 1) - \bar{Y}_{ij}(a_0; 0)$

This parameter compares the mean potential outcome of an individual assigned to treatment in a treatment cluster with treatment regimen a_1 their mean potential outcome if they were assigned to control in a cluster assigned to control (a_0).

Overall effect: $OE_{ij}(a_1; a_0) = \bar{Y}_{ij}(a_1) - \bar{Y}_{ij}(a_0)$

This parameter compares the mean potential outcome across individuals in a cluster with treatment regimen a_1 with their mean potential outcome had the cluster been assigned to control (a_0).

Distance-based spillover effects (Figure 2)

Spillover effect conditional on distance to clusters (Figure 2a): $SE_{ij}(a_1; a_0; k) = \bar{Y}_{ij}(a_1 | K_i = k) - \bar{Y}_{ij}(a_0 | K_i = k)$

This parameter compares the mean potential outcome for individuals at distance k from a cluster with treatment regimen a_1 with the mean potential outcome at distance k from a cluster with treatment regimen a_0 (all individuals assigned to control).

Spillover effect conditional on distance between clusters (Figure 2b):
$SE_{ij}(a_1; a_0; b_0; k) = \bar{Y}_{ij}(a_1; b_0 | K_i = k) - \bar{Y}_{ij}(a_0; b_0 | K_i = k)$

This parameter compares the mean potential outcome of individuals in control clusters with treatment regimen b_0 within distance k of treatment clusters with treatment regimen a_1 with those in clusters assigned to control with treatment regimen b_0 within distance k of control clusters with treatment regimen a_0.

Spillover effect conditional on treatment density (Figure 3)

Define p_i as the proportion of individuals allocated to treatment in treatment clusters. In a two-stage randomized trial, in the first stage, clusters are randomly assigned to receive a certain proportion of treatment ($E[a_i] = p_i$), including clusters with $p_i = 0$. In the second stage, individuals are randomized to treatment ($p_i > 0$) or control ($p_i = 0$) in clusters.

$$SE(p_i; p^0_i) = \bar{Y}_{ij}(p_i; 0) - \bar{Y}_{ij}(p^0_i; 0)$$

This parameter compares the mean potential outcome of an individual assigned to control in clusters with different proportions of individuals assigned to treatment (p_i vs. p_i^0, where $p_i \neq p_i^0$).

Social network spillover effect (Figure 4)

This parameter can be estimated as a trial that randomizes treatment to egos (the initially enrolled subjects) and compares the mean outcomes of alters (the persons socially connected to the egos) in the treatment vs control group.

For this parameter, we define the potential outcome for alter j as $Y_j(a_1, a_0)$, which is a function the treatment assignment of the ego (a_1) and the treatment assignment of the alter (a_0).

$$SE = E[Y_j(1; 0) - Y_j(0; 0)]$$

This parameter compares the mean potential outcome among an untreated alter socially connected to a treated ego (Y_j (a_1 = 1, a_0 = 0)) with their mean potential outcome if they were connected to an untreated ego (Y_j (a_1 = 0, a_0 = 0)).

Vaccine efficacy for infectiousness (Figure 5)

This parameter is typically estimated in studies that enrol households with an infected individual (a 'case') and at least one uninfected individual (a 'susceptible'). S is the outcome for the primary household case; Y is the outcome for the susceptible individual; and A is the treatment assignment for the case.

$$VE_I = E[Y_j A = 1; S = 1] - E[Y_j A = 0; S = 1]$$

The parameter compares the secondary attack rate among uninfected susceptible individuals in households with a vaccinated case with the rate among those in households with unvaccinated cases. Identification of this parameter requires assumptions that complicate the presentation of causal parameters, so we define a statistical parameter here and provide a causal parameter definition in the Supplement. The parameter is labelled 'VE' to be consistent with how it is presented in the vaccine literature.

individual's own treatment assignment (e.g. Y_a could indicate the potential outcome for a person with treatment assignment $A = a$). Under the partial interference assumption, the treatment assignment of each individual (j) in each cluster (i) can be summarized in a vector of treatments for n individuals: $A_i = (A_{i1},.., A_{in_i})$. Similarly, $A_{i,-j} = A_{i1},.., A_{ij-1}, A_{ij+1},.., A_{in_i}$ denotes the vector of treatments for individuals in cluster i for all individuals except for individual j. A_i can be considered a random treatment allocation regimen, and $A(n_i)$ is the set of all possible treatment allocation regimens for n_i individuals (the set of values that A_i can assume). Specific regimens within $A(n_i)$ can be denoted by a, the parameterization of the distribution of A_i for i = 1,.. ., N. For example, a_1 may include a scenario in which half of all individuals in a cluster are allocated to treatment and half are allocated to control, and a_0 may include a scenario in which all individuals in a cluster are allocated to control. In the following sections, we define individual-level average causal effects; see the Supplement for individual-level and group-level causal effects (available as Supplementary data at *IJE* online).

Cluster-level spillovers

Causal cluster-level spillover effects can be measured in a two-stage randomized trial[9] in which treatment is randomized to independent clusters in the first stage, and in the second stage, individuals within treatment clusters are assigned to treatment or control (Figure 1). This is sometimes referred to as a double-randomized trial.[11] The cluster-level spillover effect may be defined as the difference in mean outcomes among untreated individuals in treated clusters and the outcomes among individuals in control clusters.[5,6,8,9] We provide the definition of this parameter from Hudgens and Halloran[8] and Tchetgen Tchetgen.[10] We present this and the parameters in subsequent sections on the additive scale. To be consistent with the causal inference literature,[10] we reverse the order of treatment and control contrasts from those in Hudgens and Halloran,[8] which subtract potential outcomes for those assigned to treatment from those assigned to control.

Let $Y_{ij}(a_i)$ be the potential outcome for individual j in cluster i, where a_i denotes a vector of treatment assignments for individuals in cluster i. Let $a_{i,-j}$ be the vector of treatment for all individuals in cluster i except individual j. The potential outcome in this parameter, $Y_{ij}(a_{i,-j}, a_{ij})$, is a function of the individual's own treatment assignment (a_{ij}) and the treatment assignment of other individuals ($a_{i,-j}$) in the cluster. In treatment clusters, the treatment regimen (a_1) is defined such that at least one individual receives treatment, and in control clusters, all individuals are allocated to control (a_0). The treatment vector a_i can vary for a given treatment regimen a. Thus, the individual potential outcome averaging over different configurations of a_i for a given a (i.e. averaging over all possible treatment regimens) is defined as:

$$\bar{Y}_{ij}(a; a) = \sum_{s \in A(n-1)} Y_{ij}(a_{i,-j} = s; a_{ij} = a) \, Pr_a(A_{i,-j} = s \mid A_{ij} = a) \quad (1)$$

We assume that the proportion of individuals assigned to treatment (a_1) is the same among all treated clusters.

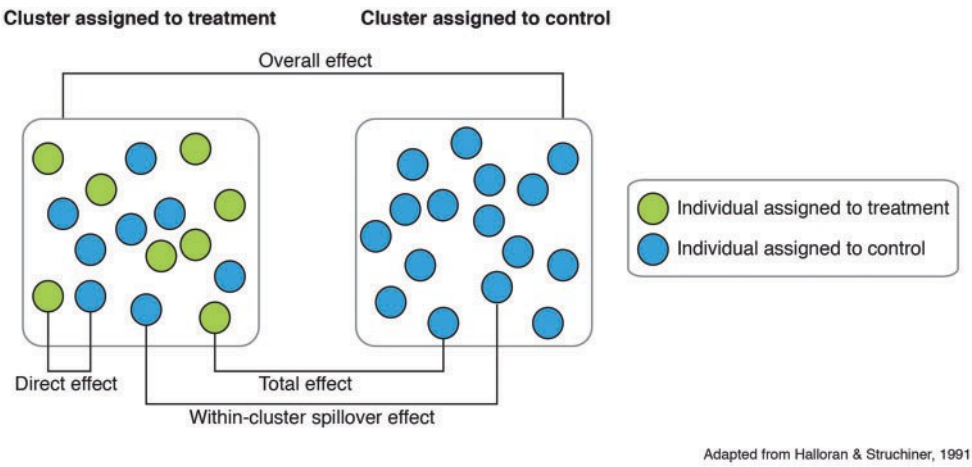


Figure 1. Cluster-level spillover effects. This spillover parameter can be estimated in a two-stage randomized trial in which clusters are randomly allocated to treatment or control and then individuals within treatment clusters are randomly allocated to treatment or control. The direct effect compares potential outcomes of individuals allocated to treatment in treatment clusters to the potential outcomes of individuals allocated to control in treatment clusters. The cluster-level spillover effect compares potential outcomes of individuals allocated to control in treatment clusters to those of individuals in control clusters. The total effect compares the potential outcomes of individuals allocated to treatment in treatment clusters to those of individuals allocated to control in control clusters. The overall effect compares the potential outcomes of all individuals in clusters allocated to treatment to those of all individuals in clusters allocated to control.

A cluster-level spillover effect measures spillovers that occur among untreated individuals in a cluster that received treatment. This effect can be defined as:

$$SE_{ij}(a_1; a_0) = \overline{Y}_{ij}(a_1; 0) - \overline{Y}_{ij}(a_0; 0) \quad (2)$$

This parameter compares the mean potential outcome of an individual assigned to control in a treatment cluster with treatment regimen a_1 with their mean potential outcome if they were assigned to control in a cluster assigned to control (a_0). For example, Chong *et al.* estimated cluster-level spillovers by randomly assigning schools to receive a sexual health education programme and then randomly assigning the programme to classrooms within intervention schools; they compared sexual health knowledge among children in control classrooms in intervention schools with that of children in control schools (Table 1, Example 1).[28]

There are several related parameters that can also be estimated in a two-stage randomized trial (Figure 1). Direct effects compare the mean potential outcome of a treated individual in a treated cluster with treatment regimen a_1 with their mean potential outcome if they were assigned to control in the same cluster ($\overline{Y}_{ij}(a_1; 1) - \overline{Y}_{ij}(a_1; 0)$); total effects compare the mean potential outcome of a treated individual in a treated cluster with treatment regimen a_1 with their mean potential outcome if they were assigned to control in control clusters ($\overline{Y}_{ij}(a_1; 1) - \overline{Y}_{ij}(a_0; 0)$); and overall effects compare the mean outcome of all individuals in treated clusters with treatment regimen a_1 with the mean outcome had the cluster been assigned to control ($\overline{Y}_{ij}(a_1) - \overline{Y}_{ij}(a_0)$) (see complete definitions in the Supplement, available as Supplementary data at *IJE* online).

The direct effect defined in the spillover literature differs from other definitions in the causal mediation literature (effect of an exposure through no intermediate variables).[29] Similarly, in the spillover literature, the term 'indirect effect' is frequently used to describe spillover effects–in the mediation literature, an 'indirect effect' is the part of an intervention's effect that is mediated through intermediate variables.

Cluster-level spillovers can be measured in studies that enrol clusters of any size. They are relatively convenient to estimate in studies with small- to medium-size clusters when the treatment status of most individuals in the cluster is known. Cluster-level spillover parameters often condition on other variables, such as eligibility to receive an intervention. For example, a comparison of outcomes among ineligible individuals in the treatment group with ineligible individuals in the control group estimates a cluster-level spillover effect (Figure 1).[30–37]

Distance-based spillovers

Spillovers can be measured as a function of distance from treated individuals or clusters. We introduce two parameters: one conditional on individuals' distance to clusters, and one conditional on distance between clusters. The first parameter measures spillover effects among individuals located a certain distance from the boundary of treatment and control clusters (Figure 2a). We define $Y_{ij}(a_i \mid K_i = k)$ as the potential outcome for individual j residing within distance k from cluster i, with treatment vector a_i. The individual average potential outcome $\overline{Y}_{ij}(a \mid K_i = k)$ is a function of the treatment regimen (a) and the individual's distance to the nearest study cluster (k). Let a_1 be a

Table 1. Examples of empirical studies estimating different spillover parameters

Example	Study design	Spillover parameter	Intervention	Outcome	Spillover group	Comparison group
1. Chong *et al.*, 2013	Double-randomized trial	Cluster-level spillover (Figure 1)	School-based sexual health education programme	Knowledge and attitudes about sexually transmitted infections and safe sex practices	Children in schools that received the programme but in classrooms that did not receive the programme	Children in schools that did not receive the programme
2. Banerjee *et al.*, 2010	Cluster-randomized trial	Distance-based spillover (Figure 2)	Immunization campaign with and without incentives	Vaccination	Individuals in randomly selected, untreated villages within 6 km of villages randomized to treatment	Individuals in villages randomized to control
3. Hawley *et al.*, 2003	Cluster-randomized trial	Distance-based spillover (Figure 2)	Insecticide-treated nets	Malaria, anaemia, child mortality	Untreated compounds 0–299 m, 300–599m and 600–899 m from treated compounds	Untreated compounds 2: 900 m from treated compounds
4. Miguel and Kremer, 2004	Cluster-randomized trial	Spillovers conditional on treatment density (Figure 3)	School-based deworming	Soil-transmitted helminth infection	Untreated students at schools for which some pupils were treated at schools within 0–3 km and 4–6 km	Untreated students at schools for which no pupils were treated at schools within 0–3 km and 4–6 km
5. German *et al.*, 2012	Randomized trial	Social network spillover (Figure 4)	Peer network intervention	Depression	Peers of individuals randomized to treatment	Peers of individuals randomized to control
6. Préziosi and Halloran, 2003	Secondary attack rate study	Vaccine efficacy for infectiousness (Figure 5)	Pertussis vaccine	Pertussis	Susceptibles living in households with treated cases	Susceptibles living in households with untreated cases

treatment regimen in which at least one individual per cluster is allocated to treatment. Again, we assume that the treatment regimen is uniform (i.e. a_1 does not vary) among treated clusters. The spillover effect conditional on distance to treated clusters can be defined as:

$$SE(a_1; a_0; k) = \bar{Y}_{ij}(a_1 | K_i = k) - \bar{Y}_{ij}(a_0 | K_i = k) \quad (3)$$

This parameter compares the mean potential outcome of an individual distance k from a cluster with treatment regimen a_1 (at least one individual allocated to treatment per cluster) with their mean potential outcome at distance k from a cluster with treatment regimen a_0 (all individuals assigned to control). Studies estimating this parameter must ensure control clusters are at a distance from treatment clusters beyond which the intervention has an effect. Hawley *et al.* estimated a similar parameter in a re-analysis of a cluster-randomized trial to measure spillovers of insecticide-treated nets on malaria and other outcomes over different distances.[38] They compared individuals assigned to control clusters who were 0–299 m, 300–599 m and 600–899 m with those who were 2: 900 m from the nearest individual in a treated cluster (Table 1, Example 3).

