Original Article

Research Synthesis Methods

Received 17 July 2015, Revised 12 September 2016, Accepted 30 September 2016 Published online 6 January 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1230

Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity

Michael Borenstein,[a]* Julian P. T. Higgins,[b] Larry V. Hedges[c] and Hannah R. Rothstein[d]

When we speak about heterogeneity in a meta-analysis, our intent is usually to understand the substantive implications of the heterogeneity. If an intervention yields a mean effect size of 50 points, we want to know if the effect size in different populations varies from 40 to 60, or from 10 to 90, because this speaks to the potential utility of the intervention. While there is a common belief that the I^2 statistic provides this information, it actually does not. In this example, if we are told that I^2 is 50%, we have no way of knowing if the effects range from 40 to 60, or from 10 to 90, or across some other range. Rather, if we want to communicate the predicted range of effects, then we should simply report this range. This gives readers the information they think is being captured by I^2 and does so in a way that is concise and unambiguous.

Keywords: I-squared; I2; I^2; heterogeneity; meta-analysis; prediction intervals; inconsistency

1. Introduction

The goal of a meta-analysis is not simply to report the mean effect size but also to report how the effect sizes in the various studies are dispersed about the mean. To report that an intervention increases scores by 50 points is only part of the picture. We need to know also if the impact is consistent, varies moderately, or varies widely, from study to study.

Researchers often use the I^2 statistic to quantify the amount of dispersion (Higgins and Thompson, 2002; Higgins *et al*., 2003). I^2 is an intuitive statistic for many reasons. It ranges from 0% to 100%, so we have a clear sense of where the heterogeneity in any given meta-analysis falls, relative to this range. The range is independent of the specific effect size and so has the same meaning for a meta-analysis of odds ratios as it does for a meta-analysis of mean differences. I^2 is largely unaffected by the number of studies in the meta-analysis, and so allows us to compare the I^2 for different analyses even if the number of studies differs. Most computer programs report I^2, and so it is readily available.

Additionally, there are widely used benchmarks for I^2. For example, I^2 values of 25%, 50%, and 75% have been interpreted as representing small, moderate, and high levels of heterogeneity. These are seen to provide a convenient context for discussing the results of any analysis. For these reasons, the use of I^2 as the primary basis for discussing how much heterogeneity is present and the use of benchmarks for interpreting the magnitude of heterogeneity have become ubiquitous in meta-analysis.

Unfortunately, the use of I^2 in this way is inappropriate. It represents a fundamental misunderstanding of what I^2 is and how it should (and should not) be used. Our goal in this paper is to explain what I^2 is, how to interpret it, and why its common use is fundamentally wrong. In place of I^2, we will discuss indices that *do* report the dispersion of true effects on an absolute scale. These are the indices that actually address the questions that people *think* are being addressed by I^2.

The intended audience for this paper is researchers, rather than statisticians. Therefore, our approach is primarily conceptual rather than mathematical.

[a]*BioStat, Inc., Englewood, NJ, USA*
[b]*School of Social and Community Medicine, University of Bristol, Bristol, UK*
[c]*Department of Statistics, Northwestern University, Evanston, IL, USA*
[d]*Department of Management, Baruch College–City University of New York, New York, NY, USA*
**Correspondence to: Michael Borenstein, BioStat, Inc., Englewood, NJ, USA.*
E-mail: Biostat100@GMail.com

2. Motivating example

We will use the attention-deficit hyperactivity disorder (ADHD) analysis (Castells *et al.*, 2011) as a motivating example. ADHD is a condition where people have trouble focusing on tasks. This is often treated with the drug methylphenidate, and this meta-analysis is a synthesis of studies where adults with ADHD were randomly assigned to receive either methylphenidate or placebo.

In this analysis, the effect size is the standardized mean difference (*d*) between the treated and control groups on a cognitive task. The mean *d* is 0.50, but we also want to know how the effect size varies across populations.

A simple thought experiment will make it clear that I^2 does not provide this information. In this analysis, I^2 is 47%. What does this tell us about the variation in effect size? Do the effects range from 0.40 to 0.60, or from 0.30 to 0.70, or across some other range? We do not know. Suppose we are told that an I^2 of 47% corresponds to a "moderate" level of heterogeneity. What is a "moderate" amount of heterogeneity in this context? We still do not know.

The fact is that I^2 cannot, and was never intended to, provide this kind of information. That will become clear when we explain what I^2 is. To do so, we need to provide some context.

3. True effects versus observed effects

In a primary study with one level of sampling, we typically treat the observed scores as identical to the true scores. For example, consider a study where we enroll a sample of students and record their scores on a test. If a student scores 40 on the test, for purposes of the analysis, we proceed as though the student's true score is 40. It follows that there is no distinction between the distribution of true scores and the distribution of observed scores. If we want to know how the scores are distributed, we compute the standard deviation of the scores. If we assume that the scores are normally distributed, then most scores fall within two standard deviations on either side of the mean.

By contrast, in a meta-analysis, we need to distinguish between an observed effect size and a true effect size in any given study. The *observed* effect size in a study is the effect size that we see in that study. It serves as the estimate of the effect size in that study's population, but invariably differs from the true effect size in that population due to sampling error. By contrast, the *true* effect size for a given study is the actual effect size in the study's population. It is the effect size that we would see if we conducted a study in that population with an infinitely large sample size, and (it follows) no sampling error.

Figure 1 displays two plots for the ADHD analysis. The left-hand plot shows the *observed* effects, while the right-hand plot shows an example of how the *true* effects might be distributed. The standard deviation of observed effects is approximately 0.30, as reflected in line [A]. This line represents an interval that extends two standard deviations on either side of the mean (−0.10 to + 1.10), a span of 120 points. By contrast, the standard deviation of the true effects is approximately 0.20, as reflected in line [B]. This line represents an interval that extends two standard deviations on either side of the mean (0.10 to 0.90), a span of 80 points.

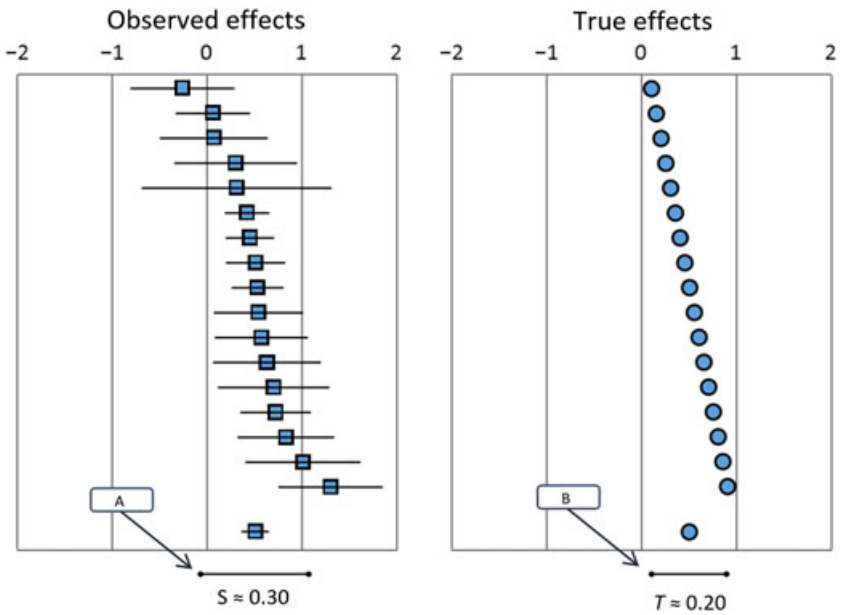

Figure 1. Observed effects for the ADHD analysis (at left) and true effects (at right). The line underneath each plot represents the mean effect plus/minus two standard deviations. The standard deviation of observed effects is around 0.30, while the standard deviation of true effects is around 0.20. The true effects do not correspond to any actual studies, but are simply meant to show one set of possible effects with a standard deviation of 0.20.

