6. STATISTICAL MODELS FOR META-ANALYSIS

6.1. Overview

Whenever we perform a meta-analysis, we do so within the framework of a statistical model. The model reflects the way the studies were sampled, and it determines how we can generalize from the results. Typically, we would choose among the following three models.

6.1.1. Random effects

The random-effects model applies when we identify a universe of studies, sample studies from that universe, and then use the results of our analysis to generalize to that universe. The word *random* reflects the assumption that the studies in our analysis are a random sample of all possible studies in this universe. The word *effects* is in the plural because the effect (the impact of the treatment) is assumed to vary from study to study.

6.1.2. Fixed effect (singular)

The fixed-effect (singular) model applies when all studies are based on the same population and are identical to each other in all material ways. The results of our analysis apply to this specific population, and cannot be generalized beyond that. The word *effect* is in the singular because all studies share a common effect size, and the word *fixed* is assumed to mean common. This model is sometimes called the common-effect model.

6.1.3. Fixed effects (plural)

The fixed-effects (plural) model applies when we identify a specific set of studies that we want to include in our analysis. Like the random-effects model, we assume that the effect size varies from one study to the next. Unlike the random-effects model, these studies are not seen as having been sampled from a larger universe. Rather, these are the only studies we care about. We will report statistics for the studies in the analysis, but will not generalize beyond them. The word *effects* is in the plural because the true effect size varies from study to study. The word *fixed* reflects the fact that these studies have not been sampled from a larger universe, but rather have been *identified* as being the only studies of interest.

While most researchers are familiar with the random-effects model and the fixed-effect (singular) model, relatively few are aware of the fixed-effects (plural) model. This model is discussed by (Hedges & Vevea, 1998) and also by Rice, Higgins, and Lumley (2017).

These issues are summarized in Table 1, which provides a framework for many of the issues discussed below.

Table 1 | Sampling frame for statistical models

		Random effects	Fixed effect	Fixed effects
Sampling frame				
	Studies *sampled* from different populations	•		
	Studies sampled from *one* population		•	
	Studies *selected* from different populations			•
Inference to				
	Universe from which studies were sampled	•		
	Specific population in the analysis		•	
	Specific studies included in the analysis			•
Advantages				
	Can generalize to a larger universe	•		
	Can assess heterogeneity in effect size	•		
Requirements				
	Enumerate the universe of relevant studies	•		
	Studies are representative of the universe	•		
	Reliable estimate of τ^2	•		

6.1.4. Three textbook cases

The following three examples serve as textbook cases for each of the three models. In each example one model clearly applies, the analysis meets all the relevant assumptions of the model, and the model will work as intended. The computations are given in Appendix I.

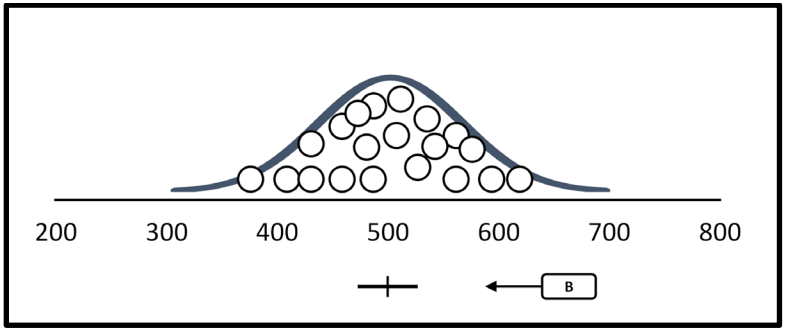

Figure 2 | Random effects| Confidence interval 60 points wide

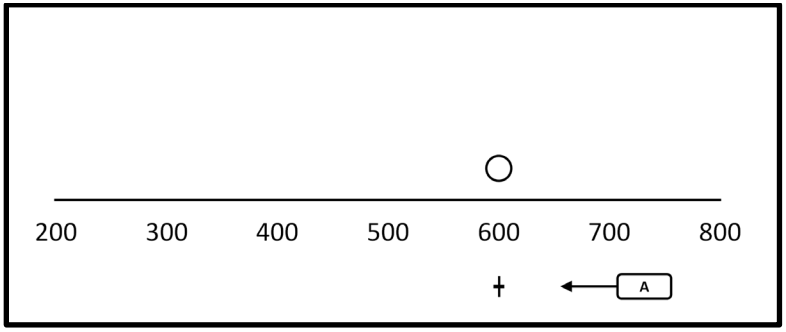

Figure 3 | Fixed effect (singular) | Confidence interval 10 points wide

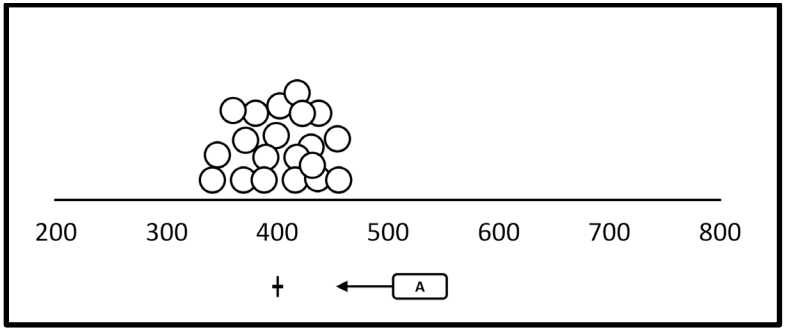

Figure 4 | Fixed effects (plural) | Confidence interval 10 points wide

6.1.5. Random effects

Suppose that we want to estimate the mean score for all high schools in a large city. We draw a random sample of 20 schools from this universe of schools, and then draw a random sample of 50 students within each of these schools (Figure 2). The 20 circles in the plot represent the true scores for the 20 schools that were included in our random sample. The key factor that makes this a random-effects analysis is the normal curve that has been superimposed on the plot. This curve reflects the fact that we have defined a universe of populations from which we will be sampling and to which we will be making an inference. We report that the mean for this universe is 500, with a confidence interval of 470 to 530. Following the convention introduced in section 5, this is labeled [B] since it is based on a random-effects analysis and applies to the universe of all comparable populations.

6.1.6. Fixed effect (singular)

Suppose that we want to estimate the mean score for a specific school, which has a selective admissions policy. We draw 20 random samples of 50 students each from this school (Figure 3). The inference is to this school only. We report that the mean for this school is 600, with a confidence interval of 595 to 605. Following the convention introduced in section 5, this is labeled [A] since it is based on a fixed-effect analysis and applies only to the one school included in the analysis. It should be clear that the mean in this school tells us nothing about the mean for all schools in the system.

The reason that there appears to be only one circle on this plot is that we are plotting *true* scores rather than *observed* scores. The *true* score is the actual mean for all students in the school. While the *observed* score will vary from study to study, the *true* score is the same for all studies and so all twenty circles fall at precisely the same point.

6.1.7. Fixed effects (plural)

Suppose that we want to estimate the mean score for 20 schools that are under the control of one specific school board. We identify these 20 schools by name, and then draw a random sample of 50 students within each of these schools. In Figure 4, the circles reflect each of the twenty schools that are included in our analysis. We report that the mean for this specific set of schools is 400, with a confidence interval of 395 to 405. Following the convention introduced in section 5, this is labeled [A] since it is based on a

fixed-effects analysis and applies only to the twenty schools included in the analysis. The word *fixed* reflects the fact that these schools have been *fixed* (or designated) as the schools of interest rather than sampled from a larger universe. The word *effects* is in the plural since each sample is estimating the effect (the mean) in a different school. Here, we limit ourselves to a specific set of schools, and the results apply only to this set. The key difference between this model and the random-effects model is the absence of a normal curve here. Since these schools are not representative of all schools in the city, the fact that the mean in these 20 schools is 400 tells us nothing about the mean for all schools in the city.

6.2. Each model is appropriate for a specific inference

The choice of a statistical model determines how much weight we assign to each study in the analysis, and this in turn affects the values computed for the summary effect size, the confidence interval, and other statistics. It is useful to understand how the values computed under each model are appropriate for that model's goals. I provide a conceptual overview here. For computational details, see Appendix I.

6.2.1. How the model affects the confidence-interval width

One goal of the meta-analysis is to compute a summary effect size, along with a confidence interval that tells us how precisely we have estimated this effect size. The confidence interval for the summary effect will tend to be relatively wide under the random-effects model, and relatively narrow under the fixed-effect and fixed-effects models. In all three examples we had twenty studies with fifty students in each, yet the confidence-interval width varied. The confidence interval was 10 points wide for the fixed-effect analysis (Figure 3) and for the fixed-effects analysis (Figure 4). By contrast, it was 60 points wide for the random-effects analysis (Figure 2).

The confidence interval is relatively narrow when we perform a fixed-effect or fixed-effects analysis, because we are making an inference only to the studies *in the analysis*. If we have a sufficient number of people in these studies, we will know the mean *for these studies* with relatively good precision. By contrast, the confidence interval is wider under the random-effects analysis because we are estimating the mean for the studies in the analysis and then using that mean to generalize to the universe of comparable studies. The leap from the studies in the analysis to the universe of comparable studies entails additional sampling error, which results in the wider interval.

6.2.2. How the model affects the estimate of the mean effect size

Under the fixed-effect model the weights assigned to individual studies may vary substantially from each other, such that large studies may dominate the analysis and small studies may be essentially ignored. By contrast, under the random-effects model the weights tend to be more moderate, such that large studies are less likely to dominate the analysis, and small studies are more likely to play a non-trivial role. Again, this follows from the logic of the intended inference.

Consider the fixed-effect analysis where all twenty studies are based on random samples from the same school and our goal is to estimate the mean in that school. If one study has very large sample size, we would *want* that study to dominate the analysis since all studies are estimating the *same* parameter (the school mean) but this study is doing so based on more information than the others. Conversely, if one study had a very small sample size, we would want to essentially ignore that study, since other studies provide more precise information about the *same* parameter.

By contrast, consider the random-effects analysis, where we are estimating the mean for all schools in a city. If one school happened to have a very large sample size, we would *not* want that school to dominate the analysis. While we might know the mean for that school precisely, there is no reason to think that the mean for this school is representative of the mean for all schools. Conversely, if one school had a very small sample size, we would not want to ignore that school. The information provided by this school may be imprecise, but it is the *only* information we have about this school, and so we need to use it when estimating the overall mean.

The fixed-effects (plural) model uses the same weights as the fixed-effect (singular) model, and so would allow a large study to dominate the analysis and small studies to be essentially ignored. In effect, this means that we assign the same weight to each person rather than each study. This would make sense if we wanted to estimate the mean for all students, rather than the mean for all schools.

6.2.3. How the model affects our ability to address heterogeneity

To this point I have been discussing how we estimate the mean effect size, and the precision of that estimate. An equally important issue is the dispersion of true effects, which we call heterogeneity. The random-effects model allows us to explore the dispersion of effects, while the fixed-effect (singular) model does not.

Consider a case where all samples are based on the same school, and we use the fixed-effect model. We cannot explore heterogeneity because, by definition, there *is no* heterogeneity in true effects. Suppose that the true mean in this school is 600. If we draw five samples from this school, the *true* mean for all five samples is 600 by definition. While the *observed* effect size will vary from one sample to the next, heterogeneity refers to the *true* effect size, and that parameter is a constant.

By contrast, consider the case where we want to make an inference to all schools in the city. Imagine that we repeat the same analysis in three cities.

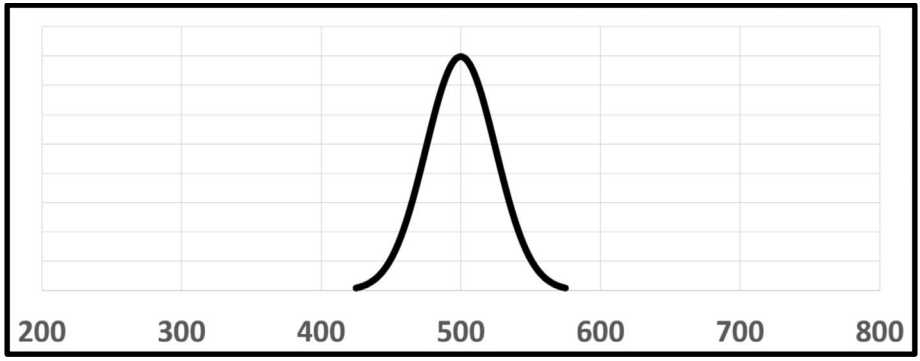

Figure 5 | Mean is 500 | Effects vary over 100 points

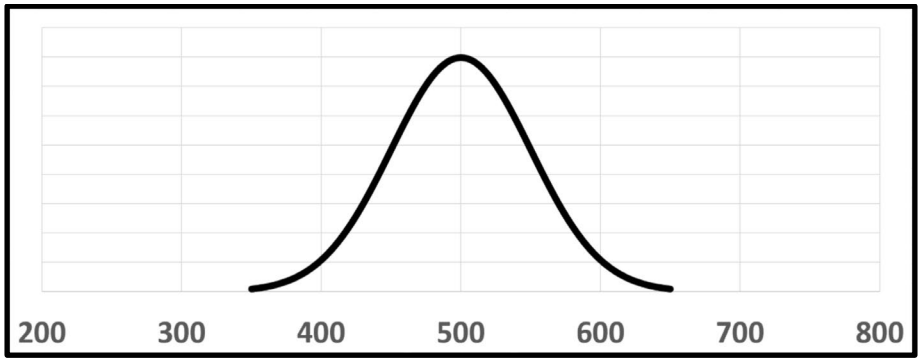

Figure 6 | Mean is 500 | Effects vary over 200 points

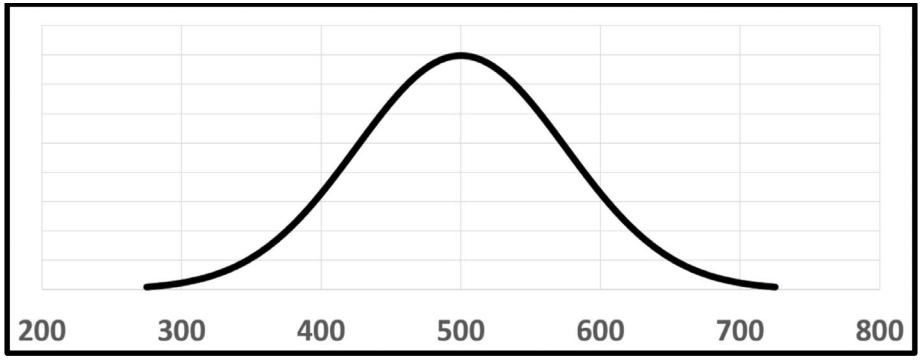

Figure 7| Mean is 500 | Effects vary over 300 points

In each city the mean score is 500, but the cities differ in the following way. In one city, 95% of all school means all fall in the range of 450 to 550 (Figure 5). In the second city, 95% of all school means fall in the range of 400 to 600 (Figure 6). In the third city, 95% of all school means fall in the range of 350 to 650 (Figure 7). While the mean is the same in all three cities, the three are obviously very different from each other. If we want to describe the distribution of school means it is imperative that we report on the dispersion as well as the mean.

