CHAPTER 20

Classifying Heterogeneity as Low, Moderate, or High

INTRODUCTION

In recent years, researchers have developed the practice of classifying heterogeneity as being low, moderate, or high based on the value of I^2. For example, some classify an I^2 value of 25% or less as low; 50% as moderate; and 75% or more as high. These classifications were proposed by one of the authors of this book, with specific reference to values that one might expect to see in meta-analyses of clinical trials in the Cochrane Database of Systematic Reviews. Unfortunately, such interpretations are widely applied inappropriately. Here we argue that the idea of classifying heterogeneity based on I^2 should be strongly discouraged.

INTEREST SHOULD GENERALLY FOCUS ON AN INDEX OF ABSOLUTE HETEROGENEITY

The first reason that we should not use this classification system is that when we talk about heterogeneity as being low, moderate, or high, we are talking about the *absolute* amount of heterogeneity. It follows that any classification scheme intended to address this goal must be based on an index that reports the *absolute* amount of heterogeneity. It could not be based on I^2 since this index does not tell us how much the effects vary. Consider the following two examples.

Holst, Petersen, Haase, Perner, and Wetterslev (2015) looked at the relationship between two treatment conditions (restrictive vs. liberal criterion for blood transfusion) and outcome. An odds ratio less than 1.0 would indicate that patients treated with

Introduction to Meta-Analysis, Second Edition.
Michael Borenstein, Larry V. Hedges, Julian P.T. Higgins, and Hannah R. Rothstein.
© 2021 John Wiley & Sons Ltd. Published 2021 by John Wiley & Sons Ltd.

the restrictive strategy were more likely to have a good outcome. The odds ratio was 0.96 with a confidence interval of 0.78 to 1.18. We will refer to this as the transfusion analysis.

Sorita *et al*. (2014) looked at the relationship between two treatment conditions (patients who presented at the hospital with acute MI during normal hours vs. off hours) and short-term mortality. The mean effect size is an odds ratio of 1.06 with a confidence interval of 1.04 to 1.09, indicating that patients who presented during off hours had a higher risk of death. We will refer to this as the off-hours analysis.

In the transfusion analysis, the I^2 statistic was 29%. In the off-hours analysis, the I^2 statistic was 75%. On that basis, many would assume that the effects varied more widely in the second analysis as compared with the first. As it happens, the opposite is true. In Figure 20.1, the top panel shows the distribution of effects for the transfusion analysis, and the bottom panel shows the distribution of effects for the off-hours analysis. When I^2 was 29%, the effects vary from 0.60 to 1.51. By contrast, when I^2 was 75%, the effects vary from 0.97 to 1.17. Thus, the *lower* value of I^2 corresponds to the *greater* amount of dispersion.

Additionally, if we were to use these values of I^2 to classify the dispersion, the dispersion in the top panel would be classified as low, while the dispersion in the bottom panel would be classified as high. This is clearly misleading since the dispersion in the top panel (classified as low) is roughly five times as wide as that in the bottom panel (classified as high).

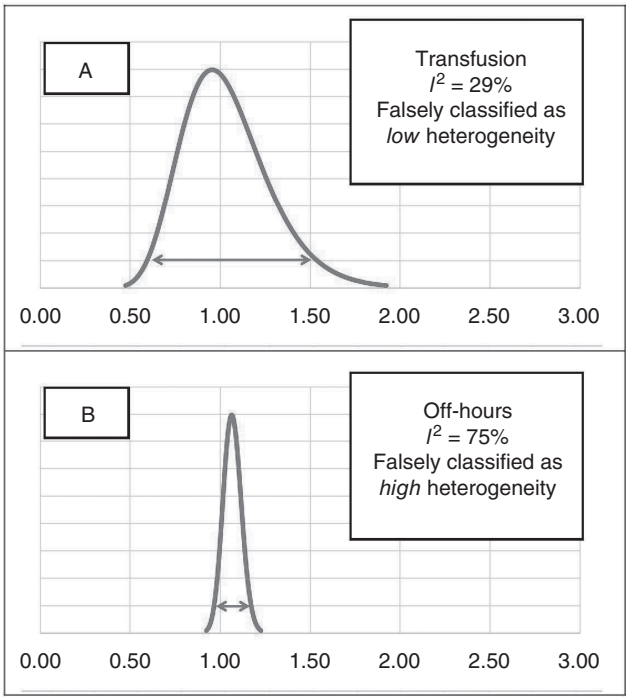

Figure 20.1 True effects for two meta-analyses.

