ELSEVIER

Journal of Clinical Epidemiology ■ (2022) ■

Journal of Clinical Epidemiology

COMMENTARY

In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies

Michael Borenstein*

Biostat, Inc, New York, NY, USA

Accepted 2 October 2022; Published online xxxx

In any meta-analysis it is important to know how the effect size varies across studies. For example, consider a meta-analysis where the mean odds ratio is 0.50. If the odds ratio consistently falls between 0.40 and 0.60, we might conclude that the treatment works equally well in all populations. If the odds ratio varies from 0.20 in some populations to 0.80 in others, we might elect to employ this treatment in all populations but would also want to know why it works better in some populations than in others. However, if the odds ratio varies from 0.10 (beneficial) in some populations to 2.5 (harmful) in others, the mean becomes largely irrelevant. In this case, we would need to determine where the treatment would be beneficial and where it would cause harm.

For this reason, all reports of a meta-analysis attempt to address heterogeneity. In medicine and epidemiology, the primary index employed to report heterogeneity is the I-squared (I^2) statistic [1,2]. While the use of I^2 for this purpose is ubiquitous, it is nevertheless a serious mistake based on a fundamental misunderstanding of this index. The I^2 statistic does not tell us how much the effect size varies. It was never intended to tell us how much the effect size varies and cannot provide that information except when I^2 is zero.

While this statement might sound surprising, a simple thought experiment should convince the reader that this is correct. Consider the following two meta-analyses.

1. The "off-hours" analysis

Sorita, et al. [3] performed a meta-analysis to see if patients suffering from a myocardial infarction were more likely to die if they arrived at a hospital during "off-hours" (the overnight shift or on weekends) as compared with patients who arrived during daytime hours. An odds ratio greater than 1.0 indicates that arriving at the hospital during off-hours is associated with increased risk. The mean odds ratio was 1.06 with a 95% confidence interval of 1.04−1.09, which tells us that patients who presented during off-hours were more likely to die on average. However, we also need to know how widely the effect size varies. Figure 1 shows three possible ways that the effect size might vary across populations. In panel A, the odds ratio consistently falls in the range of 0.96−1.17. In panel B the odds ratio in some studies is 0.73, while in others is 1.56. In panel C the odds ratio varies from 0.50 in some studies to 2.28 in others. It would be important to know which of these panels corresponds to the actual distribution of effects. If the distribution resembles panel A, we might conclude that the effect is always relatively small. On the other hand, if the distribution resembles panel B, we would conclude that there are some hospitals where presenting off-hours substantially reduces the risk of death and others where it substantially increases the risk of death. It would be imperative to identify factors that account for this difference. The same holds true to an even greater extent in panel C. For example, it could be possible that the hospitals at the left-hand side of the distribution have a cardiac team on duty at all hours, while hospitals at the right-hand side of the distribution do not.

So, which of these panels corresponds to the actual distribution of effects? In this analysis, I^2 is reported as 75%, which is typically classified as "high" heterogeneity. On that basis, most readers would assume that the distribution corresponds to panel B or C. That would be a mistake, since the distribution actually corresponds to panel A (as indicated by the check-mark in that panel).

2. The "transfusion" analysis

Holst, et al. [4] published a meta-analysis that looked at the relationship between two strategies for blood transfusion and risk of death. Under the "liberal" strategy, transfusion was employed when relatively liberal criteria were met. Under the "conservative" strategy, transfusion was employed only when more stringent criteria were met. An odds ratio less than 1.0 favors the conservative approach, while an odds ratio greater than 1.0 favors the liberal

Declaration of interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

* Corresponding author. Biostat, Inc, 473 West End Avenue, New York, NY 10024, USA. Tel.: + 1 201 541 5688; fax: +1 201 541 5526.

