9. HETEROGENEITY

9.1. Overview

9.1.1. What do we mean by heterogeneity?

The term "heterogeneity" refers to the dispersion of true effects across studies. Typically, the studies in a meta-analysis will differ from each other in various ways. Each study is based on a unique population, and the impact of any intervention will typically be larger in some populations and smaller in others. The specifics of the intervention may vary from study to study, the scale used to assess outcome may vary from study to study, and so on. Each of these factors may have an impact on the effect size. One goal of the analysis will be to determine how much the effect size varies across studies, and this is variation is called heterogeneity (Ades, Lu, & Higgins, 2005; P. Glasziou & Sanders, 2002; J. Higgins, Thompson, Deeks, & Altman, 2002; J. P. Higgins et al., 2009; Keefe & Strom, 2009; Thompson, 1994).

9.1.2. Heterogeneity in a primary study

The basic idea of heterogeneity in a meta-analysis is similar to that in a primary study. Consider a primary study to assess the distribution of math scores in a high-school class. Suppose that the *mean* score across all students in the class is 50. To understand how the students are performing we also need to ask about heterogeneity, and we typically do so by reporting the standard deviation of scores. We understand that 95% of all students will score within two standard deviations of the mean. Therefore –

A. If the standard deviation is 5 points, most students will score between 40 and 60.
B. If the standard deviation is 10 points, most students will score between 30 and 70.
C. If the standard deviation is 20 points, most students will score between 10 and 90.

These intervals are called prediction intervals. If someone asked us to predict the score for a student in the class (selected at random from the class), in case A we would predict the student would score in the range of 40 to 60, and we would be correct some 95% of the time. The same idea applies to cases B and C.

When we perform a primary study, we compute several other statistics related to heterogeneity, such as the sum of squares and the variance. These are all important statistics, but if we want to know how much the scores vary, these statistics are tangential, at best. The only statistics that directly address this question are the standard deviation and prediction interval.

9.1.3. Heterogeneity in a meta-analysis

The same ideas apply when we turn to meta-analysis. For example, consider the following.

Castells et al. (2011) conducted a meta-analysis of seventeen studies to assess the impact of methylphenidate in 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 measure of cognitive function.

In this context –

- A standardized mean difference of 0.20 would represent a trivial effect size. While this difference would be captured by the test, it is so small that the patient might not be aware of any change.
- A standardized mean difference of 0.50 would represent a moderate effect size. The patient would be aware of a clinically important change, and some co-workers might notice the change as well.
- A standardized mean difference of 0.80 would represent a large effect size. The patient would be pleasantly surprised by the improvement, and some co-workers would be likely to remark that something was different.

It turns out that the mean effect size is 0.50. *On average*, across all comparable populations, the drug increases cognitive functioning by one-half a standard deviation. But to understand the potential utility of the drug we also need to ask about heterogeneity.

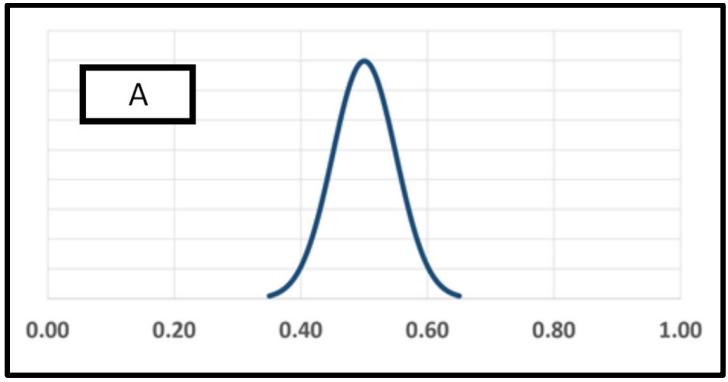

Figure 21 | Effect size varies from 0.40 to 0.60

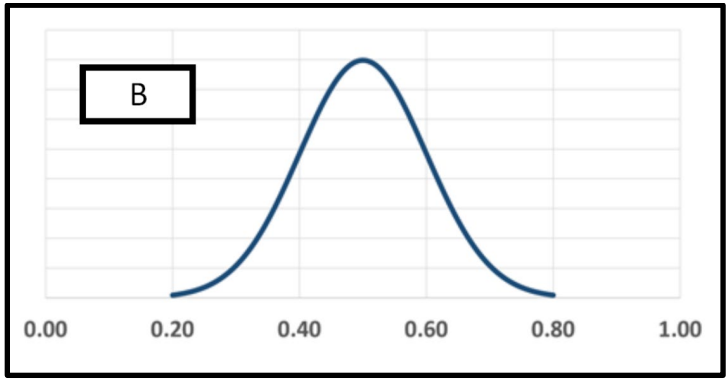

Figure 22 | Effect size varies from 0.30 to 0.70

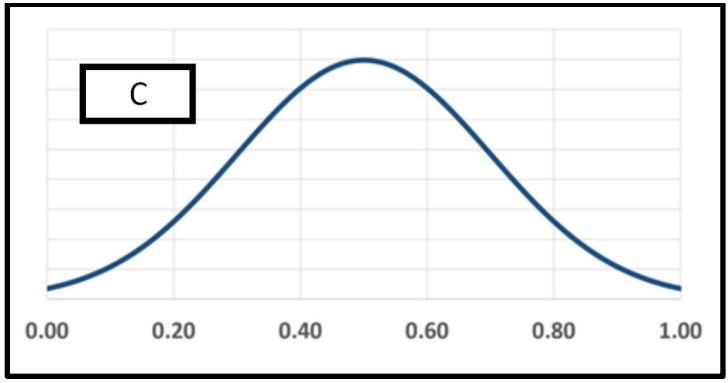

Figure 23 | Effect size varies from 0.10 to 0.90

Consider three possible results for the meta-analysis, listed here as A, B, and C. In all cases the *mean* impact is 0.50, but the consistency of the impact varies.

A. The impact is as low as 0.40 in some populations, and as high as 0.60 in others (Figure 21).
B. The impact is as low as 0.30 in some populations, and as high as 0.70 in others (Figure 22).
C. The impact is as low as 0.10 in some populations, and as high as 0.90 in others (Figure 23).

We might make the following decisions about the utility of the drug in the three cases.

A. We can expect to see pretty much the same effect in all populations.
B. The impact will vary somewhat across populations, but from a clinical perspective we can still talk about a *common* effect size.
C. The impact varies substantially across populations. It would be important to establish where the impact is trivial, moderate, and high, so that we can target this intervention more effectively. However, since the impact is always positive, we could use this intervention immediately.

These judgments are subjective. For example, we can discuss whether to recommend the intervention in case C, where the effect will be trivial in some populations. What *is* clear though, is that when we discuss the potential utility of the drug, it should be based on this type of information.

9.1.4. The sources of confusion

While basic idea of heterogeneity is the same in a meta-analysis and a primary study, there are a few technical details that differ between the two.

In a primary study (when we have one score for each subject) we typically treat the *observed* score for each subject as being the same as the *true* score for that subject. If a student scores 40 on the test, we treat 40 as being that student's true score. We compute the variance, standard deviation and prediction interval for the observed scores, and these serve also as the values for the true scores as well.

By contrast, in the case of a meta-analysis we make a distinction between the *observed* effect size and the *true* effect size for each study. The *observed* effect size is the effect size that we see in the sample. The *true* effect size is

the effect size that we *would see* if we could somehow enroll the entire population in the study. The observed effect size serves as an estimate of the true effect size but invariably falls below or above the true effect size due to sampling error.

The variance of observed effects tends to be larger than the variance of true effects. To understand why, consider what would happen if we ran five studies based on the same population, and computed the effect size in each. The *true* effect size is the same in all five studies (all studies are estimating the effect size in the same population) and so the variance of *true* effects is zero. Yet, the observed effects will differ from each other because of sampling error, and so the variance of the *observed* effects will be greater than zero. While this is most intuitive in the case when the variance of true effects is zero, it applies also when the true effects vary. The variance of observed effects tends to exceed the variance of true effects.

The ADHD analysis serves as a case in point. Figure 24 shows two plots. The inner plot shows the dispersion of *true* effects, while the outer plot shows the dispersion of *observed* effects. We *see* the outer plot, but we care about the inner plot since the inner plot tells us how much the effect size really varies across populations.

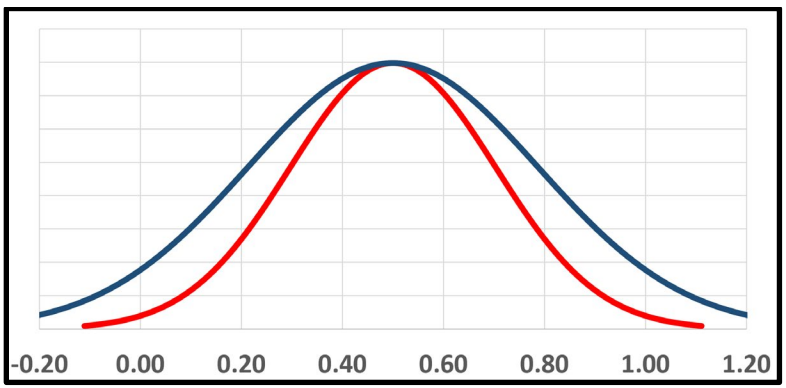

Figure 24 | Dispersion of observed effects (outer) and true effects (inner)

The heterogeneity statistics typically reported for a meta-analysis include the Q-value, a p-value, I-squared (I^2), Tau-squared (T^2), and Tau (T). The definition of each, and the relationships among them are presented in Appendix VI. The point I need to make here is that many of the statistics that are typically reported are tangential to the one issue we really care about, which is *How much does the effect size vary*. We need to be clear about what

each statistic means, and then focus on the ones that are relevant to this question.

On the pages that follow, I address various issues including the following

- Researchers sometimes assume that heterogeneity diminishes the utility of the analysis. The reality is more complicated.
- The one statistic that offers an unambiguous report of the dispersion is the prediction interval. Researchers rarely report this interval, and sometimes confuse it with the confidence interval.
- Researchers often treat the I^2 statistic as being synonymous with heterogeneity. In some cases, the I^2 statistic is used to classify heterogeneity as being low, moderate, or high. In fact, the I^2 statistic does not tell us how much the effect size varies, and the idea of classifying heterogeneity into these categories without additional context is meaningless.
- Researchers sometimes use the Q statistic or the p-value for a test of heterogeneity as indices of heterogeneity. This is a mistake.

9.2. Heterogeneity is *bad*

9.2.1. Mistake

Heterogeneity refers to the fact that the true effect size varies across studies. Some researchers believe that heterogeneity diminishes the utility of the analysis. In an extreme version of this idea, some have asserted that when the effect sizes are heterogeneous, it is a bad idea to perform a meta-analysis at all. The truth is more complicated.

9.2.2. Details

Heterogeneity is not inherently good or bad, but it does affect what we can learn from the analysis. If our goal in the analysis is to report that the intervention increases scores by a certain value, then heterogeneity is indeed a problem. In the absence of heterogeneity, we can report a common effect size that applies to all populations. In the presence of heterogeneity, there is no common effect size and so we cannot meet this goal.

However, in the presence of heterogeneity we can assess the extent of heterogeneity and report, for example, that the effect size is as low as 0.05 in some populations and as high as 0.95 in others. If this is the true state of affairs, then this should be the goal of the analysis.

9.2.3. Heterogeneity affects what we can learn from the analysis

If the between-study heterogeneity is trivial, then the meta-analysis may provide definitive information about the utility of the intervention for all comparable populations.

For example, Cannon et al. (2006) conducted a meta-analysis of studies that compared a high-dose of statins vs. a standard dose for prevention of cardiovascular events (Figure 25). The mean risk ratio was 0.849 (patients assigned to a high dose were 15% less likely to have an event), and this effect size was essentially the same for all studies. On this basis, the mean effect size is a useful indicator of the effect size for all comparable populations.

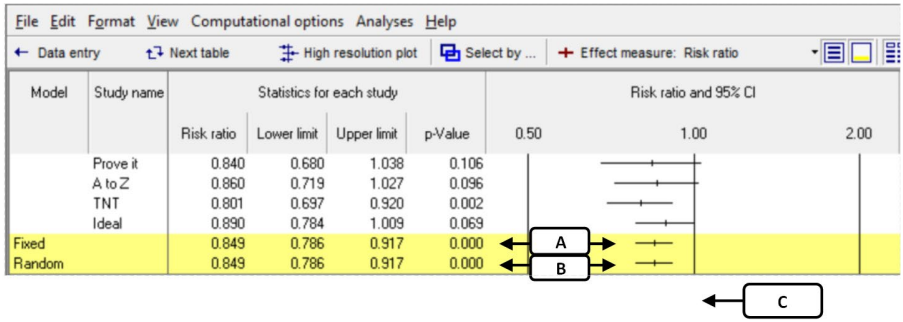

Figure 25 | High dose vs. standard dose of statins | Risk ratio < 1 favors high dose

By contrast, if the between-study heterogeneity is substantial, the meta-analysis will not be able to provide definitive information about the utility of the intervention in any given population, but it may be able to provide important information about the variation in effect size.

For example, Castells et al. (2011) conducted a meta-analysis of studies that assessed the impact of methylphenidate vs. placebo on the cognitive functioning of adults with attention deficit hyperactivity disorder (ADHD). The mean effect size was a standardized mean difference of roughly 0.50, but the effect size varied substantially across studies (Figure 28). As indicated by line [C], there were some populations where the effect size was 0.05 (which would represent a trivial effect in this context), some where it was near 0.50 (a moderate effect) and some where it was 0.95 (a very large clinical effect). In this case, the mean is not a useful indicator of the effect size we can expect to see in any given population, since the effect size in most populations falls some distance from the mean. Rather, the take-home message from this analysis might be that the treatment effect varies substantially. Therefore, we need to identify factors associated with this variation.

In this context, it would be important to clarify two related issues.

First, the suggestion that we can speak of heterogeneity as being *present* or *absent* is a misnomer, since it implies that some sets of studies are heterogeneous while others are not. In a systematic review based on studies that are pulled from the literature, especially when the studies assess the impact of an intervention, the true effect size will almost always be larger in some cases than in others. So, when we ask about the impact of heterogeneity, we are not asking about *zero* heterogeneity vs. *some* heterogeneity. Rather, we are asking about *trivial* heterogeneity vs. *substantive* heterogeneity.

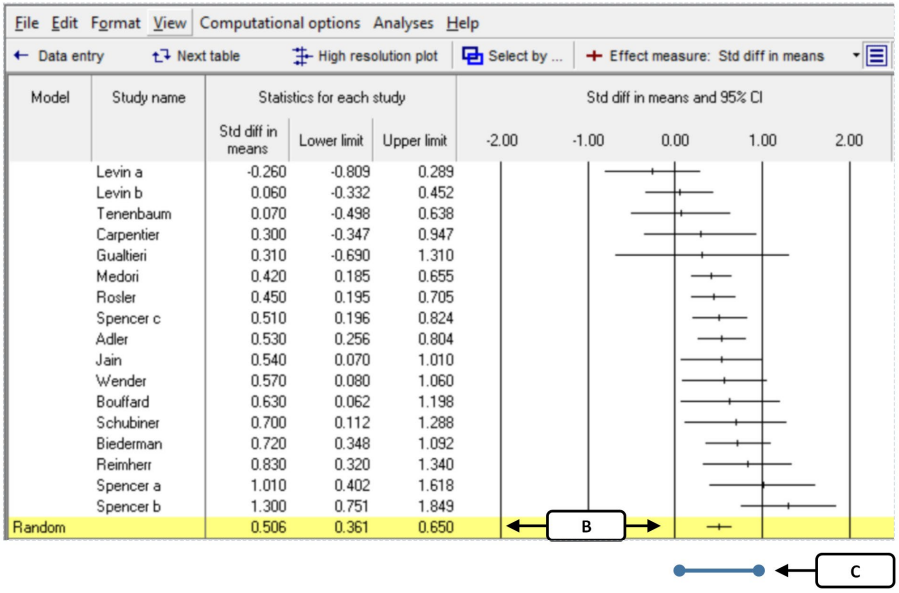

Figure 26 | Methylphenidate for adults with ADHD | Effect size > 0 favors treatment

Second, I said that when heterogeneity is trivial, the mean effect size provides definitive information about all comparable studies. This statement comes with some important caveats.

A. This refers to the *true* heterogeneity, not the *estimated* heterogeneity. The fact that heterogeneity is *estimated* as being trivial (or zero) does not necessarily mean that the true heterogeneity *is* trivial.
B. The description of heterogeneity as being *trivial* or *substantive* refers to the practical impact of the intervention rather than some statistical index. The researcher (or reader) would need to decide what amount of dispersion is of practical importance.
C. The statement that the mean effect size applies to all *comparable* studies is more useful in theory than in practice. In practice, it may not be clear what studies are comparable to those in the analysis.

9.2.4. The good folks of New Cuyama

At a conference in London to mark the 30th anniversary of the paper by DerSimonian and Laird which introduced their method for estimating heterogeneity, Dr. Laird was asked what she considered to be "too much" heterogeneity. She responded by showing the photo in Figure 27.

The good folks in the town of New Cuyama erected a sign that captured some key statistics. The population is 562, the town is 2150 feet above sea level, and the town was established in the year 1951. They summed these statistics and report the total is 4663.

Figure 27 | An example of "Too much heterogeneity"

Dr. Laird said that this would be an example where people had gone too far. But in most cases, heterogeneity is not a problem if we treat it appropriately.

Summary

The suggestion that we should not perform a meta-analysis in the presence of heterogeneity is based on the false premise that the goal of an analysis is always to estimate the *mean* effect size. In fact, the goal of an analysis is to estimate the *pattern* of effects. If the effect size is reasonably consistent across studies, we can report that the effect size is consistent and then focus on the mean. If the effect size varies across studies, we can discuss the extent of variation and what this says about the utility of the intervention. We might also try to explain some of the variation.

9.3. The prediction interval

9.3.1. Mistake

The prediction interval addresses the question *we intend to ask* when we ask about heterogeneity. It tells us how the true effect size varies across populations, and it does so on a scale that allows us to address the utility of the intervention. The mistake that researchers make is that they neglect to report this interval.

9.3.2. Details

The following examples show how the prediction interval addresses the issue of heterogeneity in a concise and intuitive format.

9.3.3. Example | Effect of methylphenidate on cognitive function in adults with ADHD

Castells et al. (2011) looked at 17 studies that evaluated the effect of methylphenidate on cognitive function in adults with ADHD (Figure 28). The effect size is the standardized mean difference (*d*). For purposes of this discussion I will assume that an effect size of 0.20 is small (it would show up on a test but the patient might not notice the change), an effect size of 0.50 is moderate (the patient would recognize that something was different), and that an effect size of 0.80 is large (colleagues would recognize the change).

The mean effect size is roughly 0.50 with a confidence interval [B] of 0.35 to 0.65. The confidence interval is an index of precision, and tells us how precisely we have estimated the mean effect size. Here, the entire confidence interval falls within the "moderate" range (as defined above), so we can report that the *mean* effect size is moderate.