Spillover effects can also be measured as a function of distance between clusters using a pair-matched, two-stage design (Figure 2b). This parameter differs from the previous one because the individuals used to measure spillover effects reside in separate clusters rather than in the areas around the boundaries of the treatment clusters. To measure this type of spillover effect, first a study pair-matches clusters separated by distance k and then randomly allocates each pair to treatment or control. Second, the study randomly selects one member from each pair to be the 'primary' cluster; in the treated pairs, the primary cluster is assigned to treatment and the other cluster is assigned to control. In practice, the pairs of control clusters may be reduced to a single control cluster unless the second is needed to achieve sufficient statistical efficiency. Individuals in clusters assigned to treatment are randomly

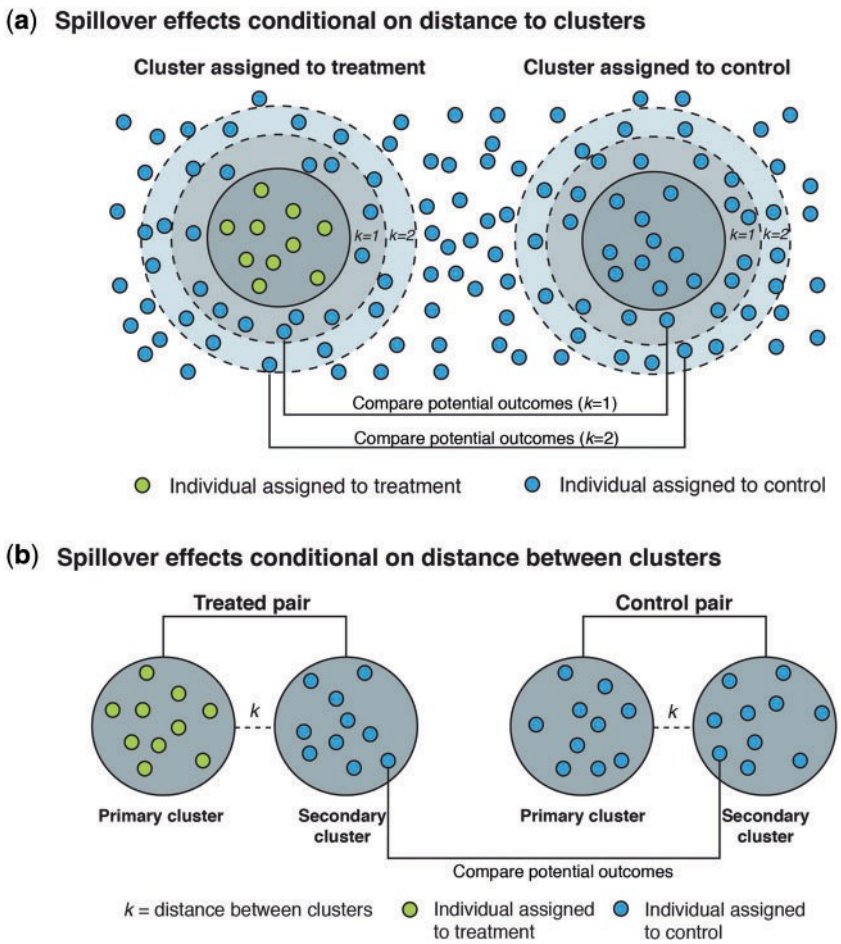


Figure 2. Distance-based spillover effects. (a) Spillover effects conditional on distance to clusters can also be estimated in a two-stage randomized trial. This parameter compares the potential outcomes of untreated individuals within distance k of treated clusters to those of untreated individuals within distance k of control clusters. (b) Spillover effects conditional on distance between clusters can be estimated in a two-stage randomized trial. In the first stage, a study pair-matches clusters separated by distance k and then randomly allocates each pair to treatment or control. In the second stage, the study randomly selects one member from each pair to be the "primary" cluster; in the treated pairs, the primary cluster is assigned to treatment and the other cluster is assigned to control. Individuals in clusters assigned to treatment are randomly assigned to treatment or control. This parameter compares potential outcomes of individuals allocated to control in secondary clusters within distance k of treated clusters to those of individuals allocated to control in secondary clusters within distance k of control clusters.

assigned to treatment or control. Let a_i be the treatment vector for primary clusters and b_i be the treatment vector for secondary clusters. Let $Y_{ij}(a_i, b_i \mid K_i = k)$ be the potential outcome for individual j in secondary cluster i with treatment vector b_i within distance $K_i = k$ from a primary cluster with treatment vector a_i. We define a as the treatment regimen for primary clusters and b as the treatment regimen for secondary clusters. The individual potential outcome averaging over different configurations of a_i and b_i for a given a and b is $\overline{Y}_{ij}(a, b \mid K_i = k)$. The spillover effect conditional on distance between clusters may be defined as:

$$SE_{ij}(a_1; a_0; b_0; k) = \overline{Y}_{ij}(a_1; b_0 \mid K_i = k) - \overline{Y}_{ij}(a_0; b_0 \mid K_i = k) \quad (4)$$

This parameter compares the mean potential outcome of a control individual in a secondary control cluster (b_0) within distance $K_i = k$ of a treatment cluster with treatment regimen a_1 with their mean potential outcome if they were assigned to a secondary control cluster (b_0) within distance $K_i = k$ of a primary control cluster (a_0). Banerjee *et al.* conducted a study with a design similar to this to measure spillovers of a vaccine promotion campaign with and without subsidies (Table 1, Example 2).[39] They assigned clusters to treatment or control and then enrolled clusters within 6 km of treated clusters to measure spillovers. Their design assumes that clusters within 6 km of treatment clusters (the study's spillover clusters) had similar baseline characteristics to control clusters; since spillover clusters were not included in the randomization, it is possible that

there were systematic differences between spillover and control clusters, which could have confounded results.

Spillovers conditional on treatment density

We have defined spillover parameters with counterfactuals indexed by a vector of treatment assignments (a_i). In some cases, spillovers are a function not of the precise allocation of treatment to specific individuals, but instead of summaries of a_i, such as the proportion of those that get treatment (p_i). Thus, we can represent the model of the counterfactual distribution as $Y_{ij}(p_i, a) = f(p_i, a)$, where f is a function of individual-level treatment assignment and low-dimensional summaries of treatment assignment among individuals in cluster i. Designs such as a two-stage randomized study can be tailored to estimate such effects. For instance, in the first stage, clusters are randomly assigned to receive a certain proportion of treatment ($E[a_i] = p_i$), and some clusters are assigned to $p_i = 0$. In the second stage, individuals are randomized to treatment or control in clusters with $p_i > 0$ (Figure 3).[8,10,13] We can define counterfactuals for individual j in cluster i as $Y_{ij}(p_i, a_{ij})$, which is indexed by two scalars: the individual treatment assignment (a_{ij}) and the average proportion treated (p_i). The spillover effect among untreated individuals ($a_{ij} = 0$) conditional on treatment density is then:

$$SE(p_i, p_i^0) = \overline{Y}_{ij}(p_i, 0) - \overline{Y}_{ij}(p_i^0, 0) \quad (5)$$

In a two-stage randomized trial, this parameter compares the mean potential outcomes of an individual assigned to control in clusters with different proportions of individuals assigned to treatment (p_i vs. p_i^0). This is equivalent to the cluster-level spillover effect on untreated individuals within the cluster (Equation 2), except for the simplifying assumption that the counterfactual is only a function of a summary of a_i, namely p_i. This effect can also be estimated among treated individuals (see the Supplement for details, available as Supplementary data at *IJE* online). Miguel and Kremer estimated this parameter in a cluster-randomized trial of a school-based deworming programme in Kenya.[3] They compared outcomes among children in untreated schools in areas with varying levels of density of treated children (Table 1, Example 4).[3] This class of parameters can also be estimated by conditioning on the proportion of treated social network nodes within a given social distance metric of each untreated individual. For example, a study of a school-based deworming programme in Kenya estimated whether child deworming was associated with the number of social links to parents whose children received deworming at school.[40] Causal interpretation of this type of parameter requires that treatment be randomized to individuals within the social network rather than to individuals within a specific geographical area.

Social network spillovers

Social networks can be used to measure spillovers as a function of social proximity. A variety of designs can be employed to estimate different social network effects. For example, causal spillovers through social networks can be measured in a design that randomizes treatment to egos (the initially enrolled subjects) and compares the mean outcomes of alters (the persons socially connected to the egos) in the treatment vs control group (Figure 4). Counterfactuals in this design are a function of a joint treatment (a_1, a_0), where the treatment assignment for the ego is a_1, and the treatment assignment for the alter is a_0. The potential outcome for the alter j connected to the ego with treatment a_1 is $Y_j(a_1, a_0)$. The social network spillover effect may be defined as:

$$SE = E[Y_j(1, 0) - Y_j(0, 0)] \quad (6)$$

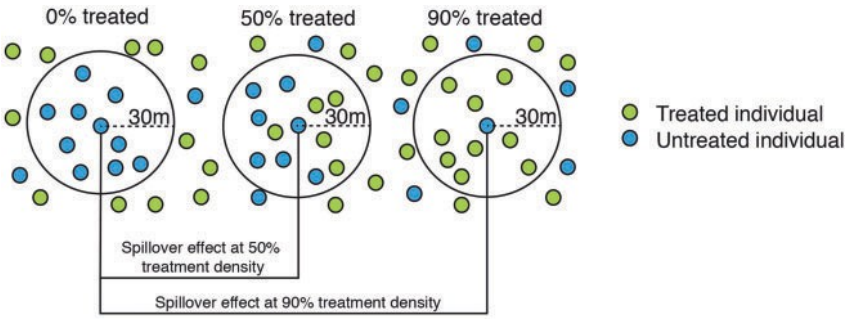


Figure 3. Spillover effects conditional on treatment density. Spillover effects conditional on treatment density can be estimated in a two-stage randomized design that randomly allocates clusters to treatment or control and then randomly allocates individuals in treatment clusters to treatment or control. This parameter compares potential outcomes of untreated individuals in clusters allocated to treatment proportion p to those of untreated individuals in clusters allocated to a different treatment proportion p'. For example, in this figure, the treatment proportion within 30m of untreated individuals varies. This parameter compares potential outcomes of untreated individuals in clusters with treatment proportion 50% and 90% to those of untreated individuals in clusters with 0% of individuals allocated to treatment (i.e. control clusters).

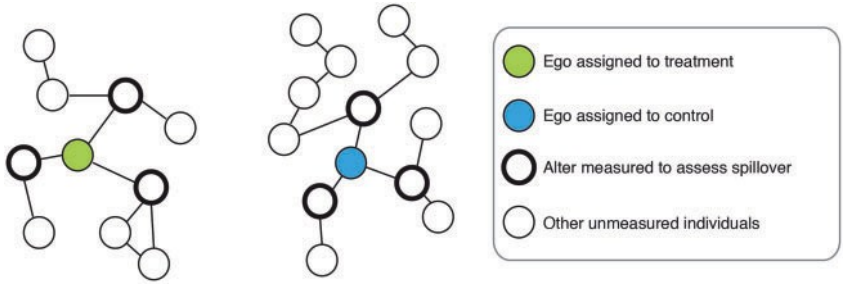


Figure 4. Social network spillover effects. Social network spillover effects can be estimated in a study that randomizes treatment to egos (the initially enrolled subject) and compares the mean outcomes of alters (the person socially connected to the ego) in the treatment vs. control group.

This parameter compares mean outcomes among untreated alters socially connected to treated egos with their mean potential outcome had the egos been untreated. It is distinct from a parameter conditioning on treatment density among social network links (above) because it focuses on alters socially connected to specific egos, isolating the spillover effect through individual peer-to-peer connections. German *et al.* estimated this parameter in a randomized study that evaluated spillover effects of a peer network intervention. They compared depression scores among peers of treated vs control participants (Table 1, Example 5).[41]

Alternative parameters can be defined to estimate spillover effects through social networks. For example, spillover effects can be estimated for certain types of social ties (e.g. direct friends, social ties in the community, village members)[42] or based on which type of social ties were targeted for intervention (e.g. villagers with the most social ties, nominated friends).[43] We presented social network spillovers among the untreated, but they can also be estimated among the treated by comparing counterfactual outcomes among treated alters connected to treated vs untreated egos. Many other types of social network spillover parameters have been described, and an overview of social network effects has been provided by Vanderweele and An.[44] We have described estimation of social network spillover effects in an individually randomized trial, but other study designs can be used to estimate spillover effects through social networks. For instance, Christakis *et al.* measured the spread of obesity[45] and smoking[46] through social networks in a large cohort study. Stochastic, actor-oriented models can also be used to estimate spillover effects in networks;[47,48] these models allow individuals to change their behaviour status and/or social ties at each time point and require strong assumptions.[44]

When estimating social network spillover effects, both observational and randomized designs face unique threats to validity. First, socially connected individuals may have correlated outcomes and may inhabit the same environment, leading to environmental confounding.[49–53] Second, an individual may sever a social tie based on their present outcome status as a result of homophily, which may bias estimates of spillover effects.[51] A randomized design can minimize homophily by randomizing individuals to peers in their social network, such as a room-mate, to ensure that potential confounders are balanced within ego and alter pairs.[54]

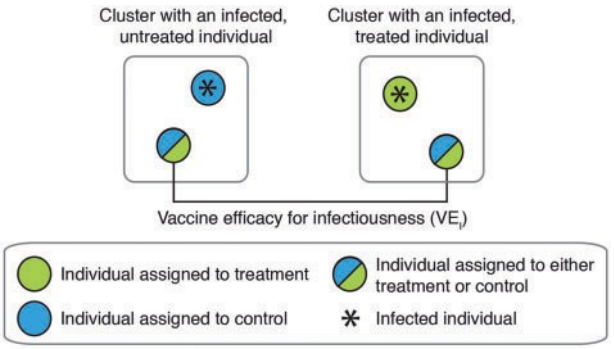


Figure 5. Vaccine efficacy for infectiousness. This parameter is typically estimated in studies that enroll households with an infected individual (a "case") and at least one uninfected individual (a "susceptible"). The parameter compares the secondary attack rate among uninfected susceptible individuals in households with a vaccinated case to the rate among those in households with unvaccinated cases. Susceptibles may be either vaccinated or unvaccinated.

Spillovers conditional on exposure to infection

Epidemiologists studying vaccines have developed parameters that condition on exposure to infected individuals to measure whether vaccines reduce transmission to uninfected individuals.[6,9,12,55–58] These parameters are typically estimated in studies that enrol households with an infected individual (a 'case') and at least one uninfected individual (a 'susceptible'). The vaccine efficacy for the infectiousness parameter (also referred to as the 'infectiousness effect'[59]) is a type of spillover parameter that compares the secondary attack rate among uninfected susceptible individuals in households with a vaccinated case, with the rate among those in households with unvaccinated cases (Figure 5; Table 1, Example 6).

Conditioning on post-treatment outcome status can introduce selection bias because individuals who become

infected may be systematically different from those who do not.[6] Under standard causal inference assumptions this parameter is not identifiable;[58] to identify this parameter, relatively strong additional assumptions are required.[58–60] For example, one must assume that susceptibles are only exposed to infection in their household, i.e. that household members interact with each other but not with other households.[6] Because these additional assumptions complicate the presentation of causal parameters, we define a parameter below that does not use counterfactuals (i.e. a statistical parameter); see the Supplement for a causal definition (available as Supplementary data at *IJE* online).

Let S be the outcome for the primary household case, Y be the outcome for the susceptible individual, and A be the treatment assignment for the case. The vaccine efficacy for infectiousness (VE) on the additive scale is:

$$VE_I = E[Y|A = 1; S = 1] - E[Y|A = 0; S = 1] \quad (7)$$

We present this parameter on the additive scale for consistency with other parameters, but it is most often estimated as a reduction in the relative risk [(1-RR) x 100%]. These parameters are similar to the cluster-level spillover effect (Equation 2) but condition on outcome status instead of treatment status. Préziosi and Halloran measured this parameter in a household secondary attack rate study of pertussis vaccination by comparing outcomes among susceptibles in households with vaccinated cases with those in households with unvaccinated cases.[61]

Designing studies to measure spillovers

In this section we discuss spillover study designs. Box 2 summarizes recommendations throughout this section.

Causal inferences about spillovers

Confounding of spillover effect estimates

Estimates of spillover effects will be confounded if there are factors associated with untreated individuals' exposure to treatment and outcomes that are not controlled for. A two-stage randomized trial is ideal for estimating cluster-level spillovers, distance-based spillovers and spillovers conditional on treatment density, because it minimizes systematic differences other than treatment assignment between treated and untreated individuals in the treatment and control arms. Individually randomized trials can generate internally valid spillover estimates for certain parameters, such as social network spillover effects, if they ensure that there is sufficient physical or social distance between untreated individuals so that some individuals can serve as a valid counterfactual. For observational studies, strategies such as propensity scores,[62] inverse probability

Box 2. Recommendations for designing studies to estimate spillover effects

Importance of theory	1. Choose a spillover parameter and study design based on theory about how the intervention's effects diffuse through a population.
	2. Especially when theory to support spillovers of an intervention is weak, use rigorous designs, such as the double-randomized design, in order to make causal inferences.
Causal inferences about spillovers	3. To estimate cluster-level spillovers, distance-based spillovers or spillovers conditional on treatment density, use a double-randomized design to maximize internal validity.
	4. If it is only possible to use a cluster-randomized design, consider using multivariate matching techniques to match untreated individuals in the control clusters to untreated individuals in the treatment clusters. Matching may improve balance for measured confounders; however, unmeasured confounding may remain, and external validity may decrease depending on the subset of the population that is matched.
	5. If a clustered study design is used, build in buffer zones between treated and control units in order to prevent contamination and ensure that there is a valid control group to serve as a counterfactual.
	6. Use individual-level outcomes to measure spillovers when possible. Group-level measurements can be useful for hypothesis generation when individual-level measurements are unavailable.
Pre-specifying analyses	7. Pre-specify the specific spillover parameter(s) to be estimated.
	8. Pre-specify the scale at which spillovers are expected and the hypothesized mechanism(s) of spillover.
	9. If the spillover parameter incorporates measurement within specific distances or areas, pre-specify distance or area definitions and provide a rationale for them based on the hypothesized strength and scale of spillovers to avoid selectively choosing cutoffs that provide favourable results. For example, describe the specific distances in which measurement will take place or describe whether measurement will occur within quantiles of the observed distance distribution.
	10. If the study protocol is registered, use the terms 'spillovers' or 'indirect effects' to refer to spillovers in the protocol because these are the most commonly used terms in the literature, and they provide a direct link to the theoretical literature on this topic.

weighting,[13,63–65] matching in the design stage,[66–68] regression discontinuity[69,70] and instrumental variables[28,29] can increase comparability between untreated individuals in proximity to treated individuals and untreated individuals not in proximity to treated individuals that serve as a control group. For example, cluster-randomized trials can measure cluster-level spillovers by matching individuals who were ineligible for treatment in treated clusters to similar individuals in control clusters and comparing outcomes.[73] However, this approach only ensures comparability on measured covariates, so unmeasured confounding may remain, and matching could potentially reduce the study's external validity.[74] We note that some of these methods, such as regression discontinuity, have yet to be applied to estimation of spillover effects; to do so would require an extension of current theory, which is appropriate for total and overall effects.