E-mail address: Biostat100@GMail.com.

https://doi.org/10.1016/j.jclinepi.2022.10.003

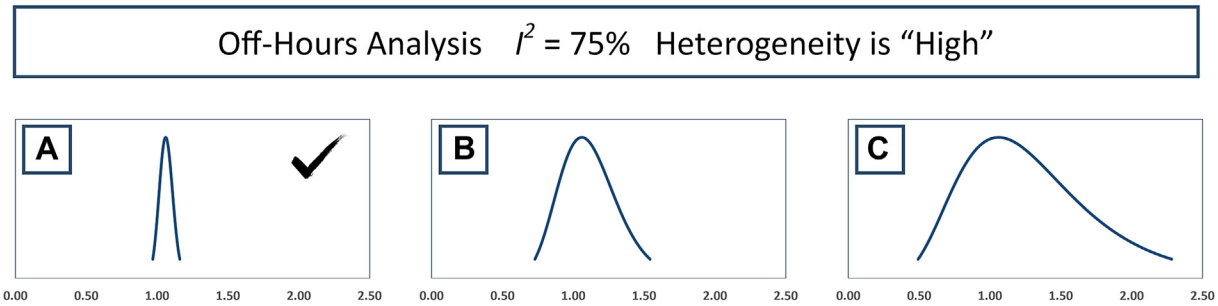

Fig. 1. Off-hours analysis – three possible distributions where I-squared is 75%.

approach. The mean odds ratio was 0.96 with a 95% confidence interval of 0.78−1.18. Thus, there is no evidence of a relationship on average. However, we also need to know how widely the effect size varies. Figure 2 shows three possible ways that the effect size might vary across populations. In panel A, the odds ratio consistently falls in the range of 0.74−1.23. In panel B the odds ratio in some studies is 0.69, while in others it is 1.33. In panel C the odds ratio varies from 0.56 in some studies to 1.63 in others. It would be important to know which of these panels corresponds to the actual distribution of effects. If the distribution resembles panel A, we might conclude that the effect is always relatively small. On the other hand, if the distribution resembles panel B we would conclude that there are some types of patients where the conservative strategy is better and others where the liberal strategy is preferable. It would be imperative to identify factors that account for this difference. The same holds true to an even greater extent in panel C.

So, which of these panels corresponds to the actual distribution of effects? In this analysis, I^2 is 29%, which is typically classified as "low" heterogeneity. On that basis, most readers would assume that the distribution corresponds to panel A or B. That would be a mistake since the distribution actually corresponds to panel C (as indicated by the check-mark).

Thus, in either analysis, I^2 did not provide useful information about the dispersion in effects. Indeed, it did not even tell us which of the two analyses had the greater amount of dispersion. In the analysis where I^2 was 29%, the effects varied over a substantially wider interval than the analysis where I^2 was 75%. If this seems odd, it's only because we think that I^2 reflects the amount of dispersion. It does not.

3. What I^2 tells us

To explain what I^2 does tell us, I need to provide some background information. In a meta-analysis, we distinguish between the true effect size and the observed effect size for each study. The true effect size is what we would see if we somehow knew the effect size in the population. The observed effect size is the effect size that we see in the sample. The latter serves as an estimate of the former but invariably differs from the former because of sampling error. For purposes of the present discussion, the relevant point is that the variance of observed effects is not the same as the variance of true effects. Specifically, the variance of the observed effects is equal to the variance of true effects plus the additional variance due to sampling error. Therefore, the distribution of observed effects tends to be wider than the distribution of true effects [5].

Once we understand that we are dealing with two distinct distributions, we might want to ask about the relationship between them. This is the function of I^2. It tells us what proportion of the observed variance reflects variance in true effects rather than sampling error. In the off-hours analysis, I^2 was 75% which tells us that 75% of the variance