The prediction interval [C] is roughly 0.05 to 0.95. The prediction interval is an index of dispersion, and tells us how widely the true effect size varies. Here, we would expect that in some 95% of all populations, the true effect size will fall in the range of 0.05 to 0.95. Using the categories outlined above, the effect size would fall between trivial and moderate in half the cases, and between moderate and large in the other half. Of note, there are no populations where the impact would be harmful. (Note that the terms moderate and large here refer to the clinical impact of the treatment and not to the extent of dispersion.)

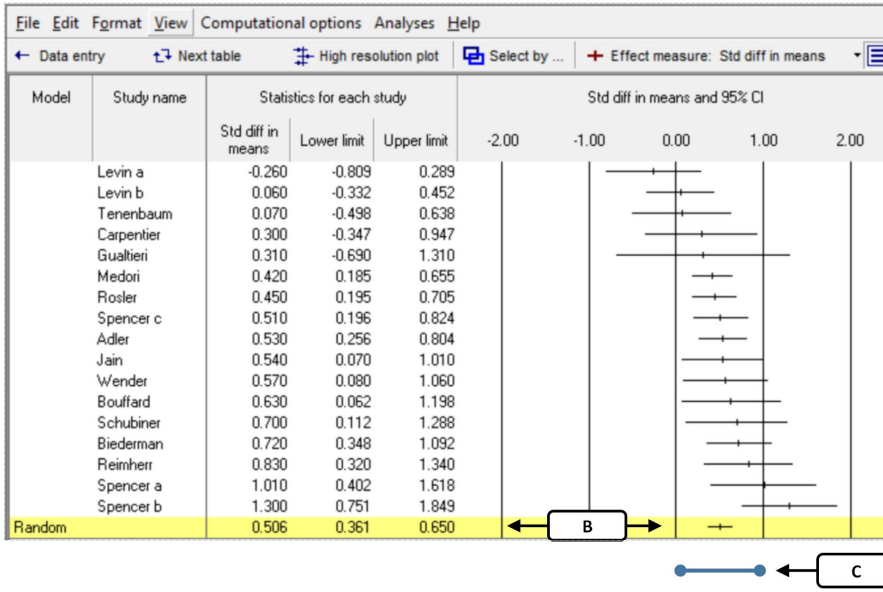

Figure 28 | Methylphenidate for adults with ADHD | Effect size > 0 favors treatment

The prediction interval allows us to address the questions that we typically have in mind when we ask about heterogeneity (Borenstein, Higgins, Hedges, & Rothstein, 2017; IntHout, Ioannidis, Rovers, & Goeman, 2016). To wit −

- Researchers typically report statistics such as Q, I^2, and T^2, but none of these tells us how much the effect size varies. Here, Q is 30.106 with 16 degrees of freedom, I^2 is 47%, and T^2 is 0.039. Based on this information, few readers would have any sense of the dispersion in effects.
- By contrast, the prediction interval reports the extent of the dispersion in the same units as the effect size. The effect size varies over roughly 90 points (in *d* units) and we understand what that means.
- Additionally, the prediction interval reports the dispersion using absolute values. It tells us not only that the effects vary over roughly 90 points, but also that the specific range of values is 0.05 to 0.95 (rather than −0.45 to +0.45, for example). The treatment is *very* helpful in some cases and *minimally* helpful in others, but there are no populations within the prediction interval where the treatment is likely to be harmful.

Based on this interval we might decide that −

- In the absence of further information, it would be reasonable to use the drug for all comparable populations.
- We should pursue additional research to identify the factors that are related to the impact of the drug. If it turns out that the drug is more effective in some populations than others, we would want to target those populations. If it turns out that the drug is more effective in certain doses than in others, we might be able to use the drug more effectively.

These types of decisions are subjective, but it should be clear that a meaningful discussion about the potential utility of the treatment would be based on the information contained in the prediction interval. By contrast, if we had simply reported Q, T^2 or I^2, the extent of dispersion would not be known, and it would not be possible to have this discussion (see section 9.5).

9.3.4. 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 29). The numbers that follow are based on our re-analysis of the data, and differ slightly from the original report due to rounding error.

The effect size index is the raw difference in mean blood pressure, with values below zero indicating a beneficial effect. The mean effect size is −0.473, with a confidence interval of −1.195 to +0.248 [B]. The confidence interval is an index of precision, and tells us how precisely we have estimated the mean effect size. Here, the confidence interval includes zero, so we cannot reject the null hypothesis that the *mean* effect size is zero.

The prediction interval [C] is roughly −4.08 to +3.13. The prediction interval is an index of dispersion, and tells us how widely the true effect size varies. When the effects vary this widely, the *mean* is largely irrelevant. This is especially true if the intervention is helpful in some cases and harmful in others. The take-home message here would be that we need to understand where the treatment is helpful, and where it is harmful.

Critically, *only* the prediction interval allows us to address the questions that we typically have in mind when we ask about heterogeneity. That is −

- The Q-value is 4084.467 with 26 degrees of freedom, I^2 is 99.363%, and T^2 is 2.933. None of these gives us any sense of the actual dispersion.

- The prediction interval reports the extent of the dispersion in the same units as the effect size (mmHg), and we understand what a range of 7 points means on this scale.
- The prediction interval reports the dispersion using absolute values. It tells us not only that the effects vary over roughly 7 mmHg, but line [C] shows that the treatment helpful (less than zero) in roughly 60% of populations and harmful (greater than zero) in the other 40%.

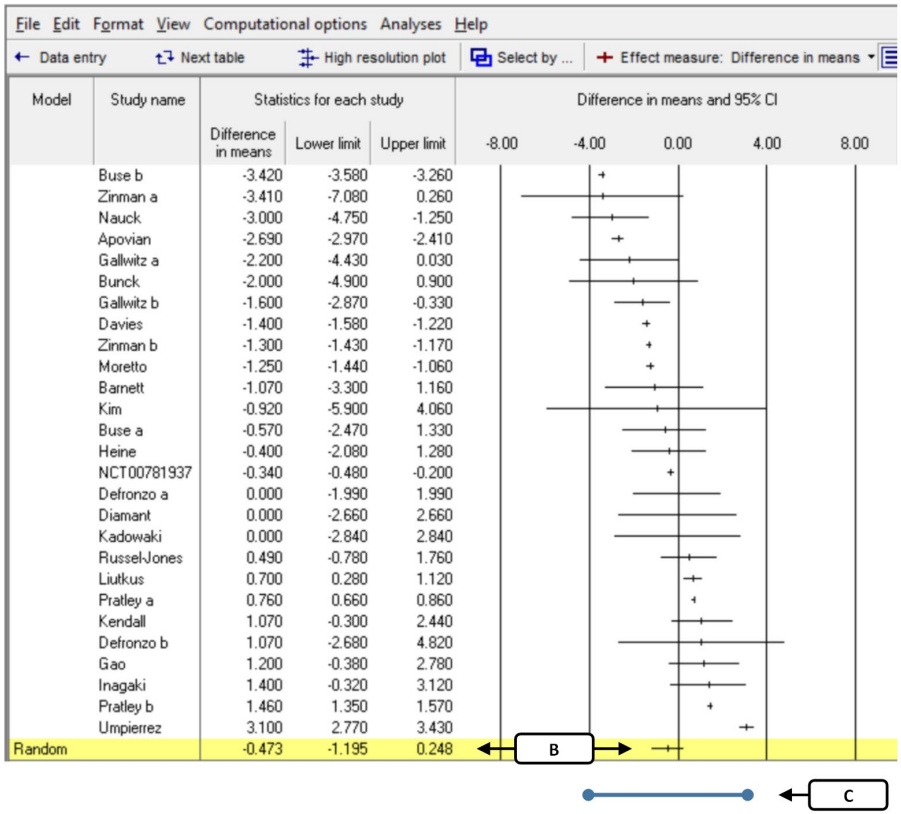

Figure 29 | GLP-1 mimetics and diastolic BP | Mean difference < 0 favors treatment

Based on this interval we might decide that this treatment is potentially useful in some cases, but we need to determine where it will be helpful and where it will be harmful. For example, it may be helpful in specific types of patients, or in specific variants of the intervention.

When we present the prediction interval, the actual extent of dispersion is clear, and we can discuss the clinical implications of this dispersion. By contrast, if we had simply reported T^2 or I^2, the extent of dispersion would not

be known, and it would not be possible to have this discussion (see section 9.5).

9.3.5. When τ^2 is estimated as zero

The prediction interval speaks to the dispersion in effects, and for that reason only applies when the estimate of the variance (T^2) is greater than zero. When the estimate of T^2 is zero, we generally would report the mean and confidence interval, but not the prediction interval.

9.3.6. Example | High dose vs. standard dose of statins

For example, Cannon et al. (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 30). The mean risk ratio of 0.849 tells us that the high dose was more effective than the standard dose in preventing the events.

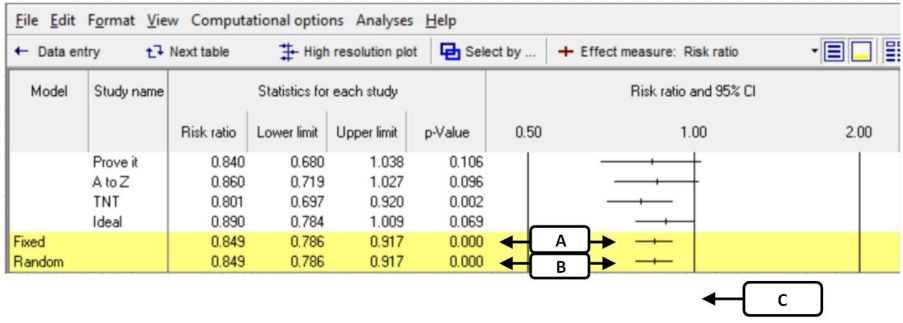

Figure 30 | High dose vs. standard dose of statins | Risk ratio < 1 favors high dose

In this analysis, τ^2, the variance of true effects, was estimated as zero. When τ^2 is estimated as zero we can generally assume that this is an underestimate and the actual value of τ^2 is positive. Nevertheless, we assume that the true variance is trivial, and proceed accordingly. Here we would report that the mean effect size in the universe of comparable populations falls in the interval 0.786 to 0.917, and that there is no evidence that the effect size varies across studies.

As always, the confidence interval is an index of precision, not an index of dispersion. The fact that the confidence interval is 0.786 to 0.917 does not tell us that the effect size varies from 0.786 in some populations to 0.917 in

others. Rather, we assume that the true effect size is roughly the same in all populations. This *common* effect size is assumed to fall somewhere in this range. Since we assume that the effect size is roughly the same for all populations, we omit the prediction interval [C].

9.3.7. Computing prediction intervals

I describe the prediction interval by reporting (for example) that the effect size ranges from 0.05 in some populations to 0.95 in others. To be clear, this is not simply a report of the lowest and highest effects. Rather, the basic approach to computing prediction intervals is to use the mean plus or minus two standard deviations, which is the same approach we would take in a primary study. However, there are some technical issues that we need to address. For *all* effect-size indices we need to expand the intervals to take account of the fact that the mean and standard deviation are estimated with error. For *some* effect-size indices we need to transform the values into another metric before computing the intervals.

In Appendix VII, I present the formulas for computing prediction intervals that address both issues. As a practical matter, it is much simpler to use a spreadsheet for the computations. This spreadsheet may be downloaded on the book's web site. This spreadsheet may be used as an adjunct to any computer program, since it requires the user to enter only four items (the number of studies, the mean effect size, the upper limit of the confidence interval, and T^2).

9.3.8. Some caveats regarding the prediction interval

All the analyses we perform as part of a meta-analysis (or any analysis, for that matter) require that some assumptions be met. If these assumptions are violated, the results may not be reliable. In the case of prediction intervals, we need to keep the following in mind.

The interval will be reasonably accurate if it is based on enough data. The minimum number of studies needed to compute a useful prediction interval would depend on the extent of heterogeneity, but would likely be at least ten in many cases (Hedges & Vevea, 1998). It would be reasonable to have more faith in the accuracy of the interval as the number of studies increases.

When computing the prediction interval, we typically assume that the effects are normally distributed. However, in practice this will not always be the case. For example, (Hackshaw, Law, & Wald, 1997) looked at the

relationship between second-hand smoke and lung cancer. On average, exposure to second-hand smoke is associated with an increased risk in lung cancer, but if we compute a prediction interval and assume that the distribution of true effects is normally distributed (in log units), we would conclude that in some small minority of cases exposure is associated with a decreased risk of lung cancer. Here, it makes more sense to assume that the distribution is truncated at a risk ratio of 1.0.

Importantly, the prediction interval applies to the universe from which the studies were drawn, and this may not be the same as the universe that we had in mind when we planned the systematic review (IntHout et al., 2016). Both the mean and the standard deviation of effects will depend on the specific mix of populations reflected in the included studies, and so will the prediction interval which is based on these statistics (see section 7.4).

The spreadsheet cited above expands the interval to take account of the imprecision of the estimate, and make it more likely that the interval covers some 95% of all populations. Since the goal of this approach is to ensure that most populations are included under the interval, it always errs on the side of expanding (rather than narrowing) the interval. As such, it may exaggerate the true extent of the dispersion.

9.3.9. The prediction interval is only a first step

The prediction interval allows us to *quantify* the extent of dispersion, but is not intended to *explain* that dispersion. When the prediction interval tells us that the impact of treatment varies substantially, we know that we need more information to use the intervention effectively. In the ADHD analysis, we need to know *where* the drug's impact is trivial and *where* it is substantial. In the GLP-1 example, we need to know *where* the treatment is helpful and *where* it is harmful. If we have enough studies in the meta-analysis, we might be able to use subgroup analysis or meta-regression to see which factors are associated with the effect size, and develop hypotheses to be tested in future research.

9.3.10. The normal curve

There is no convention for how to display the prediction interval on a plot. In this book I generally superimpose a line under the forest plot. For example, in Figure 28 the prediction interval for the ADHD analysis is displayed as a line [C] that extends from 0.05 to 0.95.

However, we also have the option of constructing a normal curve for the prediction interval, as in Figure 31, which is also based on the ADHD analysis. In this figure line [C] denotes the part of the curve which captures the effect size in some 95% of all populations. The sections of the plot to the left and right of line [C] correspond to the 5% of effects that fall outside the 95% prediction interval. Line [C] in Figure 31 is the same as line [C] in Figure 28. However, Figure 31 highlights the fact that most populations will have an effect size toward the center of the curve, with relatively few near the extremes.

The web site includes an Excel spreadsheet that can be used to create this plot. To use the plot, the user needs to enter only the mean effect size, the upper limit of the confidence interval, Tau-squared, and the number of studies. Since all programs report these values, the spreadsheet can be used as an adjunct to any software for meta-analysis.

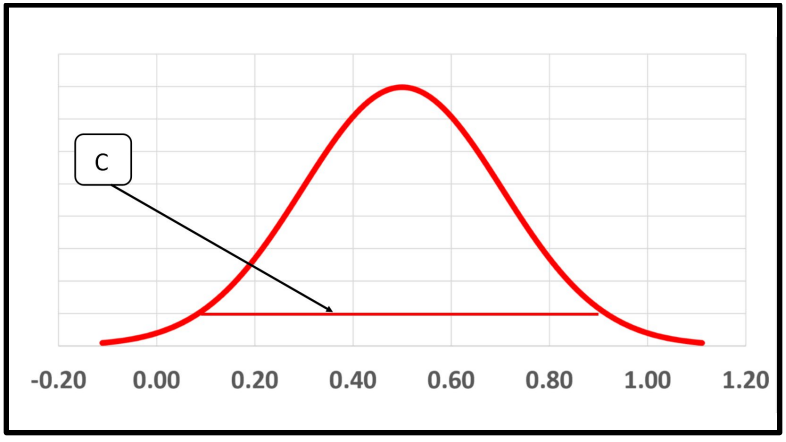

Figure 31 | Distribution of true effects and prediction interval

9.3.11. Reliability of the prediction interval

As noted above, the prediction interval will not be reliable when based on a small number of studies. To be clear, the problem of trying to estimate the prediction interval with too few studies applies also to the other indices, including T^2, T, and I^2. So, if we are concerned that we do not have enough studies, switching to one of those indices is not a useful option. Ironically, the poor precision for T^2 and I^2 has few practical problems because people do not actually use those values in any meaningful way. By contrast, the prediction interval does present information in an intuitive format, and so reporting incorrect values for this interval can have real repercussions. For

that reason, it might be best to only report the interval when we have enough studies to ensure that the estimate is reasonably precise.

Summary

When we ask about heterogeneity, what we have in mind is "What is the actual range of effects." The statistics typically reported for heterogeneity (such as I^2) do not address this question.

The one statistic that does provide this information is the prediction interval. The prediction interval tells us the range of effects in the same metric that we use for the effect size, so that we understand the range of dispersion. Critically, it tells us the range of effects on an absolute scale, so we know (for example) if the impact ranges from moderate to large, or from trivial to moderate, or from harmful to helpful.

The accuracy of the prediction interval (and all other indices of heterogeneity) depends in part on the number of studies in the analysis. When the analysis includes at least ten studies, the prediction interval is likely to be accurate enough to be useful.

A spreadsheet for computing the prediction interval is available on the book's website.

9.4. Prediction interval vs. confidence interval

9.4.1. Mistake

The summary effect in a forest plot is typically displayed as a point estimate with a confidence interval. Researchers sometimes assume that the confidence interval corresponds to the dispersion of effects. In a variant of this mistake, the forest plot will be used to display one confidence interval for the fixed-effect model and a second (wider) confidence interval for the random-effects model. Readers sometimes assume that the additional width of the random-effects confidence interval corresponds to the dispersion of effects. In either case, this is a fundamental mistake.

9.4.2. Details

The confidence interval and the prediction interval are two entirely separate indices. They address two entirely distinct issues.

When we perform a meta-effects analysis, we typically have two distinct goals.

- One goal is to estimate the *mean* effect size. The confidence interval is an index of precision, and tells us how precisely we have estimated the mean. A confidence interval of 40 to 60 tells us that the mean effect size in the universe of comparable populations falls somewhere in this range. (More accurately, in 95% of all meta-analyses the mean effect size will fall within the confidence interval).
- A second goal is to estimate the *dispersion* of effects. The prediction interval is an index of dispersion. A prediction interval of 25 to 75 tells us that the true effect size will be as low as 25 in some populations, and as high as 75 on others.

Figure 32 shows a fictional set of studies for a meta-analysis to assess the impact of tutoring. In these studies, students are randomized to receive tutoring or to a control group, and we assess their scores on a math test. The effect size is the raw difference in means between groups. The mean difference is 50 points, which tells us that the tutoring increases the mean score by this amount.

