CHAPTER 19

An Intuitive Look at Heterogeneity

INTRODUCTION

In previous chapters, we explained the meaning of the various indices employed to quantify heterogeneity. While many researchers understand the distinction between these indices in the abstract, relatively few actually put this knowledge into practice. Our goal in this chapter is to provide practical advice about how to think about heterogeneity.

The potential utility of an intervention depends not only on the mean effect size, but also on the dispersion of effects about that mean. We need to know if the intervention has essentially the same impact in all populations; or has a trivial impact in some populations and a large impact in others; or if it is harmful in some populations and helpful in others. When researchers ask about heterogeneity, they are asking which of these descriptions applies. However, the statistics typically reported

Introduction to Meta-Analysis, Second Edition.
Michael Borenstein, Larry V. Hedges, Julian P.T. Higgins, and Hannah R. Rothstein.

for heterogeneity (Q, T^2, I^2) do not directly address this question. In this chapter, we highlight the prediction interval, the statistic that reports the range of true effects. This statistic provides the information that we need, and that many think is being provided by the other statistics.

While it is important to report the relevant statistics, it is also imperative to understand the limitations of these statistics. We need a reasonable number of studies to yield reliable estimates of any statistics related to heterogeneity. This applies to I^2 and T^2 as well as to the prediction interval. An estimate of heterogeneity that is based on a handful of studies (or fewer) is not likely to be reliable.

MOTIVATING EXAMPLE

Ronksley, Brien, Turner, Mukamal, and Ghali (2011) published a meta-analysis in *BMJ* that looked at the relationship between alcohol consumption and all-cause mortality. The mean risk ratio was 0.87, which tells us that persons classified as drinkers had a lower risk of death than those classified as nondrinkers. The confidence interval is 0.83 to 0.91, and the Z-value for a test of the null hypothesis is 5.77 with a corresponding p-value of <0.001.

On this basis, we conclude that the mortality risk is 13% lower for drinkers, *on average*. However, we still need to address heterogeneity. That is, we need to know if the distribution of effects resembles panel A, B, or C in Figure 19.1. In each panel, the distribution may be summarized by means of the prediction interval, denoted by an arrow. The true effect size in 95% of all populations will fall inside that interval.

If the distribution of effects resembles panel A, we might report that the drinkers have a lower risk of death than nondrinkers in virtually all populations. At one extreme, there are some populations where the risk of death is 29% *lower* for drinkers. At the other extreme, there are a few populations where the risk of death is 5% *higher* for drinkers. As such, the relationship between drinking and mortality is relatively modest but also relatively consistent.

If the distribution of effects resembles panel B, we might report that the drinkers have a lower risk *on average*, but there is substantial variation in this relationship. At one extreme, there are some populations where the risk of death is 41% *lower* for drinkers. At the other extreme, there are some populations where the risk of death is 30% *higher* for drinkers.

Finally, if the distribution of effects resembles panel C, we might report that there is so much variation in the effect that the mean effect is of little relevance. At one extreme, there are some populations where the risk of death is 60% *lower* for drinkers. At the other extreme, there are some populations where the risk of death is 91% *higher* for drinkers.

These interpretations of the numbers are subjective, and others will characterize the implications of the heterogeneity differently. That discussion is necessary and welcome. However, to have an informed discussion about the implications of the dispersion, we must first know if the distribution resembles panel A, B, or C.