The distinction between the fixed-effect model vs. the random-effects model in this context is clear. The latter allows us to address heterogeneity while the former does not. The issue is somewhat less clear when we apply the fixed-effects (plural) model. In this case the effect size does vary across studies, but since the studies in the analysis are not seen as representative of any universe, there is no conceptual basis for discussing the heterogeneity in any larger universe. For that reason, we generally avoid discussing the extent of dispersion when working with this model.

6.2.4. How the model affects the meaning of the null hypothesis

Typically, when we perform a meta-analysis, we pose the null hypothesis that the true effect size is zero, and then test that null hypothesis. The meaning of the null hypothesis depends on the statistical model. Consider a meta-analysis to assess the impact of an intervention, where an effect size of zero would mean that the intervention had no impact.

Under the fixed-effect model we are working with one population, and the null hypothesis is that the true effect size in that population is zero. If we reject the null hypothesis, we conclude that the intervention has an impact in this population. We do not need to be concerned that the intervention is effective in some populations and not others, since there is only one population being discussed.

By contrast, under the random-effects model the null hypothesis is that the *mean* effect size in the universe of comparable populations is zero. If we reject the null hypothesis, we conclude that the intervention is effective *on average*, but we must still ask about variation in the effect. The intervention could be effective in some populations but ineffective (or even harmful) in others.

Under the fixed-effects (plural) model we are working with multiple populations, and the null hypothesis is that the mean effect size in this specific set of populations is zero. Since the mean depends on the specific mix of

populations included in the analysis, this null hypothesis is only relevant if we care about this specific set of studies, to the exclusion of all others.

Summary

Each statistical model is appropriate for a specific type of inference.

The random-effects model applies when the studies in the analysis will be used to make an inference to a larger set of comparable populations. We assume that the true effect size varies from study to study. Our goal is to estimate the mean effect size in this universe of comparable populations, and also the dispersion of effects about that mean.

The fixed-effect (singular) model applies when all studies in the analysis are based on the same population and identical to each other in all material respects. Our goal is to estimate the common effect size in this population.

The fixed-effects (plural) model applies when we want to make an inference only to the studies actually included in the analysis, and not generalize beyond them to any larger set of comparable studies.

7. MISTAKES IN CHOOSING A STATISTICAL MODEL

7.1. Overview

Earlier, I introduced three statistical models for meta-analysis, as follows.

- The random-effects model applies when the studies in the analysis are representative of a larger universe of studies. Our goal is to make an inference to that larger universe.
- The fixed-effect (singular) model applies when the studies in the analysis are all based on the same population. Our goal is to make an inference to this one population.
- The fixed-effects (plural) model applies when the studies in the analysis are based on multiple populations. Our goal is to make an inference to this specific set of populations, and not to generalize beyond them.

When the analysis is based on studies pulled from the literature, the random-effects model is almost invariably the model that we should apply. This model assumes that the studies in the analysis are representative of a universe of comparable studies, and that the results of the analysis will be generalized to that universe. Critically, this model allows us to discuss not only the mean effect size, but also the dispersion in effect size across studies. These are all key goals of the analysis.

Researchers sometimes elect to use either a fixed-effect or fixed-effects model. In sections 7.2 and 7.3, I explain why the use of these models is generally inappropriate when studies are pulled from the literature.

While the random-effects model is generally the most appropriate model for analyses where studies are pulled from the literature, the model has limitations when used for this purpose. Specifically, we will fail to meet some of the model's assumptions, and we need to understand how this affects our ability to generalize from the results. This is discussed in section 7.4.

7.2. Choosing between fixed effect (singular) and random effects

7.2.1. Mistake

When a published meta-analysis includes a discussion about the choice of a model, the choice being discussed is almost always the fixed-effect (singular) model vs. the random-effects model. The decision to use one model or the other then focuses on the question of whether there is evidence that the true effect size varies across studies. When a meta-analysis is based on studies pulled from the literature, especially when the studies assess the impact of an intervention, we can generally *assume* that the effect size varies, and the fixed-effect (singular) model is not a viable option. Therefore, this approach (using a test to look for evidence of heterogeneity) is generally a mistake.

7.2.2. Details

It is entirely legitimate to choose between the fixed-effect (singular) model and the random-effects model based on whether the effect size is the same for all studies. However, this decision must be based on our understanding of the sampling frame, and not on a statistical test.

7.2.3. Cases where the fixed-effect (singular) model applies

The textbook case of the fixed-effect (singular) model was the case where we wanted to estimate the mean score for all students in a school (section 6.1.6). We drew twenty random samples from that school and then performed a meta-analysis on those studies. In this case all studies are estimating the same parameter. If the mean for all students in the school is 600, then the true effect size for the first study (the effect size that we would see if there was no sampling error) is 600, the true effect size for the second study is 600, and so on for all studies. *By definition*, all studies are estimating the same value.

Another case where the fixed-effect model (singular) would apply is the case where a drug company draws ten random samples from one population and performs the identical clinical trial with each sample. Again, *by definition*, all studies are estimating the same parameter.

By contrast, when a meta-analysis is based on studies that are pulled from the literature, the situation is very different, as explained below.

7.2.4. Cases where the random-effects model applies

In papers where researchers choose between the fixed-effect (singular) and the random-effects model, the researchers make the decision based on a test of significance. They elect to use the fixed-effect (singular) model as the default, and then perform a significance test for heterogeneity. If the test result *is not* statistically significant, they assume that all studies are estimating the same parameter and stay with the fixed-effect model. If the test result *is* statistically significant, they assume that the true effect size varies, and switch to the random-effects model. This approach is misguided because the conclusion that all studies are (or are not) estimating the same parameter must be based on our understanding of how the studies were sampled, rather than a test of statistical significance.

When studies are pulled from the literature, and especially when these are studies that assess the impact of an intervention, each study is based on a unique population, and the impact of the intervention will vary from one population to the next. A drug might be more effective (or less effective) in populations that are older, or where the patients are generally healthier, or exercise more, or have better medical care, or live in a colder climate, and so on. An intervention to improve students' scores might be more effective (or less effective) in populations where students are more motivated, or have better resources, or have better reading skills, or have a shorter school year, and so on.

Additionally, the details of the protocol will typically vary from study to study. The dose of a drug, the duration of the intervention, the attention to detail, the training of the staff, may vary. The group that serves as a comparator may vary. The instrument that is used to measure outcome may vary. The study design might vary. The impact of these factors on the effect size might be substantial or it might be trivial, but in general it will not be zero. Once it is not zero, the fixed-effect model is no longer applicable (see for example (Borenstein, Hedges, Higgins, & Rothstein, 2010; J. P. Higgins, 2008; J. P. Higgins, Thompson, & Spiegelhalter, 2009; Lorenc et al., 2016)).

7.2.5. In context

In the examples of the school and the case of a drug company we should choose the fixed-effect model because logic tells us that the true effect size is the same in all studies. Conversely, when studies are pulled from the literature, we should choose the random-effects model because common sense tells us that the true effect size varies across studies.

The assertion that when studies are pulled from the literature, the true effect size will vary across studies, is not absolute. If the intervention really had no relation at all to the outcome, the true effect size would be the same (zero) in all studies. However, the assertion is correct in the vast majority of cases.

7.2.6. What difference does it make?

If we can use either our understanding of how the studies were sampled (on the one hand) or a significance test (on the other), why is it imperative that we use the former rather than the latter? There are two key reasons.

The first reason is that statistical tests are intended for instances when the true state of affairs is unknown. If we know the true state of affairs, we should use that knowledge.

The second reason is that the test for heterogeneity will sometimes lead to the wrong model. When we should be using the random-effects model the test may not be statistically significant, which would lead us to use the fixed-effect model. Conversely, when we should be using the fixed-effect model, the test will sometimes yield a statistically significant result, leading us to use the random-effects model.

7.2.7. Examples

Following are some examples where researchers (incorrectly) used a test for heterogeneity to choose a statistical model, and others where the researchers (correctly) relied on their understanding of the intended inference to choose a statistical model.

7.2.8. Example | PTSD in parents of children with chronic illness

Post-traumatic stress disorder (PTSD) is a mental-health disorder that develops in some people following a traumatic event. While we generally associate this with soldiers and combat, it can also develop in parents whose children suffer from a chronic illness. Cabizuca, Marques-Portella, Mendlowicz, Coutinho, and Figueira (2009) looked at the incidence of PTSD in mothers of children with chronic illnesses (Figure 8). The reviewers elected to use the fixed-effect model by default, and switch to the random-effects model only if a test provided evidence that the prevalence of PTSD varied across studies.

The studies were conducted in different countries. They included parents whose children suffered from an array of different illnesses. The time from the onset of illness to the examination varied across studies. The methods employed to diagnose PTSD varied by study. Therefore, common sense tells us that the prevalence of PTSD will vary across studies. Indeed, common sense also suggests that a key part of the analysis should be to determine *how much* the prevalence varies. On this basis, we should be using the random-effects model.

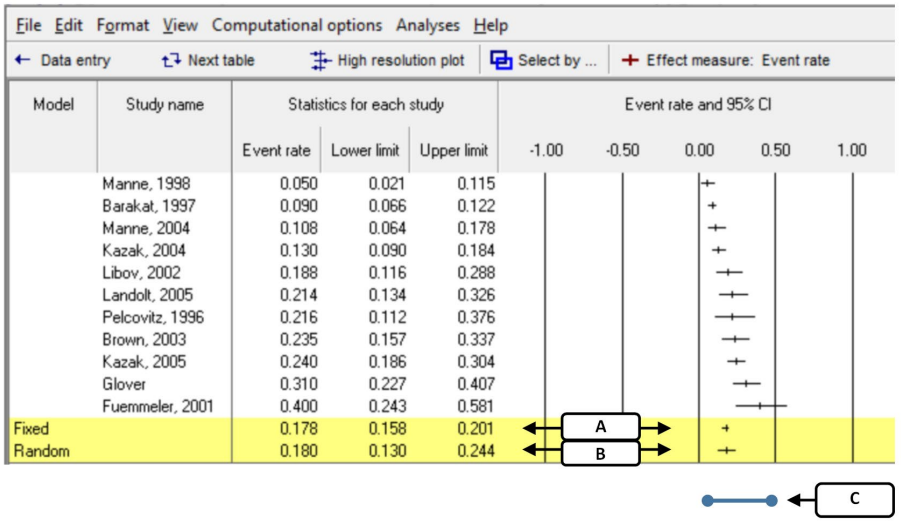

Figure 8 | Prevalence of PTSD in mothers of children with chronic illness

As it happens, the test for heterogeneity *was* statistically significant, and the reviewers *did* adopt the random-effects model. This allowed them to assess dispersion, and it turns out that the prevalence of PTSD varies from 5% in some populations to 47% in others, as indicated by the prediction interval [C].

Still, this example shows why we should use common sense rather than a significance test to select a statistical model. Had the test failed to yield a significant *p*-value, the reviewers would have stayed with the fixed-effect model, and assert that the prevalence of PTSD was precisely the same in all studies. This would subvert a key goal of the analysis (to assess heterogeneity). Also, it would defy common sense, since it is simply not plausible that the prevalence would be identical across such a disparate array of populations.

Importantly, the possibility that this approach can lead to the use of the fixed-effect model is a real problem. If we have only a few studies, it is

entirely possible to have this amount of dispersion and still have a non-significant *p*-value for the test for heterogeneity. That is the case in the next example.

7.2.9. Example | Preoperative statin therapy

Liakopoulos et al. (2008) looked at the impact of preoperative statin therapy on the incidence of stroke in patients undergoing cardiac surgery (Figure 9). The effect size index is the odds ratio, with values less than one indicating that the treatment reduced the risk of stroke.

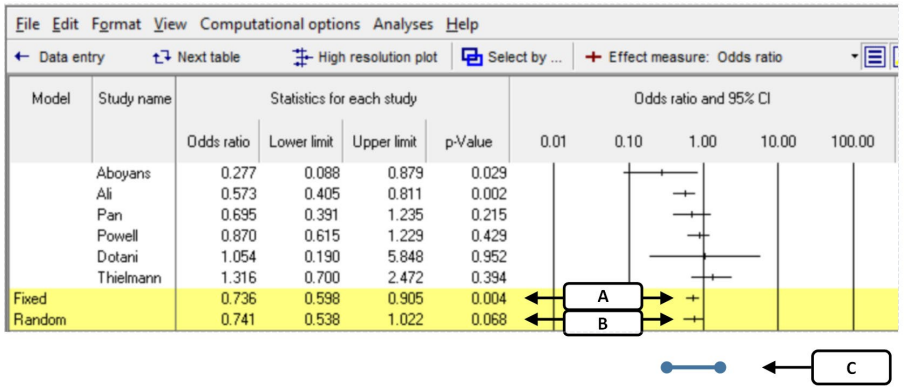

Figure 9 | Impact of preoperative statins | Odds ratio < 1 favors treatment

The correct approach would be to choose a statistical model based on our understanding of the sampling frame. In this analysis, each study was based on a unique population, and we can assume that the impact of therapy will vary by population. Additionally, the studies varied in the type of procedure (isolated CABG, isolated valve, or both) as well as the statin type, the dose, the follow-up period, and the methodological quality. One can assume that the impact of the intervention will not be precisely the same in all studies, and therefore, the fixed-effect model does not apply.

Nevertheless, the reviewers elected to use the fixed-effect model by default, and switch to the random-effects model only if the test for heterogeneity yielded a *p*-value under 0.10 and/or the I^2 value was greater than 50%. Neither of these criteria was met (the *p*-value was 0.105 and I^2 was 45%) so the reviewers applied the fixed-effect model. Figure 9 shows the results under this model, and also the results that we would have obtained using the random-effects model.

The results of the fixed-effect analysis [A] can be summarized as follows. The impact of the intervention is precisely the same for all studies included in the analysis. The effect is statistically significant so we can reject the null hypothesis and conclude that the treatment is effective in all these studies. This *common* effect is estimated as an odds ratio of 0.736.

The results of the random-effects analysis [B] can be summarized as follows. The impact of the intervention varies from one population to the next. The prediction interval [C] tells us that in some populations, the treatment *reduces* the odds of a bad outcome by as much as 68%, while in others it *increases* the odds of a bad outcome by as much as 70%. The mean effect size is not statistically significant, but in the presence of this much variation the mean is of little relevance. If the treatment is helpful in some cases and harmful in others, we need to understand where the treatment works and where it does not.

In sum, the fixed-effect analysis tells us that the treatment works, and that it works consistently. The random-effects analysis tells us that the treatment effect is stronger in some populations than in others, and in fact may be harmful in some cases. These estimates are based on only six studies, and we would want to gather additional data before reaching any firm conclusions. Nevertheless, it should be obvious that the random-effects approach yields a more plausible framework for understanding the data.