Spillover effects conditional on distance can be confounded by factors that affect whether untreated individuals live near treated individuals and as well as their outcomes. Two-stage randomized designs (Figure 2) minimize this form of confounding by balancing the distribution of treatment across space. In observational studies, investigators can select comparison areas with similar geographical features to treatment areas to minimize confounding. Sensitivity analyses can estimate the extent of possible bias in studies with unmeasured confounding or studies that randomize treatment but condition on a post-treatment variable, such as the presence of an infected case in a household, which can also lead to bias.[52]

Violations of SUTVA and the partial interference assumption

SUTVA is one of the core assumptions required to make causal inferences; but when spillovers occur, the assumption can be relaxed to allow for spillovers within but not between clusters (partial interference)[24] (Figure 6a). Two-stage randomized designs can reasonably assume partial interference, but the assumption is difficult to assert in individually randomized studies unless enrolled individuals are separated by a large physical or social distance. In cluster-randomized trials, when interventions affect outcomes in the control clusters ('contamination'), the partial interference assumption is violated (Figure 6b). In this case, spillover effects can generally be considered lower bounds of the true spillover effect under the assumption that the effect of treatment in the control group is less than or equal to its effect in the treatment group. However, if contamination causes re-composition of treatment and control units that alters transmission dynamics, treatment effects may be biased away from the null. Estimates will also be biased when spillovers occur in multiple directions—untreated individuals affect treated individuals' outcomes and treated individuals affect untreated individuals' outcomes (Figure 6c). For example, in a trial that randomizes improved latrines to a subset of households in a village, environmental contamination resulting from use of unimproved latrines by non-recipients may spread enteric infections to latrine recipients, diluting benefits from improved latrines.

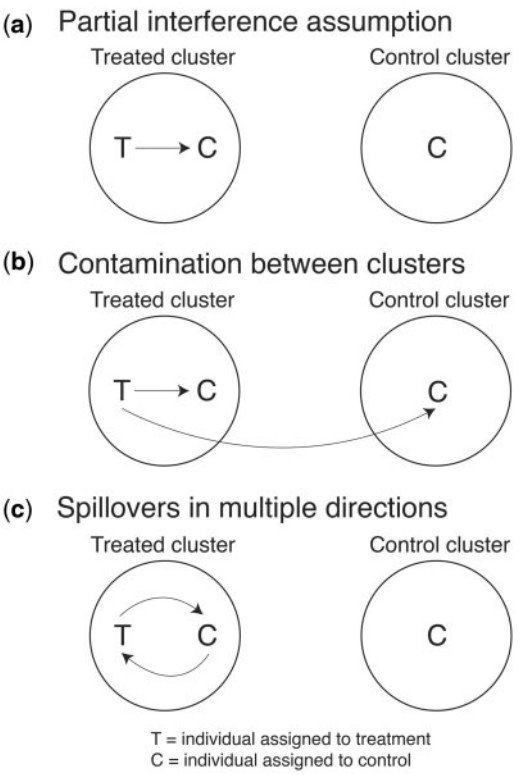


Figure 6. Schematic of spillovers within and between clusters. (a) The partial interference assumption states that there are no spillovers between clusters of individuals but allows for spillovers among individuals within the same cluster. (b) When an intervention affects individuals in a cluster assigned to control, this is often referred to as "contamination" in the cluster-randomized trial literature. This is an example of a violation of the partial interference assumption depicted in (a). (c) Spillovers may occur in multiple directions: individuals assigned to treatment may influence potential outcomes of individuals assigned to control and vice versa. When such multi-dimensional effects occur, causal inference becomes complicated.

To assess possible violations of SUTVA and the partial interference assumption, investigators can conduct Fisher's permutation test of no effect, which assumes every study unit has the same outcome under all treatment assignments; in a cluster-randomized study, if the null hypothesis is true, there is no effect of intervention and there are no spillovers between clusters.[75–77] In some cases nothing can be done to prevent spillovers between individuals or clusters–a violation of the partial interference assumption. Recent efforts have explored causal inference in this setting, which is referred to as 'general interference'. These

studies have described statistical approaches and developed software packages[78] for analyzing network data in which individuals are not considered independent.[17,64,79–90]

Causal epidemiological effects vs biological effects

Another consideration in estimating spillover effects is that the causal estimates from an epidemiological study may not be equivalent to the biological effects of intervention, even in a randomized trial.[91] Comparing the quantities estimated in challenge studies vs epidemiological studies of vaccines illustrates this difference. The biological effect of a vaccine is the reduction in risk associated with exposure to a specific volume of inoculum. Challenge studies estimate this effect by comparing risk among individuals randomly assigned to receive a known quantity of inoculum or a placebo.[91] On the other hand, epidemiological studies cannot control the level of exposure to pathogens targeted by vaccines;[6] as a result, they produce estimates that are a function of the level of population mixing and frequency of exposure to illness, both of which are typically unknown. Because most trials do not condition on the level of exposure to infection, their causal estimates will differ from the biological effect for interventions that produce spillovers. In addition, spillover and total effects estimated in trials with differing levels of baseline transmission are not directly comparable.[91] These principles may apply to studies of interventions other than vaccines, even if analogous biological effects cannot be estimated through challenge studies. For example, in a study led by German *et al.* of a peer network intervention to reduce depression,[41] peers of study participants were recruited to measure spillovers, but the level of social contact between peers was not controlled by the investigator; estimates of spillover effects were a function of the extent and type of social contact between peers, which may vary across populations.

Importance of theory

Spillover estimates will be most meaningful when the theory about how the intervention diffuses through a population informs the parameter choice and study design. For example, to measure spillovers of an intervention that aims to reduce HIV transmission, social network spillovers are likely to provide information that is more useful for public health intervention than distance-based spillovers since transmission occurs through sexual contact. For spillovers of behaviour change interventions, network[92] and diffusion theory[93,94] may inform how intervention effects diffuse over geographical areas or through social networks and whether spillovers occur evenly across a population or more strongly within subgroups. Interventions that diffuse through communication between individuals over great distances may cause spillovers over large geographical areas, even if no contact is made in person. If so, measuring spillovers through social networks is more appropriate than measuring spillovers through physical proximity since the latter may fail to capture the full spillover effect.

The strength of theory to support spillovers may inform design and analysis choices when estimating spillovers. When the causal mechanism for spillovers is not strongly grounded in theory, we encourage investigators to use a design that can effectively minimize confounding (e.g. two-stage randomization in Figure 1). However, for certain interventions such as vaccines, the biological mechanism for reducing pathogen transmission is often sufficiently clear for observational studies to provide strong inference about vaccine spillovers.[9] In some cases, even if theory to support spillovers is strong, statistical models are required to make inferences about spillovers; this may occur if there is limited variation in treatment status within levels of confounders. Statistical models can yield biased estimates if they are mis-specified or if they extrapolate beyond the observed data, so model-based estimates should always be considered carefully in light of their potential assumptions and limitations.

Individual- vs group-level measurements

Spillovers can be estimated in studies that measure outcomes either in individuals or in groups. Two-stage randomized studies and other cluster-randomized studies produce individual-outcome level data. Studies have commonly attempted to measure spillover effects by assessing how rates of illness among untreated individuals change with the proportion of individuals treated in different areas, using group-level data from trials or observational studies.[95–102] Studies with individual-level outcome measurements typically have the strongest inference because they are better able to control for both individual-level and group-level confounders, whereas group-level studies can only control for the latter. Despite these drawbacks, group-level measurements can be useful for hypothesis generation when individual-level measurements are not available.

Measuring spillovers within geographic areas

Certain types of spillovers, such as spillovers conditional on treatment density and distance-based spillovers, measure intervention coverage and outcomes within specific geographical areas (e.g. neighbourhoods). Spillover estimates are likely to be very sensitive to the way areas are defined.[103] The size or shape of the area may determine whether spillovers are detected. Pre-specification of area definition before looking at outcomes prevents investigators

from selectively using a definition that provides the most favourable result. Ideally, area definition is based on the hypothesized strength and scale of spillovers. For example, when spillover effects are expected to be weak, measurement is best within small areas where spillovers are most likely to be detected. Expected transmission dynamics may also inform area definition: a study of the cholera vaccine measured spillovers associated with immunization coverage near a shared water source, where cholera transmission was hypothesized to be the strongest.[97]

Power calculations for spillovers

Typically, spillover effects are smaller than total or overall effects of interventions, so larger sample sizes are needed in order to detect them. We recommend that investigators conduct power calculations in the study design phase, to assess whether statistical power will be sufficient to detect spillovers. Several studies have provided sample size formulas to estimate spillovers using randomized designs and variants of them.[9,16] For non-standard study designs, simulations can be used to estimate statistical power.[78]

Pre-specifying spillovers

We encourage investigators who plan to measure spillovers to pre-specify the specific spillover parameter(s), the scale at which spillovers are expected, and the hypothesized mechanism(s) of spillover be estimated in a study protocol. Pre-specification reduces the chance that the spillover parameters selected are those that detect positive spillovers, whether intentionally or not.[105] It also reduces the chance of publication bias.[106,107] We also encourage investigators to use the terms 'spillovers' or 'indirect effects' to refer to spillovers in protocols and manuscripts, because these are the most commonly used terms in the literature, and they provide a direct link to the theoretical literature on this topic.

Summary

We have defined different types of spillover effects relevant to a wide range of interventions using standardized notation to encourage estimation and reporting of spillovers by a broad range of investigators. We have also provided a general introduction to assumptions required to make causal inferences about spillover effects. Rigorous definition and study of spillover effects will improve the accuracy of estimates of the population-level impact and cost-effectiveness of interventions that have benefits beyond direct recipients.

Supplementary Data

Supplementary data are available at *IJE* online.

Conflict of interest: None declared.

References

1. Fine PE. Herd immunity: history, theory, practice. *Epidemiol Rev* 1993;15:265–302.
2. John TJ, Samuel R. Herd immunity and herd effect: new insights and definitions. *Eur J Epidemiol* 2000;16:601–06.
3. Miguel E, Kremer M. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 2004;72:159–217.
4. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO. The diffusion of microfinance. *Science* 2013;341:1236498.
5. Halloran ME, Struchiner CJ. Study designs for dependent happenings. *Epidemiology* 1991;2:331–38.
6. Halloran ME, Struchiner CJ. Causal inference in infectious diseases. *Epidemiology* 1995;6:142–51.
7. Longini IM, Sagatelian K, Rida WN, Halloran ME. Optimal vaccine trial design when estimating vaccine efficacy for susceptibility and infectiousness from multiple populations. *Stat Med* 1998;17:1121–36.
8. Hudgens MG, Halloran ME. Toward causal inference with interference. *J Am Stat Assoc* 2008;103:832–42.
9. Halloran E, Longini IM Jr, Struchiner CJ. *Design and Analysis of Vaccine Studies*. New York, NY: Springer, 2010.
10. VanderWeele TJ, Tchetgen Tchetgen EJ. Effect partitioning under interference in two-stage randomized vaccine trials. *Stat Probab Lett* 2011;81:861–69.
11. Clemens J, Shin S, Ali M. New approaches to the assessment of vaccine herd protection in clinical trials. *Lancet Infect Dis* 2011;11:482–87.
12. VanderWeele T, Tchetgen Tchetgen E, Halloran M. Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects. *Epidemiology* 2012;23:751–61.
13. Tchetgen EJT, VanderWeele TJ. On causal inference in the presence of interference. *Stat Methods Med Res* 2012;21:55–75.
14. Halloran ME. The minicommunity design to assess indirect effects of vaccination. *Epidemiol Methods* 2012;1:83–105.
15. Angelucci M, Maro VD. Program Evaluation and Spillover Effects. *J Dev Effectiveness* 2016;8(1).
16. Baird S, Bohren A, McIntosh C, Ozler B. *Designing Experiments to Measure Spillover Effects*. PIER Working Paper No. 14–006. 2014. http://ssrn.com/abstract¼2402749 or http://dx.doi.org/10.2139/ssrn.2402749 (9 June 2017, date last accessed).
17. Bowers J, Fredrickson MM, Panagopoulos C. Reasoning about interference between units: a general framework. *Polit Anal* 2013;21(1):97–124.
18. Sinclair B, McConnell M, Green DP. Detecting spillover effects: Design and analysis of multilevel experiments. *Am J Polit Sci* 2012;56:1055–69.
19. Benjamin-Chung J, Abedin J, Berger D *et al.* Spillover effects on health outcomes in low- and middle-income countries: a systematic review. *Int J Epidemiol* 2017;46:1251–76.