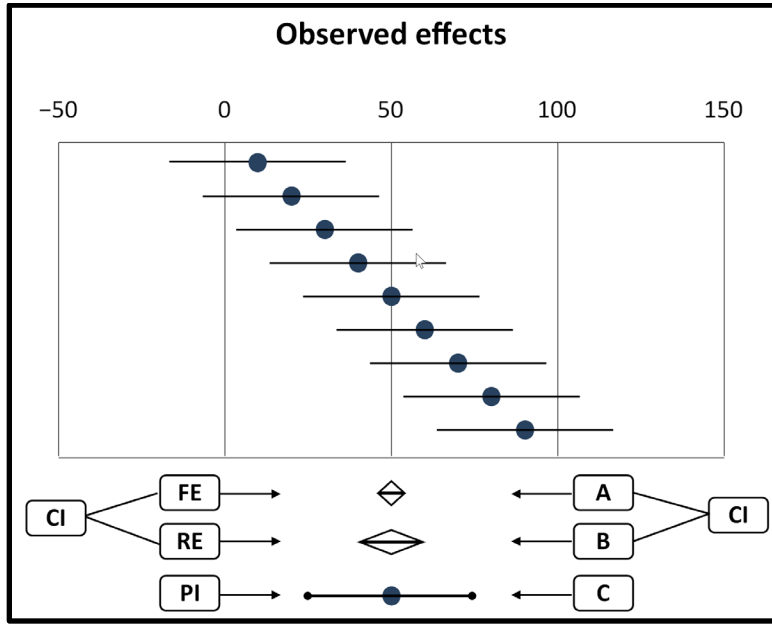

Figure 32 | Confidence intervals and prediction intervals for a fictional meta-analysis

At the bottom of the plot are two diamonds. The first diamond shows the confidence interval for the fixed-effect model, while the second diamond shows the confidence interval for the random-effects model. The first diamond has a width of 7.5 points while the second has a width of 20 points. Researchers sometimes assume that the span for the random-effects model tells us that the effects are dispersed over this (wider) range. This is incorrect – *both* diamonds speak *only* to the precision of the estimate for the mean.

- The confidence interval labeled "FE" is based on the standard error for the fixed-effect model or the fixed-effects model. If all studies are sampled from the same population (fixed effect) or if we are reporting the mean for the studies in the analysis only and not for a wider universe of comparable studies (fixed effects), in 95% of all analyses this confidence interval will include the true effect size for the population(s) in question. This interval has a width of 7.5 points. This is also labeled [A] in keeping with the conventions of this volume (see section 5).
- The confidence interval labeled "RE" is based on the standard error for the random-effects model. If the studies are sampled from different populations, and we are generalizing to the universe of comparable populations, in 95% of all analyses this confidence interval will include

the true mean effect size for the universe. This interval has a width of 20 points. This is also labeled [B] in keeping with the conventions of this volume.

The second diamond is wider than the first because it includes an additional source of sampling error. Under the fixed-effect (singular) model the error comes from the fact that we are sampling people from a specific population. Similarly, under the fixed-effects (plural) model the error comes from the fact that we are sampling people from a fixed set of populations. By contrast, under the random-effects model the error comes from the fact that we are sampling people from populations, and *additionally* sampling populations from a universe of populations. Critically, the additional width in the second diamond reflects additional error that comes from a second level of sampling. *It tells us nothing about how widely the effects are actually dispersed*.

Rather, to address the dispersion of effects we turn to the prediction interval, which is denoted as "PI". The prediction interval is 50 points wide. We expect that in some 95% of all relevant populations, the treatment will increase scores by at least 25 points to as much as 75 points. This is also labeled [C] in keeping with the conventions of this volume.

In this example I displayed the confidence intervals using a diamond rather than a horizontal line. This is the format used by many computer programs (and included as an option in CMA). However, when used for this purpose the diamond has precisely the same meaning as the simple line.

For a fixed-effect or fixed-effects analysis we would display line [A] only. For a random-effects analysis we would display lines [B] and [C] only. I display all three here for pedagogical reasons.

Below, I present examples based on real data.

9.4.3. Example | Prevalence of ADHD in patients with SUD

van Emmerik-van Oortmerssen et al. (2012) looked at prevalence of ADHD in patients with SUD (substance abuse disorder). On the plot (Figure 33) −

- The *confidence* interval for the fixed-effect model [A] tells us that the *mean* prevalence *in this set of thirty studies* falls in the range of 0.235 to 0.257.
- The *confidence* interval for the random-effects model [B] tells us that the *mean* prevalence *in the universe of comparable populations* falls in the range of 0.194 to 0.272.

- The *prediction* interval [C] tells us that the prevalence in *any single population* is as low as 0.082 in some, and as high as 0.500 in others.

In this example, the random-effects confidence interval [B] spans eight points while the prediction interval [C] spans forty-two points. Clearly, to conflate one with the other would be a serious mistake.

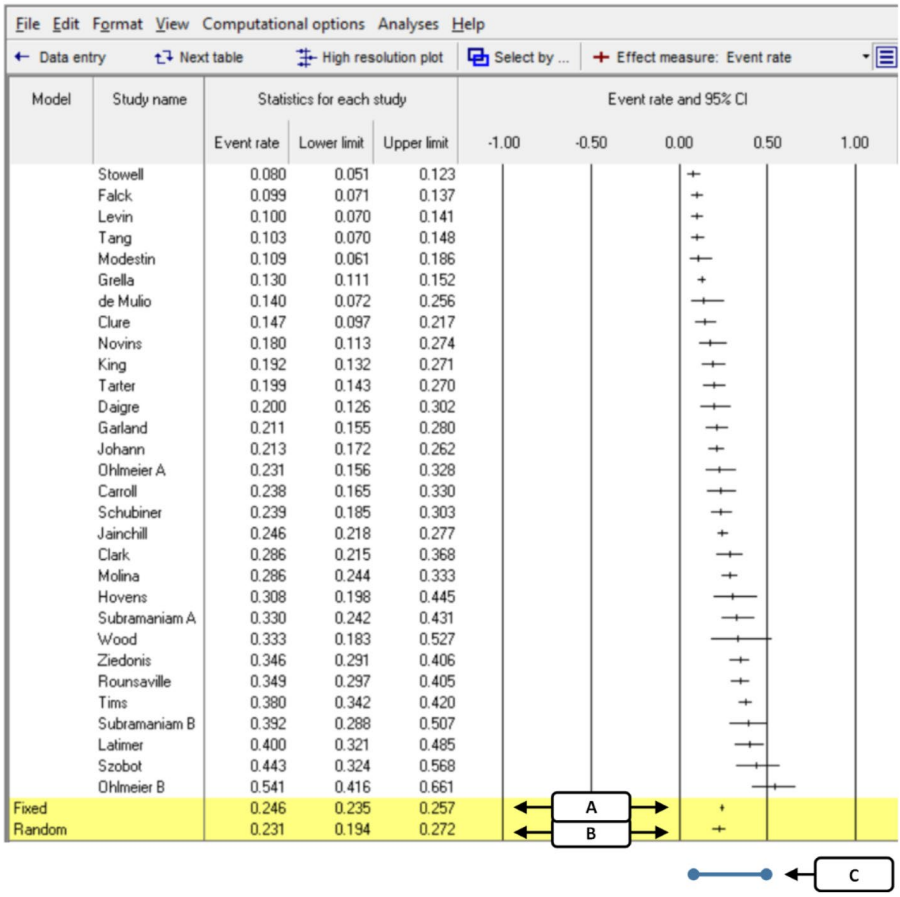

Figure 33 | Prevalence of ADHD in patients with SUD

9.4.4. Example | Augmenting clozapine with a second antipsychotic

Taylor, Smith, Gee, and Nielsen (2012) looked at the impact of augmenting clozapine with a second antipsychotic (Figure 34). The effect size index is the standardized mean difference (*d*).

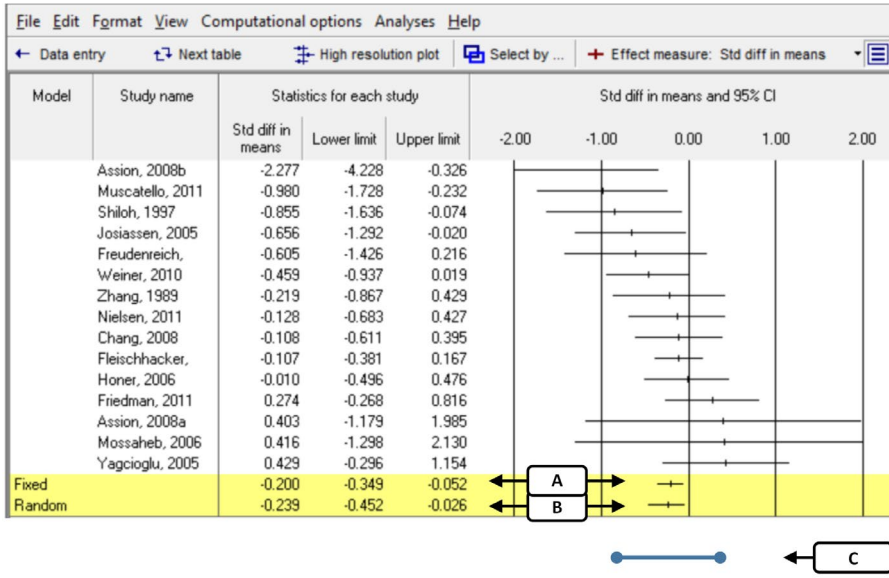

Figure 34 | Augmenting clozapine | Std mean difference < 0 favors augmentation

- The *confidence* interval for the fixed-effect model extends 0.151 on either side of the mean [A]. This tells us that the mean effect in this specific set of fifteen studies falls in the range of −0.349 to −0.052.
- The *confidence* interval for the random-effects model extends 0.213 on either side of the mean [B]. This tells us that the mean effect in the universe of comparable populations falls in the range of −0.452 to −0.026.
- The *prediction* interval extends 0.590 on either side of the mean [C]. This tells us that the effect size in any one population will could be as low as −0.83 (improving function by 0.83 units) or as high as +0.35 (harming function by 0.35 units).

We can say that the *mean* effect is "Helpful" *on average* since the *confidence* interval for the mean falls entirely to the left of zero. However, *in any single population* the effect could be either helpful or harmful since the *prediction* interval includes values on both sides of zero. What should be clear, is that the confidence interval and the prediction interval are addressing two entirely distinct issues, and to conflate one with the other would be a serious mistake.

9.4.5. 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 35). Mean differences less than zero indicate that the treatment was effective in lowering blood pressure. The numbers that follow are based on our re-analysis of the data, and differ slightly from the original report, due to rounding error.

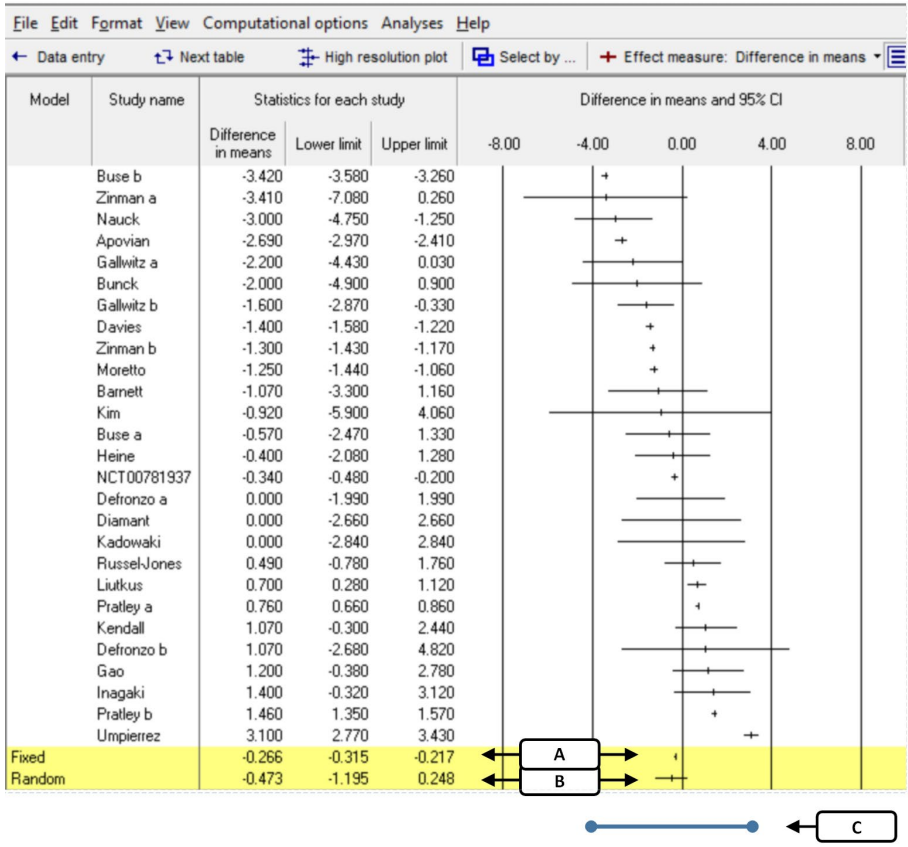

Figure 35 | GLP-1 mimetics and diastolic BP | Mean difference < 0 favors treatment

- Under the fixed-effect model the confidence interval extends roughly 0.05 units on either side of the mean [A]. This tells us tells us that we can estimate the *mean* effect *for the studies in the analysis* within 0.05 units.
- Under the random-effects model the confidence interval extends roughly 0.72 units on either side of the mean [B]. This tells us that we can estimate the *mean* effect in universe of comparable studies within 0.72 units.

- The prediction interval extends 3.65 units on either side of the mean [C]. This tells us that the effect size *in any given population* will usually fall with 3.65 units of the mean, in the range of −4.08 to +3.13.

As always, it would be a serious mistake to conflate the confidence interval with the prediction interval. These are two different indices that address two entirely different elements of the analysis.

9.4.6. Impact of additional studies

It is instructive to consider what happens to the confidence interval and to the prediction interval when we add studies to the analysis.

The confidence interval tells us how precisely we can estimate the mean effect size. As we add studies to the analysis, our estimate of the mean tends to become more precise. Therefore, the confidence interval tends to become narrower.

The prediction interval tells us how widely the treatment's effect varies from one population to another. If there are some populations where the treatment's effect is as low as 0.10 and some where the effect is as high as 0.90, then this is true regardless of how many studies we include in our sample. Therefore, as we add comparable studies to the analysis, the prediction interval tends to remain essentially unchanged (except for small changes as the estimate becomes more precise).

9.4.7. Formulas

The confidence interval is based on the mean effect size and the *standard error* of the mean effect size. By contrast, the prediction interval is based on the mean effect size and the *standard deviation* of the effect size. The confidence interval for the mean may be computed as

$$CI_M = M \pm 1.96(SE), \tag{5}$$

where M is the sample mean and SE is the standard error of the mean. By contrast, the prediction interval may be computed as

$$PI = M \pm 1.96(T), \tag{6}$$

where T is the standard deviation of the true effects.

The formula for the confidence interval (5) is the same for the fixed-effect and the random-effects model, in that both are based on the mean and the standard error of the mean. Where they differ is in the computation of the standard error (*SE*). For the fixed-effect model, the *SE* reflects sampling error based on within-study variance, whereas for the random-effects model, the *SE* reflects sampling error based on within-study variance and between-study variance. In the case where the effect size is the score in one group, the within-study variance is the same for all studies, the standard error for the fixed-effect model is

$$SE = \sqrt{\frac{V}{N}}, \tag{7}$$

and for the random-effects model is

$$SE = \sqrt{\frac{V}{N} + \frac{T^2}{k}}, \tag{8}$$

where V is the common within-study population variance, N is the sample size accumulated across studies, T^2 is the estimate of the between-study variance, and k is the number of studies in the analysis.

These formulas are useful for highlighting the difference between the fixed-effect and random-effects model, but in practice we use more general versions of these formulas as explained in Appendix II and Appendix VII.

9.4.8. Future options

While researchers sometime confuse the confidence interval with the prediction interval, there are several ways to avoid this confusion. One option for a random-effects analysis is to report both the confidence interval and prediction interval, and then explain what each one means. It would also help to include the prediction interval on the plot (as in these examples). Over the longer term, it would helpful if the research community would adopt some conventions to display both the confidence interval and the prediction interval (J. P. Higgins et al., 2009; Riley, Higgins, & Deeks, 2011).

Summary

Researchers sometimes conflate the confidence interval with the prediction interval. The confidence interval is an index of precision, that tells us how precisely we have estimated the mean effect size. The prediction interval is an index of dispersion, that tells us how widely the effect size varies across populations. The two are entirely distinct from each other.

9.5. Mistakes in using the I^2 statistic

9.5.1. Mistake

It is widely believed that the I^2 statistic tells us how much the effect size varies across studies. In some cases, this belief is codified, with I^2 values of 25%, 50%, and 75% taken to reflect low, moderate, and high amounts of dispersion. While this interpretation of I^2 is ubiquitous, it is nevertheless incorrect, and reflects a fundamental misunderstanding of this index.

9.5.2. Details

To explain what I^2 is, I need to provide some background. In a meta-analysis, we need to distinguish between the true effects and the observed effects. The true effect size in any study is the effect size that we would see if we could somehow enroll the entire population in the study, so that there was no sampling error. The observed effect size is the effect size that we see in our sample. The observed effect size serves as an estimate of the true effect size, but invariably differs from the true effect size because of sampling error.

For reasons discussed in Appendix VIII, the variance of the observed effects tends to be larger than the variance of the true effects. For example, consider the analysis represented in Figure 36. In this figure, the outside curve reflects the distribution of *observed* effects, while the inner curve reflects the distribution of *true* effects.

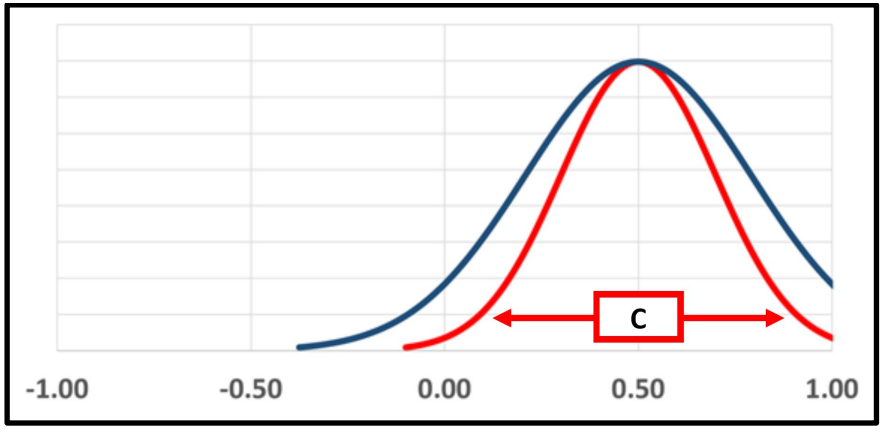

Figure 36 | ADHD Analysis – True effects vs. Observed effects

When we ask about heterogeneity, we typically intend to ask, "How much does the true effect size vary across studies?"

- The prediction interval, which corresponds to line [C] in the plot, tells us that the true effect size in 95% of all populations will fall in the approximate range of 0.10 to 0.90. This is what we have in mind when we ask about heterogeneity.
- By contrast, the I^2 statistics tells us about the relationship between the two distributions. Concretely, I^2 is 47%, which tells us that the variance of true effects (the inner curve) is 47% as large as the variance of observed effects (the outer curve). This information is relevant for other purposes, but is tangential to the question of how much the effect size varies.

I present two sets of examples to illustrate this point. The first set uses the standardized mean difference as the effect size index. The second set uses the risk ratio as the effect size index. Aside from that, the two sets of examples are parallel to each other, and the reader should feel free to focus on either one.

