References

[1] De Bac M. Se il feto prematuro è vivo va rianimato. [When the preterm fetus is born alive, resuscitation is mandatory] Corriere della Sera. Available at: http://archiviostorico.corriere.it/2008/febbraio/03/feto_vivo_rianimato_scontro_co_9_080203181.shtml (Accessed February 3, 2008).

[2] Law 127/97, *Gazzetta Ufficiale* 113, 17th May 1997 and Executive Regulations 403/98, *Gazzetta Ufficiale* 275, 24th November 1998.

[3] Fellegi IP, Sunter AB. A theory for record linkage. *J Am Stat Assoc* 1969;**64:**1183–210.

[4] Buitendijk S, Zeitlin J, Cuttini M, Langhoff-Roos J, Bottu J. Indicators of fetal and infant health outcomes. *Eur J Obstet Gynecol Reprod Biol* 2003;**111 (Suppl 1):** S66–77.

[5] Valls-i-Soler A, Carnielli V, Claris O *et al*. and Scientific Steering Committee of EuroNeoStat. EuroNeoStat: a European information system on the outcomes of care for very-low-birth-weight infants. *Neonatology* 2008; **93:**7–9.

doi:10.1093/ije/dyn185
Advance Access publication 4 September 2008

Heterogeneous views on heterogeneity

From NIKOLAOS A PATSOPOULOS, EVANGELOS EVANGELOU and JOHN PA IOANNIDIS*

The insightful and stimulating commentary by Julian Higgins[1] on our paper[2] raises several important issues that need to be clarified. First, we need to agree on nomenclature. The heterogeneity literature has been plagued by inconsistent terminology. Terms like 'heterogeneity', 'inconsistency', 'variation', 'diversity', 'between-study variance', 'variability', etc. are used interchangeably. While Higgins prefers the term 'inconsistency' for I^2, in other writings he has used the words 'variability' and 'heterogeneity' in association with this measure.[3] We believe that the term 'heterogeneity' is a nice word with roots going back to ΕΤΕΡΟΓΕΝΗΣ of Aristotle and ΕΤΕΡΟΓΕΝΩΣ of Sextus Empiricus. It can be applied to any of the popular metrics and tests, but then one simply has to specify which metric or test is exactly alluded to. 'Inconsistency' is also a nice, more recent word, but again we need to clarify what it refers to each time.

Higgins worries about 'the *post hoc* hypotheses that need to be thought up to explain why the excluded studies might be outlying or influential'. We were clear cut in our paper that this is indeed not an easy task. We believe that sensitivity analyses, as currently performed, are usually an invitation to *post hoc* data dredging with few or no rules in the game. This reduces their inferential reliability. However, this is a major reason why our proposed algorithms may offer one way to improve this free-lunch situation. There are two components to any sensitivity analysis. The first component is how it is done. The second component is how the results are interpreted. We argue that our method takes away much of the subjectivity in the first component. We do not wish to diminish the uncertainty that arises in the second component, and we wish that all meta-analysts recognize and acknowledge this uncertainty properly.

Higgins questions whether it is sensible to define a 'desired' threshold in terms of I^2 statistic. Although we agree that indeed '(some) heterogeneity is to be expected in (almost any) meta-analysis' and 'any amount of heterogeneity is acceptable, providing both that the predefined eligibility criteria for the meta-analysis are sound and that the data are correct', we believe that using thresholds to describe heterogeneity is an unavoidable consequence of the effort to translate statistical terms into real life. Higgins and colleagues have faced this problem, similarly recommending categorization of values for I^2 and assigning adjectives of low, moderate, and high heterogeneity or inconsistency.[4,5] In our article we have used these values of 50% and 25% for I^2, as traditional thresholds for large and moderate heterogeneity, respectively. This does not negate the need to recognize the major uncertainty in heterogeneity estimates,[6] but provides a standardized approach that can be applied consistently across meta-analyses.