20. Anderson R, Garnett G. Low-efficacy HIV vaccines: potential for community-based intervention programmemes. *Lancet* 1996;348:1010–13.
21. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health* 2000;21:121–45.
22. Rubin D. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Stat Sci* 1990;5:472–80.
23. Cox DR. *Planning of Experiments*. Oxford, UK: Wiley, 1958.
24. Sobel ME. What do randomized studies of housing mobility demonstrate? *J Am Stat Assoc*. 2006;101:1398–407.
25. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med* 2007;26: 2–19.
26. Hayes RJ, Moulton LH. *Cluster Randomised Trials*. Abingdon, UK: Taylor & Francis, 2009.
27. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. Hoboken, NJ: Wiley, 2010.
28. Chong A, Gonzalez-Navarro M, Karlan D, Valdivia M. *Effectiveness and Spillovers of Online Sex Education: Evidence from a Randomized Evaluation in Colombian Public Schools*. Cambridge, MA: National Bureau of Economic Research, 2013.
29. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;3:143–55.
30. Avitabile C. *Spillover Effects in Healthcare Programmes: Evidence on Social Norms and Information Sharing*. Washington, DC: Inter-American Development Bank, 2012.
31. Buttenheim A, Alderman H, Friedman J. Impact evaluation of school feeding programmes in Lao People's Democratic Republic. *J Dev Eff* 2011;3:520–42.
32. Contreras D, Maitra P. *Health Spillover Effects of a Conditional Cash Transfer Programme*. 2012. http://www.buseco.monash.edu.au/eco/research/papers/2013/4413healthcontrerasmaitra.pdf (9 June 2017, date last accessed).
33. Fitzsimons E, Malde B, Mesnard A, Vera-Hernández M. *Household Responses to Information on Child Nutrition: Experimental Evidence From Malawi*. 2012. http://papers.ssrn.com/sol3/papers.cfm?abstract_id¼2034133 (9 June 2017, date last accessed).
34. Handa S, Huerta M-C, Perez R, Straffon B. *Poverty, Inequality, and Spillover in Mexico's Education, Health, and Nutrition Programme*. 2001. http://agris.fao.org/agris-search/search.do?recordID¼US2012205787 (9 June 2017, date last accessed).
35. House JI, Ayele B, Porco TC *et al*. Assessment of herd protection against trachoma due to repeated mass antibiotic distributions: a cluster-randomised trial. *Lancet* 2009;373:1111–18.
36. Kazianga H, de Walque D, Alderman H. *School feeding Programs and the nutrition of siblings: evidence from a randomized trial in rural Burkina Faso*. Oklahoma State University, Department of Economics and Legal Studies in Business, 2009.
37. Ribas RP, Soares FV, Teixeira CG, Silva E, Hirata GI. *Externality and Behavioural Change Effects of a Non-Randomised CCT Programmeme: Heterogeneous Impact on the Demand for Health and Education*. Brasilia: International Policy Centre for Inclusive Growth, 2011.
38. Banerjee AV, Duflo E, Glennerster R, Kothari D. Improving immunisation coverage in rural India: clustered randomised controlled evaluation of immunisation campaigns with and without incentives. *BMJ* 2010;340:c2220.
39. Hawley WA, Phillips-Howard PA, Kuile FOT *et al*. Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in Western Kenya. *Am J Trop Med Hyg* 2003;68(Suppl 4):121–27.
40. Kremer M, Miguel E. The illusion of sustainability. *Q J Econ* 2007;122:1007–65.
41. German D, Sutcliffe CG, Sirirojn B *et al*. Unanticipated effect of a randomized peer network intervention on depressive symptoms among young methamphetamine users in Thailand. *J Community Psychol* 2012;40:799–813.
42. Shakya HB, Christakis NA, Fowler JH. Social network predictors of latrine ownership. *Am J Public Health* 2014;104:5.
43. Kim DA, Hwong AR, Stafford D *et al*. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *Lancet* 2015;386:145–53.
44. Vanderweele T, An W. Social networks and causal inference. *Handbook of Causal Analysis for Social Research* New York, NY: Springer, 2013.
45. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *N Engl J Med* 2007;357:370–79.
46. Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. *N Engl J Med* 2008;358:2249–58.
47. Snijders TAB. The statistical evaluation of social network dynamics. *Sociol Methodol* 2001;31:361–95.
48. Snijders TA. Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S (eds). *Models and Methods in Social Network Analysis*. New York: Cambridge, 2005, pp.215–47.
49. Cohen-Cole E, Fletcher JM. Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *BMJ* 2008;337:a2533.
50. Lyons R. The spread of evidence-poor medicine via flawed social-network analysis. *Stat Polit Policy*. 2011;2:1.
51. Noel H, Nyhan B. The 'unfriending' problem: The consequences of homophily in friendship retention for causal estimates of social influence. *Soc Netw* 2011;33:211–18.
52. VanderWeele TJ. Sensitivity analysis for contagion effects in social networks. *Sociol Methods Res*. 2011;40:240–55.
53. VanderWeele TJ, Ogburn EL, Tchetgen Tchetgen JE. Why and when 'flawed' social network analyses still yield valid tests of no contagion. *Stat Polit Policy* 2012;3:1–11.
54. Sacerdote B. Peer effects with random assignment: results for Dartmouth Roommates. *Q J Econ* 2001;116:681–704.
55. Longini IM, Koopman JS, Haber M, Cotsonis GA. Statistical inference for infectious diseases risk-specific household and community transmission parameters. *Am J Epidemiol* 1988; 128:845–59.
56. Halloran ME, Haber M, Longini IM Jr, Struchiner CJ. Direct and indirect effects in vaccine efficacy and effectiveness. *Am J Epidemiol*. 1991;133(4):323–331.
57. Halloran ME, Struchiner CJ, Longini IM. Study designs for evaluating different efficacy and effectiveness aspects of vaccines. *Am J Epidemiol* 1997;146:789–803.
58. Hudgens MG, Halloran ME. Causal vaccine effects on binary postinfection outcomes. *J Am Stat Assoc* 2006;101:51–64.
59. VanderWeele TJ, Tchetgen Tchetgen EJ. Bounding the infectiousness effect in vaccine trials. *Epidemiology* 2011;22:686–93.

60. Halloran ME, Hudgens MG. Causal inference for vaccine effects on infectiousness. *Int J Biostat* 2012;8:2.
61. Préziosi M-P, Halloran ME. Effects of pertussis vaccination on transmission: vaccine efficacy for infectiousness. *Vaccine*.2003; 21:1853–61.
62. Hong G, Raudenbush SW. Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educ Eval Policy Anal* 2005;27:205–24.
63. Perez-Heydrich C, Hudgens MG, Halloran ME, Clemens JD, Ali M, Emch ME. Assessing effects of cholera vaccination in the presence of interference. *Biometrics* 2014; 70:731–44.
64. Liu L, Hudgens MG, Becker-Dreps S. On inverse probability-weighted estimators in the presence of interference. *Biometrika* 2016;103:829–42.
65. Lundin M, Karlsson M. Estimation of causal effects in observational studies with interference between units. *Stat Methods Appl* 2014;23:417–33.
66. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
67. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics* 1985;41:103–16.
68. Diamond A, Sekhon JS. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Rev Econ Stat* 2013;95: 932–45.
69. Bor J, Moscoe E, Mutevedzi P, Newell M-L, Ba¨ rnighausen T. Regression discontinuity designs in epidemiology. *Epidemiology* 2014;25:729–37.
70. Imbens G, Lemieux T. *Regression Discontinuity Designs: A Guide to Practice*. Cambridge, MA: National Bureau for Economic Research, 2007.
71. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996;91: 444–55.
72. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;29:722–29.
73. Janssens W. Measuring Externalities in Programme Evaluation. *Tinbergen Institute Discussion Paper*. Amsterdam: Tinbergen Institute, 2005.
74. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;347:f6409.
75. Fisher R. *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd, 1925.
76. 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.
77. *WASH Benefits Bangladesh Primary Outcome Analysis Plan Update*. 2016. https://osf.io/63mna/ (9 June 2017, date last accessed).
78. Sofrygin O, van der Laan MJ. *Targeted Maximum Likelihood Estimation for Networks*. 2015. Report No.: R package version 0.1. https://github.com/osofr/tmlenet (9 June 2017, date last accessed).
79. Rosenbaum PR. Interference between units in randomized experiments. *J Am Stat Assoc* 2007;102:191–200.
80. Aral S, Walker D. Identifying social influence in networks using randomized experiments. *IEEE Intell Syst* 2011;26: 91–96.
81. Toulis P, Kao E. *Estimation of Causal Peer Influence Effects*. 2013. http://www.jmlr.org/proceedings/papers/v28/toulis13.html (10 January 2017, date last accessed).
82. Ugander J, Karrer B, Backstrom L, Kleinberg J. *Graph Cluster Randomization: Network Exposure to Multiple Universes*. 2013. https://arxiv.org/abs/1305.6979 (9 June 2017, date last accessed).
83. Eckles D, Karrer B, Ugander J. *Design and analysis of experiments in networks: Reducing bias from interference*. 2014. https://arxiv.org/abs/1404.7530 (9 June 2017, date last accessed).
84. Walker D, Muchnik L. Design of randomized experiments in networks. *Proc IEEE* 2014;102:1940–51.
85. van der Laan MJ. Causal inference for a population of causally connected units. *J Causal Infer* 2014;2:13–74.
86. Aral S, Walker D. Tie strength, embeddedness, and social influence: a large-scale networked experiment. *Manag Sci* 2014;60: 1352–70.
87. Aronow PM, Samii C. *Estimating Average Causal Effects Under Interference Between Units*. 2015. https://arxiv.org/abs/1305.6156 (9 June 2017, date last accessed).
88. Sofrygin O, van der Laan MJ. Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. *J Causal Inference*. 2017;5:35.
89. Choi D. Estimation of monotone treatment effects in network experiments. J Am Stat Assoc *In press*.
90. Basse GW, Airoldi EM. *Optimal Model-Assisted Design of Experiments for Network Correlated Outcomes Suggests New Notions of Network Balance*. 2016. https://arxiv.org/abs/1507.00803 (9 June 2017, date last accessed).
91. Struchiner CJ, Halloran ME. Randomization and baseline transmission in vaccine field trials. *Epidemiol Amp Infect* 2007; 135:181–94.
92. Valente T. Network interventions. *Science* 2012;337:49–53.
93. Rogers EM. *Diffusion of Innovations*. 4th ednn. New York, NJ: Simon and Schuster, 2010.
94. Greenberg MR. The diffusion of public health innovations. *Am J Public Health* 2006;96:209–10.
95. Ali M, Sur D, You YA *et al*. Herd protection by a bivalent-killed-whole-cell oral cholera vaccine in the slums of Kolkata, India. *Clin Infect Dis* 2013;56:1123–31.
96. Chen W-J, Moulton LH, Saha SK, Mahmud AA, Arifeen SE, Baqui AH. Estimation of the herd protection of Haemophilus influenzae type b conjugate vaccine against radiologically confirmed pneumonia in children under 2 years old in Dhaka, Bangladesh. *Vaccine*. 2014;32:944–48.

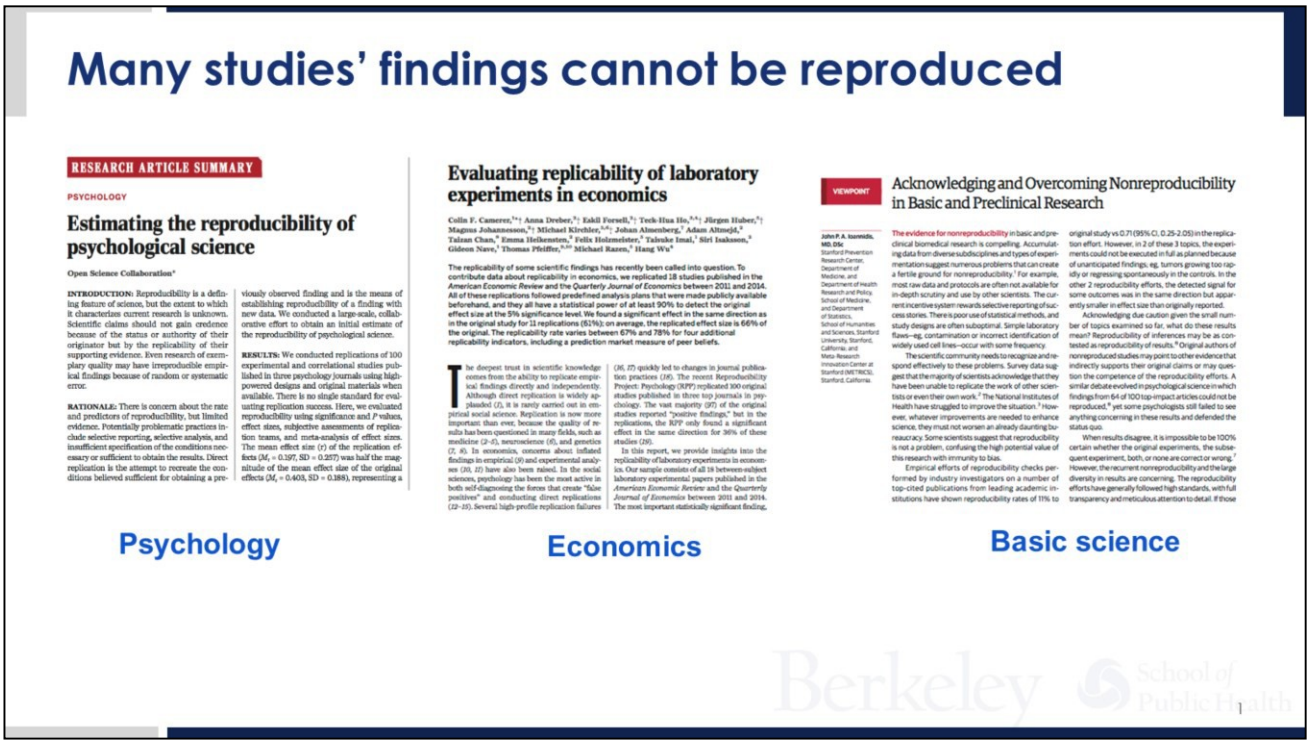
97. Emch M, Ali M, Root ED, Yunus M. Spatial and environmental connectivity analysis in a cholera vaccine trial. *Soc Sci Med* 2009;68:631–37.
98. Haile M, Tadesse Z, Gebreselassie S *et al.* The association between latrine use and trachoma: a secondary cohort analysis from a randomized clinical trial. *Am J Trop Med Hyg* 2013;89: 717–20.
99. Huq A, Yunus M, Sohel SS *et al.* Simple sari cloth filtration of water is sustainable and continues to protect villagers from cholera in Matlab, Bangladesh. *mBio*.2010;1:e00034–10.
100. Khatib AM, Ali M, von Seidlein L *et al.* Effectiveness of an oral cholera vaccine in Zanzibar: findings from a mass vaccination campaign and observational cohort study. *Lancet Infect Dis* 2012;12:837–44.
101. Root ED, Giebultowicz S, Ali M, Yunus M, Emch M. The role of vaccine coverage within social networks in cholera vaccine efficacy. *PLoS One* 2011;6:e22971.
102. Root ED, Lucero M, Nohynek H *et al.* Distance to health services affects local-level vaccine efficacy for pneumococcal conjugate vaccine (PCV) among rural Filipino children. *Proc Natl Acad Sci U S A* 2014;111:3520–25.
103. Openshaw S. Ecological fallacies and the analysis of areal census data. *Environ Plann A* 1984;16:17–31.
104. Arnold BF, Hogan DR, Colford JM, Hubbard AE. Simulation methods to estimate design power: an overview for applied research. *BMC Med Res Methodol* 2011;11:94.
105. Miguel E, Camerer C, Casey K *et al.* Promoting transparency in social science research. *Science* 2014;343:30–31.
106. Ioannidis JPA, Munafò MR, Fusar-Poli P, Nosek BA, David SP. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn Sci* 2014;18:235–41.
107. Dal-Re´ R, Ioannidis JP, Bracken MB *et al.* Making prospective registration of observational research a reality. *Sci Transl Med* 2014;6:224.

Berkeley School of Public Health

Replication, transparency & reproducibility

PHW250 B – Andrew Mertens

Andrew Mertens: This video focuses on replication, transparency, and reproducibility.



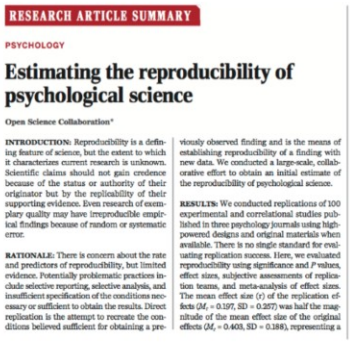
In recent years, a growing number of studies have documented evidence that prior published studies could not be reproduced, and this has occurred across disciplines. From psychology, to economics, to basic science. When I say the studies can't be reproduced, what I mean is that original studies provided certain effect estimates. For example, they would report a certain measure of association for an exposure in a disease. And then another study would try to replicate this result in another population with similar features, and they were not able to obtain a similar result.

Alternatively this could mean, that the data from the original study was analyzed and no one was able to replicate the original result. In epidemiology, we tend not to do exact replications from start to finish of prior studies, because our studies tend to involve large populations, it would be quite expensive.

In psychology though, where a large body of the-- in which the largest body of evidence has unfolded, that there are concerns about reproducibility. It's common for experiments to be done with relatively small numbers of participants, in a relatively short period of time. And as a result, it is possible to carry out an experiment that had already been done in one lab, in another lab, in a different population.

And it has been found, that many of the published findings, in psychology in particular, cannot be reproduced. This calls into question, the validity of the evidence in the published literature in these disciplines.

Many studies' findings cannot be reproduced



RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (*r*) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a

Psychology

A study attempted to replicate 100 psychology studies; they only reproduced a little over a third of them.

"If one assumes that the vast majority of the original researchers were honest and diligent, then a large proportion of the problems can be **explained only by unconscious biases.**"

Nuzzo et al., 2015. *Nature* Vol 526.

To give you a concrete example, there was one study that attempted to replicate 100 psychology studies, and they were only able to reproduce a little over a third of them. And the quote from one of the articles describing this reproducibility crisis is that, "If one assumes that the vast majority of the original researchers were honest and diligent,"-- meaning they did not engage in any fraudulent activities, to obtain desirable results, --"then a large proportion of the problems can be explained only by unconscious biases. We'll come back to what this unconscious bias could look like in a moment.

Threats to reproducibility

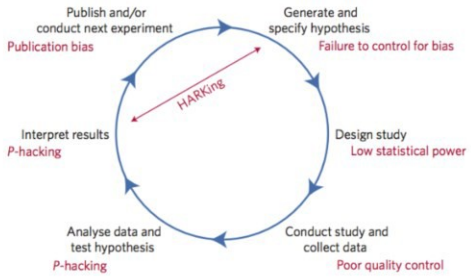


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Failure to control for bias

- **Observational studies:**
  - Unmeasured confounding
  - Measurement error
  - Etc.
- **Randomized trials:**
  - Improper randomization
  - Lack of blinding
  - Lack of allocation concealment
  - Measurement error
  - Etc.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

3

The figure on the left shows several different threats to reproducibility. I'm going to go through each of these. The first is a failure to control for bias. We spend a great deal of time talking about this in epidemiology. For example in observational studies, we're very concerned about trying to predict the extent of unmeasured confounding, or measurement error, that could occur.