9.5.3. Examples using the standardized mean difference

Castells et al. (2011) looked at 17 studies that evaluated the effect of methylphenidate on cognitive function in adults with ADHD. The effect size index is the standardized mean difference, with values greater than zero indicating that the drug increased cognitive function. The mean effect size is a standardized mean difference of 0.50, and I^2 is 47%.

Simpson, Rorie, Alper, and Schell-Busey (2014) looked at six studies that assessed the impact of interventions such as oversight to reduce corporate crime (people acting illegally on behalf of a company). The effect size index is the standardized mean difference, with values greater than zero indicating that the intervention was associated with a drop in crime. The mean effect size is a standardized mean difference of 0.10, and I^2 is 92%.

Most researchers would assume that there is less dispersion in the ADHD analysis (where I^2 is 47%) as compared with Crime analysis (where I^2 is 92%). However, it should be clear from Figure 37 that the opposite is true, since the distribution of effects for the ADHD analysis is obviously wider than the distribution of effects for the Crime analysis.

In each panel, line [C] corresponds to the prediction interval, which tells us the dispersion of true effects in the metric of the effect-size index. In the ADHD analysis (top panel) I^2 is 47% and the effects vary over 80 points. In

the Crime analysis (bottom panel) I^2 is 92% and the effects vary over 40 points. Thus, the *higher* value of I^2 corresponds to *smaller* amount of dispersion.

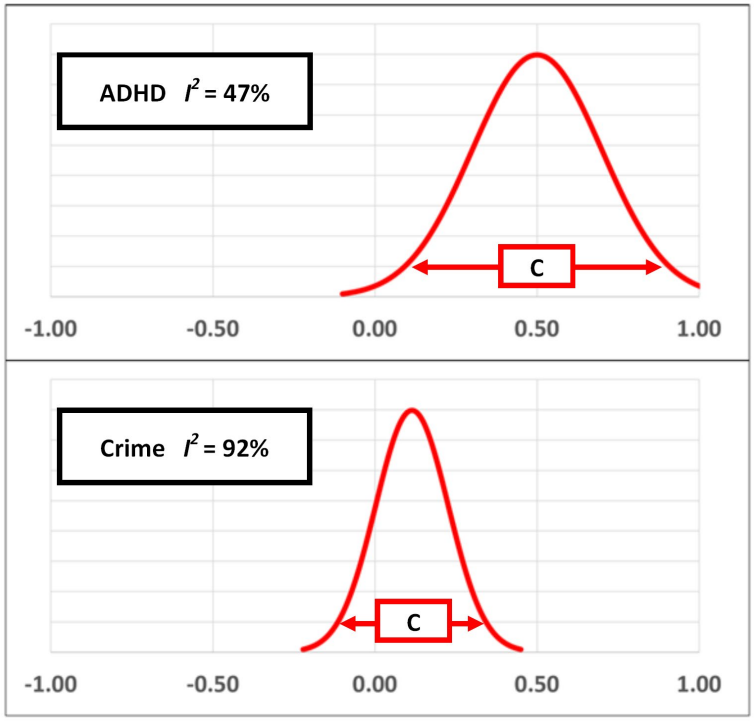

Figure 37 | Distribution of true effects for two meta-analyses

The fact that the higher value of I^2 corresponds to the smaller amount of dispersion will be confusing to researchers who assume that I^2 tells us how much the effect size varies. However, it will make sense for researchers who understand that I^2 is a proportion, not an absolute value. This becomes clear with reference to Figure 38. This is similar to Figure 37, but now each panel has two curves rather than one. The inner curve is identical to the one in the prior plot, and corresponds to the dispersion of *true* effects. But here, we have added an outer curve which corresponds to the dispersion of *observed* effects.

The top panel in Figure 38 shows the ADHD analysis. To quantify the difference between the inner and outer curves we can pick any point on the distribution and compare the width of one curve vs. the other. At line [C] the inner curve covers 77 points, whereas the outer curve covers 113 points. The ratio of inner to outer is thus 68% in linear units or 47% in squared units. This

is the meaning of I^2, which is defined as ratio of true to total variance (Appendix VIII).

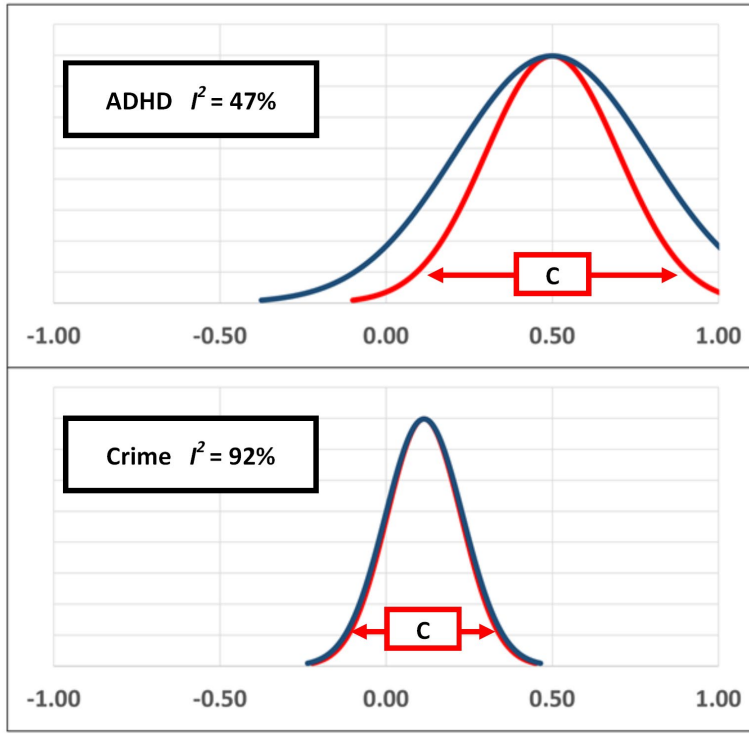

Figure 38 | I^2 and Prediction interval for two meta-analyses

Similarly, the bottom panel in Figure 38 shows the Crime analysis. To quantify the difference between the inner and outer curves we can pick any point on the distribution and compare the width of one curve vs. the other. At line [C] the inner curve covers 44 points, whereas the outer curve covers 46 points. The ratio of inner to outer is thus 96% in linear units or 92% in squared units. This is the meaning of I^2, which is defined as ratio of true to total variance (Appendix VIII).

If we want to know what *proportion* of the variance in observed effects is due to variance in true effects, the answer is provided by the ratio of the inner curve to the outer curve. In the top panel the ratio is 47% and in the bottom panel the ratio is 92%. (In the bottom panel the two lines are so close to each other, they might appear to be a single line). This is what I^2 tells us.

However, if we want to know *how much* the effect size varies, the answer is provided by the width of the inner curve in the metric of the analysis. In

the top panel the true effect size varies from roughly 0.10 in some populations to 0.90 in others, as indicated by line [C]. In the bottom panel the true effect size varies from −0.10 in some populations to +0.30 in others, as indicated by line [C]. When we are asking about the utility of an intervention, we almost invariably are interested in the *amount* of variance, not the *proportion*. As such, we are asking about the prediction interval, and not about I^2.

Finally, it might be helpful to show the relationship between these numbers and the actual forest plot for the two analyses.

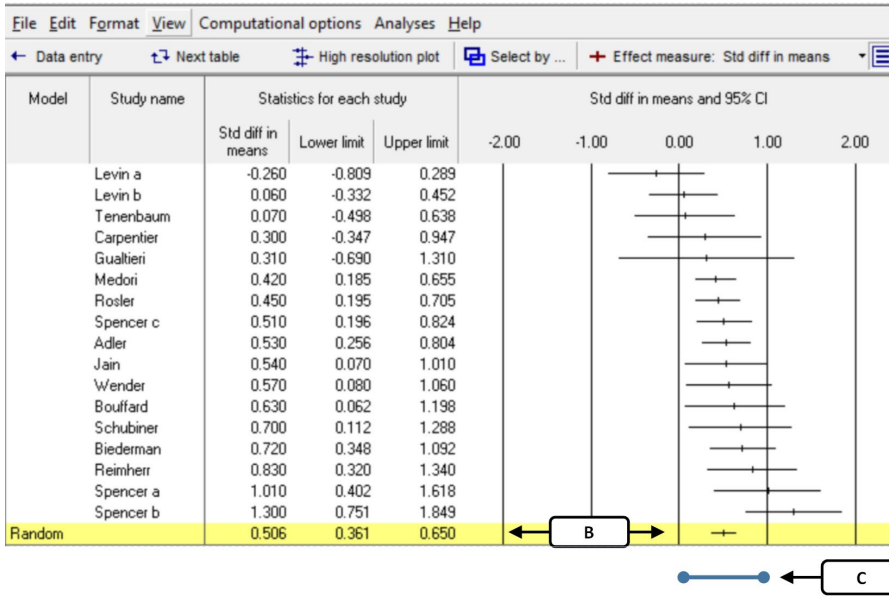

Figure 39 | ADHD analysis | Standardized difference > 0 favors treatment

Figure 39 shows the ADHD analysis. The general sense conveyed by the plot is that there is substantial dispersion in the observed effects, but also substantial sampling error (as reflected in the width of the confidence interval about most of the effect sizes). The sampling error can explain some 53% of the observed variance, and the remaining 47% reflects variance in true effects. This 47%, the *ratio* of true to total variance, is I^2. As a separate matter, if we want to know the dispersion of effects *on an absolute scale* we turn to line [C]. This corresponds to the prediction interval, and tells us that true effects vary from around 0.10 in some populations to 0.90 in others. This is the same as line [C] in the top panel of Figure 38.

Figure 40 shows the Crime analysis. The general sense conveyed by the plot is that there is only modest dispersion in the observed effects, but *very*

little sampling error in comparison. Critically, the *ratio* of sampling error to observed variance is small. The sampling error can explain only 8% of the observed variance, and the remaining 92% reflects variance in true effects. This 92%, the *ratio* of true to total variance, is I^2. As a separate matter, if we want to know the dispersion of effects *on an absolute scale* we turn to line [C]. This corresponds to the prediction interval, and tells us that true effects vary from around −0.10 in some populations to +0.30 in others. This is the same as line [C] in the bottom panel Figure 38.

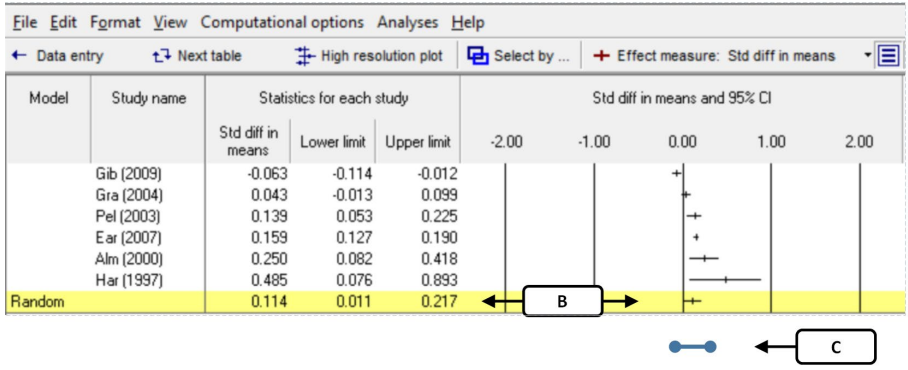

Figure 40 | Crime analysis | Standardized difference > 0 favors treatment

9.5.4. Examples using risk ratios

Immediately above, I presented two examples where the effect-size index is the standardized mean difference. Here, I will make the same points using two examples where the effect-size index is the risk ratio.

Kasapis et al. (2009) looked at eight studies that evaluated the impact of a stent implantation on the failure rate for angioplasty. The effect size is a risk ratio, with ratios below one indicating that stents reduced the risk of failure. The mean risk ratio was 0.283, and I^2 is 56%.

Lin et al. (2013) looked at the effects of no-smoking laws on the risk of acute myocardial infarction. As recently as the 1990s, most cities allowed smoking in public spaces. Over the more recent decades, governments have passed laws that prohibit smoking in restaurants, workplaces, airports, and so on. A number of studies have been performed to see if the risk of having a heart attack changed when these laws were implemented. The effect size is a risk ratio, with ratios below one indicating a reduction in events. The mean risk ratio was 0.877, and I^2 is 92%.

Most researchers would assume that there is less dispersion in the Stents analysis (where I^2 is 56%) as compared with Smoking analysis (where I^2 is

92%). However, it should be clear from Figure 41 that the opposite is true, since the distribution of effects for the Stents analysis is obviously wider than the distribution of effects for the Smoking analysis.

In each panel, line [C] corresponds to the prediction interval, which tells us the dispersion of true effects in the metric of the effect-size index. In the Stents analysis (top panel) I^2 is 56% and the effects vary over 86 points. In the Smoking analysis (bottom panel) I^2 is 92% and the effects vary over 25 points. Thus, the *higher* value of I^2 corresponds to the *smaller* amount of dispersion.

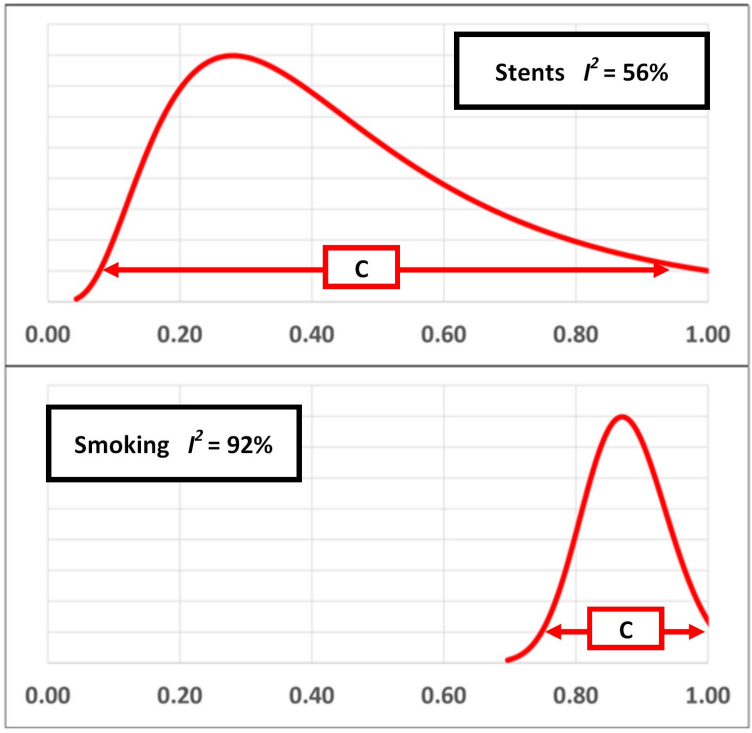

Figure 41 | Distribution of true effects for two meta-analyses

The fact that the higher value of I^2 corresponds to the smaller amount of dispersion will be confusing to researchers who assume that I^2 tells us how much the effect size varies. However, it will make sense for researchers who understand that I^2 is a proportion, not an absolute value. This becomes clear with reference to Figure 42. This is similar to Figure 41, but now each panel has two curves rather than one. The inner curve is identical to the one in the prior plot, and corresponds to the dispersion of *true* effects. But here, we have added an outer curve which corresponds to the dispersion of *observed* effects.

The top panel in Figure 42 shows the Stents analysis. To quantify the difference between the inner and outer curves we can pick any point on the distribution and compare the width of one curve vs. the other. At line [C] the inner curve covers 86 points, whereas the outer curve covers 140 points. The ratio of inner to outer in squared units in the log metric is 56%. This is the meaning of I^2, which is defined as ratio of true to total variance (Appendix VIII).

Similarly, the bottom panel in Figure 42 shows the Smoking analysis. To quantify the difference between the inner and outer curves we can pick any point on the distribution and compare the width of one curve vs. the other. At line [C] the inner curve covers 25 points, whereas the outer curve covers 27 points. The ratio of inner to outer in squared units in the log metric is 92% (Appendix VIII). This is the meaning of I^2, which is defined as ratio of true to total variance.

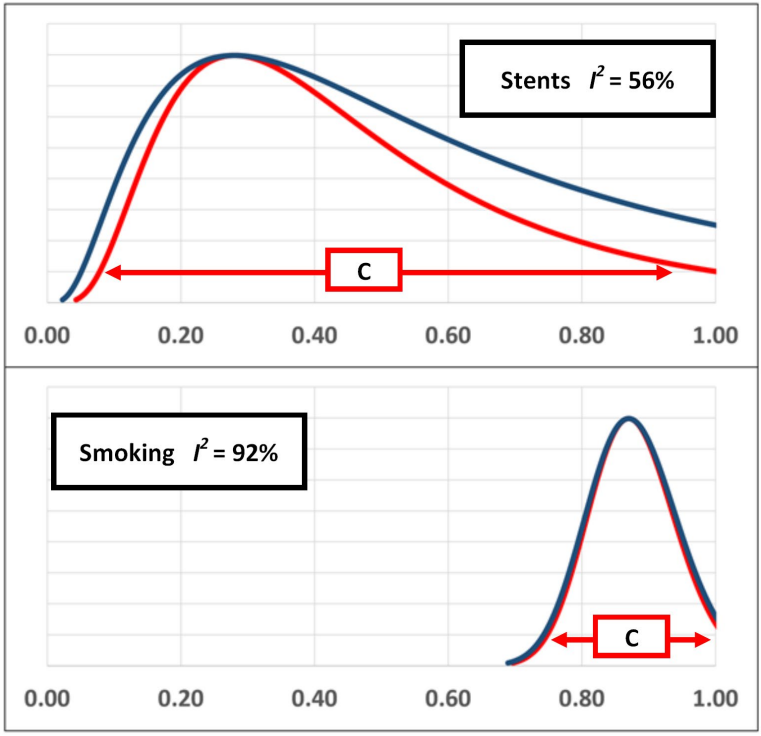

Figure 42 | I^2 and Prediction interval for two meta-analyses

If we want to know what *proportion* of the variance in observed effects is due to variance in true effects, the answer is provided by the ratio of the

inner curve to the outer curve. In the top panel the ratio is 56% and in the bottom panel the ratio is 92%. (In the bottom panel the two lines are so close to each other, they might appear to be a single line). This is what I^2 tells us.

However, if we want to know *how much* the effect size varies, the answer is provided by the width of the inner curve on the metric of the analysis. In the top panel the true risk ratio varies from roughly 0.08 in some populations to 0.96 in others, as indicated by line [C]. In the bottom panel the true effect size varies from 0.76 in some populations to 1.01 in others, as indicated by line [C]. This is what the prediction interval tells us. When we are asking about the utility of an intervention, we almost invariably are interested in the amount of variance, not the proportion. As such, we are asking about the prediction interval, and not about I^2.

Finally, it might be helpful to show the relationship between these numbers and the actual forest plot for the two analyses.