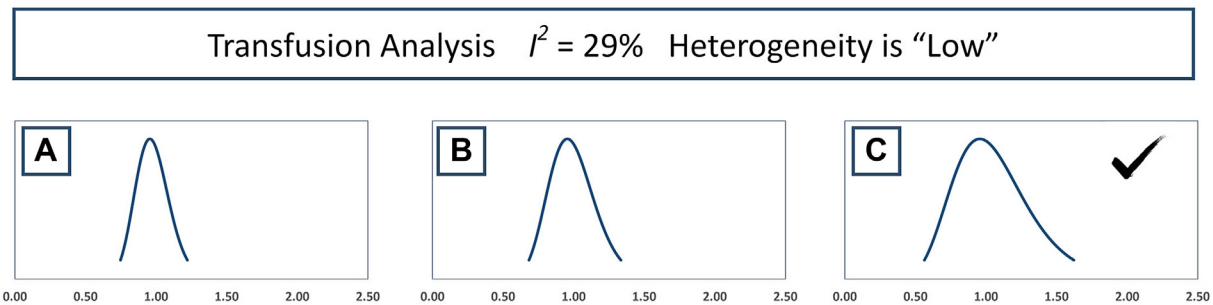

Fig. 2. Transfusion analysis – three possible distributions where I-squared is 29%.

in observed effects reflects variance in true effects. In the transfusion analysis, I^2 was 29% which tells us that 29% of the variance in observed effects reflects variance in true effects.

Critically, I^2 is a proportion, not a value on an absolute scale. It tells us what proportion of the variance in observed effects reflects variance in true effects rather than sampling error. It does not tell us how much the true effects vary. For that we would need to take account not only of I^2 but also the variance of observed effects.

In the off-hours analysis, the observed effects all fell in a narrow range. The I^2 value of 75% tells us that most of the observed dispersion was ''real'', but 75% of a small number is still a small number. By contrast, in the transfusion analysis the observed effects varied widely. The I^2 value of 29% tells us that only a small proportion of that dispersion is real. But a small proportion of a large number can still be a large number.

4. The prediction interval

If the I^2 statistic does not tell us how much the effect size varies, what statistic does provide that information?

That information is provided by the prediction interval. The prediction interval is defined as the mean plus or minus (roughly) two standard deviation of the true effects. If we can assume that the effects are normally distributed (in the relevant units) we can expect that the true effect size in 95% of all comparable populations will fall within this interval.

The prediction interval tells us how much the effect size varies using the same scale as the effect size itself. As such, it provides information about the absolute amount of dispersion, and it does so using an index that is intuitive and clear. Additionally, it frames this information in relation to the mean effect size rather than in the abstract, and thus provides information that is clinically relevant. In the off-hours analysis, instead of reporting that I^2 is 75% we would report that the odds ratio in any single population is expected to fall between 0.96 and 1.17. In the transfusion analysis, instead of reporting that I^2 is 29% we would report that the odds ratio varies from 0.56 in some populations to 1.63 in others.

When we report the prediction interval, we need to be aware of its limitations. The interval will only be accurate if based on a sufficient number of studies. Therefore, it might be best to report the interval only when this condition is met (this caveat applies also to T^2 and I^2). In computing the prediction interval, we assume that the effects are normally distributed in the relevant metric, an assumption that will not always be true. Therefore, if the interval includes effects on both sides of the null value, we should check to see if there actually are studies in the analysis that support this assumption. Additionally, we should adjust the interval to allow for the fact that the mean and the standard deviation are estimated rather than known [5]. Many computer programs for meta-analysis now offer the option to compute the prediction interval and will incorporate this adjustment [6].

5. An analogy

This paper is not intended to serve as a criticism of I^2. The I^2 statistic was intended to tell us what proportion of the observed variance reflects variance in true effects. When employed for this purpose, it is an entirely valid and useful statistic. However, when researchers ask about heterogeneity, they want to know how much the effect size varies. Therefore, they use the I^2 statistic as a surrogate for the amount of heterogeneity on an absolute scale. The I^2 statistic was never intended to be used for this purpose.

It is instructive to draw an analogy between this issue and the well-known problem that researchers sometimes conflate the *P*-value with the effect size in primary studies. The *P*-value was intended to tell us something about the viability of the null hypothesis. When employed for this purpose it is an entirely valid and useful statistic. However, when researchers ask about significance, they want to know about clinical significance rather than statistical significance. Therefore, they use the *P*-value as a surrogate for the magnitude of the effect. The *P*-value was never intended to be used for this purpose.