To understand why the standard deviation of the observed effects [A] is wider than the standard deviation of the true effects [B], consider what would happen if the true effect size was identical in all studies. While the *true* effects (at right) would all be the same, the *observed* effects (at left) would vary because of sampling error. Concretely, the expected variation of the observed effects (V_{OBS}) would be equal to the (average) error variance of the individual studies (V_{ERR}). That is,

$$V_{OBS} = V_{ERR}.$$

In the ADHD analysis, the average V_{ERR} is 0.0439. If the true effect for all studies was precisely 0.50, the expected value of the variance at left (V_{OBS}) would be 0.0439.

While this idea is most intuitive when the true effects are all identical to each other, the same idea holds true when the true effects vary. In this case, the expected variation of the observed effects is equal to variance of the true effects (T^2) plus the (average) error variance of the individual studies. That is,

$$V_{OBS} = T^2 + V_{ERR}.$$

In the ADHD analysis, where T^2 is 0.0387 and the average V_{ERR} is 0.0439, the variance of observed effects is

$$V_{OBS} = 0.0387 + 0.0439 = 0.0825.$$

These estimates of the variance for each plot lead directly to lines [A] and [B]. For the left-hand plot, the variance is 0.0825, so the standard deviation (the square root of the variance) is 0.2872, which we rounded to 0.30. Two standard deviations on either side of the mean yield a range of −0.10 to +1.10. For the right-hand plot, the variance is 0.0387, so the standard deviation is 0.1967, which we rounded to 0.20. Two standard deviations on either side of the mean yield a range of 0.10 to +0.90.

4. What is I^2?

Once we recognize that the variance of observed effects incorporates two distinct elements – the variation of true effects and variation due to sampling error, we might want a statistic that addresses the relationship between these components. This statistic is I^2, defined as

$$I^2 = \frac{V_{TRUE}}{V_{OBS}} = \frac{T^2}{V_{OBS}}.$$

Thus, I^2 is a proportion. It deals with the left-hand plot and tells us that proportion of the variance in this plot reflects variation in true effects. In the ADHD analysis,

$$I^2 = \frac{0.0387}{0.0825} = 47\%.$$

With reference to Figure 1, one way to think about I^2 is that it is based on the comparison of (1) the dispersion of observed effects and (2) the dispersion we would expect based on sampling error alone. Another way to think about I^2 is that it reflects the amount of non-overlap among confidence intervals. In the Appendix, we present a series of examples to illustrate these aspects of I^2. For purposes of the present discussion, we will focus on another way of thinking about I^2 as follows.

If I^2 tells us what proportion of the variation in observed effects is due to variation in true effects, then (by definition) it tells us what proportion of this variation would remain if we could somehow get rid of the sampling error. As such, I^2 serves as a bridge between the left-hand plot and the right-hand plot. If we start with the variance of observed effects, and multiply it by the proportion that reflects variance in true effects (I^2), we obtain the variance of true effects, at right. That is,

$$T^2 = V_{OBS} \times I^2.$$

As such, I^2 provides us with a context for describing and interpreting the plot of observed effects. If I^2 is near zero, then most of the observed variance would disappear if we were looking at the true effects. If I^2 is near one, then most of the observed variance would remain. In the ADHD analysis, I^2 is 47%, which tells us that 47% of the variance would remain.

Our key point is that I^2 is a proportion and not an absolute value. As such, it *cannot tell us* how much the effects vary. If the question is "How much do the effects vary," then the answer must be in the form of absolute values, for example, "They vary from X_1 in some populations to X_2 in other populations." The answer cannot be in the form of a proportion, such as "They vary 50%." It is not even clear what that means.

5. Prediction intervals

So, how do we compute and report these absolute values?

When we are working with a primary study, we compute the standard deviation of the scores. If the scores are normally distributed, some 95% of scores will fall within two standard deviations on either side of the mean. If the mean is 50 and the standard deviation is 20, then most scores will fall in the range of 10 to 90. This range is called a prediction interval, defined as

$$Prediction\ Interval = Mean \pm 2S,$$

where S is the standard deviation of the scores. This range is called a prediction interval, because if we were asked to predict the effect size for any one subject (randomly sampled from the population), we would predict that the score would fall in this range. And we would be correct some 95% of the time.

We can take the same approach with a meta-analysis. We need to work with the standard deviation of the true effects (T) as reflected in the right-hand plot, but the idea is the same as for a primary study. If the effects are normally distributed, we expect that the effect size in some 95% of all populations will fall within two standard deviations on either side of the mean. This range is called a prediction interval, because if we were asked to predict the effect size for any one population (randomly sampled from the same universe as those included in the meta-analysis), we would predict that the effect size would fall in this range. And we would be correct some 95% of the time.

5.1. Prediction interval for means

When the effect size is a mean difference or a risk difference, we can compute the 95% prediction interval using

$$Prediction\ Interval = Mean \pm 2T,$$

where T is the standard deviation of true effects. In the ADHD analysis (using rounded numbers), the mean effect size is a d of 0.50 and T is 0.20. The prediction interval is given by

$$LL = 0.50 - 2\times 0.20 = 0.10$$
$$UL = 0.50 + 2\times 0.20 = 0.90.$$

This tells us that some 95% of all populations will have an effect size in the range of 0.10 to 0.90.

5.2. Prediction interval for ratios

When the effect size is not a mean difference, the situation is a bit more complicated. For example, when the effect size is a risk ratio, the risk ratio is reported as a ratio, but the standard deviation (T) is reported in log units. To compute the prediction interval, we need to convert all numbers into log units, compute the interval, and then convert the numbers back into ratio units. The following study serves as an example.