And in randomized trials, we might have improper randomization, or a lack of blinding, or lack of allocation concealment, that could contribute to bias in our final results. So these are things that are well known to be sources of systematic error, that could cause us to produce a result that cannot be reproduced later on.

Threats to reproducibility

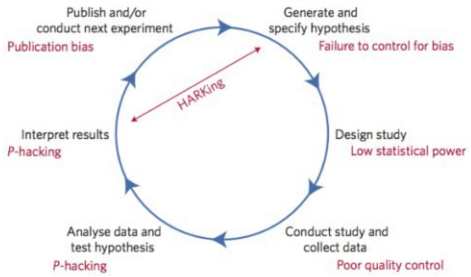


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Low statistical power

- If the sample size is too small, you may fail to detect an effect that is truly there.
- I.e., you may make a Type II error

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

Another reason is low statistical power. If the sample size is too small in a study, you may fail to detect an effect that's truly there. And this is called making a type II error. And if we make a type II error, and then we publish this finding this again is inconsistent with the truth. If the true effect exists, but we fail to detect it, what we put out in the published literature will be incorrect, and difficult to reproduce in the future.

Threats to reproducibility

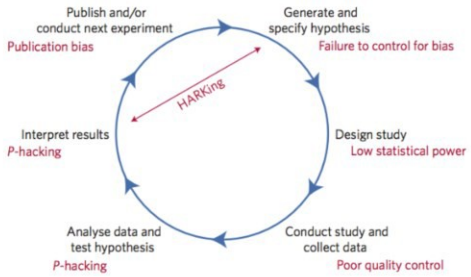


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Poor quality control

- Errors during data collection and data entry can lead to bias and misclassification that are nearly impossible to correct or to identify.
- Robust survey design and staff training is needed to prevent this.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

Poor quality control, is another major factor affecting reproducibility. There can be errors that occur during data collection and entry, and these can contribute to bias and misclassification. And they're also very difficult kinds of errors to identify and correct. What do we mean? Well, during data collection, if we use a survey that has a poor design it may trigger respondents to answer questions in certain ways that biases their answers. This is something that's very difficult to do anything about, once the survey has already been conducted.

There is also a variety of issues that can arise during data analysis, that can lead to errors in the final results. If errors are made during coding that are not caught, a study may have improper, or inaccurate conclusions about its findings.

Threats to reproducibility

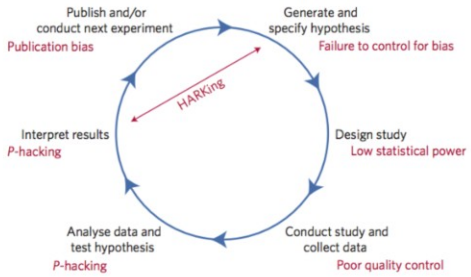


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

P-hacking

- P-hacking occurs when investigators make adjustments to their statistical model repeatedly after viewing the results in order to obtain a p-value < 0.05 or < 0.001
- Statistical significance is based on the probability that a particular result would be observed due only to chance.
- When ignoring the number of models run, eventually it is almost always possible to obtain a p-value < 0.05.
- However, this p-value is confounded by the number of tests and does not necessarily represent the true probability of a Type I error.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

P-hacking has gotten a lot of attention in recent years. This occurs when investigators have already seen the data, and while they're doing their analysis they continually make adjustments to their statistical model. And then view the results again and again, and do this until they obtain a desirable p-value of say under 0.05 or under 0.001, depending on their goal.

So statistical significance is based on p-values, and it's the probability that a particular result would be observed due only to chance. And when you ignore the number of models that have been run, with slight tweaks to hypothetically improve the model performance, it's possible that you'll eventually obtain a p-value under 0.05, simply from tweaking the model enough times. However this p-value is confounded by the number of tests that has been run. And typically when people do this, they don't correct for the number of tests that they did, or the number of models they run. And so the p-value doesn't necessarily represent the true probability of a type I error. So we'll come back to what we can do to prevent this. But this is something that we strongly discourage, this practice of changing your model over and over, once you've seen the results.

Threats to reproducibility

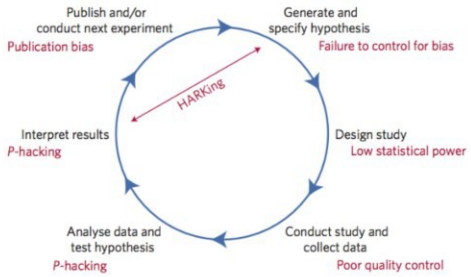


Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Publication bias

- Occurs when null or non-desirable findings are not published in the peer reviewed literature.
- If only a small percentage of all null / non-desirable findings are published, those that are published will appear not to be reproducible.
- Failure to publish such studies may lead investigators to conclude that an intervention or exposure has a more desirable impact than it truly does.

Munafò et al., 2017. DOI: 10.1038/s41562-016-0021

7

Publication bias is another major threat. It occurs when a paper presenting a study's findings, reports null, or not desirable findings. And so it's difficult to get them published in the peer reviewed literature, or they don't ever get published at all. Unfortunately, it's the case that journals are incentivized to select papers for publication that have exciting new findings, that demonstrate an effect in most cases. And when a study does not find evidence of an effect, it's much less likely it will get published, or publish in a good journal.

If only a small percentage of these null or non-desirable findings are published, those that are published will appear not to be reproducible. Even though there may be other examples out there, there just were unpublished null findings. And failure to publish such studies, may lead investigators to conclude interventions or exposures have a more desirable impact, than they really do.

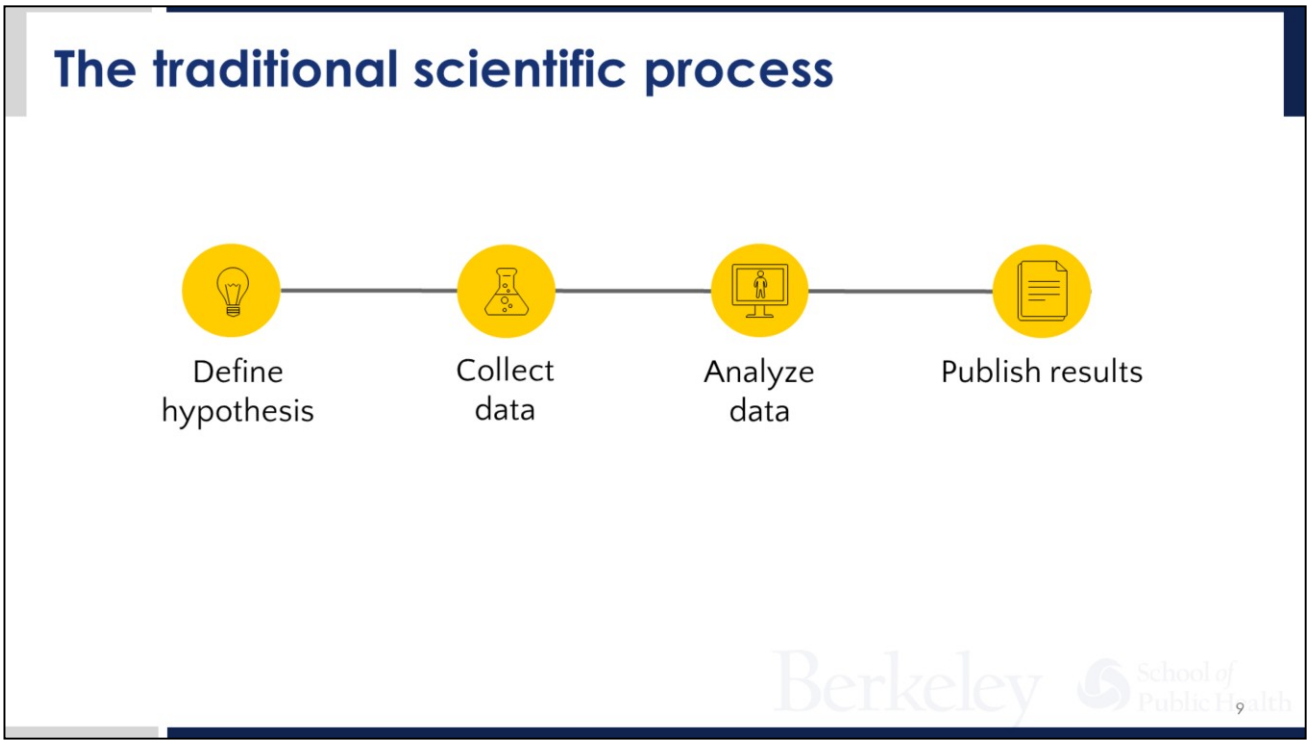
Confirmation bias is human nature

- It is natural to look for the patterns we expect to find — to confirm our biases.
- When we don't obtain a result we expect to see, we are more likely to carefully vet our code to check for errors.
    - This may mean that errors in results that confirm our expectations go unnoticed.
- Scientists are good at coming up with explanations for unexpected findings.
- How can we improve the scientific process to account for humans' natural confirmation bias?

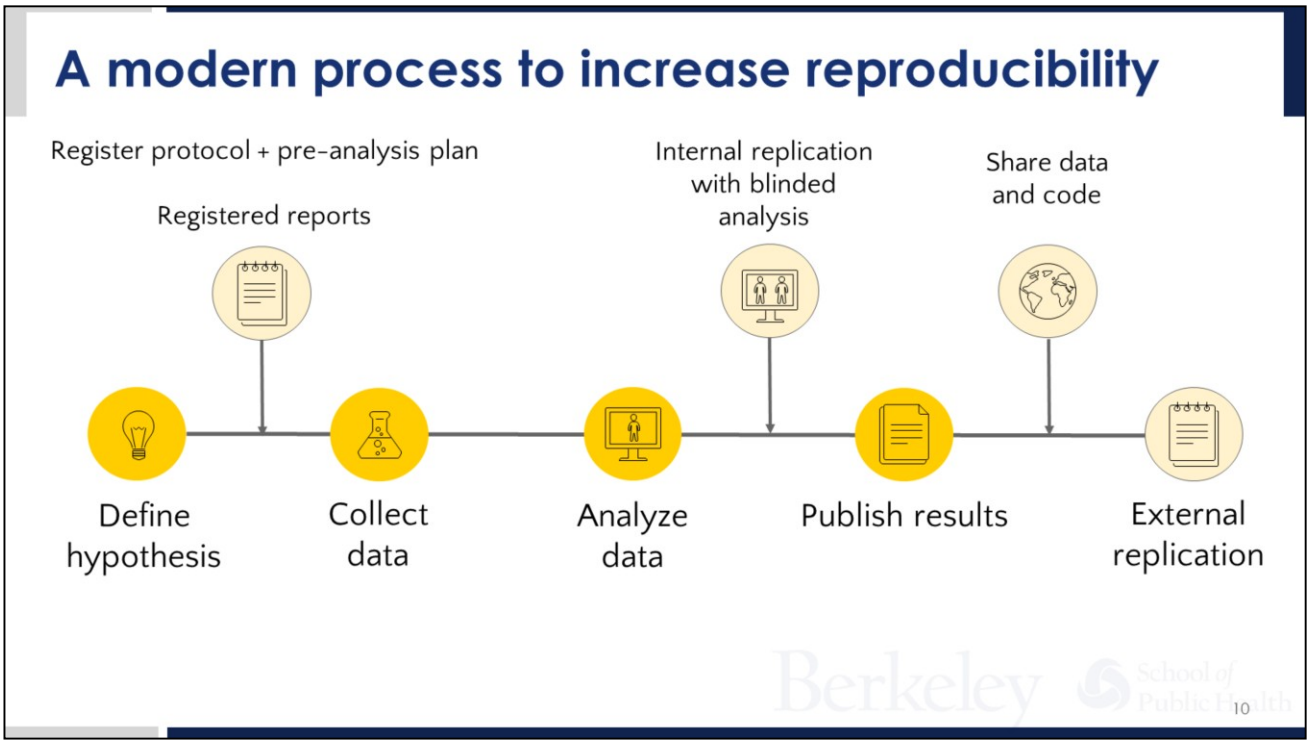
Nuzzo et al., 2015. *Nature* Vol 526.

It's natural for humans to look for patterns where we expect to find them, and to want to confirm our own biases. We can refer to this as confirmation bias. The challenge is that when we obtain results in our analysis that we don't expect to see, we're much more likely to carefully vet our code, and all the procedures in our study, to find an error.

But that means that there may be errors lurking in results that do confirm our expectations that go unnoticed. Also, when we see an unusual finding that's not consistent with our expectation, scientists are just naturally good at coming up with explanations for those possible findings-- for those unexpected findings. Since confirmation bias can be argued to be a fact of life, or just simply part of human nature, how can we improve the scientific process to account for this natural confirmation bias?

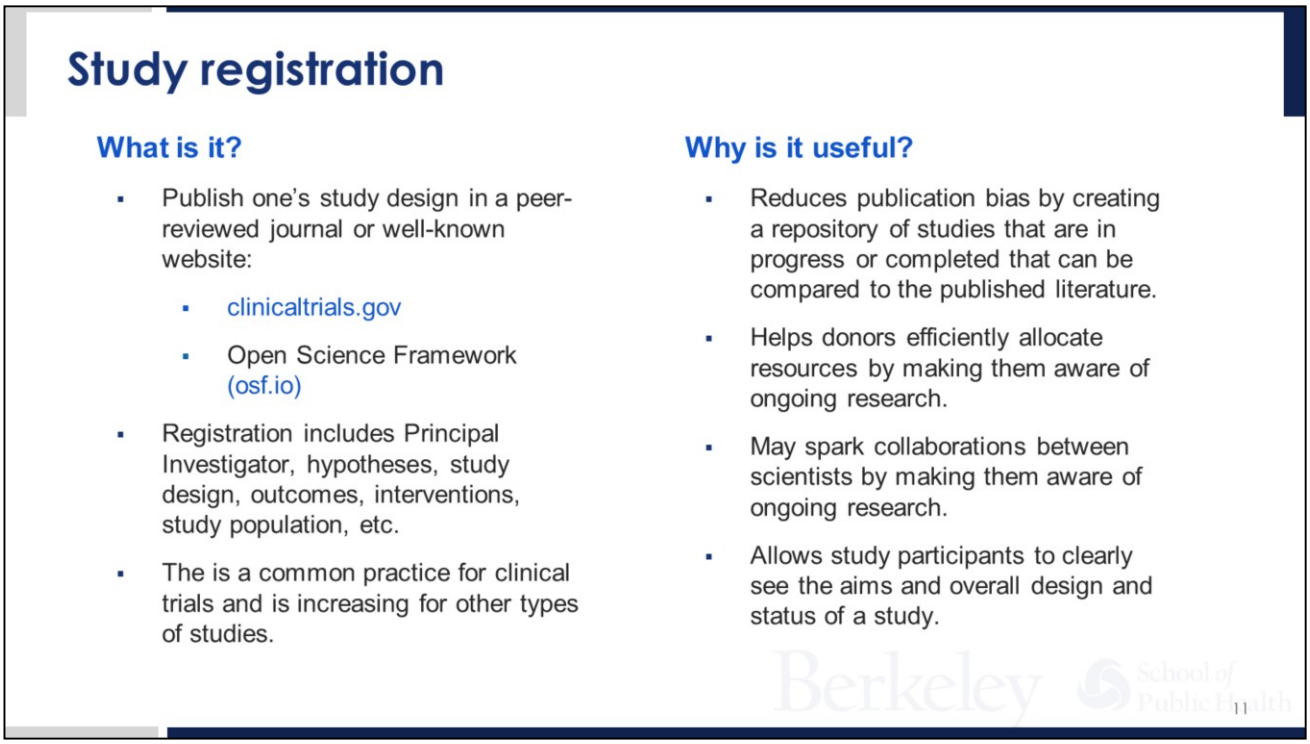


First, I'll show you the traditional scientific process. And then I'll show you some efforts to augment this process to increase reproducibility and build in some insurance against our natural tendency to confirm our own biases. So traditionally we would define our hypothesis, collect data, analyze it, and then publish our results.



A new more modern process to increase reproducibility would add additional steps shown in light yellow. So after defining our hypothesis, we would register protocol and pre-analysis plan. And we might also use something called a registered report. I'll define these in a moment. We then collect data, and analyze it. And we use internal replication with blinded analysis to minimize error prior to publication. After publishing, or at the time of publishing, we can share our data and code publicly. And then finally, our results can be externally replicated by people outside of the original study team.