7.2.10. The correct approach

In the preceding examples the reviewers employed a test of significance to choose a statistical model, which is a mistake. By contrast, in the following cases the researchers elected to use the random-effects model a priori, which is the correct approach.

7.2.11. Example | Interventions to promote physical activity

Michie, Abraham, Whittington, McAteer, and Gupta (2009) ran a meta-analysis to synthesize studies that assessed the impact of interventions to promote better physical activity and eating habits. The studies included various interventions and various outcomes. They write "A random effects model (DerSimonian & Laird, 1986) was used in the analyses to incorporate the assumption that the different studies are estimating different, yet related, treatment effects."

7.2.12. Example | Dropout rate in adult psychotherapy

J. K. Swift and Greenberg (2012) looked at the dropout rate in adult psychotherapy. They write "Given the wide range of studies that have been included in this review (the way the studies were conducted, the interventions that were used, the clients that were treated, etc.), a random-effects model was used in the calculation of the overall dropout rate and all testing of moderators and covariates."

7.2.13. Example | Impact of preference in psychotherapy

Joshua K. Swift, Callahan, Ivanovic, and Kominiak (2013) looked at the relationship between psychotherapy preference (matching patients to the type of therapy they prefer) and the utility of the therapy. They write "A random-effects model allows the true effect to vary from study to study (Borenstein et al., 2009) and was deemed more appropriate for our analyses given that significant variability was expected between studies based on how the studies were conducted and the differences in samples that were used in each study."

7.2.14. Example | Behavior-change techniques for asthma patients

Denford, Taylor, Campbell, and Greaves (2014) looked at the utility of behavior-change techniques for helping asthma patients. They write "Given the heterogeneity in intervention content, we decided in advance to pool data using a random effects meta-analysis (J. P. T. Higgins & Green, 2011)".

7.2.15. Example | Emotional congruence with children

McPhail, Hermann, and Nunes (2013) looked at the relationship between emotional congruence with children and sexual offending against children. They write, "For the meta-analyses, we used a random-effects model (REM). This model is desirable over a fixed-effects model (FEM) when conducting meta-analysis with 'real-world' or applied data (Field, 2003; Hunter & Schmidt, 2004; Overton, 1998). We assumed there was variability beyond sampling error in the current sample of effect sizes (Lipsey & Wilson, 2001; Raudenbush, 2009) due to the applied contexts of the data collection, the convenience samples used in most studies in the sample, and our intent to generalize these meta-analytic findings beyond the current sample of studies."

7.2.16. Testing for heterogeneity when the analysis is based on one population

To this point I showed that when we are working with multiple populations, we should generally use the random-effects model, and that choosing the model based on a test could lead (incorrectly) to the fixed-effect model. The same idea holds true (in reverse) when all studies are based on one population and are identical in all material respects. In this case we should be using the fixed-effect model, and choosing the model based on a test could lead (incorrectly) to the random-effects model.

Consider the fictional analysis introduced in section 6.1.6, where our goal is to estimate the mean effect size for a specific school, and we draw twenty random samples from that school. In this case the fixed-effect model applies since all studies are estimating the same value. Suppose we tested for heterogeneity and the test yielded a significant *p*-value. Indeed, if we use a criterion alpha of 0.10 when testing for heterogeneity, we would reject the null hypothesis in 10% of such analyses. How would we interpret the test? Would we conclude that the mean score for this school varies? This is not only wrong, but is actually impossible since the mean for the school is a constant.

7.2.17. Constraints of the fixed-effect model

In sum, when studies are pulled from the literature, we can generally assume that the true effect size varies across studies, and that therefore the fixed-effect model does not apply.

There is a second reason we should not use the fixed-effect model in this kind of analysis. The fixed-effect model allows us to make an inference only to the single population included in the analysis. Therefore, the researcher would need to explain what this population is and then limit the inference to this one population. If we intend to generalize the results to comparable populations, we must use the random-effects model.

Summary

If all studies are based on the same population, we should be using the fixed-effect model. If studies are based on multiple populations, we should be using the random-effects model.

The choice of a statistical model must be based on our understanding of the sampling frame, and not on a test of statistical significance. If we don't know which of these applies then we are simply playing with numbers, and have no business running the analysis at all.

7.3. Choosing between fixed effects (plural) and random effects

7.3.1. Mistake

When researchers consider which model to use for a meta-analysis, they generally assume that the two options are the fixed-effect model and the random-effects model. In fact, there is a third option, the fixed-effects model (where the word effects is in the plural).

7.3.2. Details

The fixed-effect (singular) model and the fixed-effects (plural) model are computationally identical to each other, but conceptually different. They apply the same weights to each study and yield the same results, but reflect two different views of the sampling process and intended inference.

7.3.3. The fixed-effect model

The fixed-effect (singular) model applies when all studies are estimating the identical parameter. Operationally, this would mean that all studies are based on the same population and are identical to each other in all material ways. As discussed in the prior section, when studies are pulled from the literature, we do not meet these requirements, and this is rarely a valid option.

7.3.4. The fixed-effects model

The fixed-effects (plural) model applies when we intend to make an inference to the studies actually included in the analysis, and not generalize from those to any other studies. There are no assumptions about how these studies have been sampled or selected for inclusion in the analysis. There is no assumption that these studies are similar to each other in any way.

When studies are pulled from the literature, we can elect to use this model, and it would be entirely valid. While we *can* use this model in this case, it would generally be a bad idea to do so. This model allows us to report the mean and confidence interval for the studies in the analysis, but not to generalize beyond these studies to any other studies.

The problem is that this is precisely what we would like to do. When we publish a meta-analysis that says an intervention was (or was not) effective in our set of populations, we expect that readers will apply these results to their populations. Under the fixed-effects model this is specifically prohibited. Indeed, this is the primary difference between the fixed-effects model and the random-effects model. The latter allows us to generalize from the studies in the analysis to a universe of comparable studies. The former does not.

7.3.5. Where the fixed-effects model applies

One might ask why we would ever want to use a model where the results cannot be generalized beyond the studies in the analysis. In fact, there are several places where this model applies.

This model applies in the textbook case introduced in section 6.1.7. In that example we wanted to study the performance of students in the twenty schools that were under the control of a specific administrator. We selected those twenty schools and included them in the analysis. The results apply to those schools, and to those schools only. We understand that the performance of students in these schools says nothing about other schools, and would not have any reason to generalize to other schools.

This model also applies when the results of the analysis will be submitted as part of a proceeding to obtain approval for a new drug. This model is the preferred model here is because it makes no assumptions about how the studies were sampled. Additionally, if the requirement for approval is that the mean effect for the studies *in the analysis* is statistically significant, there is no need to generalize beyond these studies, and so the model works as intended.

7.3.6. When studies are pulled from the literature

By contrast, when applied to a meta-analysis where studies are pulled from the literature, the fixed-effects model gives the correct answer to the wrong question. The answer is correct in that the mean and confidence interval will be accurate (subject to the usual sampling error) for the studies actually included in the analysis. But, in the vast majority of cases, this is the wrong question since our interest is not limited to the studies actually included in the analysis. Rather, we want to be able to generalize to comparable studies as well.

In theory, it is possible for someone to publish a meta-analysis based on this model, and for readers to limit the results to the studies that are actually

included in the analysis. On that basis, one could argue that this is a valid use of the model.

However, in practice researchers and readers will rarely not honor this limitation. The researchers may generalize from the studies in the analysis to what they see as comparable populations and methods. Even if the researchers do limit their conclusions to the studies in the analysis, readers will invariably generalize as they see fit.

For these reasons, we should generally avoid this model in favor of the random-effects model, which is discussed in the next section. To be clear, the random-effects model does not solve all the problems outlined above. However, on balance, it is usually a better fit for the intended inference.

Summary

The fixed-effects (plural) model applies when the results will be used to make an inference to the studies in the analysis, but not generalized beyond them to any larger universe of comparable studies.

When we perform a meta-analysis based on studies pulled from the literature, we almost invariably will generalize from these studies to other studies that we see as comparable, and therefore should generally not use this model.

7.4. Limitations of the random-effects model

7.4.1. Mistake

The random-effects model works as intended when a series of assumptions are met. When we pull studies from the literature, we are likely to violate some of these assumptions, and we need to understand how this affects the meaning of the results. Failure to address this issue is a mistake.

7.4.2. Details

When the analysis is based on studies pulled from the literature, the random-effects model is almost invariably the model that should be used. This model assumes that the studies in the analysis are representative of a universe of comparable studies, and that the results of the analysis will be generalized to that universe. The computation of the confidence interval and the relative weight assigned to each study reflect these goals. Critically, this model allows us to discuss not only the mean effect size, but also the dispersion in effect size across studies. These are all key goals of the analysis.

While we should be using the random-effects model for these analyses, we need to recognize that we will be violating some assumptions that are required for the model to work as intended. We need to take this into account when we interpret the results.

7.4.3. Assumptions of the random-effects model

The random-effects model works well if the following assumptions are met.

A. The universe to which we will making an inference is defined clearly and is the correct universe in the sense that it is relevant to policy.
B. The studies that were performed are a random sample from that universe.
C. The studies that we include in our analysis are an unbiased sample of the studies that were performed.
D. The analysis includes enough studies to yield a reliable estimate of the between-study variance, τ^2.

These issues build on each other. To get a reliable estimate of τ^2 in the defined universe (D) we need to have a sufficient number of cases. But we also need to assume that the studies in our analysis are a random sample of those that

were performed (C), that those performed are a random sample of the defined universe (B), and that this universe is well defined and relevant to policy (A).

The quality of the evidence provided by a meta-analysis depends in large part on the extent to which that analysis meets these assumptions. If the analysis meets these assumptions fully, the quality will tend be good. To the extent that it fails to meet some (or all) of these assumptions, the quality is likely to be poor.

This list does not include assumptions about the internal validity of the studies that were performed. That is also critically important, but applies to all statistical models and is addressed elsewhere (see section on risk of bias, Appendix IV). For the present discussion I will assume that the individual studies have low risk of bias, and our concern is whether we can generalize from these studies to the larger universe.

7.4.4. A textbook case

Consider the textbook case of the random-effects model introduced in section 6.1.5, where all assumptions of the model are fully realized. In this case we want to estimate the mean score on a specific math test for the 1,700 high schools in New York City. We draw a random sample of 20 schools from this universe of schools, and then draw a random sample of 50 students within each of these schools. This is depicted in Figure 10.

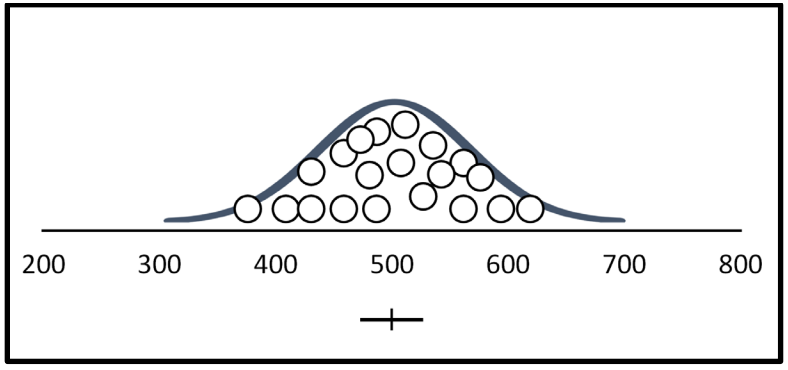

Figure 10 | Random effects| Confidence interval 60 points wide

The twenty circles in the plot represent the true scores for the 20 schools that were included in our random sample. The key factor that makes this is a random-effects analysis is the normal curve that has been superimposed on the plot. This curve reflects the fact that we have defined a universe of

populations from which we draw the samples and to which we will be making an inference.

In this example we can report that the statistical inference is of high quality since the assumptions have all been met. To wit –

A. The universe to which we will making an inference is defined as all public high schools in New York City. This is clear and unambiguous.
B. The studies that were performed are a random sample from that universe. We know that is the case, because we had a list of all 1,700 high-schools in the system and used a random process to select these twenty.
C. The studies that we include in our analysis are an unbiased sample of the studies that were performed. We know that because we know that twenty studies were performed, and all twenty of them are included in our analysis.
D. We have enough studies in our sample to yield a reliable estimate of the between-study variance.

7.4.5. When studies are pulled from the literature

By contrast, consider what happens in a typical analysis when studies are pulled from the literature. I will use the ADHD analysis as an example.

Castells et al. (2011) conducted a meta-analysis of seventeen studies to assess the impact of methylphenidate on adults with Attention Deficit Hyperactivity Disorder (ADHD). Patients with this disorder have trouble performing cognitive tasks, and it was hypothesized that the drug would improve their cognitive function. Patients were randomized to receive either the drug or a placebo, and then tested on measures of cognitive function. The effect size was the standardized mean difference between groups on the tests.

The analysis is shown in Figure 11, and it should be obvious that the effect size is smaller in some studies and larger in others. For purposes of this discussion, assume that the effect size tends to be lower in populations that employ a low dose of the drug, and higher in populations that employ a high dose of the drug.

We can use this example to highlight the differences between the textbook case and the case where studies are pulled from the literature, and show how this affects the utility of the random-effects model.

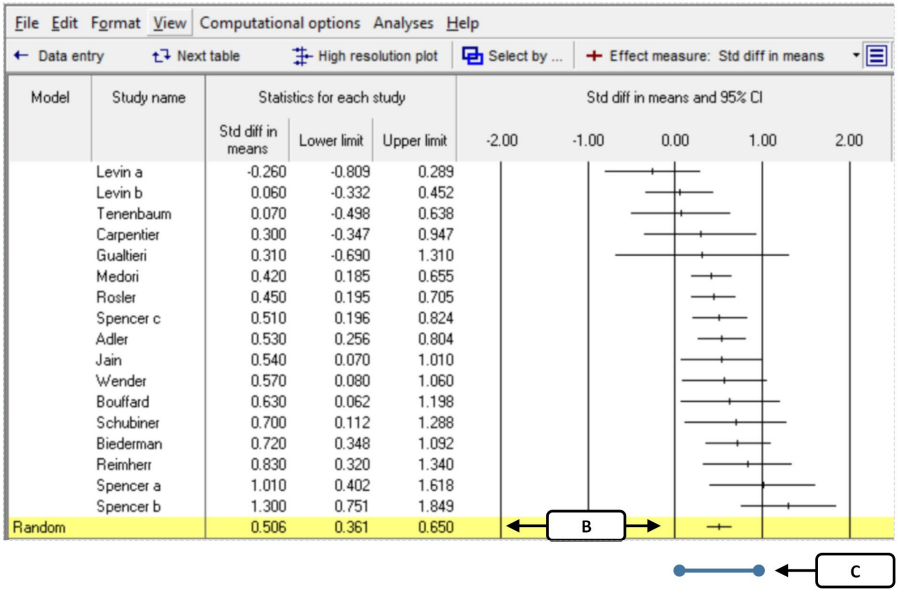

Figure 11 | Methylphenidate for adults with ADHD | Effect size > 0 favors treatment

A. We would like to think that the universe to which we are making an inference is well-defined. We might think that the universe is defined adequately by the inclusion/exclusion criteria for the review, but that is rarely the case. These criteria will not entirely define the populations and methods to be included/excluded, since there are numerous factors that could influence the magnitude of the effect and we cannot enumerate all of them. Additionally, to properly define the universe we would need to know not only (for example) that we will include studies where the dose is between 30 mg. and 80 mg. but also what *proportion* of studies will be using each dose in this range.