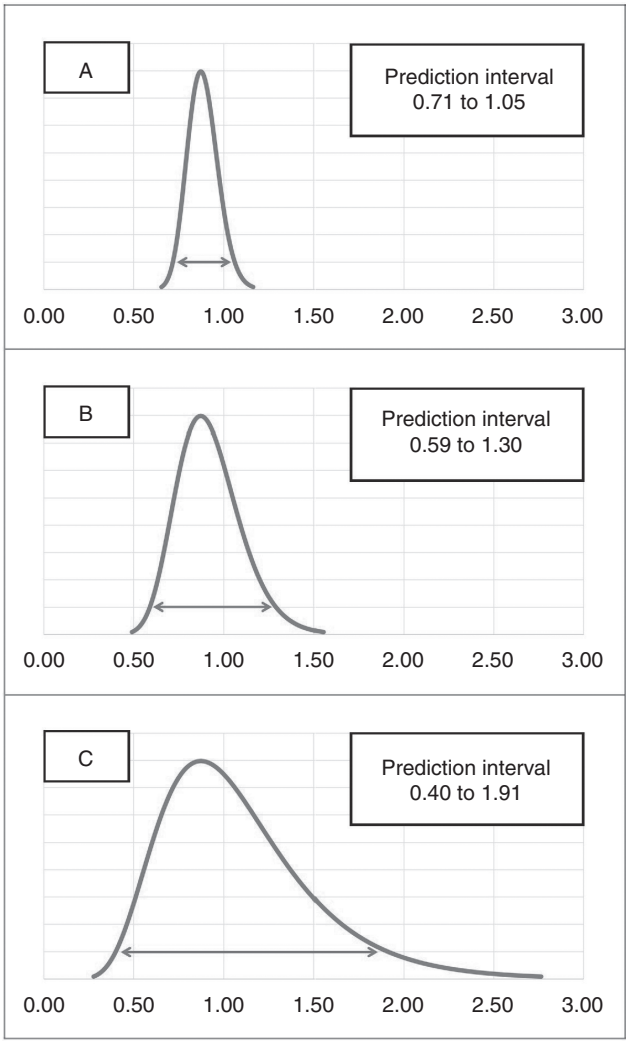


Figure 19.1 Alcohol use and mortality. Risk ratio < 1 favors drinkers. Three possible distributions of true effects.

Note that the distribution of effects is assumed to be symmetric in log units. It appears to be skewed because the plot uses the risk ratio rather than the log risk ratio on the X-axis.

THE *Q*-VALUE AND THE *p*-VALUE DO NOT TELL US HOW MUCH THE EFFECT SIZE VARIES

The statistics that most papers report for heterogeneity include the *Q*-value and the *p*-value. In the current analysis, the *Q*-value is 96.85 with 32 degrees of freedom, and

the *p*-value for a test of the null hypothesis (that the true effect size is the same in all studies) is <0.001. Based on these statistics, there is no way of knowing whether the distribution of effects resembles A, B, or C. The *Q*-value is the sum of squared deviations on a standardized scale and is driven by the number of studies and the extent of dispersion. The same applies to the *p*-value. Therefore, neither of these can serve as a surrogate for the amount of dispersion.

THE CONFIDENCE INTERVAL DOES NOT TELL US HOW MUCH THE EFFECT SIZE VARIES

The forest plot of a meta-analysis typically includes a line with the summary effect size and its confidence interval, which is sometimes displayed as a diamond. Researchers sometimes assume that the confidence interval tells us how widely the effect size varies across studies. It does not. The confidence interval speaks to the precision with which we have estimated the mean effect size. It says nothing about the dispersion in effects. See Chapter 17 for a detailed discussion of this point.

THE I^2 STATISTIC DOES NOT TELL US HOW MUCH THE EFFECT SIZE VARIES

Many researchers believe that the I^2 index tells us how much the effect size varies, but in fact, it does not. While many readers will find this statement surprising, the proof is both simple and compelling. In this analysis, I^2 was reported as 67%. Based on that, does the distribution of effects resemble panel A, B, or C? The answer is that we do not know.

There is a widespread belief that I^2 values of less than 25% represent low heterogeneity; values near 50% moderate heterogeneity; and values greater than 75% high heterogeneity. Since the I^2 value of 67% falls in the moderate to high interval, some researchers may expect that the dispersion in this case resembles panel B or C. That happens to be incorrect, since the dispersion in this case actually resembles panel A. However, the more important point is that given an I^2 value of 67%, the heterogeneity *could* resemble panel A, B, C, or an infinite number of other panels. I^2 does not tell us how much the effect size varies. It was never intended for that purpose and cannot provide that information except in special cases (Borenstein, 2019; Borenstein, 2020; Higgins, Hedges, & Rothstein, 2017; Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006; Mittlbock & Heinzl, 2006; Rucker, Schwarzer, Carpenter, & Schumacher, 2008).

WHAT I^2 TELLS US

If I^2 does not tell us how much the effect size varies, one might ask what it does tell us. To explain that, we need to provide some background.