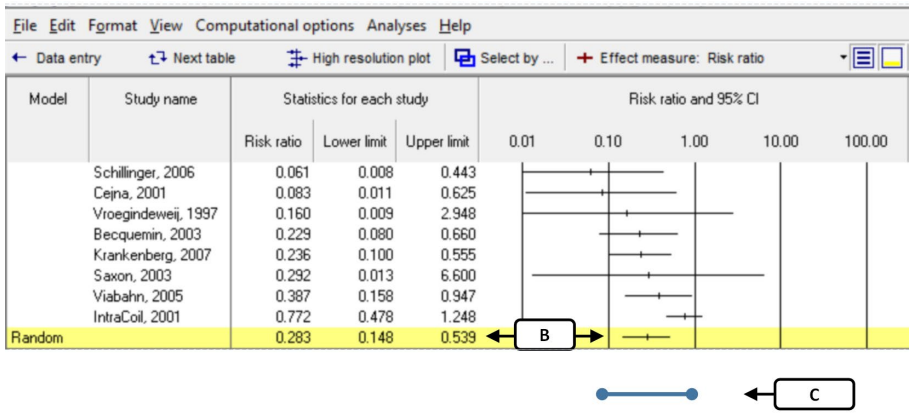

Figure 43 | Stents | Risk ratio < 1 favors treatment

Figure 43 shows the Stents analysis. The general sense conveyed by the plot is that there is substantial dispersion in the observed effects, but also substantial sampling error (as reflected in the width of the confidence interval about most the effect sizes). The sampling error can explain some 44% of the observed variance, and the remaining 56% reflects variance in true effects. This 56%, the *ratio* of true to total variance, is I^2. As a separate matter, if we want to know the dispersion of effects *on an absolute scale* we turn to line [C]. This corresponds to the prediction interval, and tells us that true effects vary from around 0.08 in some populations to 0.96 in others. This is the same as line [C] in the top panel Figure 42.

Figure 44 shows the Smoking analysis. The general sense conveyed by the plot is that there is only modest dispersion in the observed effects, but *even less* sampling error. Critically, the *ratio* of sampling error to observed variance is small. The sampling error can explain only 8% of the observed variance, and the remaining 92% reflects variance in true effects. This 92%, the *ratio* of true to total variance, is I^2. As a separate matter, if we want to know the dispersion of effects *on an absolute scale* we turn to line [C]. This corresponds to the prediction interval, and tells us that true effects vary from around 0.76 in some populations to 1.01 in others. This is the same as line [C] in the bottom panel Figure 42.

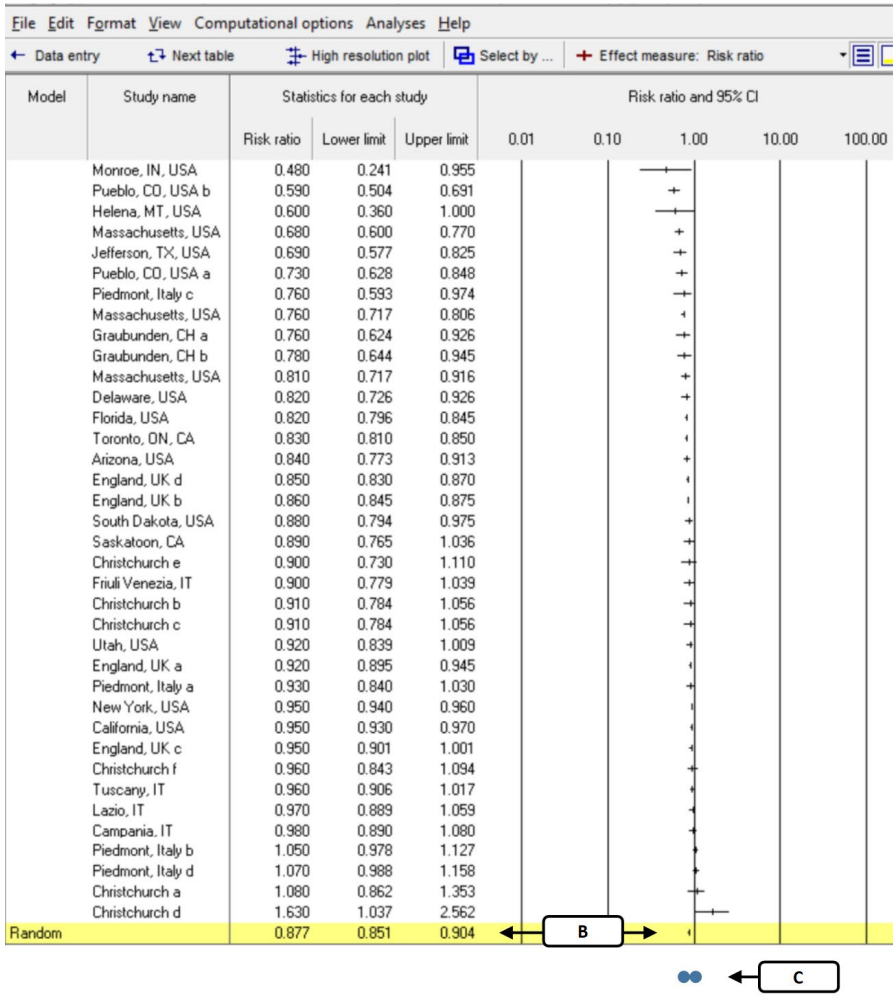

Figure 44 | Smoke-free legislation | Risk ratio < 1 indicates reduced risk

9.5.5. Words matter

The I^2 statistic is defined as being a *proportion*, not an *absolute* amount of dispersion. A proportion and an absolute amount are two different things. Nevertheless, researchers often define I^2 (correctly) as being a proportion or percentage, and then ignore this definition and speak about I^2 (incorrectly) as being an index of dispersion on an absolute scale. This is an important issue because if we paid attention to the words, we would avoid the mistake of misinterpreting I^2.

Consider the following examples.

9.5.6. Example | Drugs for ADHD

Cunill, Castells, Tobias, and Capellà (2016) looked at the impact of drugs on ADHD. They write "Between-study heterogeneity was assessed using Cochran's Q test (Cochran 1954) jointly with the I^2 index (Higgins et al. 2003), which enables the percentage of variation in the combined estimate that can be attributed to heterogeneity to be established (< 25%: low heterogeneity; 25 to 50 %: moderate; 50-75%: high; > 75%: very high)." The first part of the sentence defines I^2 as a *percentage* of variance. The part in parentheses suggests that I^2 is an index of *absolute* variance (e.g., "low heterogeneity"). These are two different things. If I^2 is the first (which it is) then logically it cannot also be the second.

9.5.7. Example | Exercise for chronic back pain

Ferreira, Smeets, Kamper, Ferreira, and Machado (2010) performed a meta-analysis that looked at the impact of exercise for chronic back pain. They write "Therefore, the [sic] I^2 provides the *percentage* [italics in the original] of total variation across studies explained by heterogeneity rather than chance (J. P. Higgins, Thompson, Deeks, & Altman, 2003). For instance, an I^2 of 0% indicates that all variability in effect estimates is due to sampling error and not to heterogeneity among trials. Conversely, an I^2 of 75% suggests that three quarters of the variability in effect estimates can be attributed to inconsistency among trials. An I^2 value of more than 75% was considered to represent high heterogeneity, an I^2 of 50% to 75% was considered to represent moderate heterogeneity, and an I^2 of less than 25% was considered to represent low heterogeneity." The word "percentage" is italicized in the original to emphasize the fact that this is a percentage, but the authors nevertheless proceed to treat the index as an absolute value. Ironically, the

focus of this paper is on the heterogeneity in effects, and so the fact that they use the wrong index to discuss heterogeneity is especially problematic.

9.5.8. In context

Hundreds of papers define I^2 as a proportion and then proceed to interpret it as an absolute value. This is the statistical equivalent of someone in a car dealership being told that they will need to pay only 80% of the usual price, and then trying to pay $80 for the car. A proportion and an absolute value are not the same thing.

9.5.9. Using the I^2 statistic correctly

While I^2 does not tell us how much the effect size varies, it is useful for the following purposes (Borenstein et al., 2017; J. P. Higgins & Thompson, 2002; J. P. Higgins et al., 2003).

- If I^2 is zero, then all the variance in observed effects is due to sampling error. The variance in true effects is estimated as zero.
- If we are looking at a forest plot, I^2 provides context for understanding that plot. If I^2 is near zero, the variance of true effects is only a small fraction of that suggested by the plot. As I^2 increases, that proportion increases.
- If we are working with a set of meta-analyses where the variance of observed effects is reasonably consistent, there will be a strong correlation between I^2 and the absolute amount of variance. Within that context, I^2 can provide information about the relative amounts of dispersion across analyses.
- The I^2 statistic is useful to statisticians who are evaluating the properties of various statistics. For example, if someone wanted to run simulations to see how statistical power is affected by the ratio of true to total variance, they could do so for various values of I^2.
- Sometimes, we do care about the proportion of variance rather than the absolute amount of variance. For example, if we have various ways of conducting studies and we want to know which have the smallest amount of sampling error, I^2 is the index that allows us to address this question.

9.5.10. Further readings

The original papers on I^2 are (J. P. Higgins & Thompson, 2002; J. P. Higgins et al., 2003). For a more detailed discussion of the issues raised in this section, see (Borenstein et al., 2017). For related papers see (Borenstein, 2019; Coory, 2010; J. P. Higgins, 2008; Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006; Ioannidis, 2008a; Patsopoulos, Evangelou, & Ioannidis, 2008; Rucker, Schwarzer, Carpenter, & Schumacher, 2008).

Summary

When we ask about heterogeneity, we intend to ask how much the true effect size varies across studies. This question is addressed by the prediction interval which tells us (for example) that the true effect size in most populations will fall in the range of 0.05 to 0.95. It is not addressed by the I^2 statistic. The I^2 statistic tells us what *proportion* of the variance in observed effects reflects variation in true effects, rather than sampling error. It does not tell us how much variation there is.

9.6. Classifying heterogeneity as low, moderate or high

9.6.1. Mistake

In some fields of research, it is common for papers that report I^2 to categorize the heterogeneity as being low, moderate or high, based on the I^2 value. This is a fundamental mistake.

9.6.2. Details

Immediately above, I showed that I^2 is a proportion, not an index of absolute dispersion. It does not tell us how much the effects vary. Since I^2 does not tell us how much the effects vary, the idea of using I^2 to create categories of dispersion is a non-sequitur.

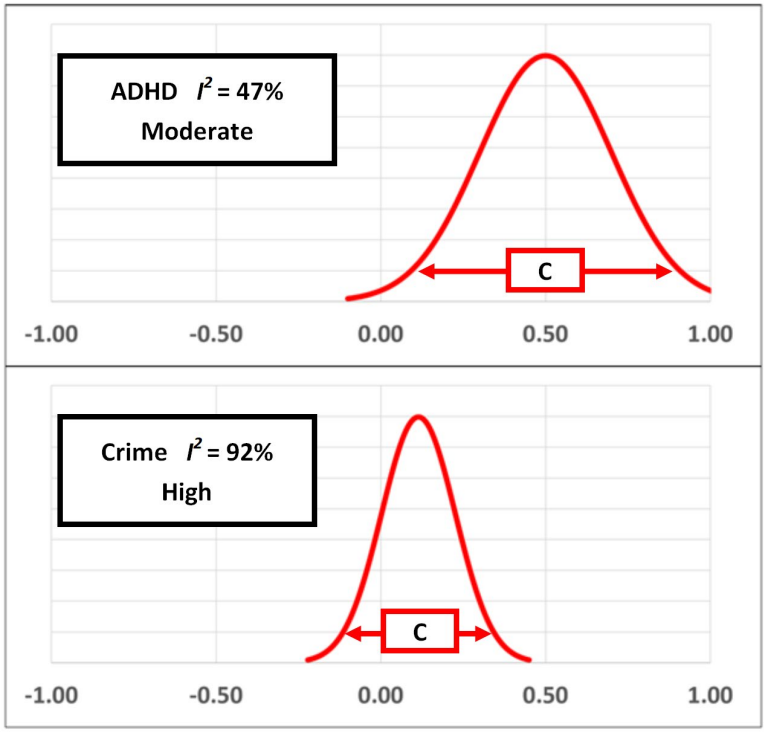

Figure 45 | Distribution of true effects for two meta-analyses

The example discussed earlier (section 9.5.3) is re-displayed in Figure 45. The top panel shows the impact of methylphenidate on the cognitive function of adults with ADHD. The bottom panel shows the impact of interventions to reduce corporate crime. In the top panel I^2 is 47% and in the bottom panel I^2 is 92%, so based on the proposed classifications we would say that the heterogeneity at the top is moderate while that at the bottom is high. This obviously makes no sense, since the dispersion in the top panel is substantially *greater* than the dispersion in the bottom panel.

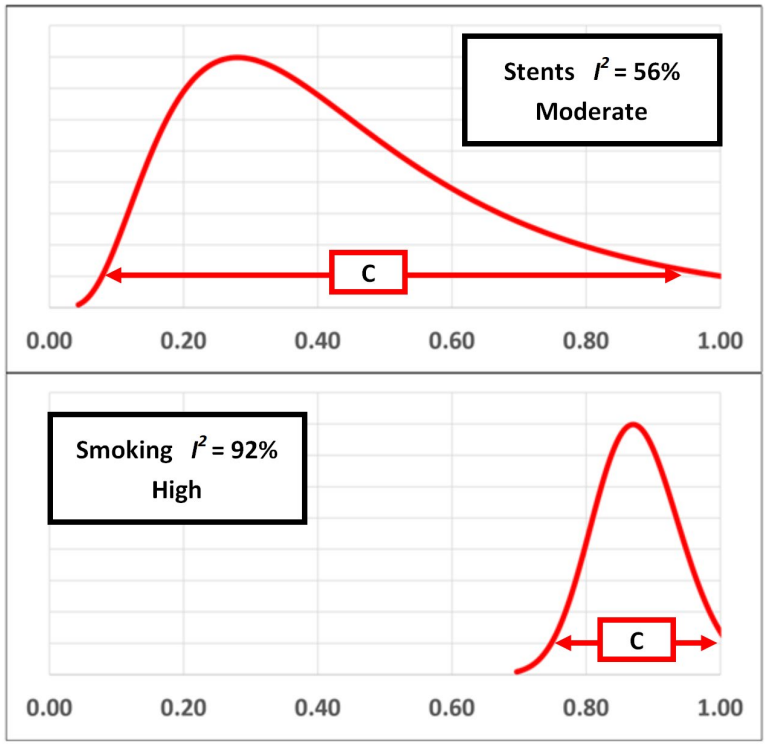

Figure 46| Distribution of true effects for two meta-analyses

Similarly, the example discussed earlier (section 9.5.4) is re-displayed in Figure 46. The top panel shows the impact of stents on the risk of failure in angioplasty. The bottom panel shows the impact of anti-smoking legislation to reduce the risk of myocardial infarction. In the top panel I^2 is 56% and in the bottom panel I^2 is 92%, so based on the proposed classifications we would say that the heterogeneity at the top is moderate while that at the bottom is high. This obviously makes no sense, since the dispersion in the top panel is substantially *greater* than the dispersion in the bottom panel.

Since I^2 does not tell us how much the effects vary, it obviously cannot be used to classify analyses as having a low, moderate, or high amount of variation. However, there is an additional point to be made. Let us assume for a moment that I^2 actually told us the *amount* of variation. What does it mean to say that a particular amount of dispersion is low, moderate, or high, unless we put that dispersion in the context of a specific outcome? Consider the following four examples.

9.6.3. Example | Allegiance to treatment

Munder, Fluckiger, Gerger, Wampold, and Barth (2012) performed a meta-analysis to see if the researchers' allegiance to one treatment vs. another would bias the outcome in studies that compared the two treatments. The effect size index is the standardized mean difference. They write "In addition, we report I^2 as another common quantitative measure of heterogeneity, which can be interpreted as the percentage of overall heterogeneity that is due to variation of the true effects. An I^2 value of 0% indicates no heterogeneity. I^2 values of 25%, 50%, and 75% can be regarded as markers of low, moderate, and strong heterogeneity, respectively (Higgins, Thompson, Deeks, & Altman, 2003)."

9.6.4. Example | Prevalence of pelvic-floor disorders

Islam et al. (2017) published the protocol for a meta-analysis to assess the prevalence of pelvic-floor disorders in women in low and middle-income countries. The effect size index is the prevalence of the disorder. They plan to use values of I^2 to classify the heterogeneity as being low, moderate, or high.

9.6.5. Example | Preventing substance abuse

Onrust, Otten, Lammers, and Smit (2016) performed a meta-analysis to assess the impact of interventions to prevent substance abuse. The effect size index is the standardized mean difference. They used values of I^2 to classify the heterogeneity as being low, moderate, or high.

9.6.6. Example | Exercises for back pain

Ferreira et al. (2010) report on a meta-analysis to assess the impact of exercises for back pain. The effect size index is the difference in means. They used values of I^2 to classify the heterogeneity as being low, moderate, or high.

9.6.7. In context

The idea of classifying the amount of heterogeneity based on I^2 would only make sense if I^2 was an index of absolute dispersion, and it is not. Therefore, the whole idea is a non-starter.

Additionally, even if the classifications were based on an index of absolute dispersion (such as T) the idea that we can have classifications of low, moderate or high variance that apply universally, makes no sense. This would require that a similar amount of variance has the same substantive meaning for an analysis of allegiance to treatment, an analysis of the prevalence of pelvic-floor disorder, an analysis of interventions to prevent substance abuse, and an analysis of the impact of exercises on back pain – among thousands of other analyses.

Indeed, the suggestion is not merely that (for example) 50% is a moderate amount of heterogeneity for risk ratios. The suggestion is that it is a moderate amount of heterogeneity for risk ratios, mean differences, prevalence, and even simple means in one-arm studies. A moment's reflection should make it clear that this idea makes no sense without additional context.

Where did these classifications originate?

When J. P. Higgins et al. (2003) proposed a link between values of I^2 and absolute amount of dispersion, that *was for a specific context*. The authors were primarily concerned with the Cochrane Database of systematic reviews, and the dispersion of observed effects tended to be reasonably consistent across analyses. In that situation, a meta-analysis with a low value of I^2 tended to have less dispersion in effects as compared with a similar analysis that had a higher value of I^2, and the labels were intended to capture this. The idea that these labels could somehow capture the amount of dispersion in analyses outside of the Cochrane database was never their intent.

Summary

The idea of using I^2 to classify heterogeneity as being low, moderate, or high makes no sense for two reasons.

First, I^2 is a proportion, not an index of absolute dispersion. It does not tell us how much variance there is.

Second, the idea that we can classify heterogeneity into categories without additional context is silly, since an amount of heterogeneity that would be considered high in one context would be considered low in another.

9.7. Using the *p*-value as index of heterogeneity

9.7.1. Mistake

Researchers typically report a test for heterogeneity as part of a meta-analysis. Some researchers assume that the test for heterogeneity speaks to the amount of dispersion in the effects. A non-significant *p*-value is interpreted as evidence that the effects are consistent, and a significant *p*-value is taken as evidence that the effects vary in some substantive way. This is a mistake.

9.7.2. Details

A meta-analysis typically includes a test for heterogeneity. The null hypothesis for this test is that there is no variation at all in true effect sizes. The test statistic (Q) along with its degrees of freedom yields a *p*-value. A significant *p*-value allows us to reject this null hypothesis, and to conclude that the effect size does vary across studies. The criterion alpha for this test is conventionally set at 0.05 in some disciplines, and at 0.10 in others (Berman & Parker, 2002; Petitti, 2001).

