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

Draft | Please do not quote

Michael Borenstein[1]

Julian P.T. Higgins[2]

Hannah R. Rothstein[3]

Larry V. Hedges[4]

[1]Biostat, Inc., Englewood, NJ, USA

[2]School of Social and Community Medicine, University of Bristol, Bristol, UK

[3]Management Department, Baruch College—City University of New York, NY, U.S.A.

[4]Department of Statistics, Northwestern University, Evanston, IL, U.S.A.

Correspondence to the first author Biostat100@GMail.com

Abstract

Researchers often use the I^2 index to quantify the dispersion of effect sizes in a meta-analysis. Some suggest that I^2 values of 25%, 50%, and 75%, correspond to small, moderate, and large amounts of heterogeneity. In fact though, I^2 is a not a measure of *absolute* heterogeneity. Rather, it tells us what *proportion* of the observed variance reflects variance in true effect sizes rather than sampling error. This distinction between an absolute number and a proportion is fundamental to the correct interpretation of I^2. A meta-analysis with a low value of I^2 *could* have only trivial heterogeneity but could also have substantial heterogeneity. Conversely, a meta-analysis with a high value of I^2 *could* have substantial heterogeneity, but could also have only trivial heterogeneity. Our goal in this paper is to explain what I^2 is, and how it should (and should not) be used in meta-analysis.

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, Thompson, Deeks, and Altman, 2003). I^2 is an intuitive statistic for many reasons. It ranges from 0% to 100%, so we have a clear sense of where any given study 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, has 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.

What we mean by heterogeneity

A simple thought experiment will make it clear that I^2 does not tell us how much the effect size varies across populations. Suppose we are evaluating an intervention that is intended to reduce the amount of time people spend recovering from a stroke. We perform a meta-analysis of studies that tested this intervention in various populations, and determine that patients in the treated group meet their goal a mean of 50 days sooner than those in the control group.

Next, we ask how much the effect size varies. That is, we want to know if the treatment effect typically falls (a) in the range of 40 days to 60 or days, or (b) in the range of 10 days to 90 days, or (c) some other range. If (a) is true, then the treatment can be applied with comparable results in all settings. If (b) is true, we may wish to use the treatment in some settings but not in others.

Now, suppose we are told that I^2 is 25%. Does the treatment effect vary as in (a) or (b)? The answer is that we do not know. It could be (a), or it could be (b), or it could be something else entirely. The reason we do not know is that I^2 is in a metric that goes from 0% to 100%. By itself, this statistic tells us nothing about the actual range of effects (Higgins, 2008; Rücker, Schwarzer, Carpenter, Schumacher, 2008; Mittlböck, Heinzl, 2006; Huedo-Medina, Sánchez-Meca, Marín-Martínez, Botella, 2006).

Rather, to distinguish between (a), (b), and (c), we need an index that quantifies dispersion as a number of days. Such an index is the standard deviation of the true effect sizes, which we will call T. We generally assume that most effects will fall within two standard deviations of the mean effect. If the standard deviation is $T = 5$ days then the treatment effect varies over 20 days, as in (a). If the standard deviation is $T = 20$ days, then the treatment effect varies over 80 days, as in (b).

Observed effects versus true effects

If I^2 does not tell us how much the treatment effect varies, then what does it tell us? To address this question we first highlight an important difference between a meta-analysis and a primary study. In a primary study, the scores that we *observe* are usually treated as the *true* scores for each subject. Therefore, the distribution of observed scores serves as the distribution of true scores. By contrast, in a meta-analysis, we need to distinguish between the observed effect size and the true effect size. The *observed* effect size is the effect size that we see in a study. It serves as the estimate of the effect size in the study's population, but invariably differs from the true effect size in that population due to sampling error. By contrast, the *true* effect size is the actual effect size in the study's population. By definition, the true effect size for a population is the effect size that we would see with an infinitely large sample size, and (it follows) no sampling error. The problem that we need to address is that the distribution of *observed* effects is not the same as the distribution of *true* effects.

Figure 1 is a forest plot of the example we introduced a moment ago. The left-hand frame shows the distribution of observed effect sizes. The mean effect is 50 days, the standard deviation is 27.4, and most effects fall in the range of −5 to 105 as indicated by the line [A] at the bottom of the plot. This would suggest that the effect size varies substantially from study to study. The right-hand frame shows the distribution of true effect sizes. The mean effect is 50 days, the standard deviation of true effects is 8.66 days, and most effects falls in the range of 33 to 67 as indicated by the line [B] at the bottom of the plot. It tells us that the effect size is reasonably consistent across the studies.