Tsertsvadze *et al*. (2009) performed a meta-analysis of 19 studies that evaluated the impact of Viagra on sexual function. Outcome was the patient's report that he was (or was not) satisfied, and the effect size index is the risk ratio. In round numbers, the mean effect size is 2.50, which means that (on average) patients treated with Viagra were 2.5 time as likely to report satisfaction as compared with patients treated with placebo. In log units, the mean effect size is 0.92, and T is 0.15. The prediction interval in log units is given by

$$LL = 0.92 - 2\times 0.15 = 0.62$$
$$UL = 0.92 + 2\times 0.15 = 1.22.$$

We then convert these values to risk ratios (see Appendix for details), which gives us 1.86 and 3.39. We expect that in some 95% of all populations, the "risk" ratio for reporting satisfaction will fall in the approximate range of 1.86 to 3.39.

5.3. Prediction interval for prevalences

Similarly, when the effect size is prevalence, the analysis may be performed using a logit transformation. In this case, prevalence is reported as a proportion, but T is reported in logit units. To compute the prediction interval, we need to convert all numbers into logit units, compute the interval, and then convert the numbers back into proportions. The following study serves as an example.

Cabizuca *et al*. (2009) performed a meta-analysis to synthesize data from eleven studies that reported prevalence of post-traumatic stress disorder (PTSD) in mothers of children with chronic illness or undergoing invasive procedures. The effect size index is prevalence, and the mean prevalence is 0.180. In logit units, the mean prevalence is −1.52, and T is 0.59. The prediction interval in logit units is given by

$$LL = -1.52 - 2\times0.59 = -2.70$$
$$UL = -1.52 + 2\times0.59 = -0.34.$$

We then convert these values to prevalence units, which gives us 0.06 and 0.42 (see Appendix for details). We expect that in some 95% of all populations, the true prevalence will fall in the approximate range of 6% to 42%.

5.4. *Prediction interval for correlations*

Finally, when the effect size is a correlation, the analysis may be performed using Fisher's *Z*-transformation. In this case, the correlation is reported as a correlation, but *T* is reported in Fisher's *Z* units. To compute the prediction interval, we need to convert all numbers into Fisher-*Z* units, compute the interval, and then convert the numbers back into correlations. The following study serves as an example.

Wright and Bonett (2002) performed a meta-analysis to synthesize data from 27 studies that reported the correlation between attitudinal commitment and job performance. The effect size index is the correlation, and the mean correlation is 0.175. In Fisher's *Z* units, the mean correlation is 0.177, and *T* is 0.119. The prediction interval in Fisher's *Z* units is given by

$$LL = 0.177 - 2\times0.119 = -0.061$$
$$UL = 0.177 + 2\times0.119 = 0.415.$$

We then convert these values to correlations, which give us −0.06 and 0.39 (see Appendix for details). We expect that in some 95% of all populations, the true correlation will fall in the approximate range of −0.06 to 0.39.

5.5. *Adjustments to prediction intervals*

This formula for the prediction interval works well when the estimates of the mean and the standard deviation are reasonably precise. In the Appendix, we show how the formula can be modified to take account of the fact that these parameters are estimated with error.

5.6. *Prediction intervals and confidence intervals*

To avoid confusion, note that the prediction interval is not the same as a confidence interval. The confidence interval is an index of precision (based on the standard *error*) that tells us how precisely we have estimated the mean effect size. As such, it is a property of the sample and strongly driven by the number of studies in the analysis. By contrast, the prediction interval is an index of dispersion (based on the standard *deviation*) that tells us how widely the effects vary across populations. As such, it is a property of the universe of populations and (it follows) not related to the number of studies in the analysis.

When we report the prediction interval, we focus attention on the issue that we really care about, when we ask about heterogeneity. We do so in a way that is unambiguous and concise. And, we are not only reporting the extent of dispersion but we are also reporting this in the context of the mean. In the ADHD example, the effects vary over 80 points. It is critical that the 80 points range from 10 to 90 (which we might call a trivial beneficial effect to a very strong beneficial effect) and not from minus 20 to plus 60 (a harmful effect to a modest beneficial effect).

6. The genesis of I^2

To summarize, I^2 is a proportion. It tells us what proportion of the variance in the left-hand plot is due to variation in real effects rather than sampling error. By contrast, *T* is an absolute value. It tells us how the effects in the right-hand plot are distributed. When we ask about heterogeneity, it is clearly the latter that we care about. So, how did we arrive at a situation where researchers focus on I^2 and use it as a surrogate for *T*?

The answer can be traced to the computational issue we introduced earlier. When we are working with a mean difference (as we are in the ADHD example), *T* is reported in the same metric as the effect size. We understand what a standard deviation of 0.20 means. And, we understand implicitly that if the mean effect size is 0.50 and *T* is 0.20, most effects will fall in the range of 0.10 to 0.90. It is likely that if all meta-analyses were based on means, researchers would have focused on *T*.

However, the situation is more complicated when we are working with other effect sizes. When we are working with a risk ratio or an odds ratio, *T* is reported in log units. Few people have an intuitive sense of what a standard deviation of 0.20 (for example) means in log units. The same idea holds true when the effect size is an estimate of prevalence, and the standard deviation might be reported in logit units. And it holds true when we are working with correlations and the standard deviation is reported in Fisher's *Z* units. Because they had no sense of *T* in these metrics, researchers started using the *Q*-statistic (Appendix) or the *p*-value associated with this statistic as indices of dispersion, and these are remarkably poor surrogates for *T*.

In an attempt to address this problem, Higgins and Thompson (2002) proposed that researchers instead use the I^2 statistic. In a field where the within-study error tends to be relatively constant, I^2 may be highly correlated with T^2 (and T). This is the case for the analyses reported in the Cochrane Database of Systematic Reviews, where the correlation between I^2 and T^2 is 0.93. In this limited case, I^2 could be used to establish some broadly defined rules-of-thumb for small, moderate, and high levels of heterogeneity, and these are proffered in a chapter in the Cochrane Handbook for Systematic reviews (Deeks *et al.*, 2008). The Handbook explained that to interpret I^2 correctly, one needed to do so in the context of the specific analysis, and that the rules of thumb would not always apply. The same point has been made by Higgins and Thompson (2002); Higgins *et al.* (2003); Higgins (2008); Higgins *et al.* (2009), among others.

If I^2 had been used as intended, it might have served as a useful surrogate for T^2 (and T) in some contexts. Unfortunately, the overwhelming majority of papers that employ I^2 do *not* use it as intended. Rather, many papers treat I^2 as though it were an index of absolute dispersion. In other words, researchers who had interpreted Q or p as reflecting the amount of dispersion came to interpret I^2 as reflecting the amount of dispersion. Thus, the impact of I^2 was not to solve the problem that numbers were being interpreted incorrectly, but merely to shift the problems associated with Q and p to a new statistic.

Indeed, it is striking how often I^2 is misinterpreted. Many papers define I^2 correctly as being a proportion and then proceed to interpret the statistic as though it were an absolute value. We deliberately avoid citing examples here, because this mistake is ubiquitous in the literature.