When we discuss a meta-analysis, we need to distinguish between *true* effects and *observed* effects. The *true* effect size in any study is the effect size that we would

observe if we could somehow enroll the entire population in the study, so that we knew the effect size with no error. By contrast, the *observed* effect size is the effect size observed in the study's sample. This serves as an estimate of the true effect size but invariably underestimates or overestimates the true effect size due to sampling error.

When we perform a meta-analysis, we work with the *observed* effect size for each study, but what we really care about is the *true* effect size for each study. As it happens, the dispersion of *observed* effects tends to exceed the dispersion of *true* effects. To understand why this is, consider what would happen if we drew ten random samples from the same population. Since all samples are estimating the same parameter (the effect size in that one population), the variance in true effects is zero by definition. Nevertheless, the variance of observed effects will be greater than zero because of sampling error. In this case,

$$V_{OBS} = V_{ERR}, \tag{19.1}$$

where V_{OBS} is the variance of observed effects, and V_{ERR} is the variance due to sampling error. The same idea applies when the variance of true effects exceeds zero. In this case, the variance of observed effects is equal to the variance of true effects plus the error variance. That is,

$$V_{OBS} = T^2 + V_{ERR}, \tag{19.2}$$

where T^2 is the variance of true effects.

For the present discussion, the key point is that we have two distinct distributions. One is based on the variance of observed effects (which we see in the forest plot). The other is based on the variance of true effects (which tells us how much the effects actually vary). And, the variance of the former is greater than the variance of the latter. It would be useful to have a statistic that gives us the relationship between the two variances. That statistic is I^2, which is defined as

$$I^2 = \left(\frac{V_{TRUE}}{V_{OBS}}\right) \times 100 = \left(\frac{T^2}{V_{OBS}}\right) \times 100 = \left(\frac{V_{TRUE}}{V_{TRUE} + V_{ERR}}\right) \times 100. \tag{19.3}$$

In words, I^2 tells us what proportion of the observed variance is attributed to the variance in true effects rather than to sampling error (Borenstein, 2019; Borenstein, Higgins, Hedges, & Rothstein, 2017; Higgins & Thompson, 2002).

Critically, I^2 is a proportion, not an absolute value. To obtain the variance of true effects (T^2), we need to multiply I^2 by the variance of observed effects. That is,

$$T^2 = I^2 \times V_{OBS}. \tag{19.4}$$

The practical implications of this formula are evident in Figure 19.2. This is based on the same example as Figure 19.1, but in this case each panel displays two distributions. The inner curve is the same curve that we saw in Figure 19.1 and reflects the dispersion of true effects, with an arrow denoting the 95% prediction interval. The outer curve represents the dispersion of observed effects.

- In panel A, the observed effects all fall within the outer curve, which is relatively narrow in this case. When we multiply this by I^2, we find that the true effects fall in the relatively narrow interval of 0.71 to 1.05, as indicated by the inner curve.