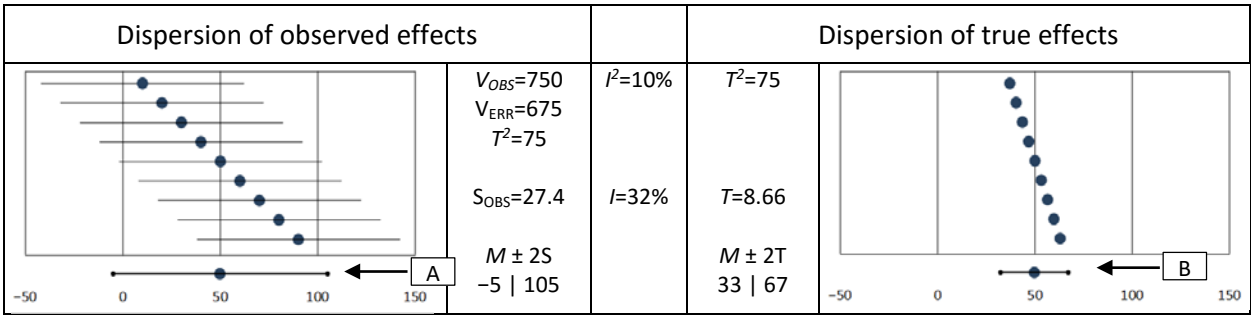


Figure 1 | Dispersion of observed effects and dispersion of true effects

Why do the observed effects (at left) vary more than the true effects (at right)? The reason is that the variance of the observed effects incorporates both the variance of true effects and also random sampling error. Concretely, if T^2 is the variance of the true effects, V_{ERR} is the variance due to sampling error (assumed here to be the same in each study), and V_{OBS} is the variance of the observed effects, then

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

In this example the variance of observed effects (at left) is 750, and it can be decomposed into variance of true effects and variance due to sampling error as

$$750 = 75 + 675.$$

If we want to know about the potential utility of the treatment, we have little interest in the left-hand frame where the effects vary over 100 days, since the dispersion we see here is partly based on random sampling error. Rather, we care about the dispersion in the right-hand frame, where the effects are confined to a range of 34 days. This is the frame that speaks to the utility of the treatment.

Since we start with the left-hand plot but we need to impute the right-hand plot, it would be helpful to have an index that speaks to the relationship between the two. One index for this is I^2. Recall that the left-hand plot is based on $T^2 + V_{ERR}$ whereas the right-hand plot is based on T^2 alone. The I^2 index quantifies this relationship as

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

or (equivalently) as

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

The first formula makes it clear that I^2 tells us what proportion of the variance in observed effects reflects variance in true effects rather than sampling error. The second formula makes it clear that I^2 gives us the ratio of the variance in the right-hand plot to the left-hand plot. It follows that if we know the variance of the observed effects and I^2 we can compute the variance of true effects using

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

If we know the standard deviation of the observed effects (S_{OBS}, the square root of V_{OBS}) and I^2 we can compute the standard deviation of true effects (T) using

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

Similarly, if we know the range of the observed effects (R_{OBS}) and I^2 we can compute the range of true effects (R) using

$$R = R_{OBS} \times \sqrt{I^2} = R_{OBS} \times I.$$

Applying these formulas to the values in Figure 1 we get

$$I^2 = \frac{T^2}{T^2 + V_{ERR}} = \frac{75}{75 + 675} = 10\%,$$

$$I^2 = \frac{T^2}{V_{OBS}} = \frac{75}{750} = 10\%,$$

$$T^2 = V_{OBS} \times I^2 = 750 \times 10\% = 75,$$

$$T = S_{OBS} \times I = 27.4 \times 32\% = 8.66,$$

and

$$R = R_{OBS} \times I = 105 \times 32\% = 34 .$$

These values are included in the center columns of Figure 1.

Where Figure 1 displayed one meta-analysis, Figure 2 displays a series of meta-analyses. For each analysis, the left-hand plot shows the observed effects while the right-hand plot shows the true effects. The center columns show what happens if we apply these formulas. The second analysis in Figure 2 is identical to the analysis we saw a moment ago.

By focusing on the left-hand column we can get an intuitive sense of why I^2 goes from 0% (at the top) to 100% (at the bottom). The observed effects (and so V_{OBS}) are identical in all rows. What differs as we move from row to row is that the error (V_{ERR}) decreases. If the observed variance is constant but the proportion due to error decreases, then the proportion attributed to variance in true effects (that is, I^2) must increase.