B. We would like to think that the studies in the analysis are a *random* sample of all studies in the universe, but that is almost never the case. Researchers who perform primary studies do not design these studies using a random process. Rather, they tend to design studies that work well for their purposes and that employ populations that are relatively easy to work with. The universe defined in (A) might include equal numbers of all doses, but the studies actually performed might favor higher doses, since these studies tend to show larger effects.

C. We would like to think that the studies included in the analysis are a random sample of all studies that had been performed, but for various reasons (including publication bias, discussed in section 11) the studies

included in the analysis might be a biased subset of the studies that had been performed.

D. We would like to think that we are able to estimate the between-study variance reliably, but that might not be the case. We need as many as twenty studies to obtain a reliable estimate of this variance, and will often have substantially fewer studies in our analysis. Additionally, we will be estimating the between-study variance for studies in the analysis, but this may be different than the value for the studies in the *intended* universe. For example, suppose that the effect size tends to be higher for studies that employed a higher dose of the drug. If the intended universe includes all doses from 30 mg. to 80 mg. but the studies in the analysis are primarily using between 60 mg. and 80 mg. the variance in our sample may be substantially smaller than the variance in the intended universe.

7.4.6. A useful fiction

In sum, when we apply the random-effects model to a meta-analysis where studies are pulled from the literature, we are engaging in a useful fiction. The model is useful because it provides a framework for thinking about the mean effect size and the dispersion in effects. But it is also something of a fiction because we are violating some (or all) of the assumptions that make the model work. We need to think of how these violations affect the results.

In the ADHD analysis the mean effect size was reported as 0.50. But, given that the mean will shift left or right depending on the mix of populations in the analysis, what universe does the mean effect size represent? We cannot say that it represents the mean in the universe that we described using inclusion/exclusion rules, since we have not met assumption (A), and do not have a sampling frame. We cannot say that it represents the mean of relevant clinical populations, since the we have not met assumption (B). For purposes of this discussion I will ignore the potential problems introduced by (C) and (D).

To get around these violations, we use language. We say that the results can be generalized to studies which are *comparable* to those in the analysis, without specifying what those studies are. This verbal sleight-of-hand yields a definition that is accurate but not useful. It is accurate since it is a tautology. The studies in the analysis are indeed comparable to studies that are comparable. But it is not useful since it does not really tell us *which* studies are comparable. That critical item is left to the judgment of the researcher or

the reader, and it may not be the same as we had intended when we planned the review.

Given that there are limitations inherent in the analysis we need to approach the results logically and see what conclusions we can draw. When we look at the entire distribution of effects, we can get a sense of the dispersion. We can then look at the mean in that context.

If there is only trivial heterogeneity among the universe of comparable studies, if follows that (a) the mean provides a useful estimate of the effect size in any given study and (b) the estimated mean will be reasonably stable regardless of which studies we happen to include in the analysis.

By contrast, if there is substantial heterogeneity among the universe of comparable studies, if follows that (a) the mean does *not* provide a useful estimate of the effect size in any given study and (b) the estimated mean will vary depending on which studies we happen to include in the analysis.

For example, in the ADHD analysis there are some combinations of factors that will lead to effects as low as 0.05, and others that will lead to effects as high as 0.95. Given the amount of heterogeneity, we should understand that the mean could shift substantially based on the particular mix of populations and methods (for example, dosage) included in the analysis. As such, the mean is not very robust.

At the same time, given the amount of heterogeneity, the mean is not terribly important. In other words, the mean is not very useful as a predictor of the effect size in any single population. Rather than focus on the mean, we need to identify factors that tell us where the effect size will be closer to 0.05 and where it will be closer to 0.95. Since the mean itself refers to a specific (and somewhat arbitrary) mix of populations, we should recognize that the test of the null hypotheses pertains to this specific mix of populations only. In that context, the test has limited value (see 10.3).

7.4.7. Transparency

If the violation of assumptions affects the kinds of conclusions we can draw from the analysis, we should explain what that means.

Many readers assume that the mean in the analysis pertains to the mean in some clearly designated universe. In the ADHD analysis we should make it clear that this is not the case. The overall mean applies to the mix of populations and treatments included in the analysis, and would shift if we included a different mix of studies.

Many readers focus on the mean effect size, and pay little attention to the dispersion in effects. In the ADHD analysis we should explain that the

true effect size in any given study could fall some distance from the mean. And, the mean itself could shift left or right, depending on the mix of studies included.

7.4.8. A narrowly defined universe

In almost any meta-analysis where studies address the impact of an intervention and are pulled from the literature, we will be violating some assumptions of the random-effects model, and therefore we need to think about the issues outlined above. However, the severity of the violations (and the potential impact) depends on several factors. Primary among these is the extent to which the universe is defined to encompass a very narrow set of studies or a more broadly defined set of studies. The ADHD analysis includes a clinically diverse set of studies and effects, but other analyses will work with a narrowly defined set of criteria.

A. When the universe is defined narrowly, it may be possible to provide a clear and comprehensive definition of the universe. Experts may be able to identify all variables that could be related to the effectiveness of the drug and set strict inclusion/exclusion criteria for these.
B. When the universe is defined narrowly, there is less concern that the studies being performed fall toward one end or the other of a distribution, since the entire distribution is narrow. We do not need to be concerned that the dosage in our studies is higher than the typical dose in the universe since the universe is limited to one dose. The same idea applies to the type of patient, the outcome, and so on.
C. When the universe is defined narrowly, the potential impact of publication bias is less than it would be in other cases. Publication bias can be based on random sampling error and also on heterogeneity in true effects. If the true effects all fall in a narrow range, there is a natural limit to how much bias can be introduced based on the latter.
D. The number of studies that we need to get a reliable estimate of the between-study variance (τ^2) depends in part on how widely the true effects vary. When the universe is defined narrowly and the within-study variance is small, we may be able to get a reliable estimate of τ^2 with only a handful of studies (see section 9.9).

In the Cochrane Database of Systematic Reviews, a substantial proportion of the meta-analyses report that the between-study variance is *estimated* as zero. While it is not likely that the true variance is actually *zero* (see section

7.2.4) it is possible that the true variance is *trivial*. In that case, the issues outlined here for a narrowly defined universe would apply.

Critically, the advantages associated with a narrowly defined universe only apply if the results are actually limited to that universe. In practice, readers may generalize beyond that universe to other populations, variants of the intervention, and so on.

7.4.9. In context

Given the problems associated with using the random-effects model when studies are pulled from the literature, some have advocated for using the fixed-effects (plural) model instead. While there are limitations to both models, there is a growing consensus that the random-effects model is generally preferable, for the following reasons.

First, the random-effects model provides the correct conceptual framework for thinking about the analysis. It explicitly acknowledges that we intend to make an inference to a wider set of studies. Even if parts of the process are ambiguous (for example, deciding which studies are comparable) it is preferable to include them in the process so that we are clear about where the model is not reliable.

Second, the random-effects model allows us to compute prediction intervals that tell us the range of true effect sizes that might be expected in comparable studies. As explained earlier, this can be an essential element in our understanding of the results.

Third, the fixed-effects model reports a confidence interval for the mean effect size for the studies *in the analysis*, and tends to be relatively small. By contrast, the random-effects model reports a confidence interval for the *universe of comparable studies*, and tends to be wider. If we will be making an inference to the universe of comparable studies, the random-effects interval is a better match for the intended inference.

Fourth, under the fixed-effects model, large studies may dominate the analysis and small studies may be effectively ignored. Essentially this means that we assign the same weight to each person rather than each study (see (Hedges & Vevea, 1998; Peto, 1987; Rice et al., 2017; "Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group," 1998). As such, the random-effects model is a better match if our intent is to make an inference to all comparable studies.

Fifth, while both models have limitations, the random-effects model has the *potential* to work well when we have enough data and a representative

sample of the intended universe. As we approach these conditions, the model *will* yield a useful estimate of the mean effect size and dispersion of effects in that universe. By contrast, the fixed-effects model is designed to make an inference only to the studies actually included in the analysis, and not to the universe of comparable studies. This will not change as the quality and quantity of the data improves.

7.4.10. Extreme cases

For the reasons discussed above, the random-effects model should generally be the model we use when studies are pulled from the literature. This model is likely to work *well enough* when the universe is defined narrowly. It may also work *well enough* when the universe is defined broadly but we have a reasonable number of studies, so we can get a general sense of the dispersion.

However, if the universe is defined broadly and we have only two or three studies (for example), the model becomes untenable. In this case we need to choose among several options.

- We can apply the random-effects model and explain that the estimates are unreliable. It would be very useful to apply the Knapp-Hartung correction (see section 7.5). This will substantially expand the width of the confidence interval, and thus clarify the extent of the uncertainty. The drawback to this approach is that the interval may be so wide that we learn almost nothing from the analysis.
- We can apply the fixed-effects (plural) model, and make it very clear that the results apply only to the studies in the analysis, and cannot be generalized beyond them to any other studies. The drawback to this approach is that readers will tend to ignore the caveat, and generalize as they see fit.
- We can display the forest plot without a summary effect (Poole & Greenland, 1999). The problem with this approach is that readers may construct an even more flawed summary of their own.

None of these options is a good one, and regardless of the option chosen, it is imperative to be transparent about the limitations of the analysis.

Summary

The random-effects model works as intended if all assumptions are met. Specifically, it will work as intended if we (A) enumerate a universe of all possible studies, (B) draw a random sample of studies from that universe, (C) ensure that the studies in the analysis are a representative sample of the studies performed, and (D) have a sufficient number of studies to yield an accurate estimate of the between-study variance.

When studies are pulled from the literature, we are likely to violate some (or all) of these assumptions. Typically, (A) it is not entirely clear what studies are included in the intended universe, (B) the studies actually performed are not a random sample from the intended universe, (C) the studies included in the analysis might not be representative of those actually performed, and (D) we may not have a sufficient number of studies to yield an accurate estimate of the between-study variance.

Therefore, when we say that the results apply to studies which are comparable to those in the analysis, it may not be clear which studies are actually comparable to those in the analysis. This will be true especially when the universe is defined broadly, and the effect size varies substantially across studies.

7.5. Knapp-Hartung adjustment

7.5.1. Mistake

Researchers generally compute a confidence interval and *p*-value using the *Z*-distribution. In most cases it would be preferable to apply the Knapp-Hartung adjustment.

7.5.2. Details

Traditionally, the confidence interval for the mean is based on the *Z*-distribution, which yields a relatively narrow interval. When we use the random-effects model it would be better to use the Knapp-Hartung adjustment (sometimes called the Hartung-Knapp-Sidik-Jonkman adjustment), which yields a wider (and more accurate) confidence interval (J. P. Higgins & Thompson, 2004; IntHout, Ioannidis, & Borm, 2014; Jackson, Law, Rücker, & Schwarzer, 2017; Knapp & Hartung, 2003; Sidik & Jonkman, 2002).

The adjustment includes two components. First, it modifies the standard error of the mean. Second, it multiplies the standard error by a factor based on the *t* distribution rather than the *Z* distribution. It is *always* a good idea to use this adjustment, since the adjusted interval is more accurate. However, it is *especially important* to use this adjustment when there are a small number of studies in the analysis and the between-study variance is non-trivial.

Consider an analysis where the effect size index is the standardized mean difference (*d*) and the standard error of the mean is 0.10. Table 2 shows how the confidence interval is affected when we use *t* rather than *Z*. Without the correction, the confidence interval width is around 0.40 regardless of the number of studies. With the correction, the width increases as the number of studies decreases. When the number of studies is 30, 10, 4, and 2 the interval width is approximately 0.41, 0.45, 0.64, and 2.54. Equivalently, the width is increased by a factor of 1.04, 1.15, 1.62, and 6.48. As noted above, there is a second part to the adjustment which involves the standard error, and which may widen the interval even further.

While this discussion has been focused on the width of the confidence interval, the same issues apply to tests of the null hypothesis (see Appendix V).

Table 2 – Impact of using t-distribution on the confidence interval width

	Number studies	Critical value	Lower Limit	Upper Limit	Width	Ratio t: Z
Z-Distribution	n/a	1.960	0.304	0.696	0.392	1.00
t-distribution	100	1.984	0.302	0.698	0.397	1.01
	30	2.045	0.295	0.705	0.409	1.04
	20	2.093	0.291	0.709	0.419	1.07
	10	2.262	0.274	0.726	0.452	1.15
	5	2.776	0.222	0.778	0.555	1.42
	4	3.182	0.182	0.818	0.636	1.62
	3	4.303	0.070	0.930	0.861	2.20
	2	12.706	−0.771	1.771	2.541	6.48

While there is a consensus among statisticians that we should always apply this adjustment when we use the random-effects model, it is used only rarely in practice, for several reasons.

- Most researchers are not aware of this adjustment.
- The adjustment is not always available in software.
- The adjustment may yield a very wide confidence interval, and may move the *p*-value to a non-significant range, which makes it less attractive to researchers who may have a vested interest in reporting a statistically significant result.

In an effort to address the second of these items, the adjustment is being added as an option in software, and hopefully will be adopted more widely in the near future. It is possible to use this adjustment now in CMA (Appendix V). IntHout et al. (2014) show how the adjustment can be implemented in Excel.

7.5.3. Limitations of the Knapp-Hartung adjustment

For the random-effects model, the Knapp-Hartung adjustment always yields better coverage than the non-adjusted value, and so it should always be used. However, it works better under some circumstances than others (IntHout et al., 2014). Details are provided in Appendix V.

The adjustment makes it more likely that the confidence interval will include the true mean for studies comparable to those in the analysis. However, it cannot adjust for the possibility that the studies in the analysis are

not representative of the intended universe. For example, suppose that the universe is defined as studies that employed a dose between 30 mg and 80 mg, and the true mean effect size for these studies is 0.50. However, most studies in the analysis employed a dose between 60 mg and 80 mg, and the true mean effect size for these studies is 0.70. The K-H adjustment makes it more likely that the confidence interval will include the value of 0.70. It cannot adjust for the fact that this is different than the mean in the *intended* universe.

Ironically, when the adjustment is most needed (when we have a small number of studies) the impact of the adjustment may be so large that the estimate of the mean effect size will be uninformative. In the example presented in 7.5.2, with twenty studies the interval would have a width of 0.42 but with two studies would have a width of 2.54. The latter interval is so wide that it tells us nothing of real value. While this is unfortunate, it represents the true state of affairs. When the between-study variance is non-trivial, an estimate of the mean effect size for the universe of comparable studies, based on two studies, is not reliable.