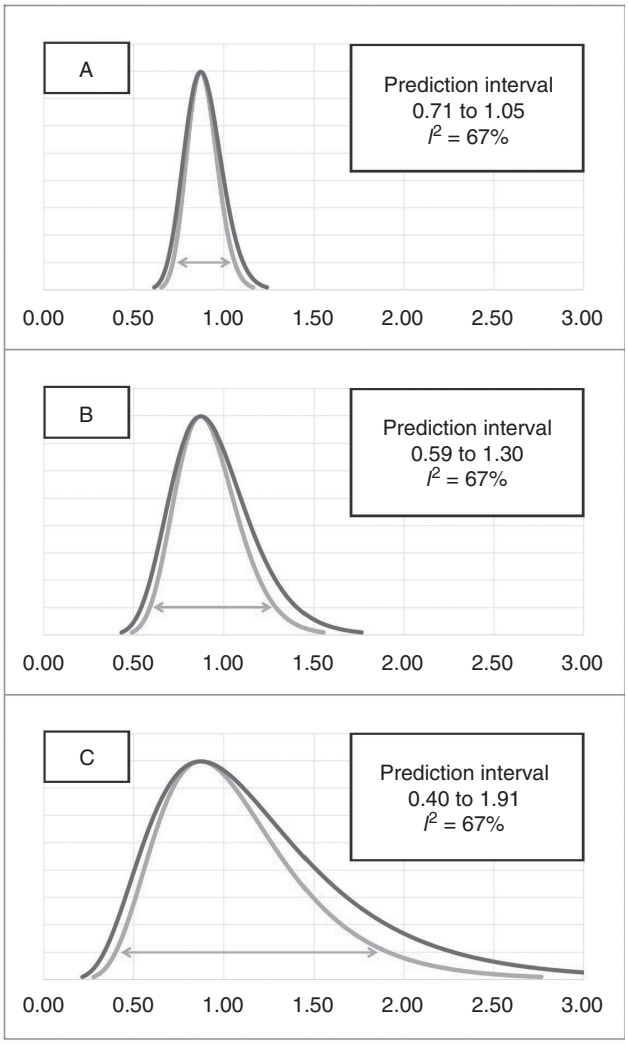


Figure 19.2 Alcohol use and mortality. Risk ratio < 1 favors drinkers. Three possible distributions of true effects (inner) and observed effects (outer).

- In panel B, the observed effects again fall within the outer curve, which is relatively wide in this case. When we multiply this by I^2, we find that the true effects fall in the relatively wide interval of 0.59 to 1.30, as indicated by the inner curve.
- In panel C, the observed effects again fall within the outer curve, which is even wider in this case. When we multiply this by I^2, we find that the true effects fall in even wider the interval of 0.40 to 1.91, as indicated by the inner curve.

THE I^2 INDEX VS. THE PREDICTION INTERVAL

If we want to know what proportion of the variance in observed effects is attributed to variance in true effects, we look at the relationship between the two curves. In all three panels, the relationship between the inner curve and the outer curve is the same, with the variance of true effects being 67% as large as the variance of observed effects. In all three cases, I^2 is 67%. This is the domain of I^2 – it addresses the *ratio* of true to total variance.

By contrast, if we want to know *how much* the effect size varies, we are asking for an absolute measure of dispersion. In panel A, the effects fall in the interval of 0.71 to 1.05. In panel B, they fall in the interval of 0.59 to 1.30. In panel C, they fall in the interval of 0.40 to 1.91. This is the domain of the prediction interval – it addresses the extent of the dispersion on an absolute scale (Borenstein, 2019, 2020; Borenstein *et al*., 2017; IntHout, Ioannidis, Rovers, & Goeman, 2016).

Since I^2 is a ratio, the I^2 value of 67% could correspond to any of these panels. As it happens, the observed effects correspond to the outer curve in panel A, and so the true effects correspond to the inner curve in panel A. Computations are presented at the end of this chapter.

THE PREDICTION INTERVAL

When we ask about heterogeneity in a meta-analysis, we want to know how much the effect size varies across studies. As discussed above, the I^2 index does not provide this information. The index that does provide this information is the prediction interval (Borenstein, 2019; Michael Borenstein *et al*., 2017; Chiolero, Santschi, Burnand, Platt, & Paradis, 2012; Graham & Moran, 2012; Guddat, Grouven, Bender, & Skipka, 2012; Higgins, Thompson, & Spiegelhalter, 2009; Riley, Higgins, & Deeks, 2011).

When we perform a random-effects analysis, we assume that the studies in the analysis are a random (or at least representative) sample of studies in some universe of interest, and our goal is to make inferences about that universe. The 95% prediction interval is the interval that includes the true effect size for 95% of all populations in that universe.

Figure 19.3 is a forest plot of the studies in the motivating example. The last line on the plot [A] shows the mean effect size of 0.87 with a confidence interval of 0.83 to 0.91. The confidence interval is an index of *precision*, and it speaks to the precision with which we have estimated the mean. In 95% of all analyses, the true *mean* for the universe of comparable studies will fall within the confidence interval. The confidence interval, shown here as a line, is often shown as a diamond. It has the same meaning in either case.