7. I^2 is not an absolute measure of heterogeneity in a meta-analysis

This is not the first paper to call attention to this issue. For example, Rücker *et al.* (2008) and Mittlböck and Heinzl (2006) employed simulations to compare the behavior of I^2 versus T^2 and came to the same conclusions that we are reporting here. While these papers called attention to the problem, the fact that the authors used simulations to compare the indices could help perpetuate the myth that I^2 and T^2 are interchangeable. Our point is that *by definition*, the indices *do not* estimate the same value. One is a proportion, while the other is an absolute value. One describes a relationship between two elements in the *sample*, while the other describes a *different* element in the *population*. Each one is the best (and only) index for a specific purpose (Higgins, 2008). To choose between I^2 and T^2, we should not be asking "which is better" but rather "what question do we want to address." Consider the following analogy.

Suppose someone has an income of $200,000 and donates 10% of her income to charity, for a total donation of $20,000. We might ask what *proportion* of her income this person donates, and the answer would be 10%. We might ask what *amount* she donated, and the answer would be $20,000. Each of these facts is correct, but the two are not interchangeable. If we want a clue to the person's priorities, we might be looking to the rate. If we are trying to locate people who donated a lot of money, we would be looking to the amount. In the latter case, if we ask "who donates a large amount" and we are told that this person donates 10%, we do not have the information we need. If her income was $50,000, then 10% amounts to $5,000, but if her income was $200,000, then 10% amounts to $20,000. Of course, if we are told that she donates 10%, and we also know her income, we can multiply one by other and compute the amount. If we have only a general sense of her income, this will be a rough approximation.

By analogy, suppose that the variation in observed effects (V_{OBS}) is 200, and that 10% of that reflects variation in true effects (I^2), so the variation in true effects (T^2) is 20. We might ask about the ratio of true to total variance in the observed effects, and the answer would be 10%. We might ask about the variance in true effects, and the answer would be 20. Each of these facts is correct, but the two are not interchangeable. If we want to know how much the true effects vary, we are looking for an amount, and not a proportion. Of course, if we are told that the proportion is 10%, and we also know the variance of observed effects, we can multiply one by other and compute the variance of true effects. If we have only a general sense of the variance of observed effects, this will be a rough approximation.

When we ask about heterogeneity in a meta-analysis, we are almost always asking about the amount, not the rate. We want to know how widely the effects vary. We want to know if the intervention will be helpful in all populations, or helpful in some and harmful in others. This information is provided by the statistics that describe the population we have illustrated in the right-hand plot – T, and the prediction interval. Statistics that describe aspects of the left-hand plot have no relevance. Thus, in the motivating examples introduced earlier

- In the ADHD analysis, the fact that I^2 is 47% tells us nothing about the range of effects. By contrast, the prediction interval tells us that in most populations, ADHD will increase the mean score by at least 0.10 standard deviations, and as much as 0.90 standard deviations.
- In the Viagra analysis, the fact that I^2 is 52% tells us nothing about the range of effects. By contrast, the prediction interval tells us that in most populations, Viagra will increase the success rate by at least 86% and as much as 339% as compared with placebo.
- In the PTSD analysis, the fact that I^2 is 85% tells us nothing about the range of prevalence. By contrast, the prediction interval tells us that the prevalence of PTSD varies from 6% in some populations to 42% in others.
- In the Commitment/Performance analysis, the fact that I^2 is 64% tells us nothing about the range of correlations. By contrast, the prediction interval tells us that the correlation between commitment and performance varies from −0.06 in some populations to 0.39 in others.

8. Should we ever report I^2?

In light of this, one might ask if I^2 has any legitimate role in a meta-analysis. We believe it does, but we need to distinguish between observed effects and true effects.

When we want to describe the plot of observed effects, the I^2 statistic serves an important and unique function. It provides context for the plot, telling us what proportion of the observed variance is likely to remain if we could somehow remove the sampling error. Because we almost invariably display the plot, it is helpful to report a statistic that provides this context.

When we want to describe the plot of true effects, the statistic of choice will always be the prediction interval. However, when we are reading an analysis that does not report the prediction interval nor the statistics we would need to compute it, I^2 can be useful. The correct use of I^2 in this case is *not* as a surrogate for the dispersion (equating a large value of I^2 with a lot of dispersion) but rather to provide context for the forest plot. For example, if I^2 is near zero, then we know that most of the dispersion in the forest plot would disappear if we could somehow remove the sampling error. Conversely, if I^2 is near one, then we know that most of the observed dispersion would remain. The approach will yield only a very rough approximation for the variance of true effects, but it is better than nothing.

9. Conclusions

When we ask about "heterogeneity" of effects, we usually are asking about the substantive or clinical implications of the heterogeneity. Because this is what researchers care about, researchers generally assume that this is what is being captured by I^2. A small value of I^2 is interpreted as meaning that the effect size is comparable across studies. A large value of I^2 is interpreted as meaning that the effect size varies substantively across studies.

In fact, I^2 does not tell us how much the effect size varies. Rather, it tells us what proportion of the observed variance would remain if we could eliminate the sampling error – if we could somehow observe the true effect size for all studies in the analysis. I^2 can be used *together* with the observed effects to give us a sense of the true effects. For example, if we are presented with a plot of the observed effects, we can use I^2 to mentally rescale the plot and get some sense of how the true effects are distributed. However, this approach yields only a rough estimate of the actual dispersion.

If we care about the range of effects, then we should report the range of effects, which we call the prediction interval. For example, we can report that the effect size varies from a *d*-value of 0.10 in some studies to 0.90 in others. This provides the information that people need, and that they *think* is being provided by I^2.

10. Spreadsheets for computing prediction intervals

An Excel™ spreadsheet to compute prediction intervals is available at www.Meta-Analysis.com/Prediction or e-mail Biostat100@GMail.com to contact the first author.

Appendix:

Part 1. Relationship among various statistics for heterogeneity

When researchers discuss heterogeneity, they typically report an array of statistics which may include Q, df, p, I^2, I, T^2, and T. Here, we outline the relationship among these statistics.

Computing Q and df

The Q-value refers to the distribution of *observed* effects. The Q-value is the sum of squared deviations of all effects about the mean, on a standardized scale. Concretely,

$$Q = \sum \left(\frac{X_i - M}{SE_{X_i}} \right)^2,$$

where X_i is the effect size in the i^{th} study, M is the mean effect size using fixed effect weights, and SE_{Xi} is the standard error of the i^{th} study (which is assumed to be known). On this scale, the value of Q we would expect to see based on sampling error alone is equal to the degrees of freedom (df), which is the number of studies minus 1.

These two numbers (Q and df) serve as the basis for all the other statistics, as follows.

Computing a *p*-value

If all studies share a common true effect size (and we knew the true standard errors), then *Q* would be distributed as chi-squared with *df* equal to the number of studies minus 1. So we could evaluate *Q* with reference to the chi-squared distribution, to get a *p*-value. If *p* is less than alpha (typically set at 0.10 for this test), we reject the null and conclude that some of the dispersion reflects variation in true effects.