As is true for all null-hypothesis significance tests, the only information provided by a significant *p*-value is that the variation in effects size is probably not zero (more correctly, if the true heterogeneity is zero, it would be unusual to see a test statistic this high or higher).

The *p*-value for the test of heterogeneity is a function of three items –

1. The estimated amount of heterogeneity
2. The precision of the individual studies
3. The number of studies

If there are many studies (and/or large studies) the *p*-value might be statistically significant even if the amount of heterogeneity is trivial. Conversely, if there are few studies (and/or small studies) the *p*-value might not be statistically significant even if the amount of heterogeneity is substantial. For this reason, the *p*-value cannot serve as a surrogate for the amount of variation.

Two examples will make this clear.

9.7.3. Example | Impact of 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 47). The effect size is the odds ratio, with values less than 1.0 indicating that the treatment was helpful. The mean effect size is 0.741, which tells us that the treatment reduces the odds of a stroke by 74% *on average*. The test for heterogeneity yields a Q-value of 9.105 with 5 degrees of freedom, and a p-value of 0.105. If someone simply looked at the non-significant p-value, they might assume that there was only a small amount of heterogeneity.

In fact, the results suggest that there may be substantial heterogeneity. The prediction interval [C] is 0.32 to 1.71, which tells us that in some populations the treatment *reduces* the odds of a bad outcome by 68%, while in others it *increases* the odds of a bad outcome by 71%.

The p-value is a function of (1) the estimated amount of dispersion (2) the number of studies and (3) the precision of those studies. In this case our best estimate is that there is substantial dispersion. However, the p-value is not significant primarily because there are only a few studies, and these are not terribly precise.

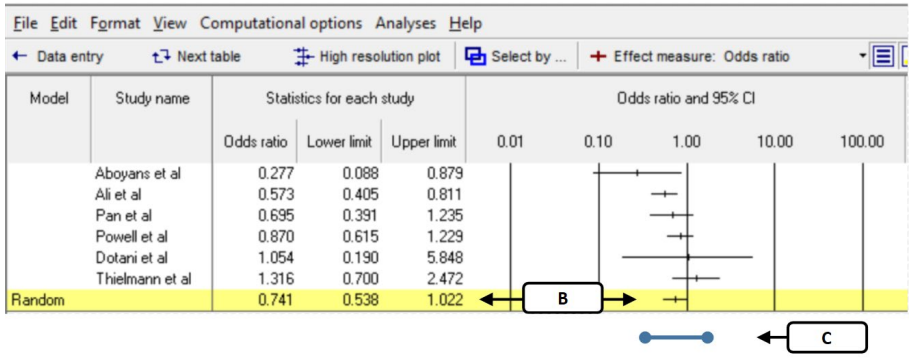

Figure 47 | Preoperative statin therapy | Odds ratio < 1 favors treatment

9.7.4. Example | Impact of smoke-free legislation

Lin et al. (2013) looked at the impact of smoke-free legislation on acute myocardial infarction (MI) (Figure 48). The mean risk ratio was 0.877, which indicates that the risk of MI was reduced on average by some 12%. The test for heterogeneity yields a Q-value of 431.106 with 36 degrees of freedom and a p-value of < 0.0000000001. If someone simply looked at the significant p-

value, they might assume that there was an exceptional amount of heterogeneity.

However, that is not the case here. In fact, the amount of heterogeneity was modest. The prediction interval [C] is 0.75 to 1.02, which tells us that in some populations, the treatment *reduces* the risk of a bad outcome by 25%, while in others it *increases* the risk of a bad outcome by 2%.

The *p*-value is a function of (1) the estimated amount of dispersion (2) the number of studies and (3) the precision of those studies. In this case the amount of dispersion is modest. The *p*-value is statistically significant primarily because of there are many studies, and many of these are precise.

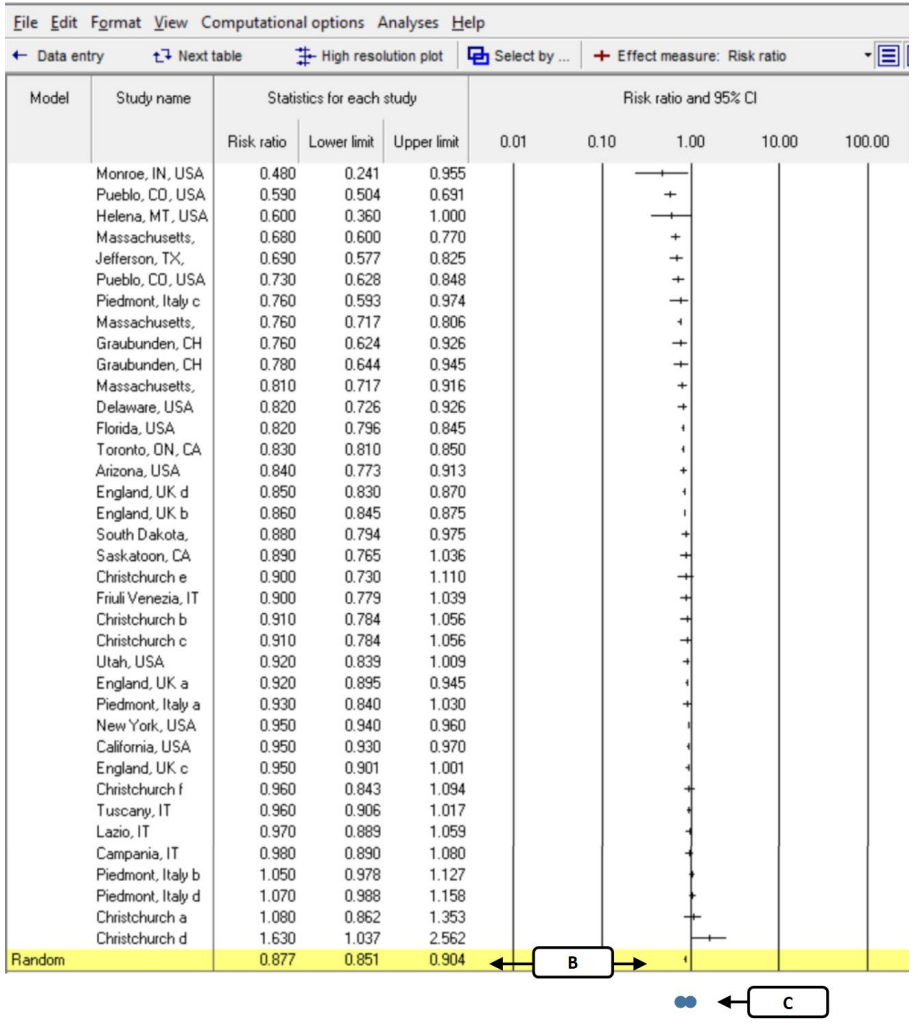

Figure 48 | Smoke-free legislation |Risk ratio < 1 indicates reduced risk

Figure 49 allows us to compare these two analyses. In this figure, the top plot corresponds to the statin analysis where the *p*-value for a test of heterogeneity is 0.105 but there the estimated dispersion is substantial. The bottom plot corresponds to the smoking analysis where the *p*-value for a test of heterogeneity is 0.0000000001 but the estimated dispersion is relatively small. Additional details are presented in Table 3.

As in these examples, the *p*-value tells us nothing about the amount of dispersion. Indeed, it does not even tell us which of two analyses had more dispersion.

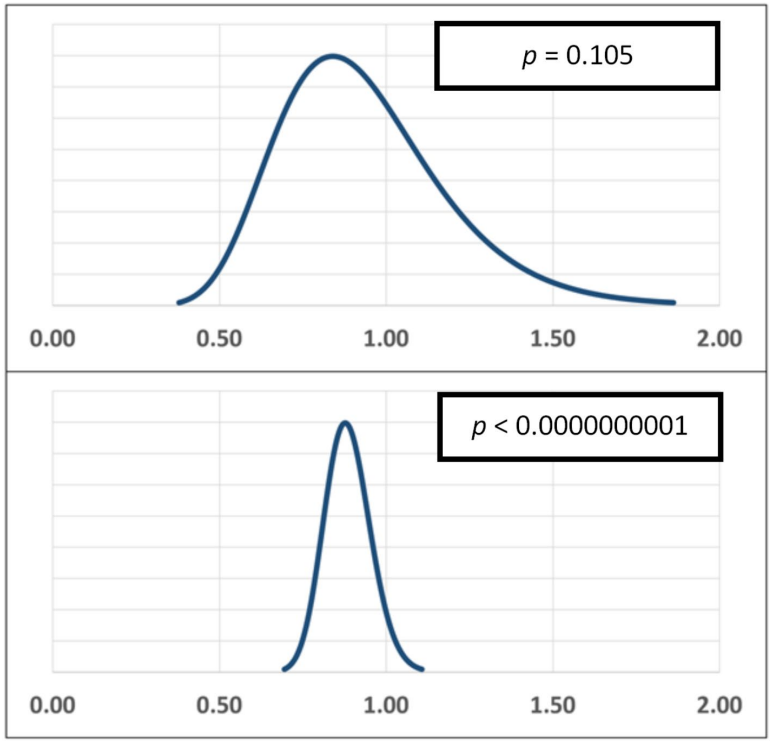

Figure 49 | Distribution of true effects for two meta-analyses

Table 3 | Heterogeneity in two analyses

Study	Index	Mean	*p*-value	Prediction Interval
Statins	Odds ratio	0.74	0.105	0.32 to 1.71
Smoking	Risk ratio	0.88	< 0.0000000001	0.75 to 1.02

Summary

The *p*-value for a test of heterogeneity is a function of (1) the estimated amount of heterogeneity, (2) the precision of the individual studies, and (3) the number of studies in the analysis.

The *p*-value may be statistically significant when the estimated heterogeneity is trivial. Conversely, the *p*-value may not be statistically significant when the estimated heterogeneity is substantial. Therefore, the *p*-value should never be used as a surrogate for the amount of heterogeneity.

9.8. Using the *Q*-value as index of heterogeneity

9.8.1. Mistake

Researchers sometimes use the *Q*-value as an index of dispersion, and assume that a large *Q*-value reflects a substantial amount of heterogeneity. This is a mistake.

9.8.2. Details

The *Q*-value is not an index of dispersion. Rather, it is simply the sum of squared deviations, on a standardized scale. The *Q*-value in a meta-analysis serves a similar function to the sum of squares in a primary study. In a primary study we compute the sum of squares as an interim step to computing the variance and the standard deviation. By itself, the sum of squares tells us nothing useful about the dispersion.

The issues here are similar to those outlined for the *p*-value in the prior section. Specifically, the value of *Q* depends on

1. The amount of observed heterogeneity
2. The precision of the individual studies
3. The number of studies

If there are many studies (and/or large studies) the *Q*-value might be high even if the amount of observed heterogeneity is trivial. Conversely, if there are few studies (and/or small studies) the *Q*-value might be low even if the amount of heterogeneity is substantial. For this reason, the *Q*-value cannot serve as a surrogate for the amount of variation.

To assume that the *Q*-value tells us something about the extent of dispersion in a meta-analysis is analogous to assuming that the sum of squares tells us something about the extent of dispersion in a primary study. In a primary study, the sum of squares (by itself) does not provide that information. In a meta-analysis the value of *Q* (by itself) does not provide that information.

The two examples in the immediately prior section (9.7) can serve here as well.

9.8.3. Example | Impact of 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 50). The effect size is the odds ratio, with values less than 1.0 indicating that the treatment was helpful. The mean effect size is 0.741, which tells us that the treatment reduces the odds of a stroke by 74% on average. The test for heterogeneity yields a *Q*-value of 9.105 with 5 degrees of freedom, and a *p*-value of 0.105. If someone simply looked at the small *Q*-value, they might assume that there was only a small amount of heterogeneity.

In fact, the results suggest that there may be substantial heterogeneity. The prediction interval [C] is 0.32 to 1.71, which tells us that in some populations the treatment *reduces* the odds of a bad outcome by 68%, while in others it *increases* the odds of a bad outcome by 71%.

The *Q*-value is a function of (1) the amount of observed dispersion, (2) the number of studies and (3) the precision of those studies. In this case, our best estimate is that there is substantial dispersion, but the *Q*-value is small primarily because there are only a few studies, and these are not terribly precise.

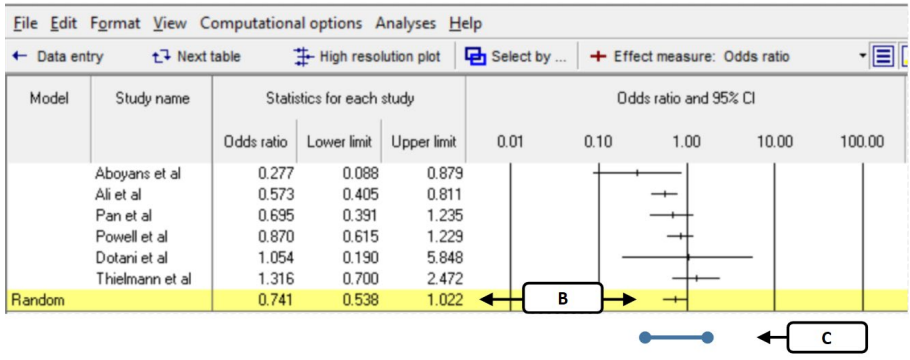

Figure 50 | Preoperative statin therapy | Odds ratio < 1 favors treatment

9.8.4. Example | Impact of smoke-free legislation

Lin et al. (2013) looked at the impact of smoke-free legislation on acute myocardial infarction (MI) (Figure 51). The mean risk ratio was 0.877, which indicates that the risk of MI was reduced on average by some 12%. The test for heterogeneity yields a *Q*-value of 431.106 with 36 degrees of freedom and a *p*-value of < 0.0000000001. If someone simply looked at the magnitude of

the *Q*-value, they might assume that there was an exceptional amount of heterogeneity.

However, that it not the case here. In fact, the amount of heterogeneity is modest. The prediction interval [C] is 0.75 to 1.02. This tells us that in some populations, the treatment *reduces* the risk of a bad outcome by 25%, while in others it *increases* the risk of a bad outcome by 2%.

The *Q*-value is a function of (1) the amount of observed dispersion, (2) the number of studies and (3) the precision of those studies. In this case, our best estimate is that there is only modest dispersion, but the *Q*-value is high primarily because there are many studies, and many of these are precise.

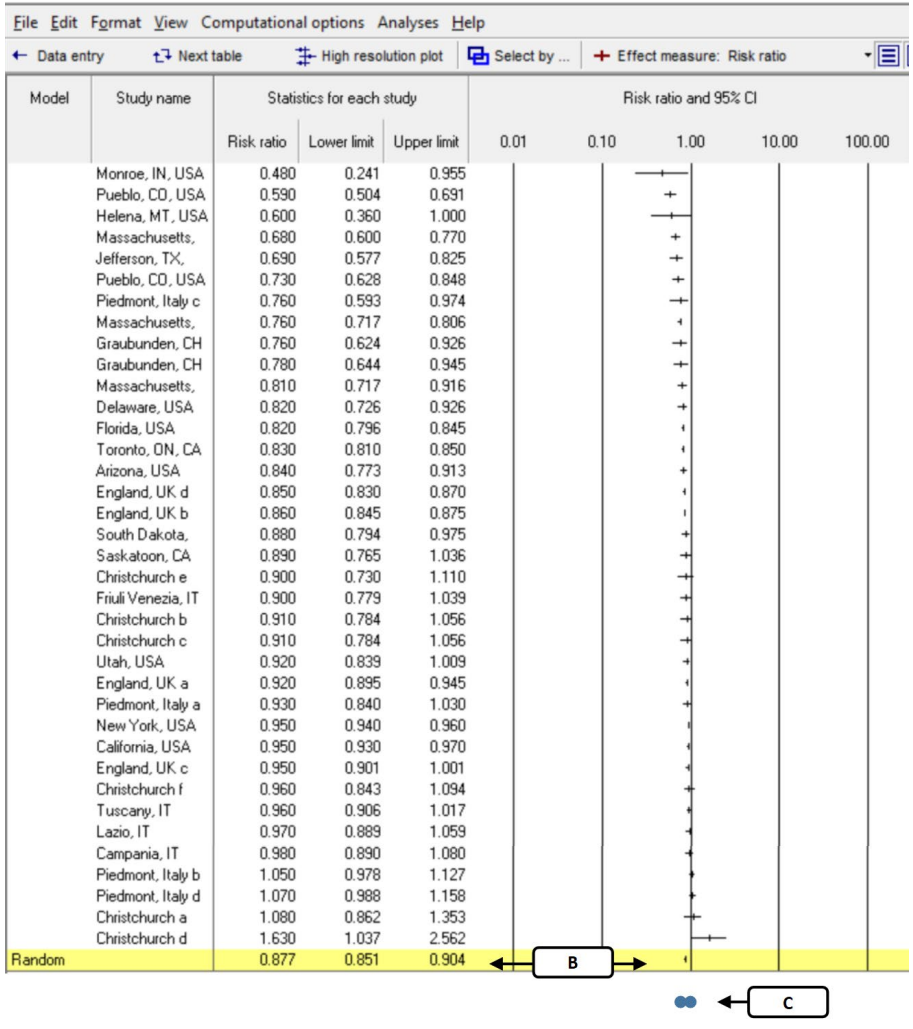

Figure 51 | Smoke-free legislation | Risk ratio < 1 indicates reduced risk

Figure 52 allows us to compare these two analyses. In this figure, the top plot corresponds to the statin analysis where the *Q*-value is 9.105 but there is substantial dispersion in effects. The bottom plot corresponds to the smoking analysis where the *Q*-value is 431.106 but the amount of dispersion is relatively small. Additional details are presented in Table 4.

It should be obvious from these examples that the *Q*-value (even when paired with its degrees of freedom) does not tell us how much the effect size varies across studies.

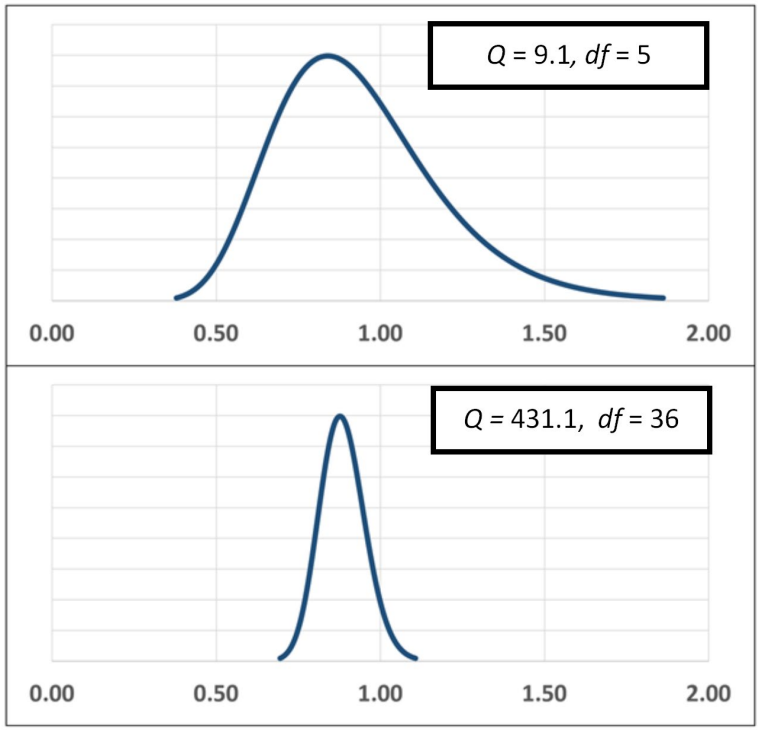

Figure 52 | Distribution of true effects for two meta-analyses

Table 4 | Heterogeneity in two analyses

Study	Index	Mean	*Q*	*df*	Prediction Interval
Statins	Odds ratio	0.74	9.1	5	0.32 to 1.71
Smoking	Risk ratio	0.88	431.1	36	0.75 to 1.02

9.8.5. *Q* does tell us one thing about the dispersion

The *Q*-value does provide one item of information about the heterogeneity. If *Q* is less than the degrees of freedom (the number of studies minus one), the variance will be estimated as zero. Conversely, if *Q* exceeds the degrees of freedom, the variance will be estimated as positive. However, that is the only information we can get directly from *Q* and the degrees of freedom. To press *Q* into service as an index of dispersion would be a mistake.