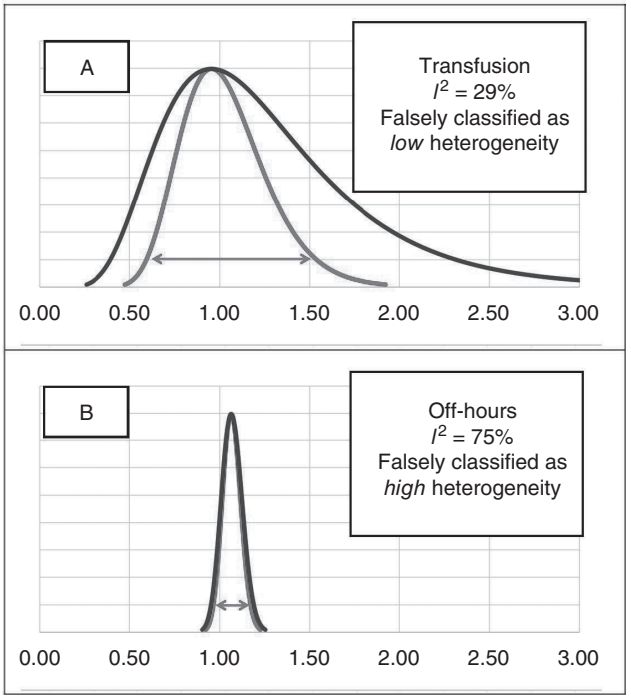

Figure 20.2 True effects (inner) and observed effects (outer) for two meta-analyses.

This is inexplicable to persons who believe that I^2 is an index of absolute dispersion. However, it makes sense to those who understand that I^2 tells us the relationship between the true effects and the observed effects. This is illustrated in Figure 20.2. In the top panel, the variance of true effects is only 29% as large as the variance of observed effects, and so I^2 is 29%. In the bottom panel, the variance of true effects is 75% as large as the variance of observed effects, and so I^2 is 75%. As it happens, the *observed* variance in the top panel is substantially greater than the *observed* variance in the bottom panel. In this case, 29% of the (relatively little) observed variance in the top panel turns out to be more than 75% of the (relatively large) observed variance in the bottom panel.

The use of the misleading classifications has serious implications. In the transfusion analysis, based on the (incorrect) classification of low heterogeneity one might assume that the mean effect size applies to all populations. In fact, though, the plot suggests that there will be some populations where the restrictive approach is much more effective, and other populations where the conservative approach is much more effective. Conversely, in the off-hours analysis, based on the (incorrect) classification of high heterogeneity, there would be some populations where the additional risk was trivial and some where it was substantial. In fact, though, the effect was relatively consistent across all comparable populations.

To be clear, these examples were not selected at random. If we were to choose two analyses at random, the analysis with the higher value of I^2 would have more variance

on average. Rather, these examples were selected to make the point that we cannot classify heterogeneity as being low, moderate, or high based on I^2. The I^2 index does not tell us how much the effects vary. Indeed, as in this example, it cannot reliably tell us which of two analyses has more dispersion.

Note that the distributions of effects are assumed to be symmetric in log units. They appear to be skewed because the plots use the risk ratio rather than the log risk ratio on the X-axis.

THE CLASSIFICATIONS LEAD THEMSELVES TO MISTAKES OF INTERPRETATION

While it would be possible to develop objective standards for what constitutes a low, moderate, or high amount of heterogeneity from a statistical perspective, the fact remains that researchers and readers understand these terms in the colloquial sense. If we report that the heterogeneity is 'low', this will be taken to mean that the impact of the intervention is consistent in a clinical or substantive sense. However, the clinical interpretation depends on the context. Variation in effects that we might consider to be trivial in one context might be clinically important in another. And even for a given context, readers will have different ideas of what 'low' or 'moderate' means.

By contrast, the prediction interval reports the actual range of effects and ensures that all readers have a common understanding of the dispersion.

CLASSIFICATIONS FOCUS ATTENTION IN THE WRONG DIRECTION

Finally, the use of a classification system encourages researchers to focus primarily on the amount of dispersion, rather than the clinical or substantive implications of the dispersion. When considering the potential utility of an intervention, we need to consider the larger picture, which requires a synthesis of the mean effect size and the heterogeneity. For example, rather than ask about the amount of variation, we might ask if the effect is consistently helpful, or if it is helpful in some cases and harmful in others.

We return to that in chapter 24

SUMMARY POINTS

- The practice of classifying heterogeneity as being low, moderate, or high based on the value of I^2 should be emphatically discouraged.
- The I^2statistic does not tell us how much the effect size varies, and therefore, a classification based on I^2 cannot tell us how much the effect size varies.
- Dispersion classified as 'low' in one meta-analysis could be substantially greater than dispersion classified as 'high' in another meta-analysis.