Computing I^2

We can define I^2 as

$$I^2 = \frac{Q - df}{Q}.$$

In the numerator, because *Q* is the total sum of squares while *df* is the sum of squares attributed to sampling error, the difference is the sum of squares due to variance in true effects. In the denominator, *Q* is again the total. So, I^2 is the ratio of true to total.

Q and *df* are on a standardized scale. To convert either of these numbers to the metric of the effect size, we would divide by *C*, a value based on the study weights. If we divide the numerator by C, we obtain T^2, and if we divide the denominator by *C*, we obtain V_{OBS}. So, we can rewrite the equation as

$$I^2 = \frac{T^2}{V_{OBS}},$$

which is the formula presented in the text. Equivalently, we could write this as

$$I^2 = \frac{T^2}{T^2 + V_{ERR}} = \frac{V_{TRUE}}{V_{TOTAL}},$$

which may be more intuitive.

Computing T^2

We can use *Q* and *df* to compute an estimate of the variance of true effects, T^2, using

$$T^2 = \frac{Q - df}{C}.$$

In this formula, the numerator is the sum of squares that reflects variation in true effects, but it is on a standardized scale. *C* is a factor based on the study weights that we applied when we standardized the deviations. Concretely,

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i},$$

where W_i is the weight for study *i*, which is $1/V_i$, the within-study error variance for that study. When we divide by *C*, we reverse that process, so that T^2 is in the same metric that was employed for the synthesis.

The standard deviation of true effects, *T*, is then

$$T = \sqrt{T^2}.$$

Note that there are other ways to compute an estimate of the variance of true effects, T^2. The method described here was proposed by DerSimonian and Laird (1986).

Part 2. Understanding I^2

In the text, we explained that I^2 is the ratio of true to total variance in the observed effects and proposed three ways to think about this ratio.

- We can compare the standard deviation of observed effects to the (average) standard error of the individual studies.
- We can look at the extent to which each study effect size is unique (the non-overlap among confidence intervals).
- We can think of I^2 as providing a bridge between the observed effects and the true effects, by telling us what proportion of the variance would remain if we could remove the sampling error.

Here, we use a series of examples to illustrate these concepts. In these examples, we assume that all studies in the meta-analysis share a common standard error, which enables us to illustrate these concepts. In any real analysis, the same concepts would apply, but the computations would be more complicated because we would need to define what we mean by V_{ERR}.

Figure 2 includes four rows, each representing a fictional meta-analysis.

I^2 reflects the relationship between the standard deviation across studies and the typical standard error of the observed effect size from an individual study

One way to think about I^2 is that it is based on the comparison of [A] the dispersion of observed effects and [B] the dispersion we would expect based on sampling error alone. We can see this by studying the left-hand column, which shows the observed effects for each meta-analysis.

Beneath each plot is a line [A] that reflects the dispersion of the observed effects. Inside each plot, each effect size is bounded by a line [B] that reflects the error with which the effect size is estimated. If all studies in an

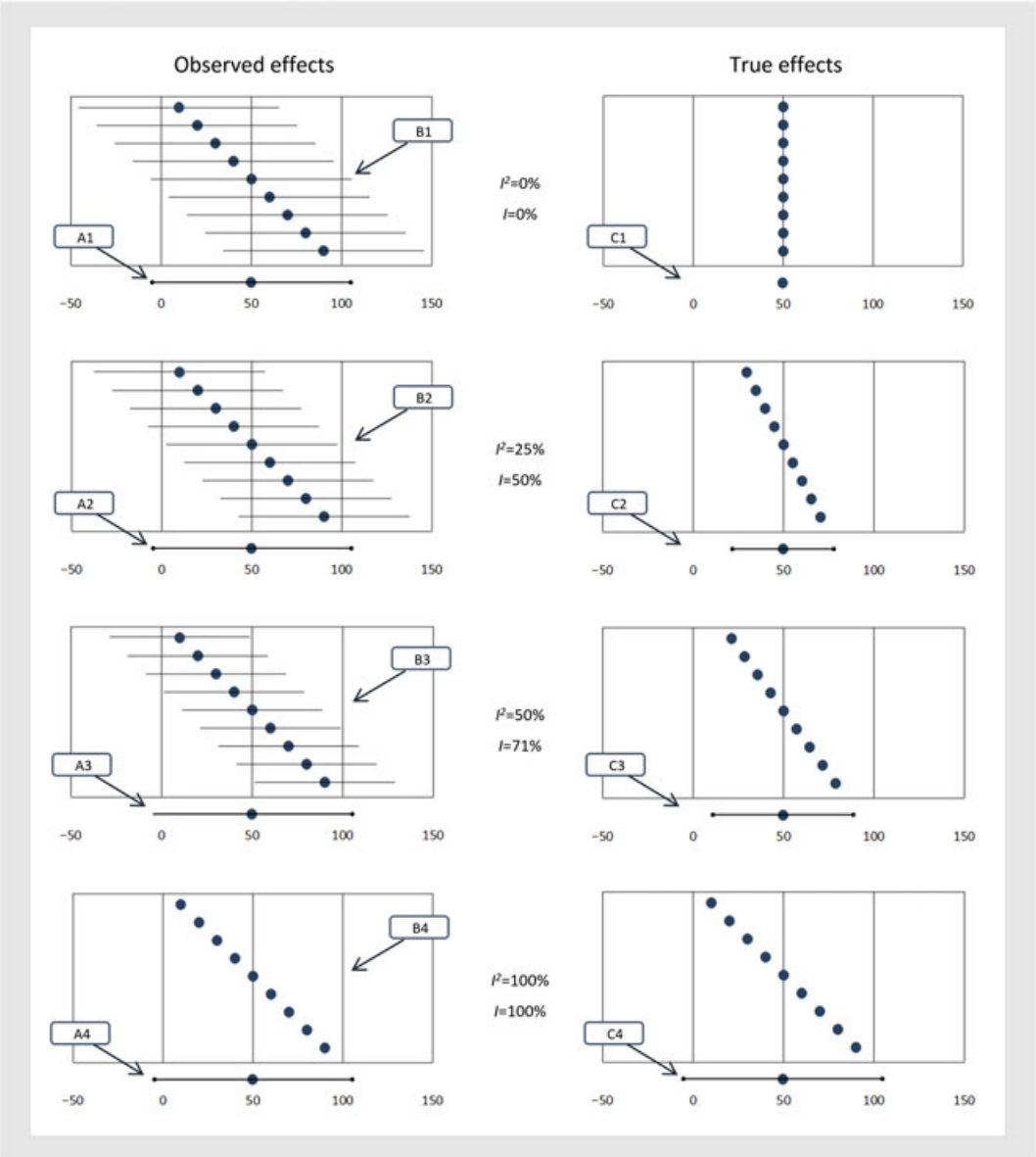

Figure 2. Four fictional meta-analyses. The left-hand column shows the observed effects. As we move from top to bottom the observed variance remains constant and the error variance decreases, so the ratio of true to total (I^2) increases from 0% to 100%. The right-hand column shows a possible distribution of true effects. When I^2 is 0% the variance of true effects is 0. When I^2 is 100% the variance of true effects is the same as the variance of observed effects.

analysis shared the same true effect size, then [A] should be the same length as [B]. So, it is the *discrepancy* between the two lines that reflects the variation in true effects.