By focusing on the right-hand column we can get an intuitive sense of what happens if the observed variance is a constant and I^2 increases. In every case we multiply the variance of observed effects by I^2 to get the variance of true effects. When I^2 is low (at the top) the variance of true effects is small and so the true effects lie close to the mean. The range of true effects (the line underneath the plot) is narrow. As we move from row to row I^2 goes up, and so the variance of true effects goes up. Effect sizes move further from the mean. The range of true effects (the line underneath the plot) widens.

By looking at the entire page we can see a clear (inverse) relationship between the error bars at left and the dispersion of true effects at right. The variance of observed effects is constant in all rows and incorporates V_{ERR} and T^2. It follows that as V_{ERR} goes down, T^2 goes up. As we move from row to row, the error bars disappear (as it were) from the left-hand plot and serve to expand the range of effects (and 95% interval) in the right hand plot.

A useful way to think about the relationship between any left-hand plot and the corresponding right-hand plot is as follows. The left-hand plot reflects the dispersion that *we actually see*. The right-hand plot reflects the dispersion that *we would see* if each study had an extremely large sample size and virtually no sampling error.

A useful way to think about I^2, is that it serves as a bridge between the left-hand plot and the corresponding right-hand plot. We can multiply the variance of observed effects (at left) by I^2 to get the variance of true effects (at right). Thus, for example, the meta-analysis in the second row has an I^2 value of 10%. The variance of the observed effects is 750, and we multiply this by 10% to get the variance of the true effects, which is 75. These details are displayed in the center columns, and they quantify the relationship between the left-hand and the right-hand plot.

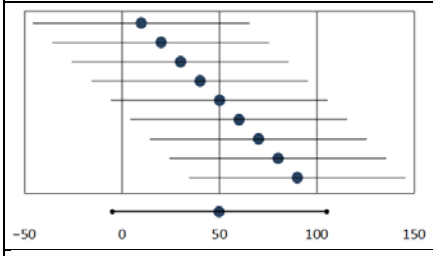
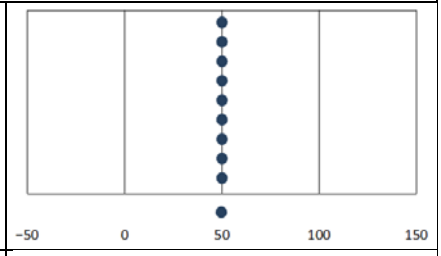
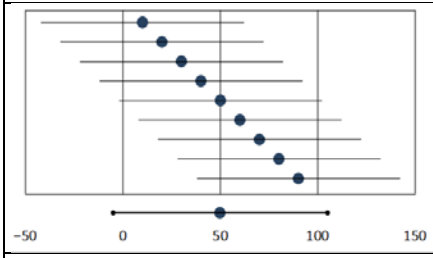
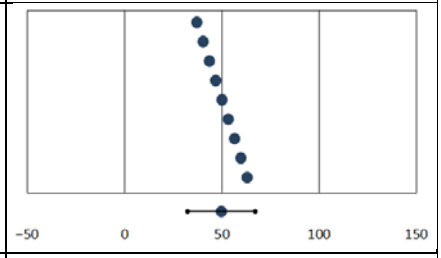
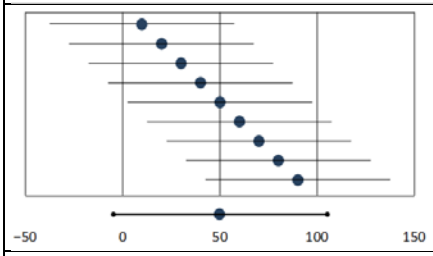
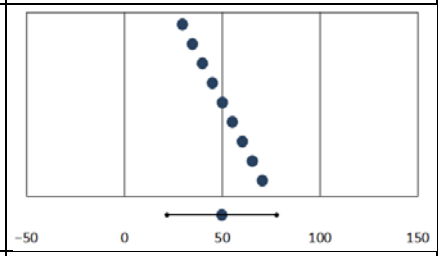
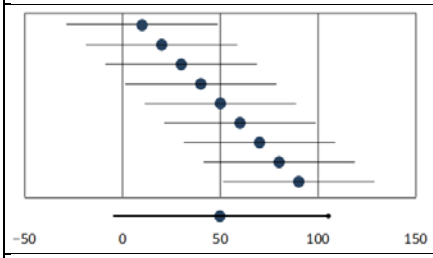
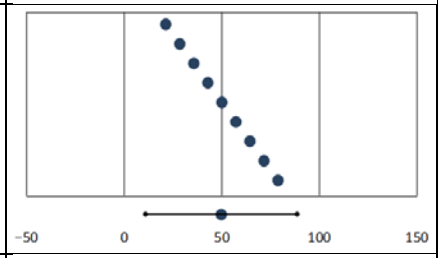
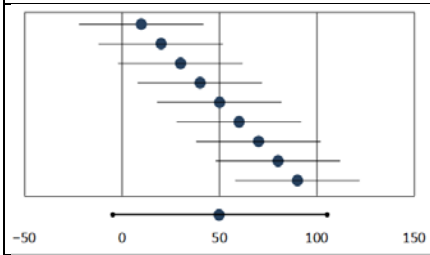
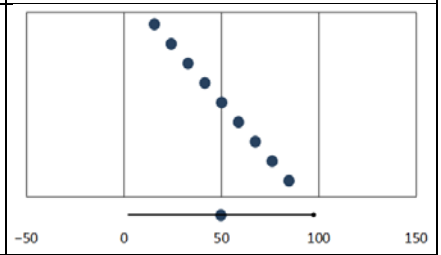
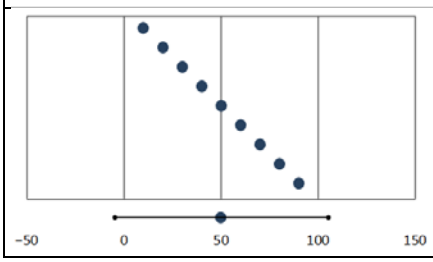
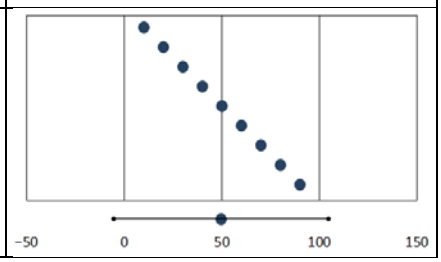
Dispersion of observed effects			Dispersion of true effects	
	V_{OBS}=750 V_{ERR}=750 T^2=0 S_{OBS}=27.4 *M* ± 2S −5 | 105	I^2=0 *I*=0	T^2=0 *T*=0 *M* ± 2T 50 | 50	
	V_{OBS}=750 V_{ERR}=675 T^2=75 S_{OBS}=27.4 *M* ± 2S −5 | 105	I^2=10% *I*=32%	T^2=75 *T*=8.66 *M* ± 2T 33 | 67	
	V_{OBS}=750 V_{ERR}=562.5 T^2=187.5 S_{OBS}=27.4 *M* ± 2S −5 | 105	I^2=25% *I*=50%	T^2=187.5 *T*=13.69 *M* ± 2T 23 | 77	
	V_{OBS}=750 V_{ERR}=375 T^2=375 S_{OBS}=27.4 *M* ± 2S −5 | 105	I^2=50% *I*=71%	T^2=375 *T*=19.36 *M* ± 2T 11 | 89	
	V_{OBS}=750 V_{ERR}=187.5 T^2=562.5 S_{OBS}=27.4 *M* ± 2S −5 | 105	I^2=75% *I*=87%	T^2=562.5 *T*=23.72 *M* ± 2T 3 | 97	
	V_{OBS}=750 V_{ERR}=0 T^2=750 S_{OBS}=27.4 *M* ± 2S −5 | 105	I^2=100% *I*=100%	T^2=750 *T*=27.39 *M* ± 2T −5 | 105	