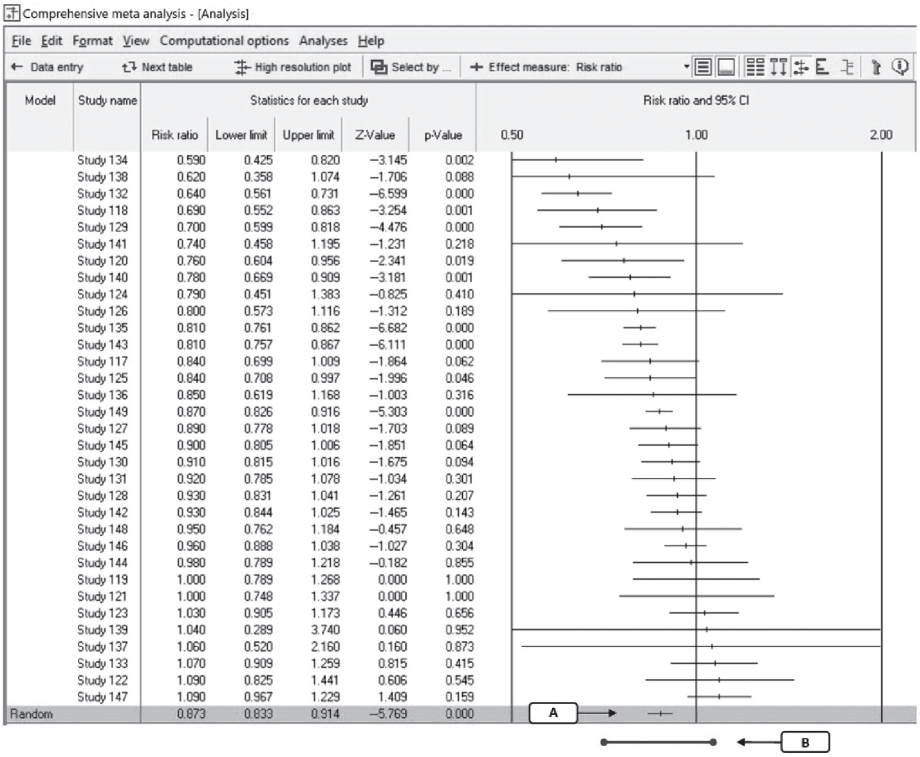


Figure 19.3 Alcohol use and mortality (Forest plot). Risk ratio < 1 favors drinkers.

The line immediately below the plot [B] shows the prediction interval of 0.71 to 1.07. The prediction interval is an index of *dispersion*, and it speaks to the heterogeneity of true effects. In some 95% of all comparable populations, the risk ratio will fall in this interval.

Researchers sometimes confuse the prediction interval with the confidence interval, but the two address entirely separate issues.

The 95% confidence interval may be estimated using

$$CI = M \pm 1.96(SE), \tag{19.5}$$

where SE is the standard *error* of the mean effect size. By contrast, the 95% prediction interval may be estimated using

$$PI = M \pm 1.96(T), \tag{19.6}$$

where T is the standard *deviation* of the effect size.

Both of these formulas are simplified versions of the formulas that we would use in practice. We use these here to highlight the difference between the confidence interval

(which is based on the standard *error*) and the prediction interval (which is based on the standard *deviation*).

In practice, one might employ the Knapp–Hartung adjustment when computing the confidence interval, as discussed in Chapter 26. Similarly, we would always recommend using the formulas discussed in Chapter 17 when computing the prediction interval. These formulas adjust the width of the interval to account for the fact that the statistics included in these formulas are estimated with error.

PREDICTION INTERVAL IS CLEAR, CONCISE, AND RELEVANT

The prediction interval is concise and unambiguous. If we report that the risk ratio varies from 0.71 in some populations to 1.07 in others, the reader understands what this means. The prediction interval is intuitive because it reports values on the same scale as the effect size. In the motivating example, the mean effect size is a risk ratio of 0.87 and the prediction interval tells us that in most comparable populations the true risk ratio will fall between 0.71 and 1.07.

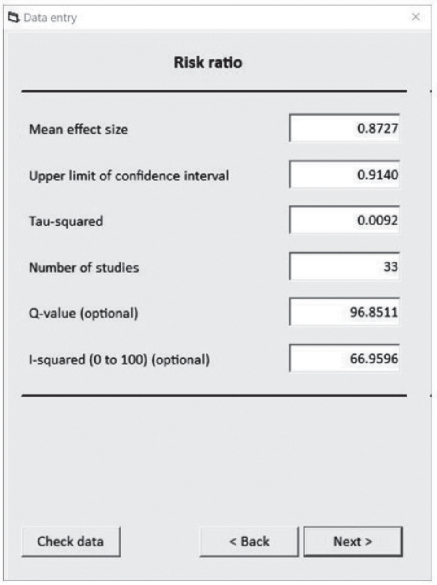
The prediction interval is on a meaningful scale. It tells us not only that the interval is a specific width, but that it ranges from one specific value to another specific value. As such, it allows us to distinguish not only between a case where the interval is 20 points wide from one where it is 40 points wide. It also allows us to distinguish between a case where those 40 points vary from trivially helpful to moderately helpful (on the one hand) vs. a case where the effects vary from harmful to helpful (on the other).