The observed effects fall at precisely the same points in all four analyses, and so line [A], which reflects the dispersion of these points, is identical in all four analyses. However, as we move from top to bottom, line [B], which reflects the error variance, narrows.

- In the first analysis, line [A1] is no wider than line [B1], and so no part of [A1] reflects variation in true effects. In this case, the ratio of true to total variance (I^2) is 0%.
- In the second analysis, line [A2] is a little wider than line [B2], and so a small part of [A2] is assumed to reflect variation in true effects. In this case, the ratio of true to total variance (I^2) is 25%.
- In the third analysis, line [A3] is a substantially wider than line [B3], and so more of [A3] is assumed to reflect variation in true effects. In this case, the ratio of true to total variance (I^2) is 50%.
- In the fourth analysis, line [B4] has essentially no width (no error), and so almost all of [A4] is assumed to reflect variation in true effects. In this case, the ratio of true to total variance (I^2) is 100%.

I^2 reflects the extent of non-overlap among confidence intervals

A second way to think about I^2 is that it reflects the amount of non-overlap among confidence intervals.

In the top analyses (where I^2 is 0%), there is a great deal of overlap among confidence intervals, and so no evidence that the effect size in any population is clearly different from the effect size in any other population. Put another way, there is no evidence that any of the dispersion represents variation in true effects (rather than sampling error). In this case, I^2 is 0%.

By contrast, in the fourth analysis (where I^2 is 100%), there is no overlap in confidence intervals, and so there is clear evidence that the effect size varies from one population to the next. Put another way, there is clear evidence that virtually all of the dispersion reflects variation in true effects (there is no sampling error to speak of). In this case, I^2 is 100%.

The same idea applies in the middle cases, but it will be harder to see. For example, in the second row, there is only minimal overlap in the confidence intervals for the first and last studies. In this case, I^2 is 25%. In the third row, there is no overlap in the confidence intervals for the first and last studies. In this case, I^2 is 50%.

I^2 provides context for the forest plot of observed effects

The final way of looking at I^2 is perhaps the most useful. It provides context for the forest plot. Because I^2 tells us what proportion of the variation in observed effects is due to variation in true effects, then (by definition) it tells us what proportion of this variation would remain if we could somehow get rid of the sampling error. Put another way, we start by looking at the plot of observed effects. I^2 tells us what that plot would look like if each study had a sample size that approached infinity, so that we were plotting the true effects. As such, I^2 serves as a bridge between the left-hand plot and the right-hand plot.

Put simply, if we start with the variance of observed effects, and multiply it by the proportion that reflects variance in true effects (I^2), we obtain the variance of true effects, at right. In the top case, this proportion is zero and the variance at right is zero. In the bottom case, this proportion is one, and the variance at right is identical to the variance at left. The two other cases fall in the middle.

We can multiply the variance at left by I^2 to obtain the variance at right,

$$T^2 = V_{OBS} \times I^2,$$

Or, we can multiply the standard deviation at left by I to obtain the standard deviation at right,

$$T = S_{OBS} \times I.$$

In Figure 2, this idea is captured by the relationship between line [A] and line [C] for each analysis. Line [A] is intended to capture some 95% of the *observed* effects. Because most effects fall within two standard deviations on either side of the mean, line [A] has a length of four times S_{OBS}. Line [C] is intended to capture some 95% of the *true* effects so line [C] has a length of 4 times T. Thus, if we know the length of [A], we can multiply that by I to obtain the length of [C]. Note that we multiply by I rather than I^2 because the standard deviation (and these lines) are in linear units rather than squared units. Concretely,

$$C = A \times I.$$

In each of the four analysis, the standard deviation of observed effects is 27.4, so most of the observed effects fall in range that covers 110 points [A1, A2, A3, A4]. In the first analysis, I is 0%, so most true effects fall in a range that covers 0 points [C1]. In the second analysis, I is 50%, so most true effects fall in a range that covers 55 points [C2]. In the third analysis, I is 71%, so most true effects fall in a range that covers 77 points [C3]. In the fourth analysis, I is 100%, so most true effects fall in a range that covers 110 points [C4].

The right-hand plot on each row is intended as an example of how the true effects would be distributed if they followed a uniform distribution with the specified mean and standard deviation. By plotting a distribution of observed effects (at left) and true effects (at right), we make it easy to compare the two plots. However, we do not mean to imply that the true effect for each study at the right corresponds to the same study at the left. Nor do we intend to suggest that we would expect the effects to follow a uniform distribution. These points apply also to Figures 3 and 4.

While Figure 2 serves to illustrate how I^2 operates, it is potentially misleading in one critical respect. If we look at Figure 2 in isolation, we might note that I^2 is strongly correlated with T and might conclude that it could serve as a useful surrogate for the range of effects. However, the reason that I^2 is strongly correlated with T for the examples in this Figure is that V_{OBS} is the same in all four of these fictional analyses.

While higher values of I^2 tend to be associated with higher values of T^2 on average, once we allow V_{OBS} to vary, the correlation between I^2 and T^2 becomes much weaker, and one can no longer serve as a surrogate for the other. To make this point, we introduce a series of analyses in Figure 3, and then compare Figure 2 with Figure 3.

The four analyses in Figure 3 follow precisely the same pattern as the four analyses in Figure 2. That is, the observed effects are identical for all four analyses, as reflected in line [A]. But as we move from top to bottom,