Figure 2 | I^2 as a link between dispersion of observed effects and dispersion of true effects | Part 1

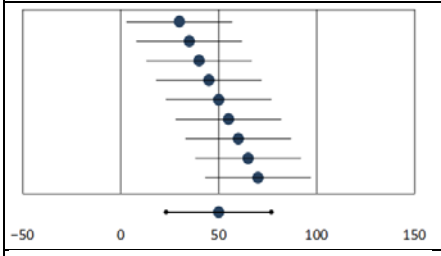
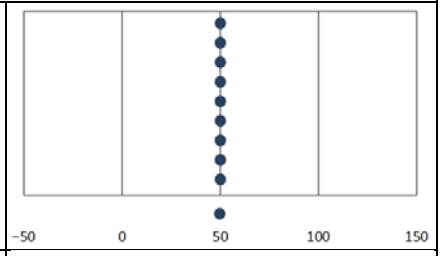
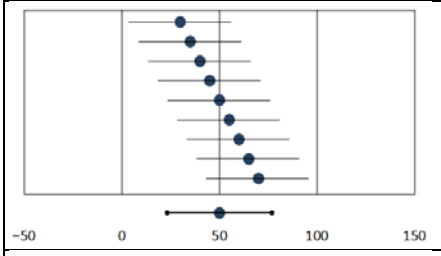
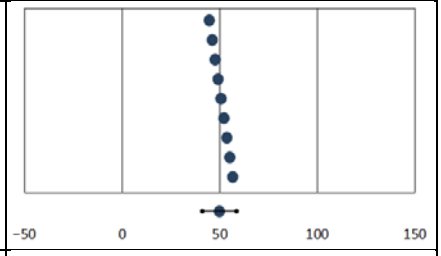
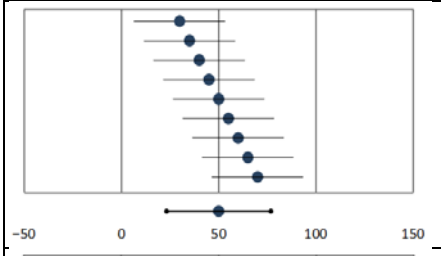
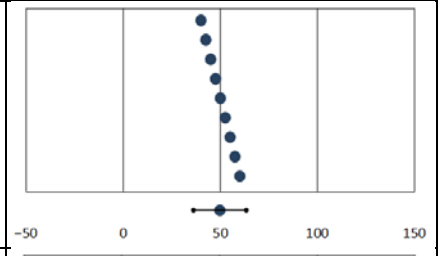
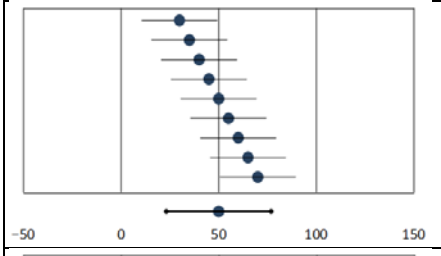
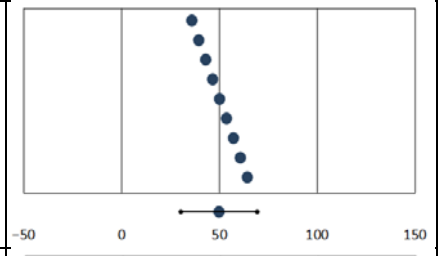
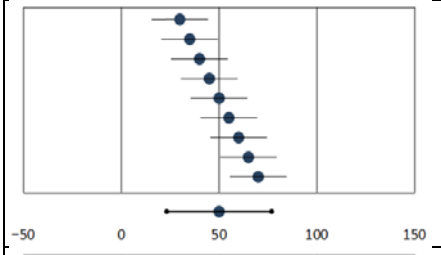
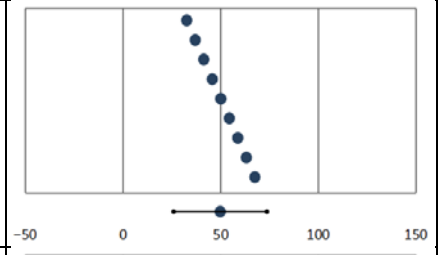
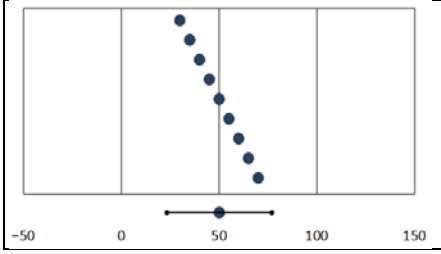
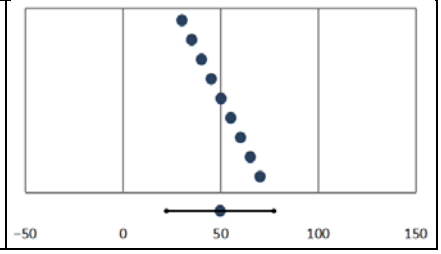
Dispersion of observed effects			Dispersion of true effects	
	V_{OBS}=187.5 V_{ERR}=187.5 T^2=0 S_{OBS}=13.69 *M* ± 2S 23 | 77	I^2=0 *I*=0	T^2=0 *T*=0 *M* ± 2T 50 | 50	
	V_{OBS}=187.5 V_{ERR}=168.75 T^2=18.75 S_{OBS}=13.69 *M* ± 2S 23 | 77	I^2=10% *I*=32%	T^2=18.75 *T*=4.33 *M* ± 2T 41 | 59	
	V_{OBS}=187.5 V_{ERR}=140.625 T^2=46.875 S_{OBS}=13.69 *M* ± 2S 23 | 77	I^2=25% *I*=50%	T^2=46.875 *T*=6.847 *M* ± 2T 36 | 64	
	V_{OBS}=187.5 V_{ERR}=93.75 T^2=93.75 S_{OBS}=13.69 *M* ± 2S 23 | 77	I^2=50% *I*=71%	T^2=93.75 *T*=9.68 *M* ± 2T 31 | 69	
	V_{OBS}=187.5 V_{ERR}=46.875 T^2=140.63 S_{OBS}=13.69 *M* ± 2S 23 | 77	I^2=75% *I*=87%	T^2=140.63 *T*=11.859 *M* ± 2T 26 | 74	
	V_{OBS}=187.5 V_{ERR}=0 T^2=187.5 S_{OBS}=13.69 *M* ± 2S 23 | 77	I^2=100% *I*=100%	T^2=187.5 *T*=13.69 *M* ± 2T 23 | 77	