Most important, the prediction interval addresses the question that we have in mind when we ask about heterogeneity. If the analysis addresses the impact of an intervention, the prediction interval provides the information that speaks to the potential utility of that intervention.

COMPUTING THE PREDICTION INTERVAL

The formula for computing the prediction interval presented above (19.6) was intended as a conceptual formula. In practice, we need to modify this formula to account for the fact that we are working with an *estimate* of the true mean effect size and an *estimate* of the standard deviation of effects. Additionally, for some effect size indices we need to transform the estimates into log units for the computations. The relevant formulas are given in Chapter 17, and worked examples are presented in Chapter 18.

In practice, one would use a spreadsheet or computer program to compute the prediction interval. Figure 19.4 shows a program which is available on the book's website (www.Introduction-to-Meta-Analysis.com). The researcher enters the number of studies in the analysis, the risk ratio, the upper limit of the confidence interval, and tau-squared. The program generates the corresponding distribution. It also generates



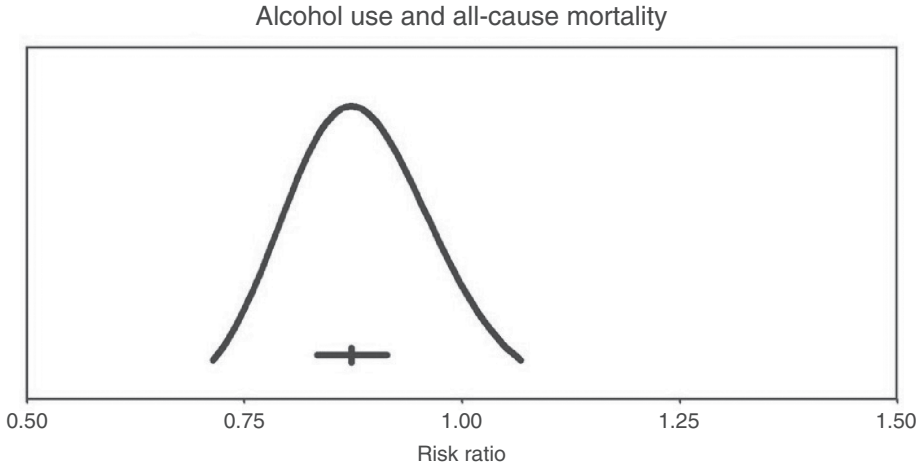


Figure 19.4 Alcohol use and mortality (true effects). Risk ratio < 1 favors drinkers.

the caption *The true effect size in 95% of all comparable populations falls in the interval 0.71 to 1.07,* which is the prediction interval.

The formulas implemented in this program are explained in this volume, and in (Borenstein *et al*., 2017; Higgins *et al*., 2009; Riley *et al*., 2011). Other approaches to computing prediction intervals are discussed in Nagashima, Noma, & Furukawa

(2019); and Wang & Lee (2019). Computational details for this example are presented at the conclusion of this section.

HOW TO USE I^2

While I^2 does not tell us how much the effect size varies on an absolute scale, it does provide other information, as follows.

- 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 can be used to compare meta-analyses of the same set of data analyzed using different effect metrics. For example, raw mean differences and standardized mean differences will be associated with different amounts of heterogeneity, but it is not meaningful to compare values of T^2 between the two scales. Because I^2 statistic has a unit-less scale, it is legitimate to compare it between the two 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.

HOW TO EXPLAIN HETEROGENEITY

Virtually all papers that report a meta-analysis include a discussion of heterogeneity which follows a standard pattern. The researchers report Q, df, and a p-value, I^2, and T^2. None of these directly addresses the question that really matters, which is 'What is the interval over which the effects vary?' Ironically, the prediction interval, the one statistic that does address this question, is rarely reported.

The paragraph that follows is based on the motivating example and can be adapted for the results section of a paper. The paragraph includes all the statistics that readers (and journal editors) expect to see, but these are annotated to make it more likely that they will be interpreted correctly. Critically, the report also includes the prediction interval.