Summary

The *Q*-value for a test of heterogeneity is a function of (1) the amount of observed heterogeneity, (2) the precision of the individual studies, and (3) the number of studies in the analysis.

The *Q*-value may be large when the estimated heterogeneity is trivial. Conversely, the *Q*-value may be small when the estimated heterogeneity is substantial. Therefore, the *Q*-value should never be used as a surrogate for the amount of heterogeneity.

9.9. Estimates of variance may not be reliable

9.9.1. Mistake

In any random-effects analysis we compute an *estimate* of the between-study variance, and that estimate will differ from the true value. While researchers are aware of this in general, many do not recognize the potential severity of the problem.

9.9.2. Details

In the textbook case of a random-effects analysis we enumerate a universe of studies, sample studies from that universe, and generalize our results to that universe. The variance of true effects in that universe is called τ^2, where we use the Greek letter to indicate that this is the parameter (the true value). We can never see that value, but (in a frequentist analysis) we estimate it using the data in our sample, and the estimate is called T^2. It is important to recognize that T^2 does not always provide a reliable estimate of τ^2.

It might help to draw an analogy to a primary study employing a between-group design. Typically, this type of primary study reports the variance and standard deviation of scores based on a sample of at least 30 participants. In some fields the typical sample size is substantially higher, but it is generally not much lower than 30. If someone tried to publish a paper for a between-group design study based on a sample size of five subjects (for example), we would (rightfully) be concerned that the statistics were not reliable.

Suppose that we perform a random-effects meta-analysis using five studies with a hundred people in each. Researchers sometime assume that the effective sample size is five hundred people. In fact, however, the estimates of the mean and variance are based on an effective sample size of (less than) five. And, just as a sample size of five people will generally not yield a reliable estimate of the between-person variance in a primary study, a sample size of five studies will generally not yield a reliable estimate of the between-study variance in a meta-analysis.

The precision with which we can estimate τ^2 is a function of the true value of τ^2, of the number of studies in the analysis, and of the error variance in those studies. If all the estimation error variances are equal to V_M and the effects are normally distributed, the exact variance of the method of moments estimator of τ^2 is given by

$$\sigma^2_{\tau^2} = \frac{2(V_M + \tau^2)^2}{k-1}, \quad (9)$$

where V_M is the within-study error variance (assumed to be the same for all studies), τ^2 is the true between-study variance, and k is the number of studies. It follows that if V_M and/or τ^2 are non-trivial, the estimate of τ^2 will have poor precision unless we have a substantial number of studies.

The same issue applies to *all* the statistics that we employ to quantify heterogeneity, including T^2, T, I^2, and the prediction interval. Thus, we cannot mitigate this problem by switching to an alternate index. When we expect that the heterogeneity is non-trivial and we have a small number of studies, the best course of action is to report the extent to which our estimates are unreliable.

Ironically, while this lack of precision affects all the statistics, the practical implications of this problem are most serious for the prediction interval. Since researchers generally misinterpret the meaning of I^2 and T^2, if we estimate these values incorrectly, there is little additional harm done. By contrast, researchers do understand the prediction interval, and if this interval is wrong, researchers may reach the wrong conclusions. For this reason, it is probably best to report the prediction interval only if it is based on at least ten studies.

Summary

We need a reasonable number of studies to estimate heterogeneity reliably. If we don't have a sufficient number of studies, all heterogeneity statistics are suspect.

9.10. Statistics for heterogeneity refer to fixed-effect model

9.10.1. Mistake

Some computer programs report statistics for Q, I^2 and T^2, on the line for the fixed-effect analysis. Researchers sometimes assume that these statistics apply to the fixed-effect analysis, and then wonder where they can find these values for the random-effects analysis. This is a mistake.

9.10.2. Details

There is only one estimate for the Q-value reported in a meta-analysis. Based on this estimate we generate various statistics, some of which apply to the fixed-effect model and some of which apply to the random-effects model.

The p-value applies to the fixed-effect model. This model requires that all studies share a common effect size, and if the p-value is statistically significant we conclude that this assumption has been violated.

While the p-value applies to the fixed-effect model, all estimates of variance (T^2, T, and I^2) apply to the random-effects model. Importantly, these estimates apply *only* to the random-effects model, since under the fixed-effect model these are all zero *by definition*.

The reason that some computer programs display these statistics adjacent to the fixed-effect estimates is because the statistics are computed using a model where T^2 is zero, and this happens to correspond to the weights used for the fixed-effect model. The decision to display these statistics in one section or another is of no consequence.

9.10.3. Example | Serotonin-Aggression relation

Duke, Bègue, Bell, and Eisenlohr-Moul (2013) ran a meta-analysis looking at the Serotonin-Aggression relation in humans. They wrote "Mean weighted effect sizes are presented for both fixed-effects and random-effects models *with estimates of heterogeneity (Q and I^2 statistics) derived from the fixed-effects model* (Italics added)." The phrase in italics is misleading, and it would be better to omit this phrase.

Summary

Researchers sometimes expect that there is one set of heterogeneity statistics for the fixed-effect model and a separate set for the random-effects model. In fact, we compute only one set of statistics. These statistics are computed using fixed-effect weights, but some apply to the fixed-effect model and others to the random-effects model.

9.11. Putting it all together

When we ask about heterogeneity in a meta-analysis, our goal is to understand the clinical or substantive implications of the heterogeneity. We need to know the if the treatment's effect is relatively consistent across studies, or if it varies substantially. We need to know if the treatment is always helpful, or if it is helpful in some populations and harmful in others.

A case in point is the impact of methylphenidate on adults diagnosed with ADHD. The mean effect is a standardized mean difference of roughly 0.50, but to understand the potential utility of this drug we need to also know how much the effect size varies. When we ask about heterogeneity, we intend to ask if the distribution of effects resembles Figure 53, Figure 54, or Figure 55. Is it the case that –

A. The impact is as low as 0.40 in some populations, and as high as 0.60 in others (Figure 53).
B. The impact is as low as 0.30 in some populations, and as high as 0.70 in others (Figure 54).
C. The impact is as low as 0.10 in some populations, and as high as 0.90 in others (Figure 55).

When we discuss the utility of the drug, this is what we have in mind. Some might suggest that the drug should be recommended for general use only if the dispersion looks like Figure 53, while others might suggest that it should be recommended immediately even if the dispersion looks like Figure 54 or Figure 55. What should be clear, though, is that this discussion should be based on the dispersion represented in these figures.

The one statistic that directly addresses this dispersion is the prediction interval. In this example the prediction interval is 0.05 to 0.95. This tells us that the effect size varies from as low as 0.05 in some populations to as much as 0.95 in others (corresponding roughly to Figure 55). The prediction interval addresses this question using the same scale as the effect size, so the information is unambiguous. It tells us not only how much the effect size varies, but also reports the interval on a meaningful scale. Not only does it tell us that the effects vary over 90 points. It also tells us that it varies from 0.05 to 0.95 rather than (for example) −0.45 to +0.45 or 0.50 to 1.40.

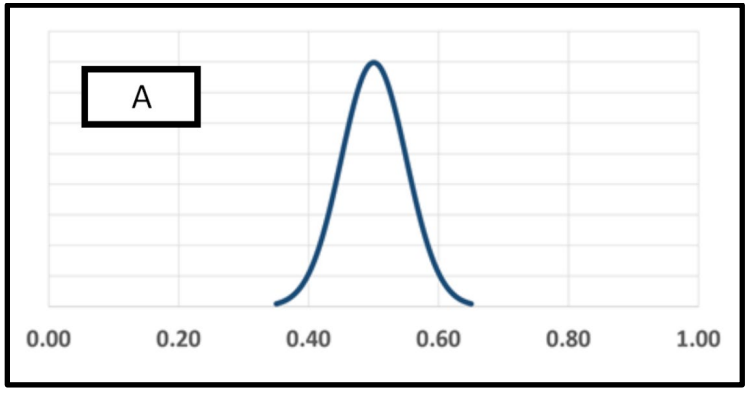

Figure 53 | Effect size varies from 0.40 to 0.60

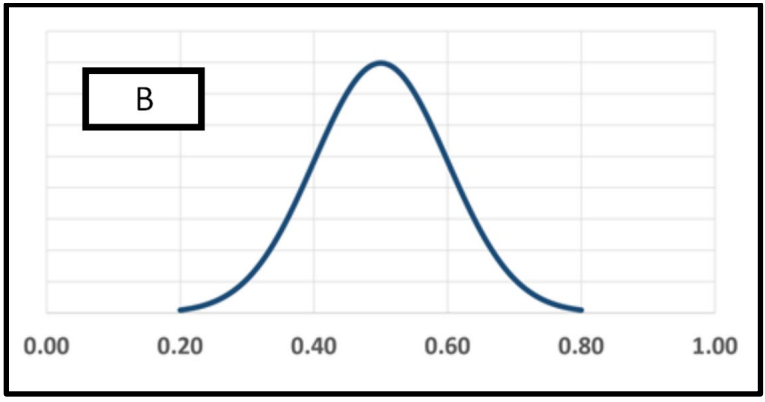

Figure 54 | Effect size varies from 0.30 to 0.70

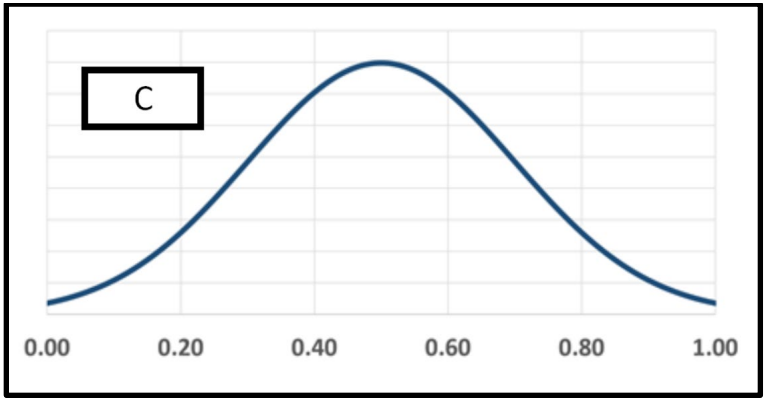

Figure 55 | Effect size varies from 0.10 to 0.90

Unfortunately, researchers rarely report the prediction interval. Rather, they typically report statistics such as Q, p, I^2, and T^2 which do not allow us to determine whether the dispersion looks like Figure 53, Figure 54, or Figure 55. Worse, researchers often push these statistics into service as surrogates for the amount of dispersion, and reach incorrect conclusions.

In some fields, the I^2 statistic has become ubiquitous as the preferred index of dispersion. This is a fundamental misinterpretation of this statistic. The I^2 statistic is a proportion, not an absolute value. It tells us what *proportion* of the observed variance reflects variation in true effects, rather than sampling error. It does not tell us how much that variance is. It makes no sense to make a recommendation about the drug based on the fact that I^2 is 47%, because that value could correspond to *any* of the three figures pictured, or to others.

This misuse of I^2 has been compounded by the fact that I^2 is commonly used to classify heterogeneity as being low, moderate, or high. This idea makes no sense for two reasons. First, the categories are based on I^2, which does not correspond to an absolute amount of dispersion. Second, the idea that we can classify heterogeneity as low, moderate, or high without additional context is silly, since an amount of heterogeneity that would be considered low in one context would be considered high in another.

Finally, it is important to recognize that estimates of T^2, and by extension estimates of all indices for heterogeneity, are often imprecise. It is probably best to report the prediction interval only when the analysis includes at least ten studies. While the imprecision affects all the indices, the practical implications of a mistake are potentially more serious for the prediction interval since this is an index that researchers would be using to make decisions.

When there is a sufficient number of studies to report a useful estimate of the prediction interval, we should report it. When we cannot report a useful estimate of this interval it would be best to omit it, and explain why.

10. MISTAKES RELATED TO SIGNIFICANCE TESTING

10.1. Overview

A common problem in the analysis of primary studies is that researchers sometimes focus on a test of statistical significance, and misinterpret the meaning of this test. The same problem applies in meta-analysis, with some additional complications. I will briefly review the problem as it applies to primary studies, and then discuss the extension of this issue to meta-analysis.

10.1.1. NHST vs. effect-size estimation in primary studies

When we perform a primary study to assess the impact of an intervention, the analysis can focus on either of two approaches.

- One is the null-hypothesis significance test (NHST). We pose the null hypothesis, that the effect size is precisely zero, and then perform a significance test. If the *p*-value is less than the criterion alpha (typically 0.05) we reject the null hypothesis and conclude that the true effect size is not precisely zero.
- The other is effect-size estimation. We report the mean effect size, and additionally the confidence interval which speaks to the precision with which we have estimated the effect size. In 95% of all studies the confidence interval will include the true effect size.

10.1.2. Cases where NHST is the preferred approach

There are cases where our intent really is to test the null hypothesis. For example, in a properly designed randomized trial to compare homeopathic compounds vs. placebo, our intent would be to test the null hypothesis that the two are equally effective. It would be important to know if one is *any* better than the other, since *any* difference greater than zero would challenge the fundamental tenets of science. In this case, the magnitude of the difference would be unimportant (Jonas, Kaptchuk, & Linde, 2003).

Another example where we really care about testing the null hypothesis is when we plan to submit the results to a regulatory agency in support of a new drug. If the criterion for approval is that we reject the null hypothesis, then we need to actually do so.

10.1.3. Cases where effect-size estimation is the preferred approach

By contrast to the above, in the overwhelming majority of analyses intended to identify the utility of an intervention, effect-size estimation is the preferred approach. The reason is simple. To assess the clinical or substantive utility of an intervention we need to know the magnitude of the effect, and not merely that the impact is not zero. Effect-size estimation addresses the former, while NHST is limited to the latter (Borenstein, 1994, 1997, 2000; Cohen, 1994; Sander Greenland et al., 2016; Harlow, 1997; Harlow, Mulaik, & Steiger, 1997; Schmidt & Hunter, 1997).

For example, suppose that we are testing the impact of tutoring for high-school students. If we were working with NHST and the test was statistically significant, we would reject the null hypothesis. This tells us *only* that the impact of the tutoring is not zero. By contrast, if we were working with effect-size estimation we would report (for example) that tutoring boosts scores by 10 points, or 20 points, or 30 points. *This* is the information that we need to assess the utility of the tutoring.

Or, suppose that we were assessing the impact of a new drug. If we were working with NHST and the test was statistically significant, we would reject the null hypothesis. This tells us *only* that the impact of the drug is not zero. By contrast, if we were working with effect-size estimation we would report (for example) that the drug reduces the risk of relapse by 5% or 10% or 20%. *This* is the information that we need to assess the utility of the drug.

10.1.4. Meta-analysis

When we move from primary studies to meta-analysis, the same issues apply. There are cases where the intent really is to test the null hypothesis of no effect, and in these cases, we should be using NHST. This would include a meta-analysis to assess the impact of homeopathy. It would also include cases where the meta-analysis is being used to gain regulatory approval, and the agency requires the use of NHST.

However, in the vast majority of cases where the goal of the analysis is to assess the impact of an intervention, we care about the magnitude of the

effect size, rather than a test of the null hypothesis. The issues are similar to those outlined for primary studies, but with some complications.

On the pages that follow, I address the following issues

- When the effect size is consistent across studies, we should almost always focus on estimating the mean effect size rather than testing the null hypothesis.
- When the effect size varies across studies, we should focus on estimating the mean effect size and additionally on estimating the *dispersion* in effect sizes. Here, estimating the mean effect size is not sufficient, and focusing on a test of the null hypothesis is especially problematic.

10.2. When the effect size is consistent across studies

10.2.1. Mistake

Researchers sometimes focus on the question of whether the analysis allows them to reject the null hypothesis of no effect. In most cases, the test of the null hypothesis is of limited relevance and the focus should be on effect-size estimation.

10.2.2. Details

Consider an analysis where the effect size is consistent across studies. In this case, the impact of the intervention is assumed to be essentially the same for all comparable populations, and so (for all intents and purposes) we are talking about a *common* effect size. (Strictly speaking, the effect size is only consistent across studies when all studies are estimating the same parameter. As a practical matter I am using the word *consistent* to apply to cases where the variance across effect sizes is small enough that the substantive or clinical impact of the variation is not important.)

The issues here are basically the same as those for a primary study. That is, the NHST addresses a question that we do not really care about (*Is the effect size precisely zero?*) while effect-size estimation addresses the question that we do care about (*What is the magnitude of the effect?*)

10.2.3. Example | Tamiflu

A case in point is the systematic review by Jefferson et al. (2014) which assessed the utility of Tamiflu for alleviating symptoms due to the flu (Figure 56). Patients with the flu were randomly assigned to receive either Tamiflu or a placebo, and researchers tracked the number of hours until the patients started to feel relief of symptoms.

The results of the analysis are shown in Figure 56. The effect size is the raw mean difference (in hours). An effect size of zero would indicate no difference between groups. An effect size to the *left* of zero would indicate that the treated group reported relief (on average) *sooner* than the control group. An effect size to the *right* of zero would indicate that the treated group reported relief (on average) *later* than the control group. The impact of the treatment was consistent across studies (the estimate of τ^2 is zero), so we can talk about a *common* effect size.

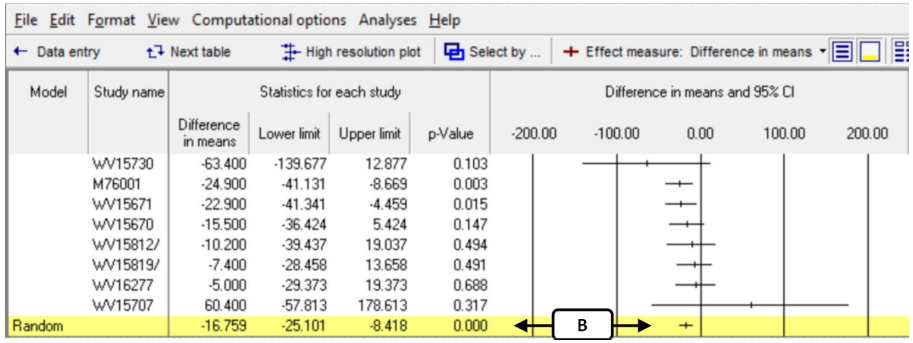

Figure 56 | Impact of Tamiflu | Raw mean difference < 0 favors Tamiflu

The difference between groups was statistically significant. The Z-value for a test of the difference is -3.938 with a corresponding p-value of < 0.001. So, we can conclude with a high level of certainty that Tamiflu *did* reduce the mean time until people began to feel a relief of symptoms.