Figure 3 | I^2 as a link between dispersion of observed effects and dispersion of true effects | Part 2

I^2 is not a measure of absolute variance

To this point, we have established that I^2 is a proportion, and not an absolute value. Nevertheless, in Figure 2 once we know I^2 we have a pretty good idea of the actual dispersion. Therefore, one might assume that I^2 can serve as a surrogate for T. Unfortunately this is not the case. While there is a strong relationship between I^2 and T in Figure 2, this is only because the variance of observed effects is the same for all rows in this figure. Once we leave the artificial constraints of this figure, the relationship no longer exists. This is evident when we consider Figure 3. This follows the same structure as Figure 2. Again, as we move from row to row, the observed variance remains the same but the error variance goes down. It follows that this yields an increase in I^2 and a corresponding increase in T^2. The difference between Figure 2 and Figure 3 is that the absolute amount of observed variance is greater in the former (V_{OBS} = 750) than in the latter (V_{OBS} = 187.5). Since we multiply I^2 by the observed variance to get the variance of true effects, for any given value of I^2 (with the exception of zero) the variance of true effects will be larger in Figure 2.

- Compare the third row in the two figures. In both cases I^2 is 25%. In Figure 2 this corresponds to a standard deviation of around 9 days, in Figure 3 to a standard deviation of around 4.5 days.
- Compare the fourth row in the two figures. In both cases I^2 is 50%. In Figure 2 this corresponds to a standard deviation of around 20 days, in Figure 3 to a standard deviation of around 10 days.
- Compare the fifth row in the two figures. In both cases I^2 is 75%. In Figure 2 this corresponds to a standard deviation of around 24 days, in Figure 3 to a standard deviation of around 12 days.

Thus, if we are told simply that I^2 is 25% or 50% or 75%, without additional context, we do not have any real sense of the actual dispersion.

Not only is I^2 a poor surrogate for the heterogeneity of true effects; it cannot reliably tell us which of two meta-analyses shows more heterogeneity in true effects. For example, suppose we are told that one meta-analysis reported an I^2 of 25% while another reported an I^2 of 75%. If both meta-analyses had comparable variances of observed effects, the comparison would be meaningful. This would be the case if both values came from Figure 2 or if both values came from Figure 3. However, suppose that the first I^2 comes a meta-analysis with more dispersion in the observed effects (Figure 2) while the second comes from a meta-analysis with less dispersion in the observed effects (Figure 3). We have extracted the relevant row from each figure (corresponding to I^2 of 25% in Figure 2 and I^2 of 75% in Figure 3) and reproduced these rows in Figure 4. In this case, the I^2 of 25% corresponds to a standard deviation of about 14 days in true effects, while the I^2 of 75% corresponds to a slightly smaller standard deviation of about 12 days in true effects. Thus, the first meta-analysis (where I^2 is 25%) has *more* variance than the second (where I^2 is 75%). Someone using I^2 as an index of absolute variance would get this backwards, and conclude that there is less variance in the first. Indeed, someone using benchmarks of 25%, 50%, and 75% might assign a label of “Small” variance to the meta-analysis with *more* variance, and “Large” variance to the meta-analysis with *less* variance.