HOW MUCH DOES THE EFFECT SIZE VARY ACROSS STUDIES?

The Q-statistic provides a test of the null hypothesis that all studies in the analysis share a common effect size. If all studies shared the same effect size, the expected value of Q would be equal to the degrees of freedom (the number of studies minus 1). The Q-value is 96.851 with 32 degrees of freedom and $p < 0.001$. We can reject the null hypothesis that the true effect size is the same in all these studies. The I^2 statistic is 67%, which tells us that 67% of the variance in observed effects reflects variance in true effects rather than sampling error. T^2, the variance of true effect sizes, is 0.009 in log units. T, the standard deviation of true effect sizes, is 0.096 in log units. If we assume that the effects are normally distributed (in log units), we can estimate that the prediction interval for the risk ratio is 0.71 to 1.07. The true effect size for any single population will usually fall in this range.

CAVEATS

All heterogeneity statistics will only be reliable if certain assumptions are met. In particular, we need to have a sufficient number of studies, and these studies must be a random sample of the intended universe. We also assume that the effects are normally distributed on the relevant scale. There is no consensus on what would be a sufficient number of studies to yield reliable estimates, but ten studies would be a useful minimum in most cases. With fewer than ten studies, T^2 (which feeds into the prediction interval) is estimated erratically and may give rise to prediction intervals that are inappropriately narrow or unhelpfully wide. While this caveat applies to I^2 and T^2 as well as the prediction interval, it is of particular import for the prediction interval since researchers understand what that interval means and will actually use it in discussing the utility of an intervention.

CONCLUSION

When we ask about heterogeneity in effects, we intend to ask how much the effect size varies across studies. We want to know the extent of the variation – *Does the effect size vary over 10 points or 50 points?* We also want to know the limits of the variation on an absolute scale – *Is the intervention always helpful, or is it helpful in some cases and harmful in others?*

The I^2 index has become the primary index for reporting heterogeneity in a meta-analysis and is widely interpreted as telling us how much the effect size varies across studies. However, this interpretation is fundamentally incorrect. The I^2 index is a proportion, not an absolute amount. It tells us what proportion of the variance in observed effects is attributed to variance in true effects. It does not tell us how much the effect size varies across studies.

The index that does tell us how much the effect size varies across studies is the prediction interval. This index is intuitive and concise. It reports the interval using the

same scale as the effect size itself. It gives us not only the width of the interval but the limits, so we know if the intervention is consistently helpful, or if it may be harmful in some cases. This statistic addresses the issue that we care about, and that many researchers *think* is being addressed by I^2. The inclusion of this interval in reports of heterogeneity will allow for a more informed discussion of the potential utility of any intervention and should be made common practice.

FURTHER READING

The original papers on I^2 are Higgins and Thompson (2002); Higgins, Thompson, Deeks, and Altman (2003). For a more detailed discussion of the issues raised in this section, see Borenstein *et al*. (2017).

(For related papers, see Borenstein, 2019, 2020; Coory, 2009; Higgins, 2008; Higgins *et al*., 2009; Huedo-Medina *et al*., 2006; IntHout *et al*., 2016; Ioannidis, 2008a; Riley *et al*., 2011; Rucker *et al*., 2008).

SUMMARY POINTS

- When we ask about heterogeneity, we want to know how the effect size varies across studies. The statistics typically reported for heterogeneity do not provide this information.
- This information is not provided in a useful format by the Q-value, nor by T^2. There is a widespread belief that the I^2 index in a meta-analysis tells us how much the effect size varies across studies, but this belief is fundamentally incorrect.
- The only statistic that directly reports this information is the prediction interval. The prediction interval tells us how much the effect size varies. It tells us whether the intervention is consistently helpful, or helpful in some populations but harmful in others. This is the information that we need to make informed decisions about the potential utility of the intervention.

THE MEANING OF I^2 IN FIGURE 19.2

The purpose of Figure 19.2 is to show how I^2 reflects the relationship between the inner curve (true effects) and the outer curve (observed effects). In this example, I^2 is 67%, which tells us that the ratio is the two variances if 0.67. However, it may not be clear how we see this in the plot. The computations for Panel A in Figure 19.2 are given in Table 19.1. In practice, we would bypass these computations and compute the prediction interval directly. This section is intended only to explain the relationship among the indices.