So now I'm going to go through and define all of these different terms from these new additions to the scientific process.

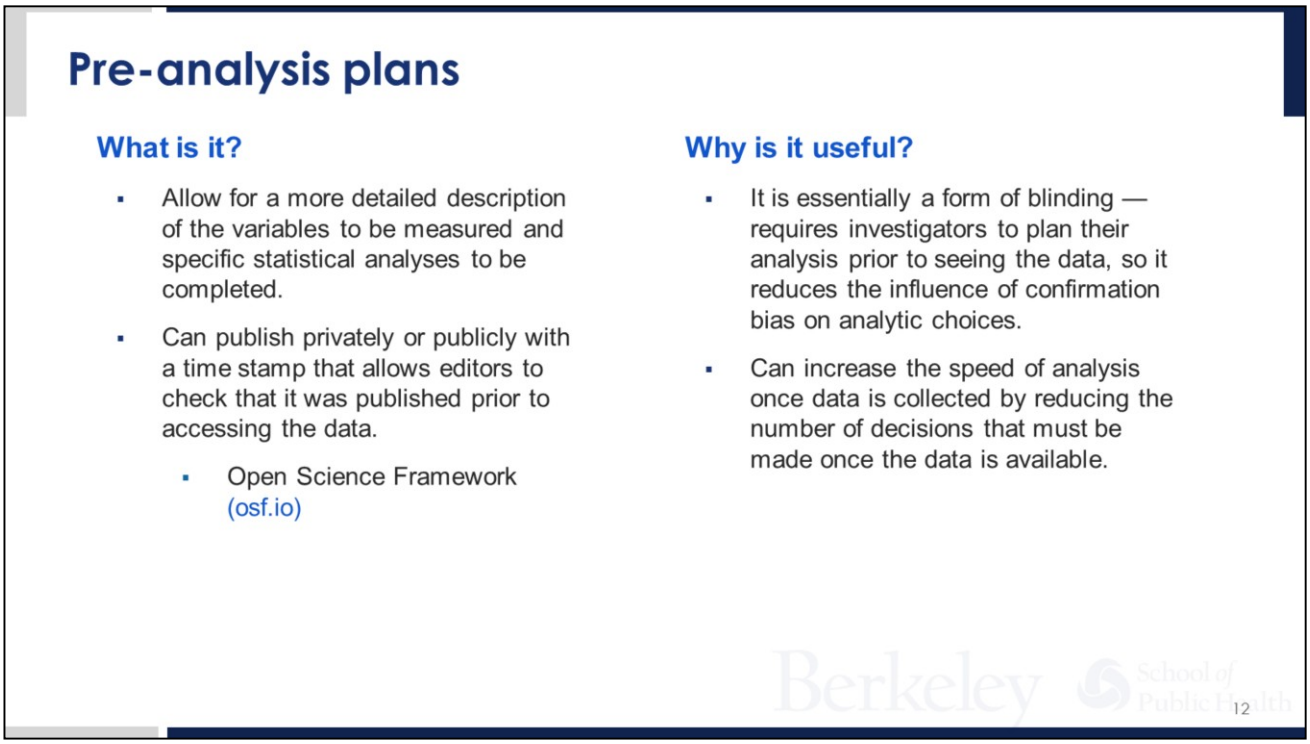


The first is study registration. This involves publishing a study design with relevant details in a peer reviewed journal or on a well-known web site, such as clinicaltrials.gov or Open Science Framework. This will include key information such as, the principal investigators name and contact information, the study hypotheses, design, outcomes, interventions, specific populations to be enrolled, et cetera.

And this is something that's been very common for clinical trials for a number of years, but only recently has become a standard for other types of studies as well, largely due to this reproducibility crisis.

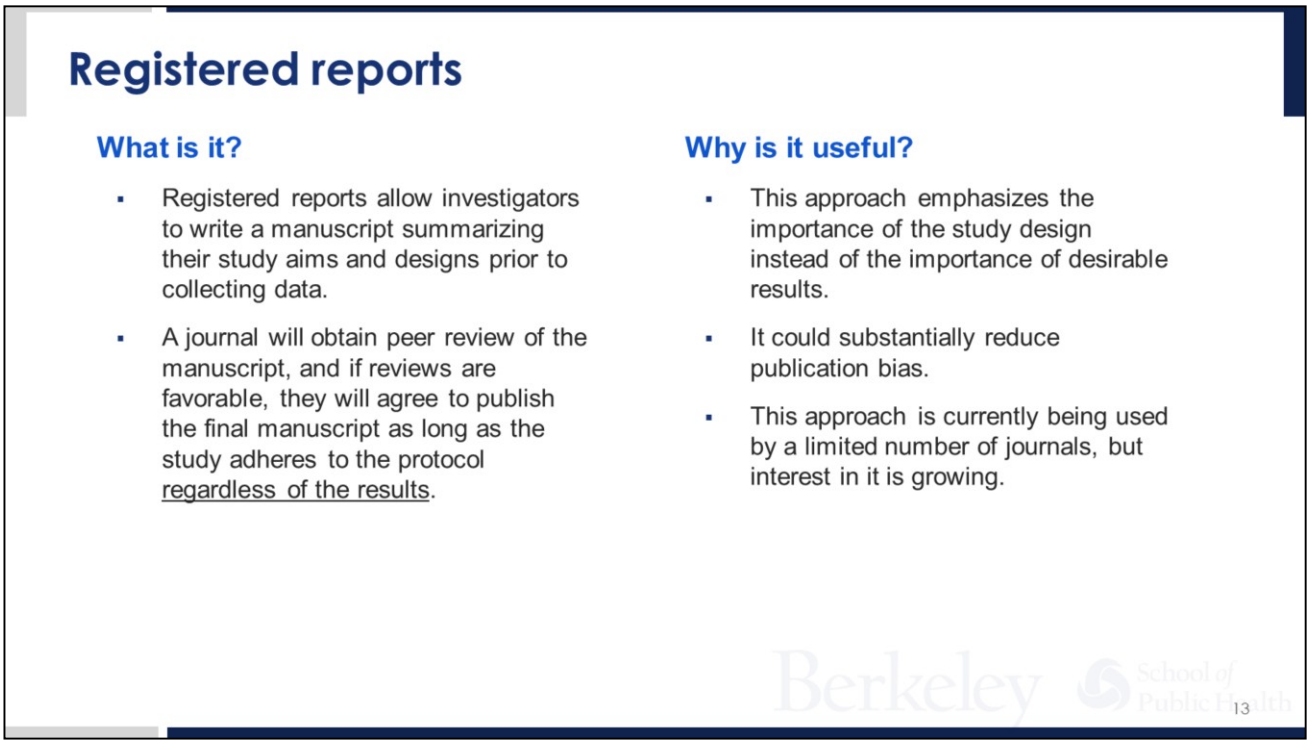
Study registration can reduce publication bias by creating a repository of studies that are in progress, or completed, that can be compared to the published literature. So if you know that a certain trial has been underway for a couple of years, and perhaps you've been waiting for its results, you know the study might have completed its data collection, because typically these websites require updates about the status of the study. And if you don't see a publication in a journal for presenting the study's results, you can contact the principal investigator of the study to inquire about the status. This is one way to reduce publication bias.

It also helps donors efficiently allocate resources, by making them aware of ongoing research. And it also can spark collaborations between scientists, by making them aware of research. Finally, it has a benefit for study participants because it allows them to go and see the aims, and overall design, and status of a study that they're participating in.



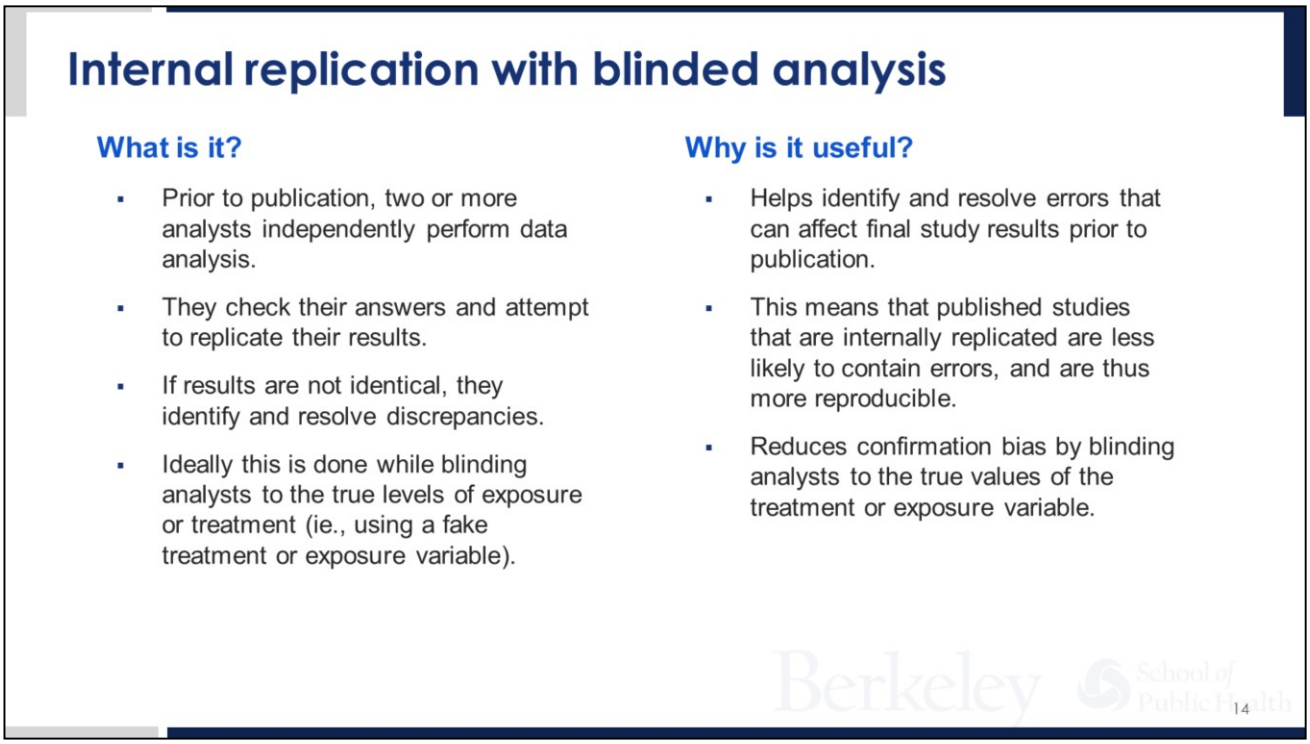
Pre-analysis plans build on study registration, and allow for a more detailed description of the particular variables that are collected and measured, and the specific statistical analyses to be completed. So this could include the covariates that are measured, how confounding is addressed, if any matching was done, the specific statistical models to be used, et cetera. The analysis plan can be published privately or publicly, and sometimes this is-- and it can often be done with a time-stamp that would later allow you, at the publication stage, to have an editor check that the pre-analysis plan was published prior to accessing the data. And an example of a place where you can do this is the Open Science Framework.

These plans are really useful, and they're essentially a form of blinding in that they require investigators to plan their analysis before they see the data. And so it reduces the influence of confirmation bias on your analytic choices. It can also really increase the speed of a data analysis after data is collected, because many decisions have already been made. And so once the data is in hand, one can simply follow the pre-analysis plan.



Registered reports are a relatively new tool that allow investigators to write a manuscript summarizing their study objectives, hypotheses, and design, before they conduct their study and collect data. The idea is that a journal would send out this article for peer review, and if the reviews are favorable they'll agree to publish the final manuscript, including all the results, as long as the study adhered to the original protocol, and was conducted well, regardless of the results. So this is a big deal, because it means that you could have a guarantee, that even if you had a null finding or an undesirable finding, as long as you did the study that you set out to do, that was already peer reviewed with that journal, they would be committed to publishing the final paper.

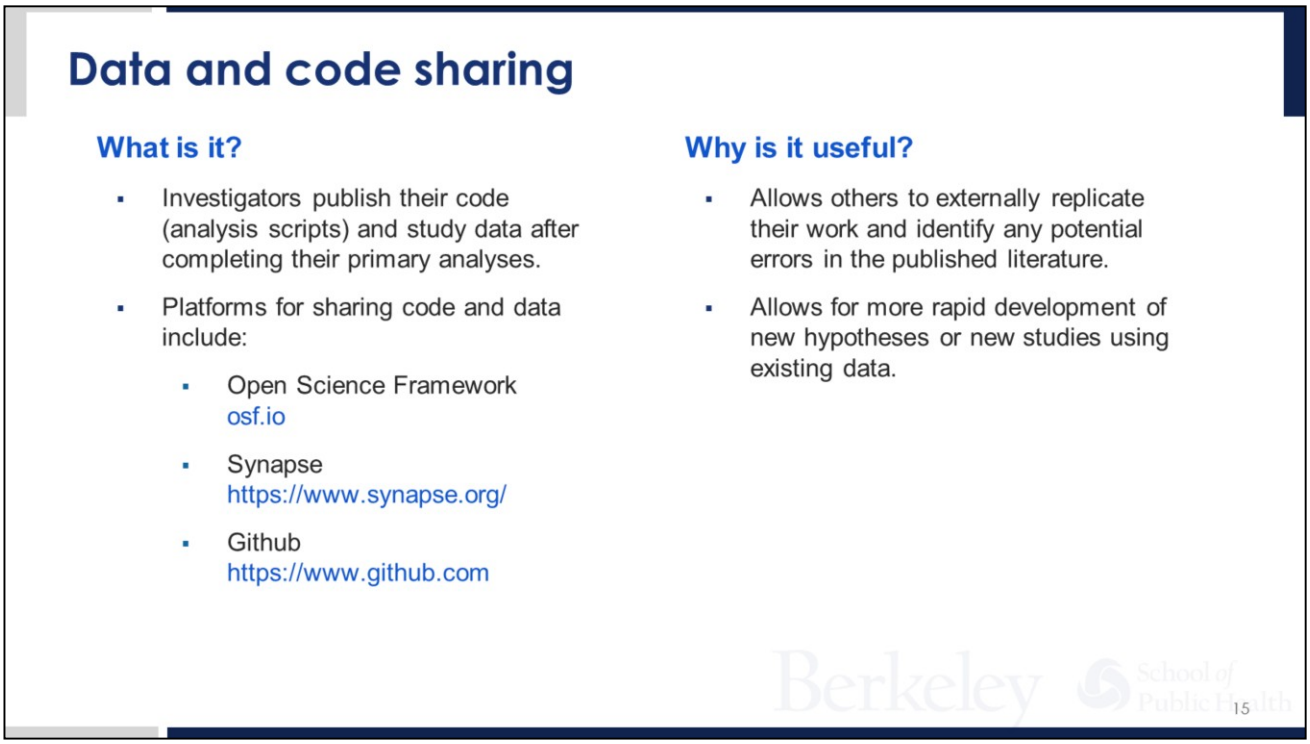
This approach is useful, because it emphasizes the importance of study design instead of finding desirable results. And this pressure to obtain desirable results, is partly what has created this reproducibility crisis. It also can substantially reduce publication bias, because of the commitment to publish results, regardless of what they are. And it's currently a process that's in use in a limited number of journals, but there is growing interest in registered reports.