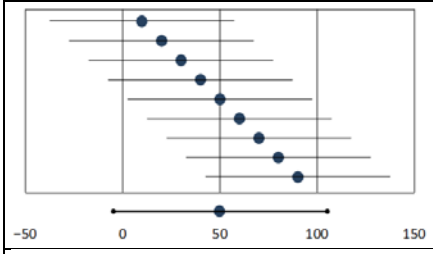
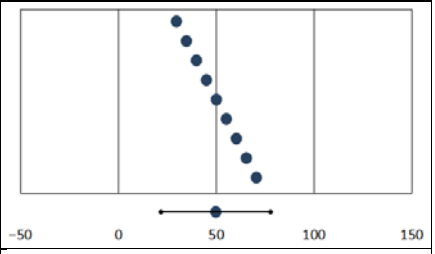
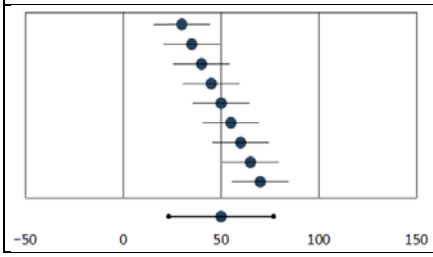
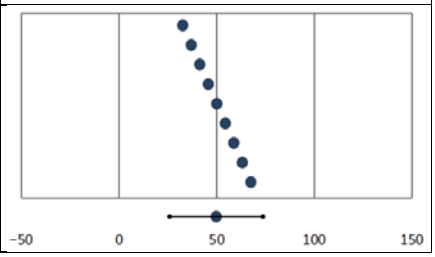
Dispersion of observed effects			Dispersion of true effects	
	V_{OBS}=750 V_{ERR}=562.5 T^2=187.5 S_{OBS}=27.4 *M* ± 2S −5 | 105	I^2=25% *I*=50%	T^2=187.5 *T*=13.69 *M* ± 2T 23 | 77	
	V_{OBS}=187.5 V_{ERR}=46.875 T^2=140.63 S_{OBS}=13.69 *M* ± 2S 23 | 77	I^2=75% *I*=87%	T^2=140.63 *T*=11.859 *M* ± 2T 26 | 74	

Figure 4 | Relationship between I^2 and dispersion of true effects for two sets of studies

How I^2 *should* be used

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 re-scale the plot and get some sense of how the true effects are distributed. While we *could* re-scale the plot using I^2, it is more intuitive to work with *I* (the square root of I^2), since this allows us to think in linear units rather than squared units.

For example, consider Figure 2, where the observed effects always range over 110 days. If I^2 is 10%, then *I* is 32% and the true effects range over 34 days (row 2). If I^2 is 25% then *I* is 50% and the effects range over 54 days (row 3). If I^2 is 75% then *I* is 87% and the true effects range over 94 days (row 5).

Similarly, consider Figure 3, where the observed effects always range over 54 days. If I^2 is 10%, then *I* is 32% and the true effects range over 18 days (row 2). If I^2 is 25% then *I* is 50% and the effects range over 28 days (row 3). If I^2 is 75% then *I* is 87% and the true effects range over 48 days (row 5).

To be clear, all of this is intended as a back-of-the-envelope approximation. If we are presented with a forest plot and I^2, we can use the two together to get a sense of the absolute dispersion. If we do plan to use I^2, then this is the way to use it. However, none of this is optimal. It would be better to have access to more direct estimates of the absolute dispersion. We turn to that now.

Range of effects

Immediately above we showed how I^2 can be employed, along with the forest plot, to address the question "How much does the effect size vary from study to study?" For the person reading a meta-analysis that provides limited information, this might be the best option. However, for the person reporting the meta-analysis, a much better approach is to report the amount of dispersion directly. This can be achieved by presenting a range within which a particular proportion of effect sizes, say 95% of them, are expected to lie. An approximate interval for this, as well as a prediction interval for the true effect in a future similar study, are briefly described in the Appendix.

When we report the actual range of effects (rather than or in addition to I^2) we are accomplishing two important things. First, we are shifting from a proportion to an absolute value. To say that I^2 is 10% tells us nothing about the absolute amount of dispersion. By contrast, if we reported that the true effects range over 20 days, then we have a clear sense of the dispersion in meaningful units. Second, we are reporting the dispersion *in the context of the mean effect size*. To report that the true effects range over 60 days gives us the amount of dispersion, but to understand the substantive impact of that dispersion we need to know the actual range of effects. A 60-point range of effects has a very different impact when centered on a mean of 50 as compared with a mean of 20. In the first case the effects range from 20 to 80. The effects are all positive, but range from small to moderate. In the second case the effects range from minus 10 to 50. The treatment is harmful in some populations and helpful in others. When we report the actual range, we are providing all of this information in a clear and unambiguous way.

Some important caveats

There are two issues that we have not addressed because they are not related to the relationship between I^2 and T^2. However, they are important to the discussion of heterogeneity. First, the estimates of I^2 and T^2 are not likely to be accurate (or even close to accurate) unless the number of studies in the meta-analysis is substantial (Ioannidis, 2007). Second, we compute a range assuming that the true effects are normally distributed about the mean. The validity of this assumption certainly varies from one domain to another. Finally, since our goal in this paper was to explain the distinction between I^2 and T^2 we used examples that allowed us to work with simplified versions of some formulas which are correct only when all studies in the meta-analysis are the same size. More general versions of these formulas are presented in the Appendix.

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. I^2 tells us about the extent of inconsistency of findings across studies in the meta-analysis, and reflects the extent to which confidence intervals from

the different studies overlap with each other. The extent of this overlap tells us nothing about the actual study to study dispersion in effects. 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 re-scale 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 actual range of effects, then we should report the actual range of effects. For example, we can report that the effect size varies from 40 days to 60 days. This provides the range in the actual metric of the effect size. This provides the information that people need, and that they *think* is being provided by I^2.