When we are working with risk ratios, data are converted to natural log units and all computations are performed in this metric. Therefore, most columns in the table are

Table 19.1 Relationship between observed effects and true effects in Figure 19.2, Panel A.

	Log units				Risk ratio units
	Mean	Variance	Standard deviation	Interval	Interval
True	−0.136154	$T^2 = 0.009222$	$T = 0.096031$	−0.324375 to 0.052067	0.722979 to 1.053446
Observed	−0.136154	$S^2 = 0.013772$ $I^2 = 66.9596\%$	$S = 0.117356$ $I = 81.8288\%$	−0.366172 to 0.093864	0.693383 to 1.098410

in these units. After the intervals are computed, they are converted back to risk ratio units as in the right-most columns.

The I^2 index is a ratio of variances,

$$I^2 = \frac{T^2}{S^2} = \frac{0.009222}{0.013772} = 0.669596 \approx 67\%. \tag{19.7}$$

To move from the variance of the outer curve to the variance of the inner curve, we would use

$$T^2 = S^2 \times I^2 = 0.013772 \times 0.669596 = 0.009222. \tag{19.8}$$

On that basis, many researchers expect that the inner curve will be 67% as wide as the outer curve. In fact, though, the ratio applies to the *variance* of the two distributions, which is a squared metric. By contrast, the distributions are in linear units, and so based on standard deviations rather than variances. Rather than work with the squared metric (I^2), we work with the linear metric (I).

The I index is a ratio of standard deviations,

$$I = \frac{T}{S} = \frac{0.096031}{0.117356} = 0.818288, \tag{19.9}$$

or simply

$$I = \sqrt{I^2} = \sqrt{.669596} = 0.818288 \approx 82\%. \tag{19.10}$$

To move from the standard deviation of the outer curve to the standard deviation of the inner curve, we would use

$$T = S \times I = 0.117356 \times 0.818288 = 0.096031. \tag{19.11}$$

This is what we see in the plot – the standard deviation of the inner curve is 82% as large as the standard deviation of the outer curve.

To compute the 95% interval for observed effects we use

$$OBS_{LL} = M - 1.96(S) = -0.136154 - 1.96(0.117356) = -0.366172 \tag{19.12}$$

$$OBS_{UL} = M + 1.96(S) = 0.136154 + 1.96(0.117356) = 0.093864. \tag{19.13}$$

We then convert these log values into risk ratio units, using

$$OBS_{LL} = \exp(-0.366172) = 0.693383 \tag{19.14}$$

$$OBS_{UL} = \exp(0.093864) = 1.098410. \tag{19.15}$$

To compute the 95% interval for the inner curve at the location of the arrow, we use the same formula, but substitute the standard deviation of true effects (T) for the standard deviation of observed effects (S). Concretely

$$PRED_{LL} = M - 1.96(T) = -0.136154 - 1.96(0.096031) = -0.324375 \quad (19.16)$$

$$PRED_{UL} = M + 1.96(T) = -0.136154 + 1.96(0.096031) = 0.052067. \quad (19.17)$$

We then convert these log values into risk ratio units, using

$$PRED_{LL} = \exp(-0.324375) = 0.722979 \quad (19.18)$$

$$PRED_{UL} = \exp(0.052067) = 1.053446 \quad (19.19)$$

The distribution of observed effects is a hypothetical distribution that allows us to illustrate the meaning of I^2. In real life, such a distribution would only exist if the error variance was identical for all studies, which is never the case.

The formulas used here to compute the prediction interval are only intended for the purpose of illustrating the relationship between the curves. For that purpose, we wanted to use a formula that isolates the difference between the distribution of true effects vs. observed effects. By contrast, to compute the prediction interval in practice, we would use the formulas presented earlier that use the t distribution rather than the Z distribution, and that take into account the error variance in estimating the mean. In this example, the prediction interval based on the correct formulas (0.71 to 1.07) is only slightly wider than the one based on the naïve formulas (0.72 to 1.05). This is true in this example because we have a substantial number of studies and a precise estimate of the mean. However, it would be a mistake to generalize from this example and assume that we can always use the naïve formulas. Often, the difference between the naïve formula and the correct formula will be substantial, and so we should always use the latter.