Higgins argues in favour of using τ^2, the estimate of between-study variance, rather than I^2 in our paper, because I^2 depends also on the within-study precisions. Actually I^2 has become popular as a measure primarily due to the groundbreaking work of Higgins.[3,4] I^2 is one of the most commonly reported heterogeneity (or inconsistency) metrics, while the between-study variance τ^2 is rarely reported in the medical literature. I^2 has an intuitive interpretation, and it is comparable across meta-analyses with different numbers of studies or different types of effect metrics, whereas τ^2 is difficult both to understand and compare, according to Higgins' writings.[2] Therefore, we focused on I^2 in our paper. However, the algorithms that we have proposed are not applicable only to I^2. These are general methods that can be used with any kind of metric, e.g. τ^2. If another metric may be useful to apply more widely, we

* Corresponding author. Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece. E-mail: jioannid@cc.uoi.gr

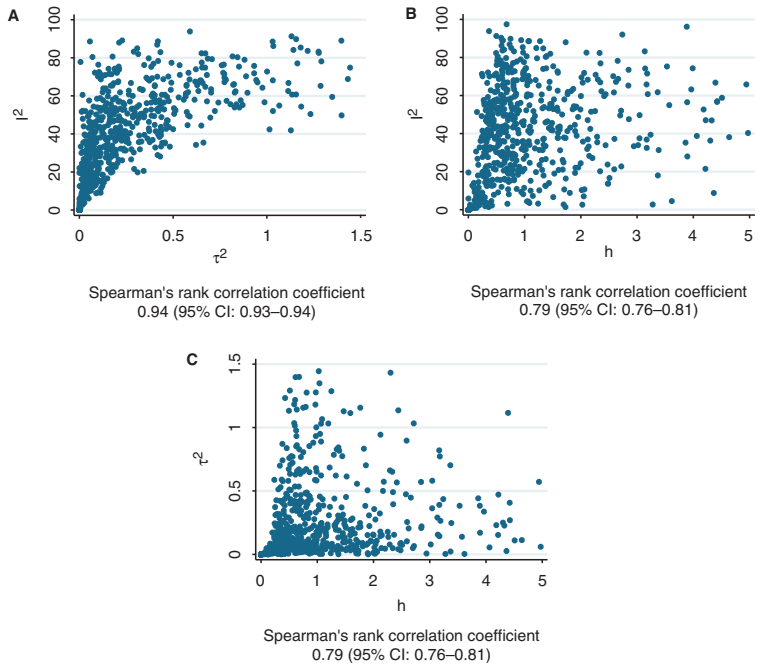

Figure 1 Correlation between (A) I^2 and τ^2 (B) h and τ^2 and (C) h and I^2 in the meta-analyses of the Cochrane database

would rather suggest h, the ratio of τ over the absolute value of the normalized summary effect (e.g. log odds ratio).[7] This h is not to be confused with yet another heterogeneity metric, H capital, which is the square root of the chi-squared heterogeneity statistic divided by its degrees of freedom.[2] The major problem with τ^2 for an epidemiologist is that it means almost nothing when seen in isolation. The same τ^2 value could be huge or negligible depending on what the summary effect is, and what impact that between-study variance has in shaping the upper and lower bounds of the summary effect confidence interval. We are in the process of implementing the sensitivity analysis algorithm on other heterogeneity metrics and we will release the new software module when it is properly beta-tested.

Regardless, at a practical level, τ^2 and I^2 tend to be largely concordant, when examined across many meta-analyses. In Figure 1 we illustrate the correlation between I^2 and τ^2 in the Cochrane meta-analyses database ($n = 1011$ meta-analyses) used in our article: the rank correlation coefficient is as high as 0.93. For comparison, the correlation coefficients for h against τ^2 and for h against I^2 are both 0.79 (Figure 1).

Higgins uses also a simulated example to illustrate why I^2 is not a sensible metric. He notes that I^2 behaves differently than τ^2 when there are different within-study errors among the studies. This is expected since these two metrics, although highly correlated, are not interchangeable. Specifically, the sequential algorithm is used to demonstrate that the drop in I^2 is not correlated with the drop in τ^2, rather τ^2 increased in the intermediate steps till the goal ($I^2 = 0\%$) is achieved. However, in that same example, using the combinatorial algorithm one can find a combination of four studies (D, E, F, G) whose exclusion results in an I^2 value of 0% (95% CI 0–73%) and also τ^2 of 0. The fact that the two algorithms give such different results reflects the complex and persistent inconsistency of this peculiar simulated meta-analysis. This is visible even in the forest plot. We argue that I^2 and τ^2 alone do not suffice to describe this complexity, and our sensitivity algorithms offer additional information.