References

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

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

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

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

Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J: Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychological Methods 2006; 11: 193-206.

Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ 2007; 335: 914-6.

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

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

Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I2 in assessing heterogeneity may mislead. BMC Medical Research Methodology 2008, 8:79.

Appendix

Our goal in this paper was to explain the distinction between I^2 and T^2. To this end we used a series of unrealistically simple examples so that we could focus on the conceptual issues. In any real meta-analysis the computations would be more complicated than those presented above. Here, we outline these more general computations.

Formula for computing I^2

We used examples where the sampling error variance was identical for all studies. In this case the true variance is simply the observed variance minus the (constant) error variance. In any real analysis the sampling error variance will differ from one study to the next, and so the computations must be based on weighted sums of squares rather than variances.

The formula for I^2 is normally given as

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

where Q is the sum of squared deviations (of each effect size from the mean effect size) on a standardized scale (where each deviation is divided by the standard error of the corresponding study). This formula works for any meta-analysis.

In the hypothetical case where the error variance is identical for all studies, we can work with the sums of squares rather than the standardized sums of squares. In this case the formula would be

$$I^2 = \frac{SS_{OBS} - SS_{ERR}}{SS_{OBS}}.$$

Finally, if we divide each element in this formula by the degrees of freedom we get

$$I^2 = \frac{V_{OBS} - V_{ERR}}{V_{OBS}}.$$

When the error variance is identical for all studies (as in our example), these three formulas are all functionally equivalent. We used this example so that we could use the variance-based version of the formula, which allowed us to show the link between I^2 and T^2 more clearly. In any real meta-analysis we would use the Q–based version to compute I^2, but the concepts discussed in this paper still apply.

Formula for computing an approximate range of effects

When the effect size is a mean difference (as in our example) the standard deviation of true effects (*T*) is in the same metric as the effect sizes (here, days). In this case we can compute the approximate range of effects as the mean plus/minus 2*T*. For some effect sizes we need to convert the effect size to another metric before performing the computations. This includes ratios (where computations are performed in log units), correlations (Fisher Z units), and prevalence (logit units).

Suppose the mean risk ratio is 0.6065 and *T* (in log units) is 0.1000. We would transform the mean effect size into log units to get −0.5000. In log units, the mean plus/minus 2*T* gives us

$$LL = -0.5000 - 2 \times 0.1000 = -0.7000$$
$$UL = -0.5000 + 2 \times 0.1000 = -0.3000$$

Then we convert these back into the original metric using

$$LL = \exp(-0.7000) = 0.4966$$
$$UL = \exp(-0.3000) = 0.7408$$

Prediction intervals

We computed the approximate range of effects as the mean plus/minus 2*T*. This gives us the relationship between I^2 and *T*, which is the theme of this paper. However, in reporting the results of a meta-analysis we would want to report the prediction interval (Higgins, Thompson and Spiegelhalter, 2009; Riley, Higgins and Deeks, 2011). This differs from the approximate range in two ways. First, rather than multiply *T* by 2, we multiply it by a number that takes into account the uncertainty with which *T* is estimated. Second, rather than assume that we know the mean effect size accurately, we expand the interval to allow that the effect size is estimated with error.

Consider the case where the analysis includes nine studies, the mean difference is 50, the error variance of the mean difference is 20, T^2 is 100, and *T* is 10. The approximate range is given by

$$LL = M - 2T$$
$$UL = M + 2T$$

which is

$$LL = 50 - 2 \times 10 = 30$$
$$UL = 50 + 2 \times 10 = 70$$

By contrast the prediction interval is given by

$$LL' = M - t_{(df)}\sqrt{V_M + T^2}$$
$$UL' = M + t_{(df)}\sqrt{V_M + T^2}$$

which is

$$LL' = 50 - 2.365\sqrt{20 + 100} = 24.097$$
$$UL' = 50 + 2.365\sqrt{20 + 100} = 75.093$$

In this equation, 2.365 is the *t*-value corresponding to the 95% interval for 7 *df*. The relevant *df* is the number of studies minus 2.

This prediction interval tells us that if we were to select a population at random from the same universe, and run the same study, in 95 of 100 cases the true effect size in that study would fall in the range of approximately 24 to 75. In this example the prediction yields a span that is about 10 days wider than the simple range. The difference between the prediction interval and the simple range tends to be more pronounced when the number of studies is small.

If the effect size is a ratio (or other metric that requires a transformation), the prediction interval is computed in the transformed metric and then converted back to the original metric.