Summary

The confidence interval for the mean effect size in random-effects analysis is too narrow when based on the Z-distribution. It would be better to use the Knapp-Hartung adjustment, which yields a wider (and more accurate) interval. The adjustment applies both to the confidence interval and to the test of the null hypothesis for the mean effect.

The magnitude of the adjustment depends on the number of studies in the analysis. When the analysis includes many studies, the adjustment will tend to be relatively modest. When the analysis includes only a few studies, the adjustment will tend to be substantial.

7.6. Meta-analysis in legal applications

7.6.1. Mistake

In analyses where the random-effects model cannot be used, researchers sometimes assume that they must use the fixed-effect (singular) model. In fact, they have the option of using the fixed-effects (plural) model, which is likely to be a better fit for the data.

7.6.2. Details

The key advantage of the random-effects model is that it allows us to generalize from the studies in the analysis to the universe of comparable studies. However, as discussed earlier, it is not entirely clear what studies are comparable to those in the analysis. If our goal is to publish the analysis, we may be willing to accept some ambiguity about which studies are comparable. However, when the analysis is being used as part of a legal proceeding, this ambiguity is not acceptable. Therefore, it may be necessary to apply the fixed-effects model. Here, the results apply only to the studies in the analysis, and we make no assumptions about how these studies came to be included in the analysis nor about what studies might be comparable to them.

7.6.3. Fixed-effect model (singular) vs. fixed-effects model (plural)

Researchers using a meta-analysis for a legal application sometimes choose the fixed-effect (singular) model because they are not aware of the fixed-effects (plural) model. In fact, the latter may be a much better option (Rice et al., 2017).

In those cases where we really expect that all studies are estimating the same parameter, the fixed-effect (singular) model is appropriate. But when the studies are not estimating the same parameter, there is no need to suggest that they are. One can simply report the same statistics, based on the fixed-effects model. The fixed-effects (plural) model applies when we intend to make an inference to the studies in the analysis, and not generalize beyond them to any universe of comparable studies. There is no assumption that the studies are estimating the same parameter.

The two models are conceptually different from each other, but computationally identical to each other. They use the same weights and yield the same results.

Summary

When the analysis will be submitted to a regulatory agency or as evidence in a legal proceeding, the random-effects model is generally not tenable since it requires that we make assumptions about the sampling process, and about what studies are comparable to those in the analysis. Therefore, for these purposes we generally use the fixed-effect or fixed-effects model. These models allow us to make an inference only to the studies in the analysis, and we make no assumptions about what other studies might be comparable.

The fixed-effect and fixed-effects model use the same weights. We use the singular label (effect) when all studies are based on the same population and identical to each other in all material respects. We use the plural label (effects) otherwise. The numbers are the same in both cases.

7.7. Putting it all together

When a meta-analysis is based on studies pulled from the literature, the random-effects model is almost invariably the model that should be used. This model assumes that the studies in the analysis are representative of a universe of comparable studies, and that the results of the analysis will be generalized to that universe. The computation of the confidence interval and the relative weight assigned to each study reflect these goals. Critically, this model allows us to discuss not only the mean effect size, but also the dispersion in effect size across studies. These are all key goals of the analysis.

However, there are typically problems in using this model. The model works properly if we have a clear understanding of the universe to which we will be making an inference, if the studies actually performed are a random sample from this universe, if the studies in the analysis are an unbiased sample of the studies that were actually performed, and if we have a sufficient number of studies to estimate the various parameters reliably. When studies are pulled from the literature, we may fall short on many of these assumptions.

An important issue here is the extent to which the effect size is consistent or varies across studies. If the true variation across studies is trivial our estimate of the mean will be robust, since it will fall within a narrow range regardless of the mix of populations included in the analysis. By contrast, if there is substantial variation in the true effect size, the mean effect size will depend on the specific mix of populations included in the analysis. Critically, this all depends on the *true* variation in effects, not the *estimated* variation. As discussed in section 9.9, the estimate may not be reliable.

Researchers sometimes assume there is such a thing as a *correct* model, and that their task is to identify that model. We need to recognize that there may be no model that will yield an entirely correct answer. Rather, each model has limitations. This idea was expressed by (Poole & Greenland, 1999) when they wrote "We see either type of summary [based on a fixed-effect or a random-effects model] as having only a minor role in a well-done meta-analysis, unless all the studies are very similar in their methods, in their populations at risk, in their exposure contrasts, and in their results. Unfortunately, random-effects summarization seems to have become one of the statistical methods, like significance testing, that tend to be applied to epidemiologic data ritualistically and without much thought for important features of the data they might conceal."

No model is able to serve as a talisman that will magically yield the correct answer. Rather, we need to decide which model works best for the

intended inference, and then consider the limitations of that model when interpreting the results.

8. ISSUES AND MYTHS ABOUT STATISTICAL MODELS

8.1. Overview

In section 7, I discussed which issues we should consider when choosing a statistical model for a meta-analysis of studies that are pulled from the literature. In practice, researchers sometimes take account of other issues when deciding which model to use. For the most part these issues should not play a role in the choice of a model. However, since researchers often ask about these issues, I address them here. These include the following.

- Some researchers believe that the random-effects model assigns equal weight to all studies, or too much weight to small studies.
- Some researchers perform an analysis using both fixed-effect and random-effects models and then compare the results.
- Researchers sometimes suggest that we should use the random-effects model because it is more conservative than the fixed-effect model.
- Researchers sometimes suggest that we should use the fixed-effect model because it has better statistical power than the random-effects model.

In general, I will refer to the fixed-effect (singular) model in this discussion, since this is the model most researchers have in mind when they address these issues. However, the discussion applies to the fixed-effects (plural) model as well.

8.2. Random-effects model assigns equal weight to all studies

8.2.1. Mistake

Some researchers have said that they used the fixed-effect model because the random-effects model gives equal weight to all studies, and they want to give more weight to larger studies. This is a mistake.

8.2.2. Details

The relative amount of weight assigned to each study is important because it affects the estimate of the overall mean as well as the estimate of heterogeneity. In general, we would want to assign more weight to larger studies, and if it were true that the random-effects model assigned equal weights to all studies, this would be a cause for concern. However, this assertion is simply incorrect.

For example, Figure 12 shows the impact of an intervention to reduce alcohol abuse in students (Smedslund et al., 2017). For purposes of this illustration, the studies are sorted based on the weight assigned to each, from low to high.

The two columns at right show the relative weight assigned to each study under either model. Under the fixed-effect model [D], the relative weights vary from a low of 0.42% to a high of 20.47%. Under the random-effects model [E] the relative weights vary from a low of 1.09% to a high of 12.26%. So, the suggestion that the random-effects model assigns the *same* weight to all studies is incorrect. Under the random-effects model, the weights *do* vary. They just do not vary as much as they vary under the fixed-effect model. The ratio of the highest weight to the lowest weight under the fixed-effect model is around 50:1, and under the random-effects model is around 10:1.

Toward the top, where the studies are small, the relative weights are *higher* in the right-hand column (random) than the left (fixed). Toward the bottom, where the studies are large, the relative weights are *lower* in the right-hand column (random) than the left (fixed). The principle is that the weights are more *moderate* under random effects. The relative weight for small studies is pulled up, while the relative weight for large studies is pulled down.

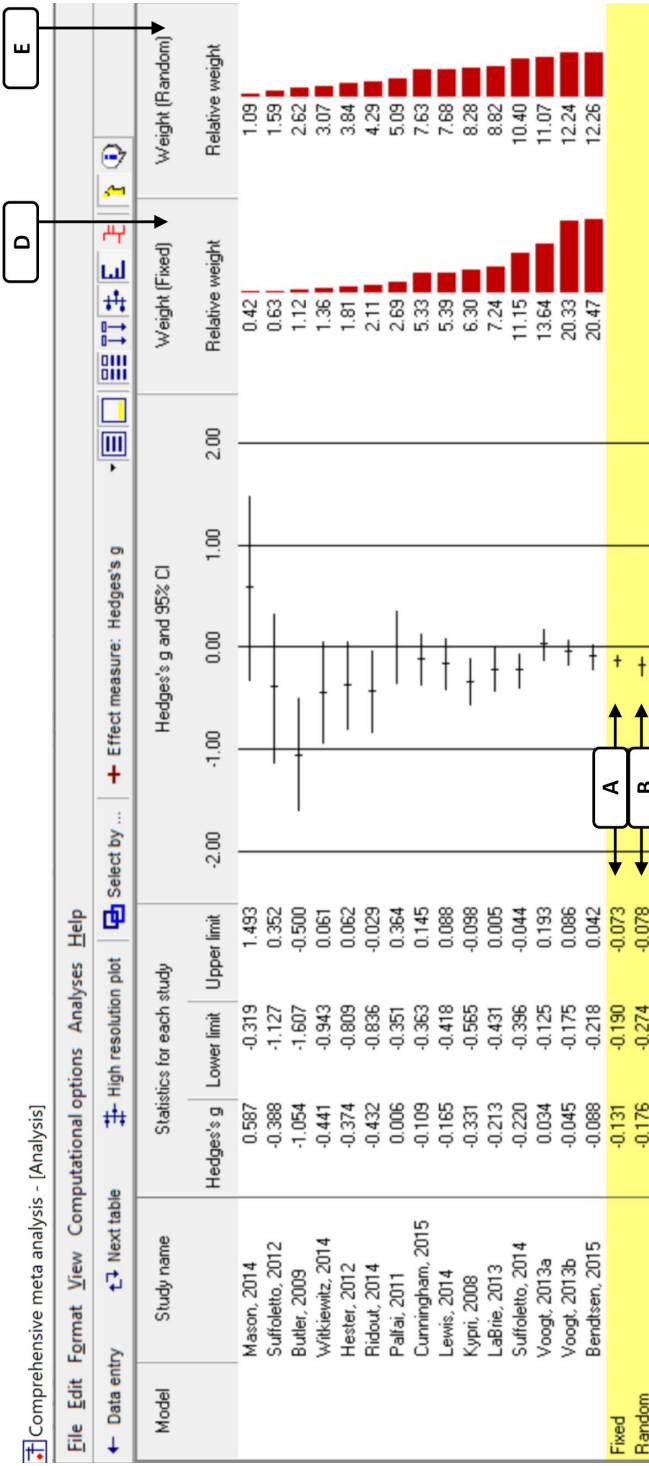

Figure 12 | Reduce alcohol abuse | Standardized mean difference < 0 favors treatment

It helps to review the reason we assign more modest weights under the random-effects model, and more extreme weights under the fixed-effect model. Consider the fictional analyses shown in Figure 13. Each analysis is based on five studies, and each circle represents the *true* effect size in one study.

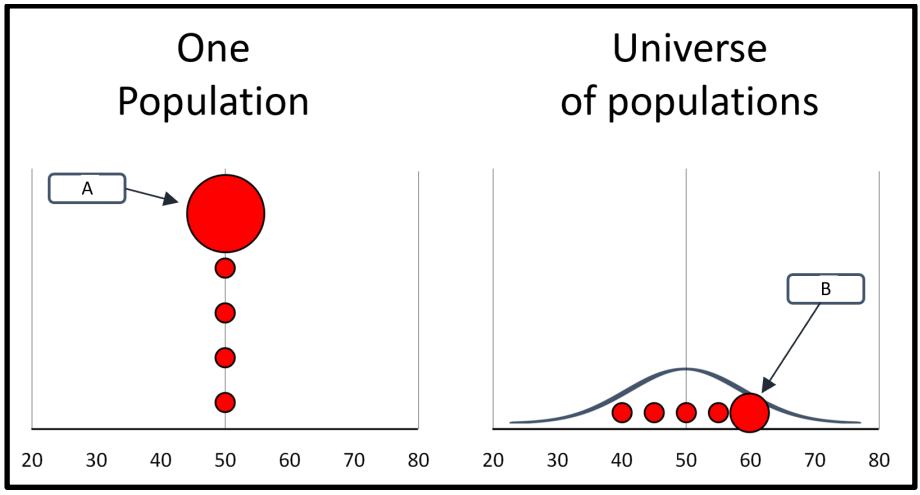

Figure 13 | Large studies are more likely to be dominant under fixed-effect model

The left-hand plot in Figure 13 corresponds to the fixed-effect model. All studies have been sampled from the same population, where the true effect size is 50 points. Each circle represents the *true* effect size in one study, and by definition the true effect size for all the studies is 50. One study [A] includes ten times as many people as the others, and so is able to estimate this parameter with one-tenth the error variance as the others. It is assigned ten times as much weight as the others, as suggested by the area of the circle (which is proportional to the relative weight assigned to that study).

The right-hand plot in Figure 13 corresponds to the random-effects model. Studies have been sampled from a universe of populations where the *mean* effect is 50 points, but the true effect size in any population could be as low as 30 or as high as 70. If one study happens to have a very large sample size, we do not want that study to dominate the analysis. We might know the effect size for that population precisely, but this is only one population among many. For example, one study [B] includes ten times as many people as the others. This study is able to estimate the effect *in this specific population* with a relatively small amount of error but there is no reason to think that this study provides a precise estimate of the mean for all studies in the universe to which

we are making an inference. Therefore, this study is assigned twice as much weight as the others (rather than ten times as much), as suggested by the area of the circle.

The weights mentioned in these examples (ten-fold, two-fold) are only for purposes of illustration, and the precise weights will depend on any number of factors. There will be cases where the weights under the random-effects model are very similar for all studies, and there will be cases where the weights vary more than they did in this example. The key point is that the weights are calibrated to match the intended inference. If there is a case where the weights are similar for all studies, that is because those weights are the ones that will yield the best (most efficient) estimate of the mean effect size in that case. See Appendix III for details.

Summary

The assertion that the random-effects model assigns the same weight to all studies is incorrect. Rather, the weights are calibrated precisely to take account of both within-study and between-study variance. This is appropriate when our goal is to make an inference to the universe of comparable studies.

8.3. Random-effects model gives too much weight to small studies

8.3.1. Mistake

Small studies tend to have more of an impact on the summary effect-size under the random-effects model, and less of an impact under the fixed-effect model. Researchers who are concerned that small studies may be of poor quality will sometimes switch to the fixed-effect model to minimize the impact of the small studies. This is generally a bad idea.

8.3.2. Details

While the rationale for the weights is based solely on the goal of minimizing the variance due to *sampling error*, some researchers thought that they could capitalize on these weights to address an *entirely separate* issue. Specifically, there is sometimes a concern that small studies may suffer from poor quality (Nuesch et al., 2010). Some researchers proposed that if we switch to the fixed-effects model, these studies will be assigned smaller weights in the analysis, which would limit their impact.

The idea of switching to the fixed-effect model to minimize the impact of poor-quality studies is a bad idea, for several reasons.