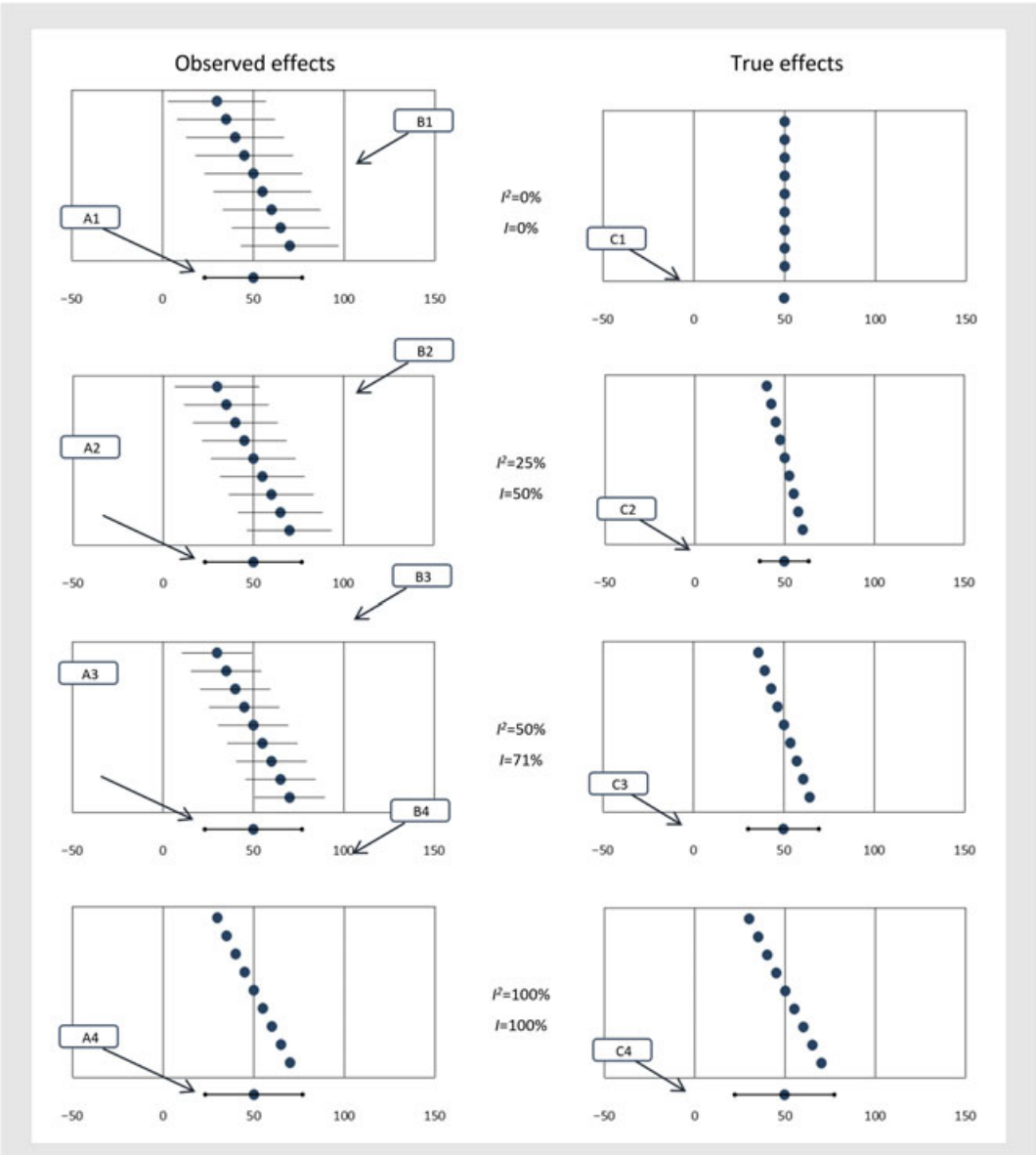

Figure 3. Four fictional meta-analyses. These follow the same pattern as the analyses in Figures 2. The difference is that the variance of observed effects here is one-fourth as large, and so the variance of true effects for any row is one-fourth as large. Similarly, the standard deviation of observed effects is one-half as large, and so the prediction interval for any row is one-half as large. (This does not apply to row-1, where I^2 is 0%.)

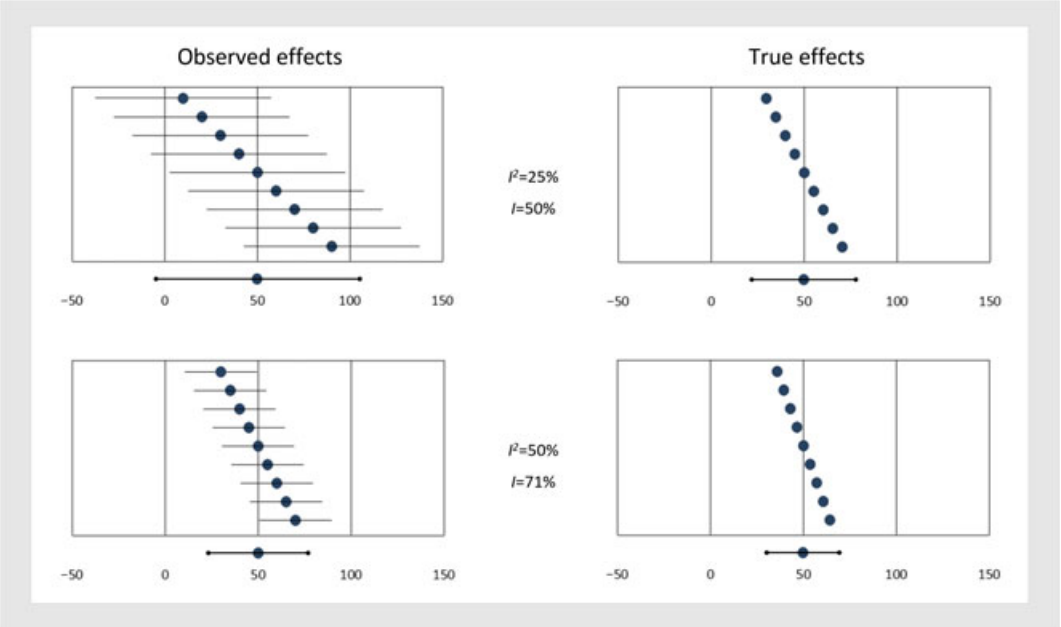

Figure 4. This Figure collates Row-2 from Figure 2 and Row-3 from Figure 3. In the top row, the variance of observed effects is 750, I^2 is 50%, and the prediction interval is 55 points wide. In the second row, the variance of observed effects is 187.5, I^2 is 25%, and the prediction interval is 39 points wide. Thus, the larger value of I^2 corresponds to the smaller range of effects.

line [B], which reflects the error variance, narrows. The I^2 values for the four rows are 0%, 25%, 50%, and 100%. All of the points we made for Figure 2 apply here as well.

The difference between the two figures is that whereas the variance of observed effects in Figure 2 was 750, the variance of observed effects in Figure 3 is 187.5. Therefore, when we compare any row in Figure 3 to the corresponding row in Figure 2,

- The variance of observed effects is one fourth as large, and so the variance of true effects is one fourth as large.
- The standard deviation of observed effects is half as large, and so the standard deviation of true effects is half as large.

For example, the second row in each figure has an I^2 value of 25%. In Figure 2, this corresponds to a prediction interval of 55 points, but in Figure 3, it corresponds to a prediction interval of 27 points. Similarly, the third row of each figure has an I^2 value of 50%. In Figure 2, this corresponds to a prediction interval of 78 points, but in Figure 3, it corresponds to a prediction interval of 39 points. The key point is that I^2 by itself does not tell us how much the effects vary. Any value of I^2 (except for zero) can reflect virtually any range of effects.

It should be clear that when V_{OBS} is allowed to vary, I^2 cannot be used as a surrogate to tell us how much the effects vary on an absolute scale. In fact, when V_{OBS} is allowed to vary, I^2 does not even tell us how much the effects vary on a relative scale.