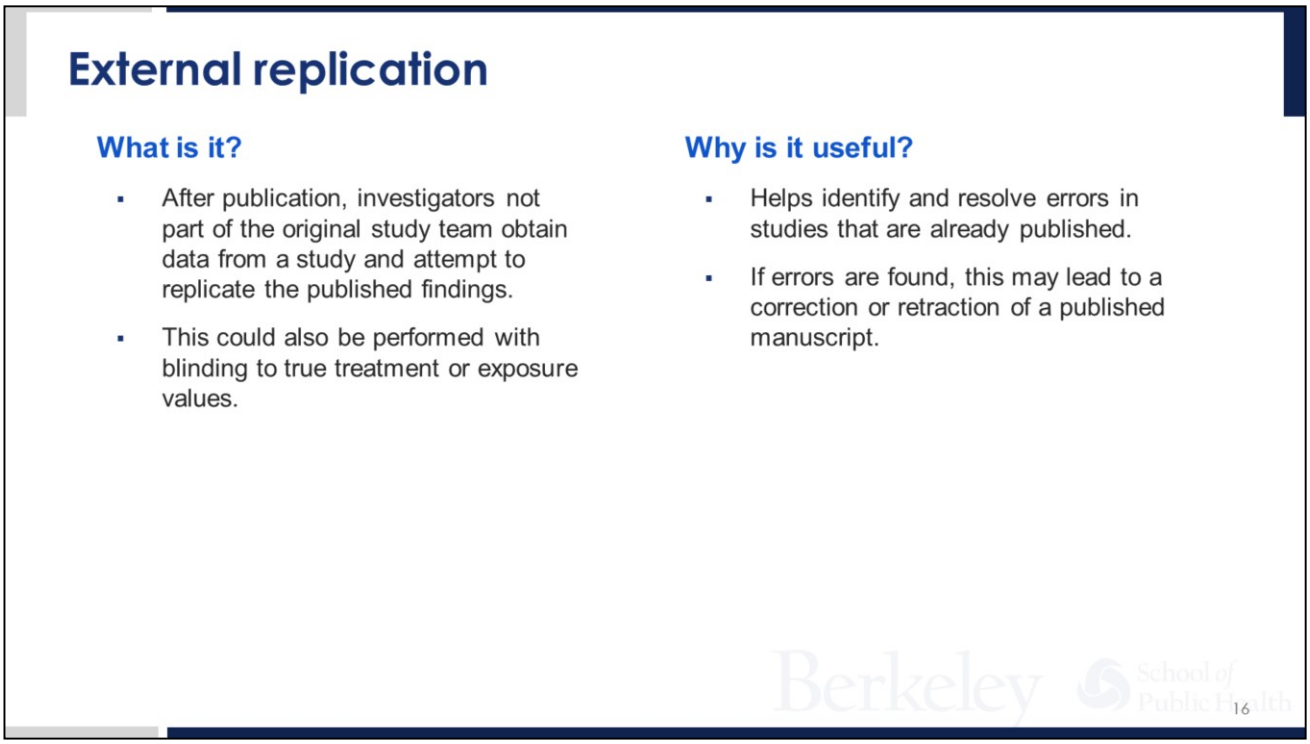
Internal replication with blinded analysis, means that prior to publication two or more analysts work independently to do the data analysis. So they do not share their code with each other. They work on their own, and they attempt to get the exact same answer in order to replicate their results. They check with each other as they complete their analyses, and if their results are not identical, they have to identify the discrepancy and resolve it.

And ideally this is done while blinding analysts to the true exposure or treatment levels. So they can use a fake treatment or exposure variable that scrambled the real values. And this means that while they're doing the analysis, they don't actually produce meaningful findings. And so this further reduces confirmation bias, because the results they're getting while doing a blinded analysis, are essentially fake. Once they are fully replicated, they can swap the fake treatment or exposure variable, with the true one, and replicate the analysis again with a real values.

This is a very useful process, because it can identify and resolve errors that would have been very difficult to catch, and might have crept into the final publication of results. And this means that published studies that were internally replicated, are much less likely to contain errors, and are much more reproducible. And again, blinding during analysis can greatly reduce confirmation bias.

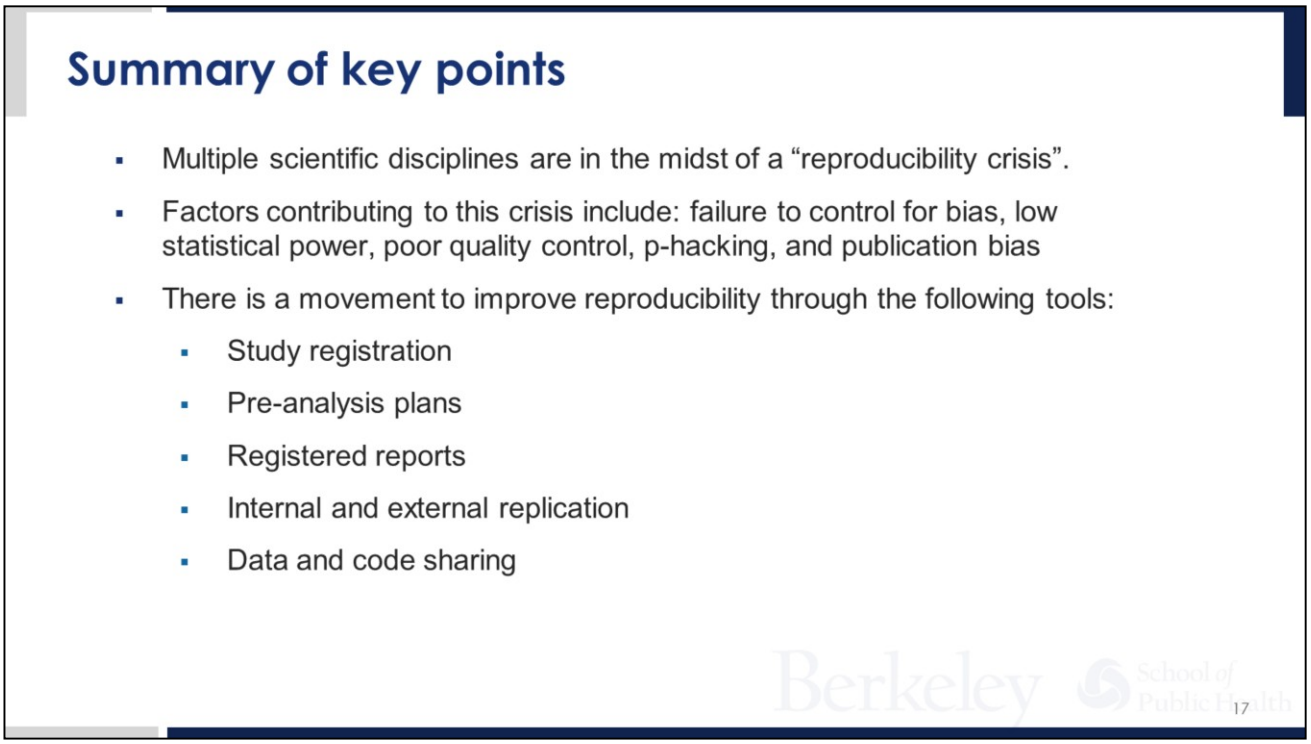


Data and code sharing can be done after a study is published, or at the time of publication. And what this essentially means, is that study investigators publish their code and data online. And they can do this on Open Science Framework, on Synapse, or on GitHub. This is useful, because it allows others to externally replicate work. And I'll come to that in the next slide. And it also can help people identify any errors in the published literature when they're working with the public code and data. It also allows for more rapid development of new hypotheses and new studies, because existing data is published quickly. And it can spark new ideas in outside investigators not participating in the original study.



And finally, external replication is done after publication. And investigators, who were not part of the study team originally, obtain data from the team and attempt to replicate the published findings. And so this could be performed with blinding to the true treatment or exposure values as well, as I mentioned for internal replication. And this is useful because it can help identify and resolve errors, in studies that are already published. And if those errors are found it may lead to corrections of published papers, or retractions of them.

And it's worth briefly mentioning that there have been some very high profile cases of well reputed, highly impactful studies, that have not been able to be externally replicated. And so this is something that's of growing practice, and is a huge motivator for trying to improve reproducibility upstream, before errors get into the published literature.



To summarize, in multiple scientific disciplines we're currently in the midst of a reproducibility crisis. And factors that are contributing to this include a failure to control for bias, low statistical power, poor quality control, P-hacking, and publication bias. There is a movement to improve reproducibility through the following tools, study registration, pre-analysis plans, registered reports, internal and external replication, and data and code sharing.

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets PPV = $(1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

Table 1. Research Findings and True Relationships

Research Finding	True Relationship		
	Yes	No	Total
Yes	$c(1 - \beta)R/(R + 1)$	$c\alpha/(R + 1)$	$c(R + \alpha - \beta R)/(R + 1)$
No	$c\beta R/(R + 1)$	$c(1 - \alpha)/(R + 1)$	$c(1 - \alpha + \beta R)/(R + 1)$
Total	$cR/(R + 1)$	$c/(R + 1)$	c

DOI: 10.1371/journal.pmed.0020124.t001

more likely true than false if $(1 - \beta)R > \alpha$. Since usually the vast majority of investigators depend on $\alpha = 0.05$, this means that a research finding is more likely true than false if $(1 - \beta)R > 0.05$.

What is less well appreciated is that bias and the extent of repeated independent testing by different teams of investigators around the globe may further distort this picture and may lead to even smaller probabilities of the research findings being indeed true. We will try to model these two factors in the context of similar 2 × 2 tables.

Bias

First, let us define bias as the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced. Let u be the proportion of probed analyses that would not have been "research findings," but nevertheless end up presented and reported as such, because of bias. Bias should not be confused with chance variability that causes some findings to be false by chance even though the study design, data, analysis, and presentation are perfect. Bias can entail manipulation in the analysis or reporting of findings. Selective or distorted reporting is a typical form of such bias. We may assume that u does not depend on whether a true relationship exists or not. This is not an unreasonable assumption, since typically it is impossible to know which relationships are indeed true. In the presence of bias (Table 2), one gets PPV = $([1 - \beta]R + u\beta R)/(R + \alpha - \beta R + u - u\alpha + u\beta R)$, and PPV decreases with increasing u, unless $1 - \beta \leq \alpha$, i.e., $1 - \beta \leq 0.05$ for most situations. Thus, with increasing bias, the chances that a research finding is true diminish considerably. This is shown for different levels of power and for different pre-study odds in Figure 1.

Conversely, true research findings may occasionally be annulled because of reverse bias. For example, with large measurement errors relationships are lost in noise [12], or investigators use data inefficiently or fail to notice statistically significant relationships, or there may be conflicts of interest that tend to "bury" significant findings [13]. There is no good large-scale empirical evidence on how frequently such reverse bias may occur across diverse research fields. However, it is probably fair to say that reverse bias is not as common. Moreover measurement errors and inefficient use of data are probably becoming less frequent problems, since measurement error has decreased with technological advances in the molecular era and investigators are becoming increasingly sophisticated about their data. Regardless, reverse bias may be modeled in the same way as bias above. Also reverse bias should not be confused with chance variability that may lead to missing a true relationship because of chance.

Testing by Several Independent Teams

Several independent teams may be addressing the same sets of research questions. As research efforts are globalized, it is practically the rule that several research teams, often dozens of them, may probe the same or similar questions. Unfortunately, in some areas, the prevailing mentality until now has been to focus on isolated discoveries by single teams and interpret research experiments in isolation. An increasing number of questions have at least one study claiming a research finding, and this receives unilateral attention. The probability that at least one study, among several done on the same question, claims a statistically significant research finding is easy to estimate. For n independent studies of equal power, the 2 × 2 table is shown in Table 3: PPV = $R(1 - \beta^n)/(R + 1 - [1 - \alpha]^n - R\beta^n)$ (not considering bias). With increasing number of independent studies, PPV tends to decrease, unless $1 - \beta < \alpha$, i.e., typically $1 - \beta < 0.05$. This is shown for different levels of power and for different pre-study odds in Figure 2. For n studies of different power, the term β^n is replaced by the product of the terms β_i for $i = 1$ to n, but inferences are similar.

Corollaries

A practical example is shown in Box 1. Based on the above considerations, one may deduce several interesting corollaries about the probability that a research finding is indeed true.

Corollary 1: The smaller the studies conducted in a scientific field, the less likely the research findings are to be true. Small sample size means smaller power and, for all functions above, the PPV for a true research finding decreases as power decreases towards $1 - \beta = 0.05$. Thus, other factors being equal, research findings are more likely true in scientific fields that undertake large studies, such as randomized controlled trials in cardiology (several thousand subjects randomized) [14] than in scientific fields with small studies, such as most research of molecular predictors (sample sizes 100-fold smaller) [15].

Corollary 2: The smaller the effect sizes in a scientific field, the less likely the research findings are to be true. Power is also related to the effect size. Thus research findings are more likely true in scientific fields with large effects, such as the impact of smoking on cancer or cardiovascular disease (relative risks 3–20), than in scientific fields where postulated effects are small, such as genetic risk factors for multigenetic diseases (relative risks 1.1–1.5) [7]. Modern epidemiology is increasingly obliged to target smaller

Table 2. Research Findings and True Relationships in the Presence of Bias

Research Finding	True Relationship		
	Yes	No	Total
Yes	$(c[1 - \beta]R + uc\beta R)/(R + 1)$	$c\alpha + uc(1 - \alpha)/(R + 1)$	$c(R + \alpha - \beta R + u - u\alpha + u\beta R)/(R + 1)$
No	$(1 - u)c\beta R/(R + 1)$	$(1 - u)c(1 - \alpha)/(R + 1)$	$c(1 - u)(1 - \alpha + \beta R)/(R + 1)$
Total	$cR/(R + 1)$	$c/(R + 1)$	c

DOI: 10.1371/journal.pmed.0020124.t002

effect sizes [16]. Consequently, the proportion of true research findings is expected to decrease. In the same line of thinking, if the true effect sizes are very small in a scientific field, this field is likely to be plagued by almost ubiquitous false positive claims. For example, if the majority of true genetic or nutritional determinants of complex diseases confer relative risks less than 1.05, genetic or nutritional epidemiology would be largely utopian endeavors.

Corollary 3: The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true. As shown above, the post-study probability that a finding is true (PPV) depends a lot on the pre-study odds *(R)*. Thus, research findings are more likely true in confirmatory designs, such as large phase III randomized controlled trials, or meta-analyses thereof, than in hypothesis-generating experiments. Fields considered highly informative and creative given the wealth of the assembled and tested information, such as microarrays and other high-throughput discovery-oriented research [4,8,17], should have extremely low PPV.

Corollary 4: The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true. Flexibility increases the potential for transforming what would be "negative" results into "positive" results, i.e., bias, *u*. For several research designs, e.g., randomized controlled trials [18–20] or meta-analyses [21,22], there have been efforts to standardize their conduct and reporting. Adherence to common standards is likely to increase the proportion of true findings. The same applies to outcomes. True findings may be more common when outcomes are unequivocal and universally agreed (e.g., death) rather than when multifarious outcomes are devised (e.g., scales for schizophrenia outcomes) [23]. Similarly, fields that use commonly agreed, stereotyped analytical methods (e.g., Kaplan-Meier plots and the log-rank test) [24] may yield a larger proportion of true findings than fields where analytical methods are still under experimentation (e.g., artificial intelligence methods) and only "best" results are reported. Regardless, even in the most stringent research designs, bias seems to be a major problem. For example, there is strong evidence that selective outcome reporting, with manipulation of the outcomes and analyses reported, is a common problem even for randomized trails [25]. Simply abolishing selective publication would not make this problem go away.

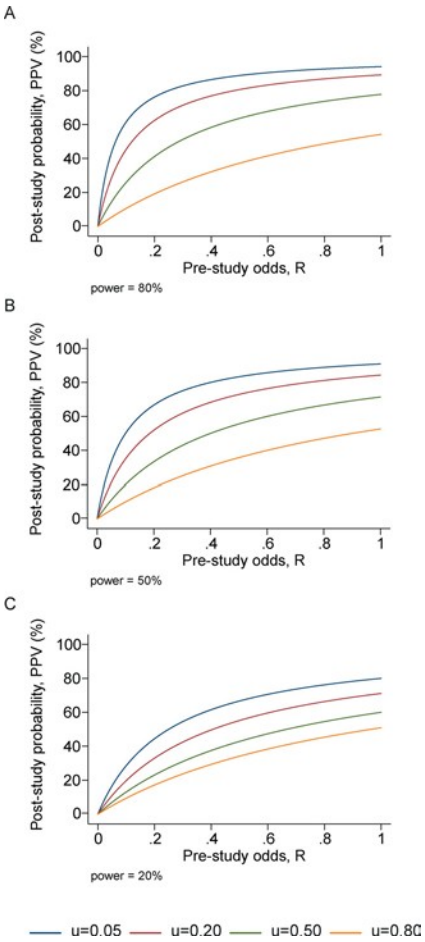
Corollary 5: The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true. Conflicts of interest and prejudice may increase bias, *u*. Conflicts of interest are very common in biomedical research [26], and typically they are inadequately and sparsely reported [26,27]. Prejudice may not necessarily have financial roots. Scientists in a given field may be prejudiced purely because of their belief in a scientific theory or commitment to their own findings. Many otherwise seemingly independent, university-based studies may be conducted for no other reason than to give physicians and researchers qualifications for promotion or tenure. Such nonfinancial conflicts may also lead to distorted reported results and interpretations. Prestigious investigators may suppress via the peer review process the appearance and dissemination of findings that refute their findings, thus condemning their field to perpetuate false dogma. Empirical evidence on expert opinion shows that it is extremely unreliable [28].