- First, there is no reason to assume, as a matter of course, that small studies are of poor quality. They *may* be of poor quality, but they also may be of equal (or better) quality than the larger studies.
- Second, if we wanted to weight studies based on quality, we would want to use weights that reflected the quality of these studies. To suggest that the fixed-effect weights somehow match the study quality is to assert (for example) that a study with 20 people has one-fifth the quality of a study with 100 people. It would be hard to even say what that means, let alone argue that the ratio is correct.
- Third, even if switching to the fixed-effect model somehow addressed the problem of poor studies being of low quality, it would still not be an acceptable solution, because if we make this switch, we are fundamentally changing the goals of the analysis. The random-effects model allows us to generalize the results to the universe of comparable studies. If we switch to the fixed-effects model, the results apply only to the studies in the analysis, and cannot be generalized beyond them.

Additionally, if we switch to the fixed-effect model we can no longer assess dispersion in the effects.

Perhaps most importantly, this approach does not allow us to isolate the reason that the results change when we switch models. If the results change when we switch models, we might assume that the change is due to the reduced the impact of small studies, but that might not be the case. A change in results *could be* due to a reduction in the impact of small studies, but could also be due to other (unrelated) factors as in the examples below.

In any event, there are other options that *do* allow us to address the possibility of poor quality in small studies, without switching to an alternate statistical model.

- One option is to run an analysis comparing the effect size in small studies vs. large studies. This would address the specific issue of sample size, rather than confounding it with a host of other issues.
- Another option would be to run an analysis comparing studies with high risk vs. low risk of a specific bias (such as selective outcome reporting). This actually looks at the issue we care about, which is the risk of bias, rather than study size, which is treated as a surrogate for this bias.

The following example illustrates these points.

8.3.3. Example | Impact of GLP-1 mimetics on blood pressure

Katout et al. (2014) looked at the impact of GLP-1 mimetics on diastolic blood pressure (Figure 14). They decided a priori to apply a random-effects model because they expected clinical heterogeneity in the effects. Then, they ran a sensitivity analysis using a fixed-effect model to ensure that the initial results were not overly affected by the impact of the small studies. The numbers that follow are based on our re-analysis of the data, and differ slightly from the original report due to rounding error.

Based on the random-effects analysis [B], the mean effect size is −0.473 mm Hg. By contrast, based on the fixed-effect analysis [A], the mean effect size is −0.266 mm Hg. The reviewers conclude that "The effect was more modest using fixed-effects model". This suggests that the effect size had been larger for the small studies, and when these studies are down-weighted, the effect size decreases. However, a closer examination of the data shows that this is incorrect.

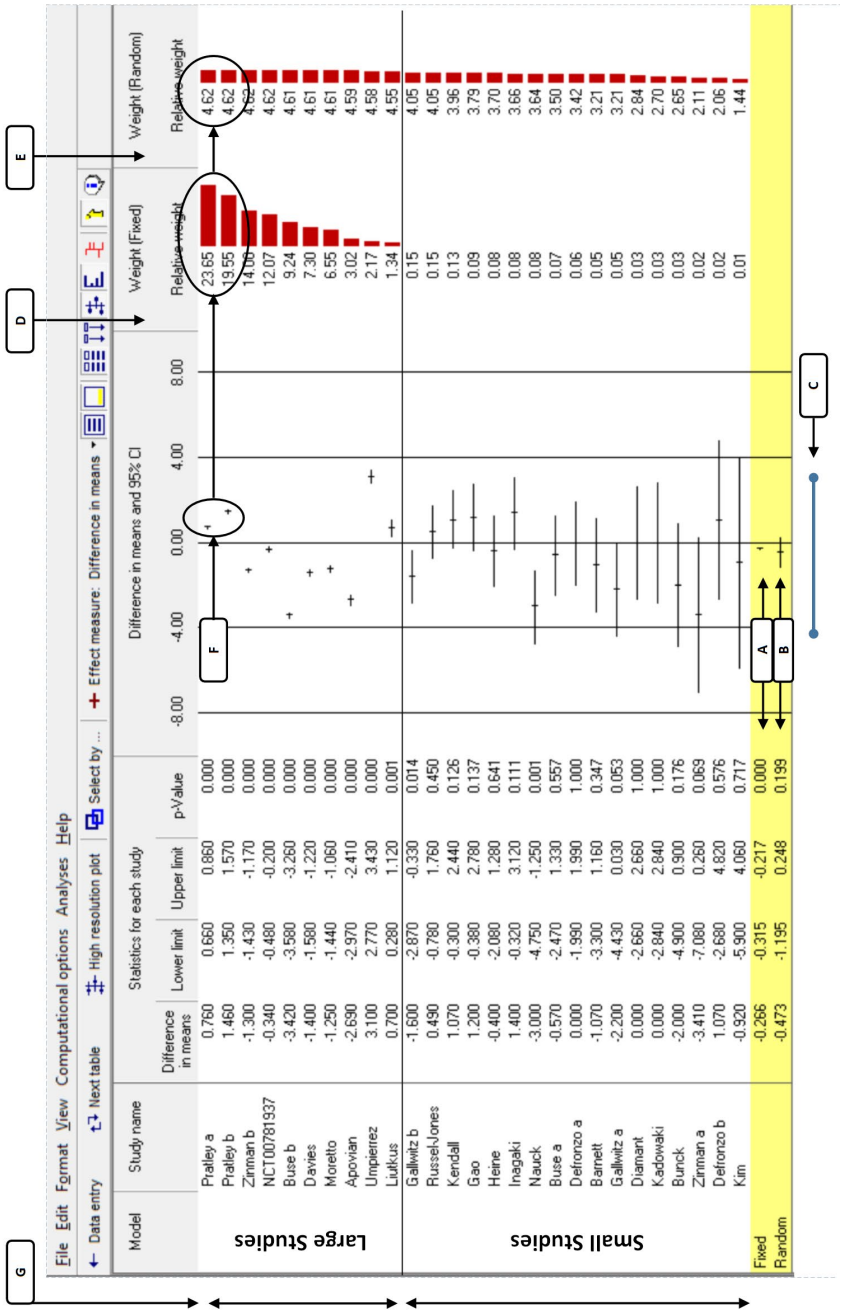

Figure 14 | GLP-1 mimetics and diastolic blood pressure | Raw mean difference < 0 favors treatment

In Figure 14, the studies are sorted with the largest studies at the top and the smaller ones at the bottom. For purposes of this discussion I will divide the studies into ten large and seventeen small studies as labeled in column [G].

Columns [D] and [E] show the relative weight assigned to each study under the fixed-effect and random-effects models. Under the random-effects model [E] the seventeen small studies are assigned some 54% of the weight. By contrast, under the fixed-effect model [D] the seventeen small studies together are assigned less than one percent of the weight – in effect, the small studies have been removed from the analysis.

As noted, the reviewers reported that when they switched to the fixed-effect model the effect size became more modest. The implication is that the small studies tended to have larger effects, and when these studies are (essentially) removed, the effect size based on the remaining studies shifts closer to zero. While this narrative is plausible, it is not actually what happened here.

In fact, the reason that the effect size shifted when we moved to the fixed-effect model was *not* that small studies tended to have larger effects. As we will see shortly, the mean effect size in the small studies was precisely the same as the mean effect size in the larger studies. Rather, the reason that the effect size shifted to the right was because of a re-alignment of weights *among the four largest studies*. Under the random-effects model, the two largest studies (together) [F] were assigned some 9% of the total weight. By contrast, under the fixed-effect model these two studies were assigned some 43% of the total weight, and came to dominate the analysis. As it happens, these two studies showed the treatment to be harmful (to the right of zero), and thus pulled the mean effect size toward the right. This is the reason for the "more modest" effect. So, the shift that the authors attributed to less influence of small studies was actually due to a shift of influence *among* the largest studies (which presumably were all of high quality).

To this point, the example shows why it is a bad idea to switch to the fixed-effect model to down-weight the smaller studies. Once we change models, weights may re-align in ways that we never intended. As in this example, we may assume that the shift is due to one factor when it is actually due to something else entirely.

Additionally, there are other lessons to be learned from this example.

Under the random-effects model we *would not* reject the null hypothesis. Under the fixed-effect model we *would* reject the null hypothesis. The reason we reject the null hypothesis under the fixed-effect model (even though the effect size is *smaller* under this model), is that when we adopt this model, we

omit the between-study variance from the error term, and the width of the confidence interval drops by 93%, from 1.50 units as indicated by line [B], to 0.10 units as indicated by line [A]. The computation of this interval for the fixed-effect model is based on the assumption that all studies are being pulled from the same population, and as such is clearly incorrect. The test of significance that allows us to reject the null hypothesis is also based on this assumption is therefore irrelevant.

The drug shows substantial benefit in some populations (lowering BP by 4.08), and substantial harm in others (increasing BP by 3.13) as indicated by line [C]. When there is this much variation in effect size, the key finding should be that the impact of GLP-1 mimetics on blood pressure varies substantially across populations. Under the random-effects model this would be the message. By contrast, if we switch to the fixed-effect model we are asserting that there is no variation in effect size, and therefore we cannot even discuss the variation.

Ironically, the issue highlighted in the paper, that switching to the fixed-effect model yielded a more moderate effect (and presumably showed that the effect size was larger in the small studies) is misleading in itself. Technically, the effect size did shift by 0.21 points. But given that the effect sizes range over 6.52 points, a shift of this size is so trivial as to be meaningless. This is clear if we compare lines [B] vs. [A]. The point estimates for the two are so close that they appear to be the same.

Fortunately, there are other ways of addressing the concern about small studies. One option is to divide the studies into subgroups (Large vs. Small) and run an analysis to compare the effects in the two subgroups, as in Figure 15. This analysis *isolates* the factor of sample size, rather than conflating it with a host of unrelated issues. In this example, the mean effect size is virtually identical in both sets of studies. The mean effect size is −0.441 for the large studies [B1], and −0.444 for the small studies [B2].

8.3.4. Study quality vs. risk of bias

Since researchers often suggest switching to the fixed-effect model as a way of addressing the assumption that small studies tend to be of poor quality, I felt that it was important to address this issue. However, I should note that the idea of classifying studies as having some level of *quality* has largely fallen out of favor. Researchers now recognize that it is preferable to look at specific items (for example, improper randomization, selective reporting of results) that could lead to the study results being biased, rather than using a general assessment of quality.

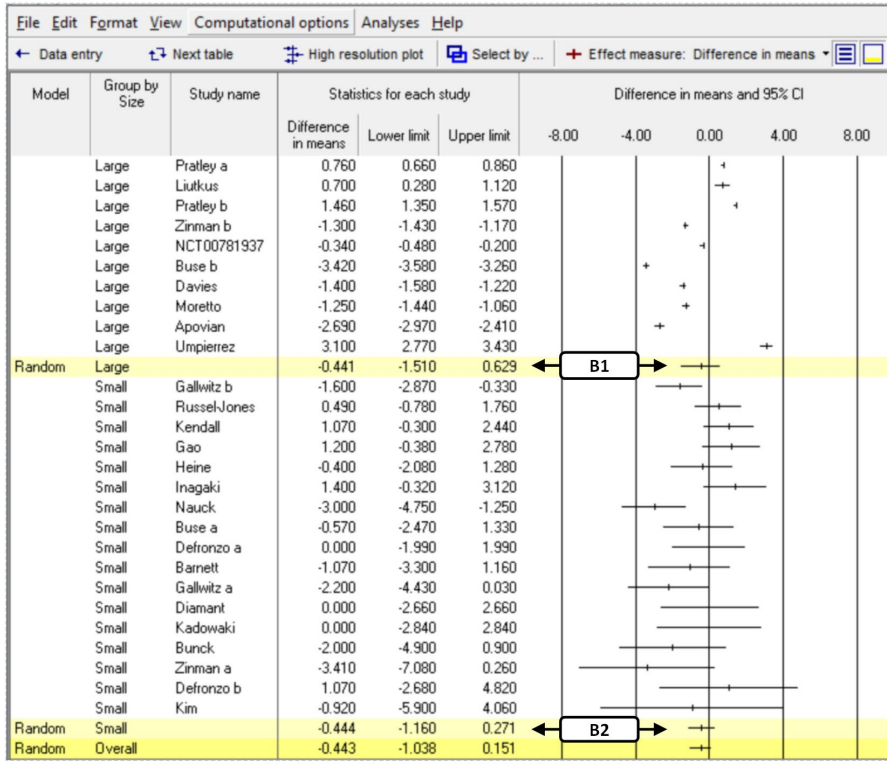

Figure 15 | GLP-1 mimetics and BP | Large studies (top) and small studies (bottom)

This is typically addressed in a risk of bias table (J. P. T. Higgins & Green, 2011). This table is used to assess the likelihood that each study suffers from a particular kind of bias, such as inadequate controls for the randomization, or selective reporting of the outcome data. This framework allows us to assess the potential for each type of bias (such as improper randomization protocols) in any study, rather than simply assuming that small studies suffer from these biases while larger studies do not (S. Greenland & O'Rourke, 2001; Juni, Witschi, Bloch, & Egger, 1999). See Appendix IV for details.

Summary

The idea of shifting to a fixed-effect model in order to down-weight small studies is a poor idea. First, we should not assume that the small studies are of poor quality. Second, even if we did make this assumption, this approach is seriously flawed since it may affect the analysis in ways that we never intended. Specifically, (A) it may re-align weights among the large studies, (B) it reduces the standard error, which may invalidate the test of significance and the confidence interval, and (C) it does not allow us to discuss variation in effects.

In any event, if we are concerned about the quality of the small studies there are approaches which allow us to isolate this factor and address it directly. We can focus on specific types of bias rather than the ambiguous construct of *study quality*. We can then address the potential impact of these specific biases, rather than conflate them with each other under the "small study" label.

8.4. Comparing results from the two models

8.4.1. Mistake

Researchers sometimes run an analysis using both fixed-effect and random-effects models, and then compare the two results. While some have advocated for this approach as a kind of sensitivity analysis, it is not clear what purpose this actually serves.

8.4.2. Details

When a researcher runs the analysis under both models, the intent is generally to see what would happen if we changed some element of the analysis. For example, suppose that we are using a random-effects analysis, and we know that under this model small studies may have a substantial impact. To see what would happen if we minimize the impact of the small studies, we re-run the analysis using a fixed-effect model. It turns out that the results shift, and we assume this is because we have minimized the impact of the small studies.

This is a variant of the prior issue (section 8.3). There, researchers thought that the fixed-effect model might be better because it down-weighted small studies. Here, they are comparing the two models without a specific rationale. In either case, the problem is that when we shift models, we set in motion a whole array of changes, and some of these changes might be inconsistent with our goals.

Consider the following example.

8.4.3. Example | Impact of educational programs

Lauer et al. (2006) looked at the impact of educational programs that were offered to at-risk students outside of the usual school hours. Figure 16 shows the studies in the analysis, sorted by effect size from low to high. The researchers ran the analysis using both models. The summary effect size under random effects [B] is 0.170, and under fixed effect [A] is 0.089. So, when we move to the fixed-effect model the effect size drops by roughly 50%.