For example, compare the second row of Figure 2 (where V_{OBS} was 750) with the third row of Figure 3 (where V_{OBS} was 187.5). To facilitate this comparison, we have excerpted these rows into Figure 4. In the top row, I^2 is 25%, and the prediction interval is 55 points wide. In the bottom row, I^2 is 50%, and the prediction interval is 39 points wide. Thus, the *larger* value of I^2 corresponds to the *smaller* range of effects. If we apply the benchmarks of "small" for 25% and "moderate" for 50%, the I^2 value with "small" heterogeneity has more variance than the one with "moderate" heterogeneity.

Part 3. Computing prediction intervals

In the text, the formula we used for the prediction interval was

$$Interval = M \pm 2T.$$

If we are primarily interested in the width of the interval (which was our goal in this paper), this formula is the one to use. This formula yields a correct interval when the estimates of the mean and the standard deviation are correct.

In any real analysis, these values are estimated with error, and if we want to take account of that error, the appropriate formula (Higgins *et al*., 2009; Riley *et al*., 2011) is

$$Interval = M \pm t_{(df)} \sqrt{V_M + T^2}.$$

This formula includes three adjustments to the simple formula.

- First, we have replaced T with the square root of T^2. This is the identical value, but this format allows us to combine two variance components in the next step.
- Second, we have added the variance of the mean (V_M) to account for the fact that the true mean may be lower or higher than M.
- Third, we have replaced the factor of 2 with the critical t-value for df, to account for the fact that the standard deviation is being estimated. The degrees of freedom for t can be taken to be the number of studies minus two.

3.1. Prediction intervals for means

In the ADHD analysis, the effect size index is the standardized mean difference, d. The mean effect size is 0.5058, the variance of M is 0.0054, T^2 is 0.0387, and the number of studies is 17. The prediction interval is given by

$$LL = 0.5058 - 2.1314\sqrt{0.0387 + 0.0054} = 0.0582$$
$$UL = 0.5058 + 2.1314\sqrt{0.0387 + 0.0054} = 0.9534.$$

We expect that in some 95% of all populations, the true effect size will fall in the approximate range of 0.06 to 0.95.

3.2. Prediction intervals for ratio

In the Viagra analysis, the mean effect size is 2.4975. In log units, the mean effect size is 0.9153, VM is 0.0024, T^2 is 0.0223, and the number of studies is 19. The prediction interval in log units is given by

$$LL = 0.9153 - 2.1098\sqrt{0.0223 + 0.0024} = 0.5837.$$
$$UL = 0.9153 + 2.1098\sqrt{0.0223 + 0.0024} = 1.2469$$

We then convert these values to risk ratios using

$$LL = \exp(0.5837) = 1.7927$$
$$UL = \exp(1.2469) = 3.4795.$$

This tells us that in some 95% of all populations, the true effect size will fall in the approximate range of 1.8 to 3.5.

3.3. Prediction intervals for prevalence

In the PTSD analysis, the mean prevalence is 0.1797. In logit units, the mean prevalence is −1.5184, V_M is 0.0389, T^2 is 0.3460, and the number of studies is 11. The prediction interval in logit units is given by

$$LL = -1.5184 - 2.2622\sqrt{0.3460 + 0.0389} = -2.9219$$
$$UL = -1.5184 + 2.2622\sqrt{0.3460 + 0.0389} = -0.1149.$$

We then convert these values to prevalence units using

$$LL = \frac{\exp(-2.9219)}{\exp(-2.9219) + 1} = 0.0511$$
$$UL = \frac{\exp(-0.1149)}{\exp(-0.1149) + 1} = 0.4713.$$

This tells us that in some 95% of all populations, the true prevalence will fall in the approximate range of 5% to 47%.

3.4. Prediction intervals for correlations

In the Commitment/Performance analysis, the mean correlation is 0.1754. In Fisher's Z units, the mean correlation is 0.1772, V_M is 0.0009, T^2 is 0.0142, and the number of studies is 27. The prediction interval in Fisher's Z units is given by

$$LL = 0.1772 - 2.0595\sqrt{0.0142 + 0.0009} = -0.0758$$
$$UL = 0.1772 + 2.0595\sqrt{0.0142 + 0.0009} = 0.4303.$$

We then convert these values to correlations units using

$$LL = \frac{\exp(2\times - 0.0758) - 1}{\exp(2\times - 0.0758) + 1} = -0.0757$$
$$UL = \frac{\exp(2\times 0.4303) - 1}{\exp(2\times 0.4303) + 1} = 0.4056.$$

This tells us that in some 95% of all populations, the true correlation will fall in the approximate range of −0.08 to 0.41.

3.5. Importance of the adjustments

In these examples, the "correct" formula had only a modest impact on the prediction interval as compared with the naïve formula, but this will not always be the case. In particular, if the number of studies is small, the adjustment will be substantial. When the number of studies is small, the interval may be so wide, as to be uninformative. In this case, the take-home message should be that we need more data. We cannot obtain a useful estimate of the standard deviation in a meta-analysis with three studies, any more than we can obtain a precise estimate of the standard deviation in a primary study with three subjects.

References

Cabizuca M, Marques-Portella C, Mendlowicz MV, Coutinho ES, Figueira I. 2009. Posttraumatic stress disorder in parents of children with chronic illnesses: a meta-analysis. *Health Psychology* **28**(3): 379–88.

Castells X, Ramos-Quiroga JA, Rigau D, Bosch R, Nogueira M, Vidal X, Casas M. 2011. Efficacy of methylphenidate for adults with attention-deficit hyperactivity disorder : a meta-regression analysis. *CNS Drugs* **25**(2): 157–169.

Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 9: analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Chichester (UK): John Wiley & Sons, 2008.

DerSimonian R, Laird NM. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**: 177–188.

Higgins JPT. 2008. Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology* **37**: 1158–60.

Higgins JPT, Thompson SG. 2002. Quantifying heterogeneity in meta-analysis. *Statistics in Medicine* **21**: 1539–58.

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. 2003. Measuring inconsistency in meta-analyses. *BMJ* **327**: 557–60.

Higgins JPT, Thompson SG, Spiegelhalter DJ. 2009. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A* **172**: 137–159.

Mittlböck M, Heinzl H. 2006. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Statistics in Medicine* **25**: 4321–4333.

Riley RD, Higgins JPT, Deeks JJ. 2011. Interpretation of random effects meta-analyses. *BMJ* **342**: d549.

Rücker G, Schwarzer G, Carpenter JR, Schumacher M. 2008. Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology* **8**: 79.

Tsertsvadze A, Fink H, Yazdi F, MacDonald R, Bella A, Ansari M, Garritty C, Soares-Weiser K, Daniel R, Sampson M, Fox S, Moher D, Wilt T. 2009. Oral phosphodiesterase-5 inhibitors and hormonal treatments for erectile dysfunction: a systematic review and meta-analysis. *Annals of Internal Medicine* **151**: 650–661.

Wright TA, Bonett DG. 2002. The moderating effect of employee tenure on the relation between organizational commitment and job performance: a meta-analysis. *Journal of Applied Psychology* **87**(6): 1183–1190.

Supporting information

Additional supporting Information may be found in the online version of this paper at the publisher's website.