The analogy can be taken one step further. The fact that the *P*-value was employed as a surrogate for the effect size millions of times made this practice widely accepted but did not make it correct. Similarly, the fact that hundreds of thousands of papers employ I^2 as a surrogate for the amount of heterogeneity does not make this practice correct. I^2 is defined as a proportion, not an absolute amount. This is a definition, not an opinion [7].

6. Conclusion

To understand the potential impact of a treatment we need to understand how the effect size varies across studies. When a paper uses the I^2 value to quantify heterogeneity readers do not have any way of knowing how much the effect size varies. They will have a vague idea (at best) or a completely wrong idea (as in the examples cited earlier). One can imagine a group of clinicians discussing how to address the risks associated with off-hours presentation, based on the fact that I^2 was 75%. Each participant would have their own idea of what the dispersion actually looked like. The group would be trying to reach a consensus on what to recommend, when each member would be working with their own set of facts. Some would be thinking that the distribution resembled plot B in Figure 1, while others would be thinking that it resembled plot C. This by itself would be problematic since these two distributions might lend themselves to different strategies.

In this case, there is the additional irony that none of the discussants would be correct since the distribution actually resembles plot A.

To have an informed discussion about the potential utility of an intervention we need a statistic that reflects the distribution of true effects, and the prediction interval is the only statistic that serves this purpose. It provides the information that researchers and clinicians are asking for when they ask about heterogeneity—the information that they believe (incorrectly) is being provided by other statistics such as I^2. The prediction interval should be reported as a part of any meta-analysis where we have a sufficient number of studies to estimate it reliably [8].

In 1997, John C. Bailar III published an editorial entitled ''The Promise and Problems of Meta-Analysis'' [9]. He expressed a concern that researchers would not properly address heterogeneity, and that ''any attempt to reduce the results to a single value, with confidence bounds, is likely to lead to conclusions that are wrong, perhaps seriously so''. As it turns out, Bailer was prescient. Twenty-five years after that editorial, virtually all meta-analyses report the I^2 statistic. Since this does not provide the information that researchers would need to properly address heterogeneity, they resort to focusing on the mean with its confidence interval. They may reach conclusions that are wrong, perhaps seriously so [10].

By contrast, when we report the prediction interval, we do know the distribution of effects, and we can therefore take this into account when discussing the utility of the intervention. For example, we might report that the intervention is always clinically useful, or that it is clinically useful in some populations but not in others, or that it is beneficial in some populations but harmful in others. By making this discussion possible, the prediction interval can change the framework for how we think about the results of a meta-analysis.

For a PDF about heterogeneity and free software to compute prediction intervals, use the link to supplemental materials or visit www.Meta-Analysis.com/ClinEpid.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.10.003.

References

[1] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21:1539–58.
[2] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60.
[3] Sorita A, Ahmed A, Starr SR, Thompson KM, Reed DA, Prokop L, et al. Off-hour presentation and outcomes in patients with acute myocardial infarction: systematic review and meta-analysis. BMJ 2014;348:f7393.
[4] Holst LB, Petersen MW, Haase N, Perner A, Wetterslev J. Restrictive versus liberal transfusion strategy for red blood cell transfusion: systematic review of randomised trials with meta-analysis and trial sequential analysis. BMJ 2015;350:h1354.
[5] Borenstein M, Higgins JP, Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Res Synth Methods 2017;8(1):5–18.
[6] Borenstein M. Research note: in a meta-analysis, the I^2 index does not tell us how much the effect size varies across studies. J Phys 2020;66(2):135–9.
[7] Higgins JP. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. Int J Epidemiol 2008;37:1158–60.
[8] IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016;6(7):e010247.
[9] Bailar JC 3rd. The promise and problems of meta-analysis. N Engl J Med 1997;337:559–61.
[10] Borenstein M. Common mistakes in meta-analysis and how to avoid them. New Jersey: Biostat, Inc.; 2019.