For purposes of obtaining approval from a governmental agency, the p-value might be the key statistic. However, from the point of view of a researcher or clinician, it would be a mistake to focus on this p-value. Rather, we also need to know the clinical utility of this intervention. The mean difference between groups was 16.759 hours, and the reviewers use this to assess the clinical utility of the drug. They point out that the mean time to relief for the control group was about a week, and that a difference of 17 hours should be seen in this context.

For recent developments in the discussion of NHST, see (Amrhein, Trafimow, & Greenland, 2019; Sander Greenland, 2019; Trafimow, 2019; R. L. Wasserstein & Lazar, 2016; Ronald L. Wasserstein, Schirm, & Lazar, 2019).

Summary

When the effect size is consistent across studies, the situation in a meta-analysis is basically the same as it is in a primary study. When the analysis concerns the utility of an intervention, the NHST paradigm addresses a question that we do not really care about (*Is the effect size precisely zero?*) while effect-size estimation addresses the question that we do care about (*What is the magnitude of the effect?*)

For decades, people in research had spoken about the *controversy* between NHST and effect-size estimation. This was the subject of hundreds of papers and several books. Today, there is widespread consensus that we should generally focus on effect-size estimation when evaluating the impact of an intervention. While guidelines now recommend (or mandate) this shift for most analyses, the transition in primary studies has been slow. In many cases, the discussion section of a paper is still driven by the *p*-value rather than the size of the effect (Rothman, 2010, 2016; Stang, Deckert, Poole, & Rothman, 2017; Stang, Poole, & Kuss, 2010; VanderWeele, 2010a, 2010b).

In the case of meta-analysis, the focus on effect-size estimation should be more natural, since the meta-analysis is built around the size of the effect. Nevertheless, there are still many cases where the researchers revert to a focus on the *p*-value. We need to avoid what Rozeboom (1960) called this *primitive tendency* of focusing on a test of the null hypothesis, and focus instead on the size of the effect.

10.3. When the effect size varies across studies

10.3.1. Mistake

When the effect size is consistent across studies, the issues in a meta-analysis are similar to those in a primary study. A test of the null hypothesis addresses an issue that we do not really care about, while an estimate of the effect size addresses the issue that we do care about. The shift from the former to the latter solves the problem. By contrast, when the effect size varies across studies, the situation in a meta-analysis is more complicated.

10.3.2. Details

Both approaches, null hypothesis significance testing (NHST) and effect-size estimation, focus exclusively on the *mean* effect size. One asks *What is the mean effect size?* while the other asks *Is the mean effect size zero?* but they both focus on the mean.

When there is only one effect size, this makes sense. By contrast, when the effect size varies across studies, the mean will be of limited importance. In this case, we need to focus also on the dispersion in effects.

10.3.3. The null hypothesis may not apply to any specific population

When we are working with a single primary study, there is only one population involved, and it makes sense to test the null hypothesis that the effect size in that population is zero. Similarly, when we are working with a fixed-effect analysis we are working with one population, and it makes sense to test the null hypothesis that the effect size in that population is zero.

By contrast, consider what happens when we move to a random-effects analysis and the true effect size varies across studies. If we reject the null hypothesis, we know that the *mean* effect size is not zero, but the mean is not representative of the actual effect size in any population. Consider a case where the intervention is helpful in some populations but harmful in others. What does it tell us that the *mean* effect size is (or is not) significantly different from zero?

10.3.4. The null hypothesis applies to a specific mix of populations

A second issue is that when the true effect size varies across studies, the mean effect size in a meta-analysis will depend on the mix of populations that are included in that analysis. For example, if the effect size tends to be larger in studies that employed a higher dose of a drug, the effect size will shift to the left if most studies employed a relatively low dose, and will shift to the right if most studies employed a relatively high dose.

In this context, we need to consider what null hypothesis is actually being addressed by the test of significance. The test addresses the null hypothesis that the mean effect size in this specific mix of populations (and the universe of comparable populations) is zero. Since the mix of populations is (a) somewhat arbitrary, and (b) not well defined, it is not always clear what a test of the null hypothesis can tell us. Put simply, if the mean effect size for this mix of populations is not zero, but the mean effect size for an alternate mix of populations may be zero, why do we care about *this* mix and not some other?

10.3.5. Example | ADHD

Consider the ADHD example, discussed earlier (Figure 57). Recall that the mean effect size was 0.506 [B], but the effect size varied from as low as 0.05 in some populations to as high as 0.95 in others [C]. In this case, the mean effect size obviously depends on the specific mix of populations included in the analysis. Depending on that mix, the mean could move left or right.

For example, suppose that the effect size tends to be higher in studies that employed a larger dose of the drug. Suppose further that the inclusion/exclusion criteria limit the analysis to studies that employed a dose in the range of 30 to 80 mg. While the inclusion/exclusion criteria may define a range of doses, it will not generally specify the *proportion* of studies with each dose. If the analysis includes primarily studies that employed a dose near 30 mg, the mean effect might be 0.30, whereas if the analysis includes primarily studies that employed a dose near 80 mg, the mean effect might be 0.80. As such, the null hypothesis is being tested for a universe of populations that is not well defined, and in any event is of no particular interest.

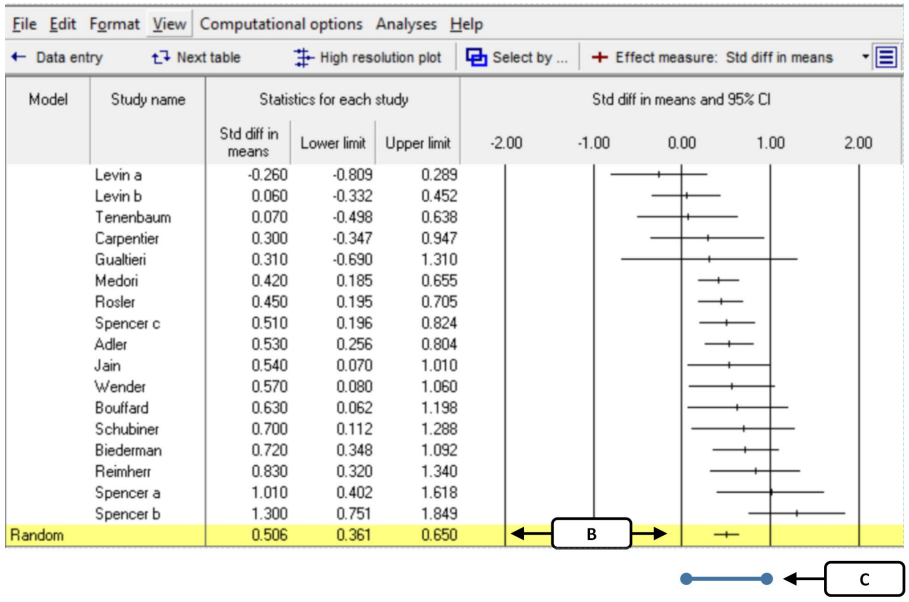

Figure 57 | ADHD Analysis – Forest plot

10.3.6. Example | Clozapine

Or, consider the analysis by (Taylor et al., 2012) which looked at the impact of augmenting clozapine with a second antipsychotic (Figure 58). The effect size index is a standardized mean difference, with scores below zero indicating an improvement. The authors report that the mean effect size is −0.239, with a 95% confidence interval of −0.452 to −0.026 [B], so *on average* the treatment improves outcome by around one-fourth of a standard deviation. Additionally, they characterize this as a small benefit. All of this is correct, and if the impact of treatment had been consistent across studies, these numbers would indeed capture the magnitude of the effect. (These results are for the random-effects model, which was incorrectly labeled in the original paper.)

However, the prediction interval is −0.84 to +0.34 [C], which means that there are some populations where the treatment *improves* response by 0.84 standard deviations and others where it *hurts* response by 0.34 standard deviations.

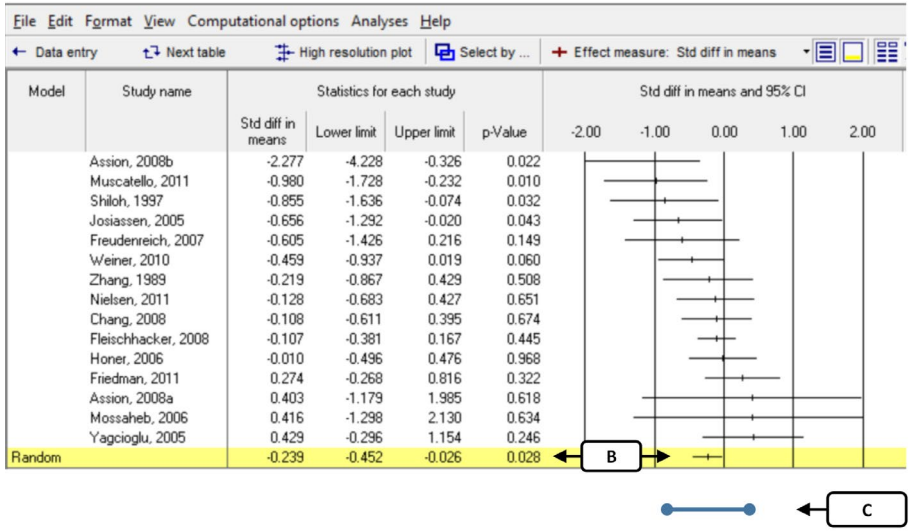

Figure 58 | Augmenting clozapine | Std mean difference < 0 favors augmentation

Here, the NHST framework is *especially* problematic. When the effect size varies this much, the mean effect should be of secondary interest. The focus of the report should be that the treatment is very effective in some populations, moderately effective in others, and harmful in others. In this situation, it is not clear why we would want to ask if the treatment is helpful "on average", and so the use of NHST is *especially* difficult to justify.

10.3.7. Example | Juvenile Drug Courts

In some parts of the United States, juveniles who have been charged with crimes related to illegal drugs may be tried in criminal court or in drug court. The drug court is a separate court where judges have more latitude than the judges in criminal courts. For example, they may be able to sentence the defendant to community service rather than prison. Tanner-Smith, Lipsey, and Wilson (2016) compared the impact of juvenile drug courts vs. standard courts for preventing recidivism. An odds ratio greater than 1.0 indicates that the drug courts did better, with juveniles more likely to stay out of trouble.

The *p*-value for a test of the null hypothesis is 0.578 [B], so there is no evidence that the drug courts are effective *on average*. However, it is clear from the plot that the effect varies substantially across studies. As indicated by the prediction interval [C], there are some studies where the drug courts do better than the controls, some where they do as well as the controls, and some where they do worse than the controls.

If the treatment is helpful in some cases and harmful in others, the question of whether the mean effect size is zero, is largely irrelevant. Unlike in a primary study, simply switching to effect-size estimation is not sufficient since this still addresses only the mean. Rather, the focus must shift to the *dispersion* in effects. We would report that the drug courts *increased* the odds of recidivism by 77% in some populations, while in others they *reduced* it by 73% [C]. In this context, the mean effect size is only a footnote.

In this example, the authors discussed the heterogeneity in effects, and then looked for relationships between the effect size and various risk factors. This is the correct approach.

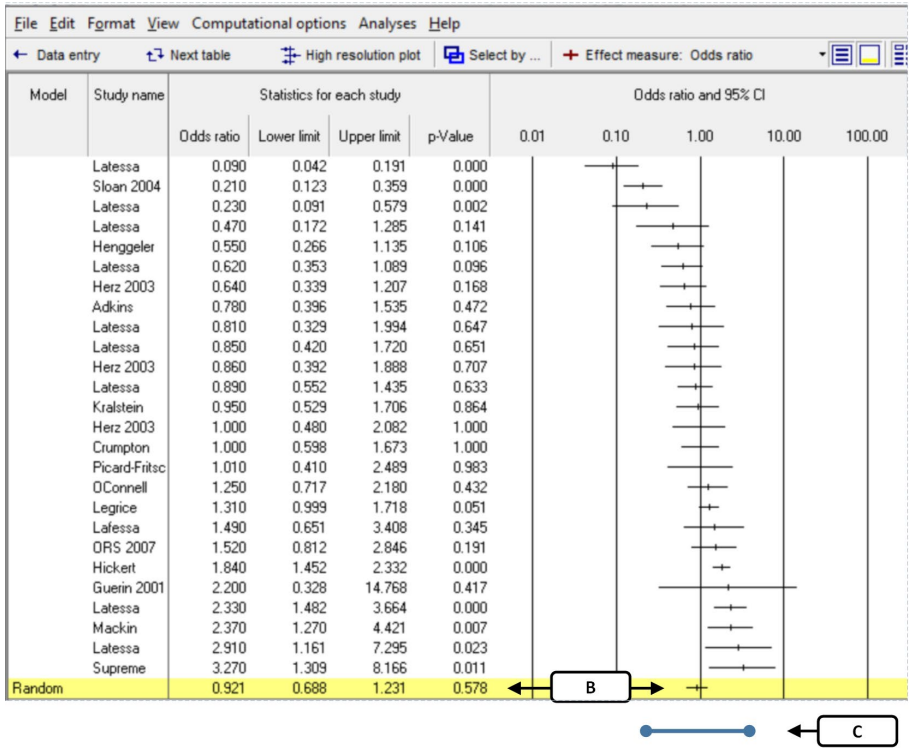

Figure 59 | Drug courts vs. standard courts | Odds ratio > 1 favors drug courts

10.3.8. In context

What distinguishes the Tamiflu analysis in section 10.2.3 (on the one hand) from the ADHD analysis, the Clozapine analysis, and the Drug Courts analysis (on the other) is the following. If the effect size is consistent across studies, as it was in the Tamiflu analysis, the mean effect size applies to each

population. Therefore, it makes sense to focus on the mean effect. By contrast, if the effect size varies substantially, as it does in the other analyses, the mean is not the effect size in *any* population (let alone *all* populations). Therefore, we need to report not only the mean effect size, but also the variation in effects. And, we need to be clear that the mean applies to the specific mix of populations included in the analysis, which may not be the same as the universe of populations to which we had intended to make an inference.

All of this applies whether or not the confidence interval for the *mean* effect excludes the null effect size. In the ADHD analysis we conclude that methylphenidate is effective *on average*, but we still need to address the fact that the impact may be trivial in some populations and substantial in others. In the clozapine example we conclude that the augmentation is effective *on average*, but we still need to address the fact that the effect varies (and in some populations may be harmful). In the Drug Courts analysis we cannot conclude that the courts are effective *on average*, but we still need to address the fact that the impact varies. If the intervention is effective in some populations and harmful in others, the mean is of little relevance.

Summary

When the effect size varies from one population to the next, the mean effect size and test of the null hypothesis will depend on the specific mix of populations included in the analysis. Since this mix is somewhat arbitrary, it is not always clear why we should care about this specific null hypothesis.

Additionally, when the effect size varies across populations, the effect size in any given population may fall some distance from the mean effect size, and so the mean (and a test of the null hypothesis that the mean is zero) may have very limited utility.

If we report that the effect is statistically significant, the take-home message is that the treatment works. However, if the treatment is helpful in some cases and harmful in others, this is not the message we want to be sending.

Similarly, if we report that the effect is *not* statistically significant, the take-home message is that the treatment may not work. Again, if the treatment is helpful in some cases and harmful in others, this is not the message we want to be sending.

Rather, if the treatment effect varies substantially, we need to shift our focus away from the mean effect size and focus on the dispersion in effect size. The test of significance deals only with the mean, and thus puts our focus on the wrong issue.

10.4. Significant effect may be harmful in some populations

10.4.1. Mistake

Researchers sometimes assume that if a treatment effect is clinically helpful and statistically significant, that treatment will be helpful in all populations. This is a mistake.

10.4.2. Details

Researchers sometimes ask, "If the treatment is helpful and the effect is statistically significant, how is it possible that there are studies where the treatment is harmful?" The answer is that the statistical significance refers only to the *mean* effect size. The treatment is helpful *on average*, but there could be some populations where the treatment is harmful.

It may be helpful to draw an analogy to a primary study where we assess the math score for all students in a class. We report that the mean score is 50 with a confidence interval of 40 to 60. We also report that there are some students who score as low as 10. The fact that the confidence interval is 40 to 60 speaks only to the *mean* effect and says nothing about the distribution of scores. We understand that the class mean can be 50, and there may still be students who score as low as 10 points (or even lower). By analogy, the *mean* impact of an intervention across all populations in a meta-analysis may be 0.50 with a confidence interval of 0.40 to 0.60. It is still entirely possible that the effect size *in any single population* could be less than zero.

While this answer fully addresses the question of how the intervention can be harmful in some populations when the mean effect size is positive and statistically significant, we need to recognize that *the intuition behind the question is correct*. People who ask this question are actually having an "Aha" moment, and recognizing a key problem with the way that significance tests are used.

Summary

The significance test addresses the *mean* effect size. If the effect size is consistent, the effect in all populations is the same, and so it makes sense to focus on the mean. By contrast, if the effect size varies, the effect in each population is unique. While the treatment may be helpful *on average*, it could be harmful in any given population. Therefore, when the effect size varies across studies, it is imperative that we consider the extent and implications of the dispersion.

10.5. Putting it all together

In primary studies, tests of statistical significance address the question "Is the effect size precisely zero?" whereas effect-size estimation addresses the question "What *is* the effect size?" While there are some circumstances where we need to test the null hypothesis of no effect, in the overwhelming majority of cases we are concerned with the second question, and so this is the approach that we should be using.

A meta-analysis where the effect size is consistent across studies is similar to the primary study in this regard. While there are specific circumstances where we need to test the null hypothesis of no effect, in the overwhelming majority of cases we should focus on estimating the size of the effect. Almost all researchers would agree with this point in the abstract, but sometimes revert to the significance test in practice. In presenting and/or discussing the results, it would be helpful to focus consistently on the size of the effect and the clinical or substantive implications of that effect.

When the between-study variance is non-trivial, the situation is more complicated. Here, the mean effect size (and the test of the null) will be affected by the specific mix of populations that happen to be included in the analysis. Put bluntly, whether we can reject the null hypothesis may depend on the particular mix of populations included in the analysis. Since this mix is somewhat arbitrary, it is not always clear why we would want to estimate the mean or test the null hypothesis.

Additionally, when the effect size varies substantially across populations, the effect size in any given populations may fall some distance from the mean, and so the mean has limited utility. Rather, we need to focus on the dispersion in effects.

If we report that the effect is statistically significant, the take-home message is that the treatment works. If the treatment is helpful in some cases and harmful in others, this is not the message we want to be sending. Similarly, if we report that the effect is *not* statistically significant, the take-home message is that the treatment may not work. Again, if the treatment is helpful in some cases and harmful in others, this is not the message we want to be sending. Rather, if the treatment effect varies substantially, we need to shift our focus away from the *mean* effect size and focus on the dispersion in effect size.