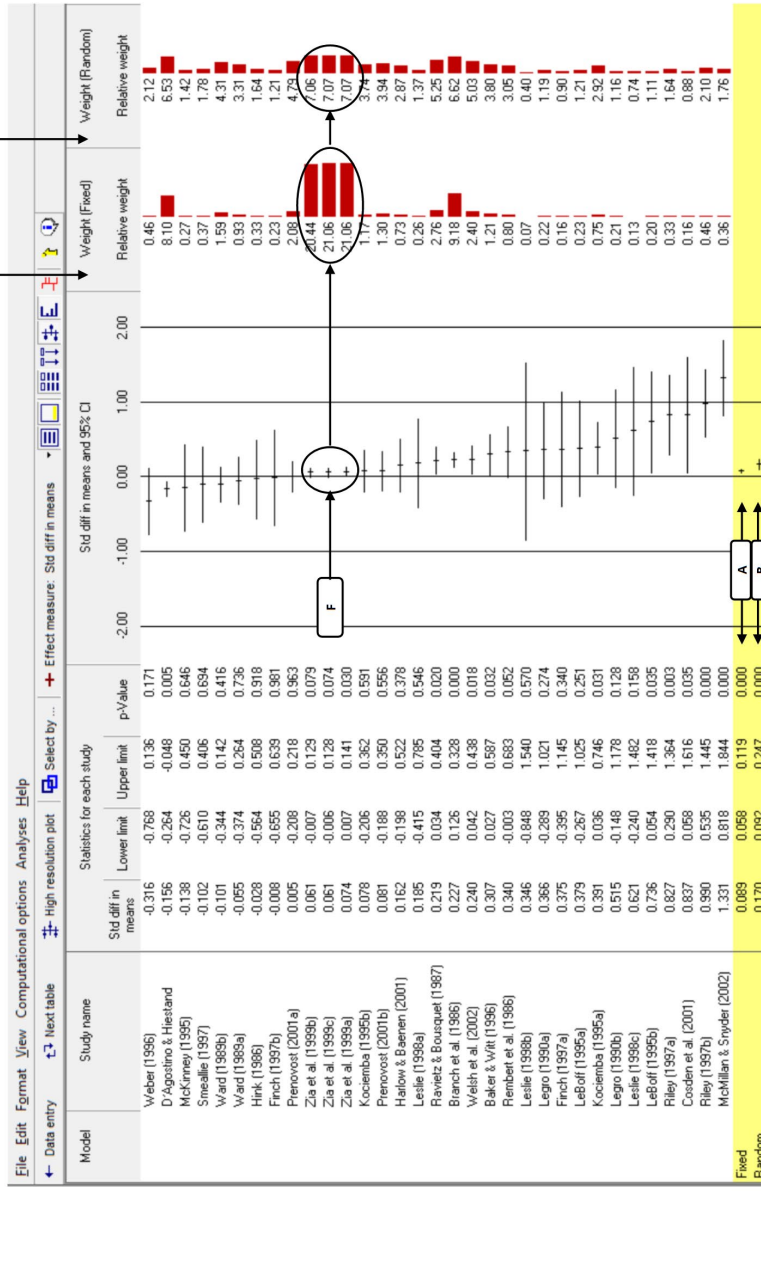

Figure 16 | Out of school programs | Standardized mean difference > 0 favors programs

When we shifted from random effects to fixed effects, a sequence of events was put into place. Prominent among these,

(1) The impact of small studies was minimized. As a group, the fifteen smaller studies were assigned some 18 percent of the total weight under random effects but less than 4 percent of the weight under the fixed-effect model.
(2) The impact of the three largest studies [F] was increased. As a group, these three studies were assigned 21 percent of the weight under random effects (column E) but roughly 63 percent of the weight under fixed effect (column D).

The intent of the sensitivity analysis had been to see how the results would change if we minimized the impact of the smaller studies. For that purpose, we intended to implement (1), but we inadvertently also implemented (2). That is, not only did we assign less weight to the small studies, but we also recalibrated the weights among the large studies so that three of these dominate the analysis. That had not been our intent.

When we shifted to the fixed-effect model, the confidence interval width decreased from 16 points as indicated by line [B], to 6 points as indicated by line [A]. The narrow width makes sense if our intent was to make an inference only to the studies in the analysis. It makes no sense if our intent was to make an inference to a larger universe of studies.

An additional problem is that if we adopt the fixed-effect model we cannot address dispersion in effects. As indicated by line [C], the effect size varies from −0.138 (a harmful effect) in some populations to +0.478 (a helpful effect) in others. This is a critically important part of the analysis. By contrast, if we switch to the fixed-effect model we need to assume that there is no dispersion in effects.

In sum, while our intent in switching to the fixed-effect model is to minimize the impact of small studies, we also re-weight the large studies, alter the width of the confidence interval, relinquish our ability to discuss dispersion in effects, and change the null hypothesis being addressed by the test of significance. It makes sense for all these items to shift in unison based on the intended inference, as reflected in the choice of a statistical model. If we want to make an inference to a wider universe, we want to implement *none* of these. If we want to make an inference to the studies in the analysis, we want to implement all of these. By contrast, if we want to assess the impact of the small studies, it makes more sense to isolate that issue and look at that issue alone. Some ways of doing so are discussed in section 12.1.

There may be cases where a researcher really does intend to ask two distinct questions – one about the studies in the analysis and another about the universe of comparable studies. In this case it would make sense to run a separate analysis for each question using the corresponding model. However, these cases are relatively rare. If we are pulling studies from the literature, we almost invariably intend to generalize to a wider universe, and so the random-effects model applies.

Summary

Some have recommended comparing the results for the fixed-effect model and the random-effects model as a kind of sensitivity analysis.

The problem with this approach is that when we change models, we change an entire array of elements, including the weights assigned to all studies, the confidence-interval width, the ability to address heterogeneity in effects, and the null hypothesis addressed by the significance test.

We take one approach to *all* these issues if we intend to limit our inference to the studies in the analysis, and adopt a fixed-effects model. We take another approach to *all* these issues if we intend to make an inference to the universe of comparable studies, and adopt a random-effects model. We cannot modify one of these elements without also modifying the others.

8.5. Random-effects model is more conservative

8.5.1. Mistake

It is common for researchers to say that the random-effects model is more conservative than the fixed-effect model, because the random-effects model is less likely to yield a statistically significant result. This belief stems from the fact that the random-effects model yields a larger standard error than the fixed-effect model. However, this statement is misleading.

8.5.2. Details

It is true that the random-effects model is less likely to yield a statistically significant result than the fixed-effect model, but it is misleading to suggest that this makes the analysis more conservative. To say that it is more conservative implies that both analyses are addressing the same question and giving different answers. In fact, the two analyses are addressing different questions.

The fixed-effect model allows us to estimate the mean effect size for the population included in the analysis. The random-effects model allows us to estimate the mean effect size in the universe of comparable populations. The standard error for this model is larger because it relates to the estimate of a *different* parameter. To say that the random-effects model is more conservative than the fixed-effect model is analogous to saying that we need more time to go from New York to California than we need to go from New York to Boston. This is (usually) true, but not relevant if our goal is to get to California.

8.5.3. Example | Statin use and bladder cancer

Zhang et al. (2013) used meta-analysis to synthesize data from eight studies that looked at the relationship between statin use and bladder cancer (Figure 17). They write "In the present meta-analysis, significant heterogeneity was observed among all studies …. Therefore, a random-effect (sic) model, which provides a more conservative standard error and a larger confidence interval, was chosen over a fixed-effect model to determine the pooled RR (risk ratio) estimates in our meta-analysis."

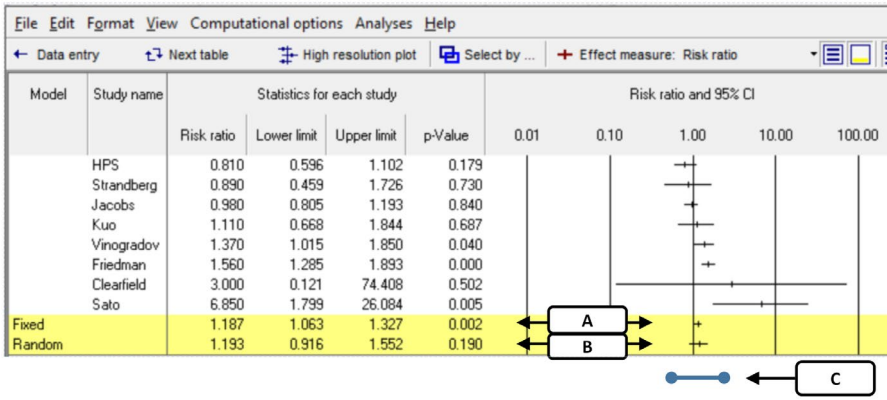

Figure 17 | Statin-use and bladder cancer | Risk ratio > 1 indicates increased risk

The authors were correct to choose the random-effects model, but their suggestion that this model is more conservative is misleading for several reasons.

The suggestion that the random-effects model is more conservative rests on the fact that the fixed-effect model [A] would allow us to reject the null hypothesis (p=0.002) whereas the random-effects model [B] would not allow us to reject the null hypothesis (p=0.190).

The idea that we can classify one model as being more conservative than another might make sense if both models were addressing the same question, but one did so with a larger error term. However, that is not the case here. The fixed-effect model tests the hypothesis that the intervention has no effect (on average) for the studies in the analysis, while the random-effects model tests the hypothesis that the intervention has no effect (on average) for the universe of comparable studies. Rather than thinking of one model as being more conservative than the other, we should think of one as being appropriate for testing one null hypothesis and the other as appropriate for testing a *different* null hypothesis.

Critically, the idea that we can think of a model as conservative based on whether or not it allows us to reject the null hypothesis assumes that the goal of the analysis is to test the null hypothesis. That may be one of the goals, but an equally (or more) important goal is to assess heterogeneity in effects. The random-effects model allows us to do so, whereas the fixed-effect model does not. In this example the prediction interval [C] is 0.553 to 2.572, suggesting that the use of statins may be associated with a 45% *reduced* risk in some populations, and a 257% *higher* risk in others.

8.5.4. Failure to reject the null hypothesis may not be conservative

The idea that wider confidence intervals are more conservative depends on the goals of the analysis. If a statistically significant effect would tell us that a treatment *is helpful,* it would be conservative to require a *higher* standard of proof (based on wider confidence intervals). However, if a statistically significant effect would tell us that a treatment is *harmful*, it would be conservative to require a *lesser* standard of proof (based on more narrow confidence intervals). In the current example a significant difference would tell us that the treatment could be associated with *increased* risk of cancer, and so the random-effects model could be seen as anti-conservative.

8.5.5. Random-effects model may not be conservative

The common wisdom that the random-effects model is more conservative hinges on the fact that this model will yield a wider confidence interval than the fixed-effect model (assuming T^2 is non-zero). Since the confidence interval is wider under random effects, for any given effect size the confidence interval is more likely to include the null effect (and we are less likely to reject the null hypothesis).

While the common wisdom is true *on average*, it does not apply in every case. The key lies in the phrase *for any given effect size* in the prior paragraph. If the effect size is the same under both models, then we are more likely to reject the null hypothesis under random effects. However, the effect size is generally *not* the same under both models. It may be substantially smaller or *greater* under the random-effects model. If the effect size happens to be substantially *greater* under the random-effects model, we might reject the null hypothesis under the random-effects model even when we do not reject it under the fixed-effect model.

The following example is a case in point.

8.5.6. Example | Water chlorination and cancer

Figure 18 shows the results of a meta-analysis as computed by Poole and Greenland (1999) based on data reported by Morris (1995). The studies in the analysis are epidemiologic studies on the relation of water chlorination to rectal cancer. A risk ratio greater than 1.0 indicates that water chlorination was associated with increased risk of cancer.

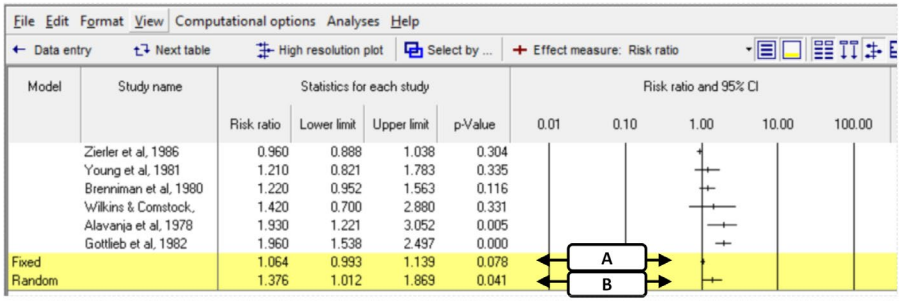

Figure 18 | Chlorination and Cancer | Risk ratio > 1 shows increased risk

For a fixed-effect analysis [A] the risk ratio is 1.064 with a confidence interval of 0.993 to 1.139. A test of the null hypothesis yields a *Z*-value of 1.761 and *p*-value of 0.078. For a random-effects analysis [B] the risk ratio is 1.376 with a confidence interval of 1.012 to 1.869. A test of the null hypothesis yields a *Z*-value of 2.039 and *p*-value of 0.041.

In this example the standard error of the effect size is substantially larger under random effects, which is what researchers have in mind when they assert that this model is more conservative. However, it is also possible (as was true in this example) that the effect size will be further from the null value under the random-effects model. Here, the random-effects results would lead us to reject the null hypothesis while the fixed-effect results would not.

Poole and Greenland (1999) cite this study to make the same point being made here. They also note that since these studies are observational, a causal relationship between chlorination and cancer should not be assumed.

Summary

The concept of conservative vs. anti-conservative focuses attention on the *mean* effect size. However, when the effect size varies, our attention should be focused also on the dispersion in effects. A key advantage of the random-effects model is that it allows us to take account of this dispersion.

As a separate issue, it is misleading to refer to the statistical models as being more conservative or less conservative. What really matters is which model yields the *correct* confidence interval for the intended inference. If all studies are based on the same population (or if we want to make an inference only to the studies included in the analysis), one confidence interval is correct. If we want to make an inference to the universe of comparable studies, a different confidence interval is correct.

8.6. Fixed-effect model has better statistical power

8.6.1. Mistake

Statistical power is the likelihood that an analysis will yield a statistically significant effect, allowing us to reject the null hypothesis of no effect. It is common for researchers to say that they have elected to use the fixed-effect model because it has better statistical power than the random-effects model. This is misleading.

8.6.2. Details

This is a variant of issue addressed in section 8.5. It is technically true that the fixed-effect model has better statistical power than the random-effects model, but this is irrelevant to the issue at hand. If we had two analyses that were estimating the *same* parameter and one did so with better precision (and better statistical power) than the other, we would want to use the more powerful test. However, that is not what is happening here.