Corollary 6: The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Table 3. Research Findings and True Relationships in the Presence of Multiple Studies

Research Finding	True Relationship		
	Yes	No	Total
Yes	$cR(1-\beta^n)/(R+1)$	$c(1-[1-\alpha]^n)/(R+1)$	$c(R+1-[1-\alpha]^n-R\beta^n)/(R+1)$
No	$cR\beta^n/(R+1)$	$c(1-\alpha)^n/(R+1)$	$c([1-\alpha]^n+R\beta^n)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

DOI: 10.1371/journal.pmed.0020124.t003

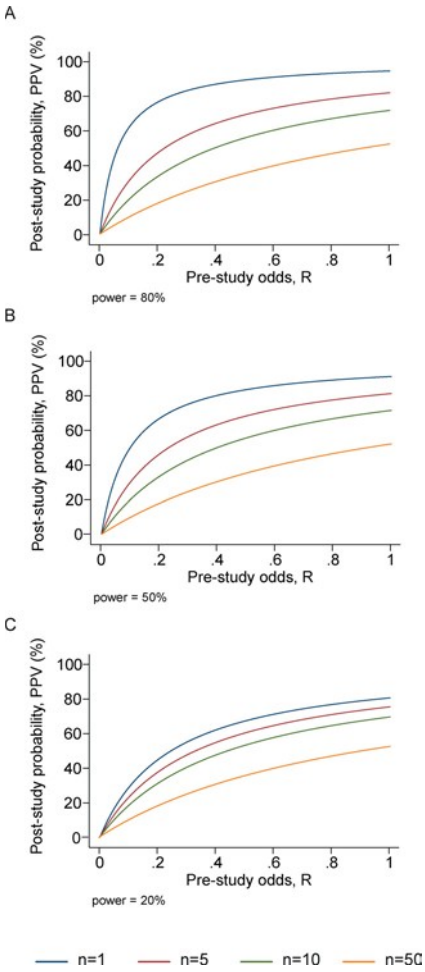


DOI: 10.1371/journal.pmed.0020124.g001

Figure 1. PPV (Probability That a Research Finding Is True) as a Function of the Pre-Study Odds for Various Levels of Bias, *u*

Panels correspond to power of 0.20, 0.50, and 0.80.

This seemingly paradoxical corollary follows because, as stated above, the PPV of isolated findings decreases when many teams of investigators are involved in the same field. This may explain why we occasionally see major excitement followed rapidly by severe disappointments in fields that draw wide attention. With many teams working on the same field and with massive experimental data being produced, timing is of the essence in beating competition. Thus, each team may prioritize on pursuing and disseminating its most impressive "positive" results. "Negative" results may become attractive for dissemination only if some other team has found a "positive" association on the same question. In that case, it may be attractive to refute a claim made in some prestigious journal. The term Proteus phenomenon has been coined to describe this phenomenon of rapidly



DOI: 10.1371/journal.pmed.0020124.g002

Figure 2. PPV (Probability That a Research Finding Is True) as a Function of the Pre-Study Odds for Various Numbers of Conducted Studies, *n*

Panels correspond to power of 0.20, 0.50, and 0.80.

Box 1. An Example: Science at Low Pre-Study Odds

Let us assume that a team of investigators performs a whole genome association study to test whether any of 100,000 gene polymorphisms are associated with susceptibility to schizophrenia. Based on what we know about the extent of heritability of the disease, it is reasonable to expect that probably around ten gene polymorphisms among those tested would be truly associated with schizophrenia, with relatively similar odds ratios around 1.3 for the ten or so polymorphisms and with a fairly similar power to identify any of them. Then $R = 10/100{,}000 = 10^{-4}$, and the pre-study probability for any polymorphism to be associated with schizophrenia is also $R/(R + 1) = 10^{-4}$. Let us also suppose thatthe study has 60% power to find an association with an odds ratio of 1.3 at $\alpha = 0.05$. Then it can be estimated that if a statistically significant association is found with the *p*-value barely crossing the 0.05 threshold, the post-study probability that this is true increases about 12-fold compared with the pre-study probability, but it is still only 12×10^{-4}.

Now let us suppose that the investigators manipulate their design, analyses, and reporting so as to make more relationships cross the $p = 0.05$ threshold even though this would not have been crossed with a perfectly adhered to design and analysis and with perfect comprehensive reporting of the results, strictly according to the original study plan. Such manipulation could be done, for example, with serendipitous inclusion or exclusion of certain patients or controls, post hoc subgroup analyses, investigation of genetic contrasts that were not originally specified, changes in the disease or control definitions, and various combinations of selective or distorted reporting of the results. Commercially available "data mining" packages actually are proud of their ability to yield statistically significant results through data dredging. In the presence of bias with $u = 0.10$, the post-study probability that a research finding is true is only 4.4×10^{-4}. Furthermore, even in the absence of any bias, when ten independent research teams perform similar experiments around the world, if one of them finds a formally statistically significant association, the probability that the research finding is true is only 1.5×10^{-4}, hardly any higher than the probability we had before any of this extensive research was undertaken!

alternating extreme research claims and extremely opposite refutations [29]. Empirical evidence suggests that this sequence of extreme opposites is very common in molecular genetics [29].

These corollaries consider each factor separately, but these factors often influence each other. For example, investigators working in fields where true effect sizes are perceived to be small may be more likely to perform large studies than investigators working in fields where true effect sizes are perceived to be large. Or prejudice may prevail in a hot scientific field, further undermining the predictive value of its research findings. Highly prejudiced stakeholders may even create a barrier that aborts efforts at obtaining and disseminating opposing results. Conversely, the fact that a field is hot or has strong invested interests may sometimes promote larger studies and improved standards of research, enhancing the predictive value of its research findings. Or massive discovery-oriented testing may result in such a large yield of significant relationships that investigators have enough to report and search further and thus refrain from data dredging and manipulation.

Most Research Findings Are False for Most Research Designs and for Most Fields

In the described framework, a PPV exceeding 50% is quite difficult to get. Table 4 provides the results of simulations using the formulas developed for the influence of power, ratio of true to non-true relationships, and bias, for various types of situations that may be characteristic of specific study designs and settings. A finding from a well-conducted, adequately powered randomized controlled trial starting with a 50% pre-study chance that the intervention is effective is eventually true about 85% of the time. A fairly similar performance is expected of a confirmatory meta-analysis of good-quality randomized trials: potential bias probably increases, but power and pre-test chances are higher compared to a single randomized trial. Conversely, a meta-analytic finding from inconclusive studies where pooling is used to "correct" the low power of single studies, is probably false if $R \leq 1{:}3$. Research findings from underpowered, early-phase clinical trials would be true about one in four times, or even less frequently if bias is present. Epidemiological studies of an exploratory nature perform even worse, especially when underpowered, but even well-powered epidemiological studies may have only a one in five chance being true, if $R = 1{:}10$. Finally, in discovery-oriented research with massive testing, where tested relationships exceed true ones 1,000-fold (e.g., 30,000 genes tested, of which 30 may be the true culprits) [30,31], PPV for each claimed relationship is extremely low, even with considerable

standardization of laboratory and statistical methods, outcomes, and reporting thereof to minimize bias.

Claimed Research Findings May Often Be Simply Accurate Measures of the Prevailing Bias

As shown, the majority of modern biomedical research is operating in areas with very low pre- and post-study probability for true findings. Let us suppose that in a research field there are no true findings at all to be discovered. History of science teaches us that scientific endeavor has often in the past wasted effort in fields with absolutely no yield of true scientific information, at least based on our current understanding. In such a "null field," one would ideally expect all observed effect sizes to vary by chance around the null in the absence of bias. The extent that observed findings deviate from what is expected by chance alone would be simply a pure measure of the prevailing bias.

For example, let us suppose that no nutrients or dietary patterns are actually important determinants for the risk of developing a specific tumor. Let us also suppose that the scientific literature has examined 60 nutrients and claims all of them to be related to the risk of developing this tumor with relative risks in the range of 1.2 to 1.4 for the comparison of the upper to lower intake tertiles. Then the claimed effect sizes are simply measuring nothing else but the net bias that has been involved in the generation of this scientific literature. Claimed effect sizes are in fact the most accurate estimates of the net bias. It even follows that between "null fields," the fields that claim stronger effects (often with accompanying claims of medical or public health importance) are simply those that have sustained the worst biases.

For fields with very low PPV, the few true relationships would not distort this overall picture much. Even if a few relationships are true, the shape of the distribution of the observed effects would still yield a clear measure of the biases involved in the field. This concept totally reverses the way we view scientific results. Traditionally, investigators have viewed large and highly significant effects with excitement, as signs of important discoveries. Too large and too highly significant effects may actually be more likely to be signs of large bias in most fields of modern research. They should lead investigators to careful critical thinking about what might have gone wrong with their data, analyses, and results.

Of course, investigators working in any field are likely to resist accepting that the whole field in which they have spent their careers is a "null field." However, other lines of evidence, or advances in technology and experimentation, may lead eventually to the dismantling of a scientific field. Obtaining measures of the net bias in one field may also be useful for obtaining insight into what might be the range of bias operating in other fields where similar analytical methods, technologies, and conflicts may be operating.

How Can We Improve the Situation?

Is it unavoidable that most research findings are false, or can we improve the situation? A major problem is that it is impossible to know with 100% certainty what the truth is in any research question. In this regard, the pure "gold" standard is unattainable. However, there are several approaches to improve the post-study probability.

Better powered evidence, e.g., large studies or low-bias meta-analyses, may help, as it comes closer to the unknown "gold" standard. However, large studies may still have biases and these should be acknowledged and avoided. Moreover, large-scale evidence is impossible to obtain for all of the millions and trillions of research questions posed in current research. Large-scale evidence should be targeted for research questions where the pre-study probability is already considerably high, so that a significant research finding will lead to a post-test probability that would be considered quite definitive. Large-scale evidence is also particularly indicated when it can test major concepts rather than narrow, specific questions. A negative finding can then refute not only a specific proposed claim, but a whole field or considerable portion thereof. Selecting the performance of large-scale studies based on narrow-minded criteria, such as the marketing promotion of a specific drug, is largely wasted research. Moreover, one should be cautious that extremely large studies may be more likely to find a formally statistical significant difference for a trivial effect that is not really meaningfully different from the null [32–34].

Second, most research questions are addressed by many teams, and it is misleading to emphasize the statistically significant findings of any single team. What matters is the

Table 4. PPV of Research Findings for Various Combinations of Power (1 − β), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

1 − β	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.
RCT, randomized controlled trial.
DOI: 10.1371/journal.pmed.0020124.t004

totality of the evidence. Diminishing bias through enhanced research standards and curtailing of prejudices may also help. However, this may require a change in scientific mentality that might be difficult to achieve. In some research designs, efforts may also be more successful with upfront registration of studies, e.g., randomized trials [35]. Registration would pose a challenge for hypothesis-generating research. Some kind of registration or networking of data collections or investigators within fields may be more feasible than registration of each and every hypothesis-generating experiment. Regardless, even if we do not see a great deal of progress with registration of studies in other fields, the principles of developing and adhering to a protocol could be more widely borrowed from randomized controlled trials.

Finally, instead of chasing statistical significance, we should improve our understanding of the range of R values—the pre-study odds—where research efforts operate [10]. Before running an experiment, investigators should consider what they believe the chances are that they are testing a true rather than a non-true relationship. Speculated high R values may sometimes then be ascertained. As described above, whenever ethically acceptable, large studies with minimal bias should be performed on research findings that are considered relatively established, to see how often they are indeed confirmed. I suspect several established "classics" will fail the test [36].

Nevertheless, most new discoveries will continue to stem from hypothesis-generating research with low or very low pre-study odds. We should then acknowledge that statistical significance testing in the report of a single study gives only a partial picture, without knowing how much testing has been done outside the report and in the relevant field at large. Despite a large statistical literature for multiple testing corrections [37], usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding. Even if determining this were feasible, this would not inform us about the pre-study odds. Thus, it is unavoidable that one should make approximate assumptions on how many relationships are expected to be true among those probed across the relevant research fields and research designs. The wider field may yield some guidance for estimating this probability for the isolated research project. Experiences from biases detected in other neighboring fields would also be useful to draw upon. Even though these assumptions would be considerably subjective, they would still be very useful in interpreting research claims and putting them in context. ◆

References

1. Ioannidis JP, Haidich AB, Lau J (2001) Any casualties in the clash of randomised and observational evidence? BMJ 322: 879–880.
2. Lawlor DA, Davey Smith G, Kundu D, Bruckdorfer KR, Ebrahim S (2004) Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? Lancet 363: 1724–1727.
3. Vandenbroucke JP (2004) When are observational studies as credible as randomised trials? Lancet 363: 1728–1731.
4. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: A multiple random validation strategy. Lancet 365: 488–492.
5. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29: 306–309.
6. Colhoun HM, McKeigue PM, Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. Lancet 361: 865–872.
7. Ioannidis JP (2003) Genetic associations: False or true? Trends Mol Med 9: 135–138.
8. Ioannidis JPA (2005) Microarrays and molecular research: Noise discovery? Lancet 365: 454–455.
9. Sterne JA, Davey Smith G (2001) Sifting the evidence—What's wrong with significance tests. BMJ 322: 226–231.
10. Wacholder S, Chanock S, Garcia-Closas M, El ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. J Natl Cancer Inst 96: 434–442.
11. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847–856.
12. Kelsey JL, Whittemore AS, Evans AS, Thompson WD (1996) Methods in observational epidemiology, 2nd ed. New York: Oxford U Press. 432 p.
13. Topol EJ (2004) Failing the public health—Rofecoxib, Merck, and the FDA. N Engl J Med 351: 1707–1709.
14. Yusuf S, Collins R, Peto R (1984) Why do we need some large, simple randomized trials? Stat Med 3: 409–422.
15. Altman DG, Royston P (2000) What do we mean by validating a prognostic model? Stat Med 19: 453–473.
16. Taubes G (1995) Epidemiology faces its limits. Science 269: 164–169.
17. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286: 531–537.
18. Moher D, Schulz KF, Altman DG (2001) The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357: 1191–1194.
19. Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, et al. (2004) Better reporting of harms in randomized trials: An extension of the CONSORT statement. Ann Intern Med 141: 781–788.
20. International Conference on Harmonisation E9 Expert Working Group (1999) ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. Stat Med 18: 1905–1942.
21. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, et al. (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 354: 1896–1900.
22. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, et al. (2000) Meta-analysis of observational studies in epidemiology: A proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. JAMA 283: 2008–2012.
23. Marshall M, Lockwood A, Bradley C, Adams C, Joy C, et al. (2000) Unpublished rating scales: A major source of bias in randomised controlled trials of treatments for schizophrenia. Br J Psychiatry 176: 249–252.
24. Altman DG, Goodman SN (1994) Transfer of technology from statistical journals to the biomedical literature. Past trends and future predictions. JAMA 272: 129–132.
25. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG (2004) Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. JAMA 291: 2457–2465.
26. Krimsky S, Rothenberg LS, Stott P, Kyle G (1998) Scientific journals and their authors' financial interests: A pilot study. Psychother Psychosom 67: 194–201.
27. Papanikolaou GN, Baltogianni MS, Contopoulos-Ioannidis DG, Haidich AB, Giannakakis IA, et al. (2001) Reporting of conflicts of interest in guidelines of preventive and therapeutic interventions. BMC Med Res Methodol 1: 3.
28. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC (1992) A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 268: 240–248.
29. Ioannidis JP, Trikalinos TA (2005) Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. J Clin Epidemiol 58: 543–549.
30. Ntzani EE, Ioannidis JP (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment. Lancet 362: 1439–1444.
31. Ransohoff DF (2004) Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 4: 309–314.
32. Lindley DV (1957) A statistical paradox. Biometrika 44: 187–192.
33. Bartlett MS (1957) A comment on D.V. Lindley's statistical paradox. Biometrika 44: 533–534.
34. Senn SJ (2001) Two cheers for P-values. J Epidemiol Biostat 6: 193–204.
35. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, et al. (2004) Clinical trial registration: A statement from the International Committee of Medical Journal Editors. N Engl J Med 351: 1250–1251.
36. Ioannidis JPA (2005) Contradicted and initially stronger effects in highly cited clinical research. JAMA 294: 218–228.
37. Hsueh HM, Chen JJ, Kodell RL (2003) Comparison of methods for estimating the number of true null hypotheses in multiplicity testing. J Biopharm Stat 13: 675–689.