If we use the fixed-effect model we are testing the null hypothesis that the common effect size for the one population included in the analysis is zero. Similarly, if we use the fixed-effects model we are testing the null hypothesis that the mean effect size for the specific studies included in the analysis is zero. By contrast, if we use the random-effects model we are testing the null hypothesis that the mean effect size in the *universe of comparable populations* is zero.

Typically, we want to address the null hypothesis about the universe of comparable studies, and the power for addressing that question is what matters. The fact that the power to address *another* question might be higher, is simply not relevant.

Summary

The fixed-effect model tests the null hypothesis that the effect size for the one population in the analysis is zero. The fixed-effects model tests the null hypothesis that the mean effect size for the specific studies in the analysis is zero. The random-effects model tests the null hypothesis that the mean effect size for the universe of comparable studies is zero. We need to decide which null hypothesis we intend to test, and then use the corresponding model.

8.7. When τ^2 is estimated as zero

8.7.1. Mistake

When the between-study variance (τ^2) is estimated as zero, the fixed-effect model and the random-effects model will yield identical estimates for all statistics. When this happens, researchers sometimes report that they are using the fixed-effect model. This is a mistake.

8.7.2. Details

In this discussion I will use τ^2 to refer to the true value of the between-study variance, and T^2 to refer to the estimate of this value. Both are pronounced tau squared.

Under the fixed-effect model, the weight assigned to each study is

$$W_i = \frac{1}{V_i}, \tag{1}$$

where W_i is the weight for study i and V_i is the error variance for study i. Under the random-effects model, the weight assigned to each study is

$$W_i = \frac{1}{V_i + T^2}, \tag{2}$$

where T^2 is the estimate of the between-study variance. It follows that when T^2 is zero, the two formulas will yield the same results.

When a researcher has adopted the random-effects model, but then discovers that T^2 is zero and the results are identical to the fixed-effect model, the researcher will sometimes assert that they have switched to the fixed-effect model. There are two ways of interpreting this statement, but in either case the idea that they had switched to the fixed-effect model is incorrect.

When the researcher says that they have switched to the fixed-effect model they may mean that their results match the results from the fixed-effect model. While it is true that the results of the random-effects analysis are the same as those from the fixed-effect analysis, this does not mean that they have changed to another model. They are still using the random-effects model. It just so happens that the fixed-effect model will yield the same result.

Alternatively, when the researcher says that they have switched to the fixed-effect model, they might mean that if τ^2 is zero, then (by definition) all

studies are estimating the same parameter, and so the fixed-effect model applies. This is a more sophisticated view, but also incorrect. If the between-study variance (the true value of τ^2) was actually zero, then this argument would be valid. But, when we are working with studies that assess the impact of an intervention and the studies are pulled from the literature, the *true* value of τ^2 is likely to be positive even if the *estimate* is zero. Therefore, the assumption that we are working with multiple populations remains intact.

8.7.3. What difference does it make?

If the numbers are the same under both models, what does it matter if we say we are using the fixed-effect model or the random-effects model? It matters because the model determines how we can generalize the results.

If we are using the random-effects model we can generalize the results to the universe of comparable studies, which had been (and should remain) our goal. When τ^2 is estimated as zero we would report the mean effect size, and would also report that the effect size is reasonably consistent across comparable populations. By contrast, if we (incorrectly) switched to the fixed-effect model, then the results would be limited to the one population included in the analysis. Then, we would not be able to generalize to a wider universe of populations, which had been the original goal.

The following examples serve to illustrate this idea.

8.7.4. Example | High-dose vs. standard-dose statins

Cannon, Steinberg, Murphy, Mega, and Braunwald (2006) used a meta-analysis to synthesize data from four studies that compared the impact of a high dose vs. a standard dose of statins in preventing cardiovascular events (Figure 19). The mean risk ratio of 0.849 tells us that the high dose was more effective than the standard dose in preventing the events. In this analysis, τ^2 was estimated as zero, as evidenced by the fact that the effect size and confidence interval are identical for both statistical models, [A] and [B].

The analysis employs the random-effects model, which allows us to generalize the results to comparable populations. The fact that τ^2 is estimated as zero tells us that the effect size may be reasonably consistent across these populations. It would be a mistake to suggest that we have somehow switched to the fixed-effect model. First, that would suggest that we are working with one population, which is not true. Second, it would prevent us from generalizing to comparable populations, and thus subvert the intent of the analysis.

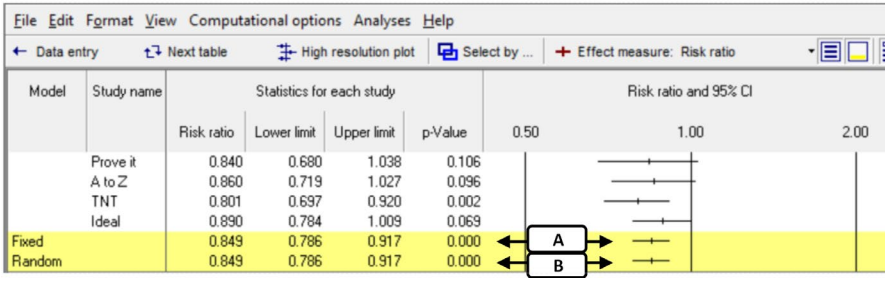

Figure 19 | High-dose vs. standard dose of statins | RR < 1 favors high dose

8.7.5. Example | Volunteer tutoring programs

Ritter, Barnett, Denny, and Albin (2009) looked at the effectiveness of volunteer tutoring programs (Figure 20). They write "The decision to use a fixed-effects model or random-effects model is based on the homogeneity analysis. The analyses of the overall effects and of the six key outcomes revealed *Q* statistics that were not large enough to allow us to reject the null hypothesis of homogeneity. That is, the variability across effect sizes did not exceed what would be expected based on sampling error (Lipsey & Wilson, 2001). Therefore, we employed a fixed-effects model for data synthesis in our study."

As always, the statistical model should reflect our goals for the systematic review. It is something that we establish as part of the protocol, not something that we look for in the data. Since the studies are based on multiple populations, and vary in other material ways, we can assume that the effect size varies across studies, and we should be using the random-effects model rather than the fixed-effect model. Critically, the random-effects model also allows us to generalize to comparable populations.

As it happens, in this analysis the *estimate* of the between-study variance (T^2) is zero. Therefore, the fixed-effect model and the random-effects model yield precisely the same numbers. However, the meaning of the numbers depends on the model.

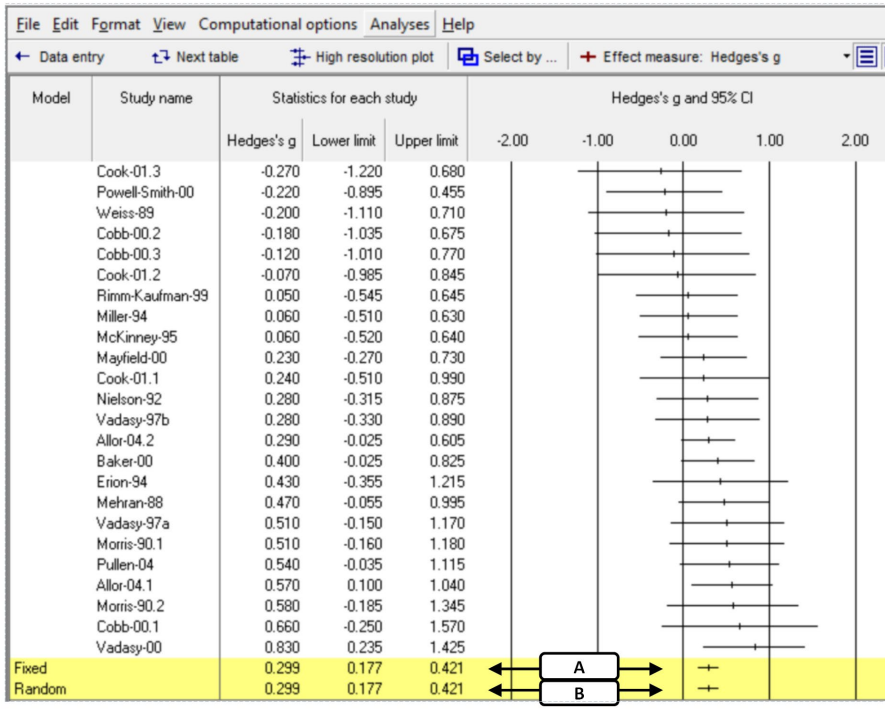

Figure 20 | Volunteer tutoring programs | Effect size > 0 favors programs

Under the fixed-effect model we would report that the mean effect size for this one population is 0.299 with a confidence interval of 0.177 to 0.421. This is a bit awkward since it is not clear what "this population" represents. And, we could not extrapolate from this population to any others.

Under the random-effects model we assume that these studies are representative of a universe of comparable studies. In that universe of studies, the mean effect size is 0.299 with a confidence interval of 0.177 to 0.421. We do not need to explain what this one population is, since we acknowledge that we are working with multiple populations. And we can extrapolate from these populations to the universe of comparable populations, which is almost invariably our intent.

Summary

The statistical model is not simply an issue of numbers. Rather, it reflects the intended inference, and provides a framework for understanding the results. If the studies in the analysis will be used to make an inference to a larger group of comparable studies, the random-effects model reflects this goal. The fact that a different model would have yielded the same numbers is irrelevant.

8.8. Switching models will have major impact on results

8.8.1. Mistake

Researchers sometimes assume that the decision to use one model or the other will have a major impact of the results. The reality is more nuanced.

8.8.2. Details

Researchers often prefer the fixed-effect model because it yields better power than the random-effects model (Rice et al., 2017). Their logic takes the form of "I would rather use the fixed-effect model because it has much better power than the random-effects model. It would be a shame to apply the random-effects model to address a small amount of between-study variance, since this will have a major impact on the test of the main effect. So, I will assert that the between-study variance is zero and stay with the fixed-effect model."

Implicit in this idea is the assumption that if we use random effects rather than fixed effect, there will be a substantial change in the results. Therefore, it may be useful to point out that the switch from fixed to random does not automatically result in a major loss of power or substantially wider confidence intervals. Rather, the difference depends on the between-study variance and the number of studies. In cases where the between-study variance is estimated as zero, the results of the random-effects analysis will be identical to those of the fixed-effect model, and so there is no loss of statistical power. When the variance is estimated as relatively trivial, the confidence interval under the random-effects model will expand slightly, and the likelihood of rejecting the null hypothesis will change only slightly.

It is only as the between-study variance becomes substantially larger that the confidence interval expands substantially, and the likelihood of rejecting the null hypothesis changes markedly. However, in these cases the researcher should recognize that using the fixed-effect model would be a serious mistake, and so the idea of staying with this model is especially problematic.

An important exception to this rule is when there are only a few studies and the analysis includes the Knapp-Hartung-Sidik-Jonkman adjustment (see section 7.5). This will substantially increase the width of the confidence interval and substantially decrease the likelihood of the rejecting the null hypothesis of no effect.

To be clear, I am not suggesting that it would ever be acceptable to use the fixed-effect model because it yields better power. As discussed in section 8.6, this should not be a consideration when choosing a model. Rather, in recognition of the fact that some researchers do make that argument, my goal here is point out that the use of the random-effects model may not have as much of an impact as they fear.

Summary

Researchers are sometimes reluctant to use the random-effects model because that are concerned that it will substantially decrease the power for testing the null hypothesis. This is not necessarily the case, since the impact of the model on the standard error depends on a number of factors.

8.9. Meta-analyses with large *N* will have good power

8.9.1. Mistake

There is a common belief that all meta-analyses have excellent statistical power to test the null hypothesis of no effect. The reality is more complicated.

8.9.2. Details

Statistical power is the likelihood that a study will yield a statistically significant result. For example, we might say that "If the treatment boosts the mean score by 10 points, the study has power of 90% to reject the null hypothesis". Power is a function of the size of the effect, the criterion alpha, and the precision of the estimate.

In the case where we use the fixed-effect model or the fixed-effects model, the precision of the estimate is given by

$$SE_M = \sqrt{\frac{V}{N}}, \tag{3}$$

where V is the within-study population variance and N is the sample size accumulated across studies. In these cases, power *will* tend to be high. However, these are generally not the statistical models we should be using. By contrast, in a case where we use the random-effects model, the precision of the estimate is given by

$$SE_M = \sqrt{\frac{V}{N} + \frac{T^2}{k}}, \tag{4}$$

where T^2 is the estimate of the between-study variance and k is the number of studies.

The mistake that researchers make is to assume that increasing N will increase the precision of the estimate. However, it should be clear from this equation that this assumption is incorrect. There are two terms under the radical, and *they operate independently of each other*. Increasing the N will affect the first term, but have no impact on the second. If T^2 is non-zero, the only way to reduce this component of the error term is to increase the number of studies. If the number of studies is low and T^2 is non-trivial, then power can remain low, no matter how large N might be.

The formulas presented here are conceptual only since they require that V is the same for all studies. For more information see Appendix II.

Summary

Under the random-effects model, statistical power is usually driven by the number of *studies*, not the number of *people*. When the between-study variance is substantial, the only way to get good power will usually be to include a large number of studies.

8.10. Putting it all together

The choice of a statistical model should be based on the issues discussed in section 7. However, since researchers sometimes make a decision based on other issues, I addressed these issues here.

Researchers sometimes suggest that it would be preferable to use the fixed-effect model since this assigns less weight to small studies, and these studies are assumed to be of poor quality. This is generally a bad idea for two reasons. First, we should not assume that poor studies are of poor quality. Second, when we change the weights, we change the estimate not only of the mean, but also of the confidence interval, *p*-value, and other statistics, in ways that we may not anticipate.

Researchers sometimes perform an analysis using both models to see if the results shift. It is not always clear what purpose this serves. The fixed-effects model yields information about the specific studies included in the analysis, while the random-effects model yields information about the universe of comparable studies. If we care about one question, it is not clear why we would want to know the answer to the other.

Some researchers suggest that they prefer to use the fixed-effects model because it has better statistical power than the random-effects model. There is a reason that the fixed-effects model has better power – that is because it is estimating the mean for the studies in the analysis, and not for the universe of comparable studies. The idea that we can take a model which addresses a different question and apply it to our question is simply incorrect.

The common theme in all these issues is that they are based on a misunderstanding of the role of the model. If we intend to make an inference about the one population included in the analysis, we adopt a set of weights which make sense for this goal. If we intend to make an inference about the specific studies included in the analysis but not generalize to any other studies, we adopt a set of weights which make sense for this goal. If we intend to make an inference about the universe of studies which are comparable to those in the analysis, we adopt a set of weights which make sense for this goal. This is all captured by the model, and affects the estimate of the mean, the width of the confidence interval, the null hypothesis being tested, and what we can say about dispersion in effects. In all cases we need to know the estimates *for our intended inference*. The fact that the estimates for another inference are different, is not relevant.