To illustrate this, let us compare the meta-analysis simulated by Higgins (Figure 2A) vs another meta-analysis where, starting from the same data, all the individual effect sizes are coined to be $\geqslant 0$ (Figure 2B). The new simulated meta-analysis has an I^2 value of 84% (95% CI 65–90%) and τ^2 of 0.028, values almost identical to the ones in Higgins' example. The gross differences between these two meta-analyses can be seen even inspecting their forest plots, but both I^2 and τ^2 have very similar values. Applying the sequential algorithm to our simulated example, I^2 becomes 0% (95% CI 0–75%) and τ^2 becomes 0 with omission of a single study (study D). This example illustrates that meta-analyses with the same I^2 and τ^2 may require a very different number of studies to be omitted to decrease I^2 to a certain level or 0%. The underlying heterogeneity cannot be described or quantified by a single metric. We therefore recommend that routinely it

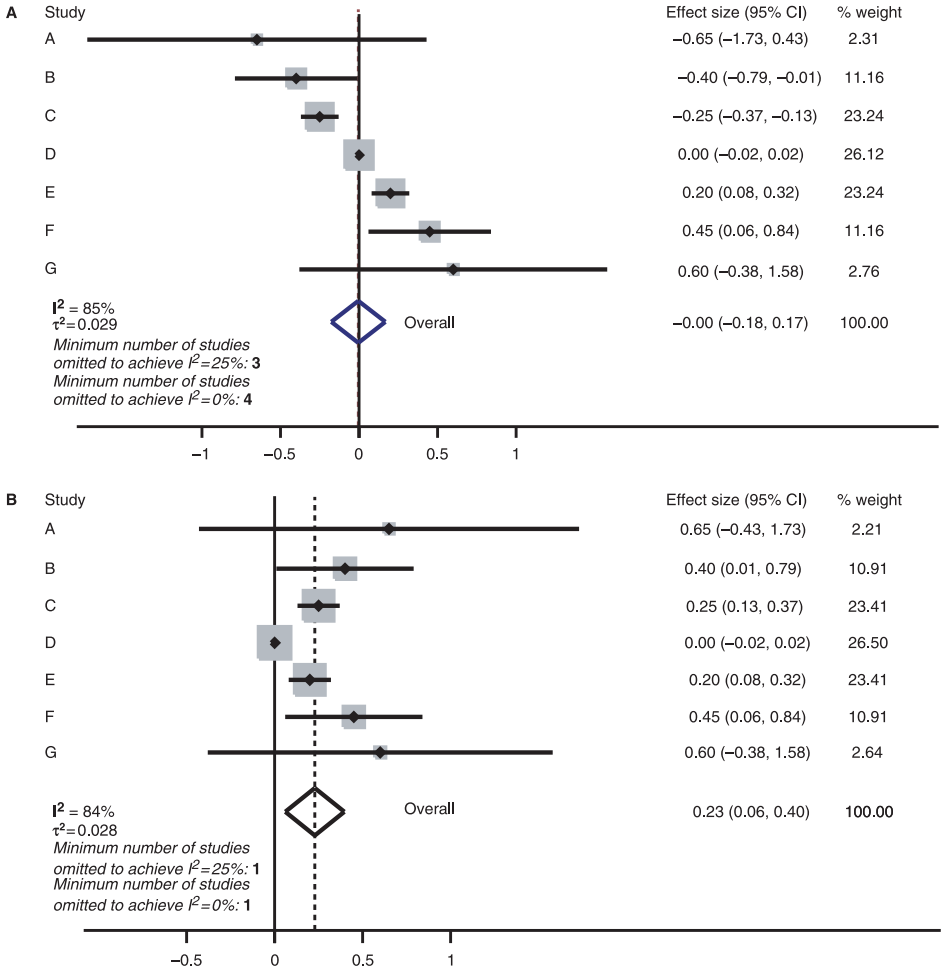

Figure 2 Simulated meta-analyses. (A) is the same as the second example of Higgins, while (B) is a simulated example where all effect sizes have been coined to be $\geqslant 0$, while otherwise the data are identical to (A). Note that even though the two meta-analyses have very similar I^2 and τ^2, they require a different number of studies to be omitted to diminish heterogeneity

may be worthwhile reporting this information besides just I^2 and/or τ^2 or any other heterogeneity metric.

References

[1] Higgins JP. Heterogeneity in meta-analyses should be expected and appropriately quantified. *Int J Epidemiol* 2008;**37:**1158–60.

[2] Patsopoulos NA, Evangelou E, Ioannidis JPA. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. *Int J Epidemiol* 2008;**37:**1148–57.

[3] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;**21:**1539–58.

[4] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Br Med J* 2003;**327:**557–60.

[5] Higgins J, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0* [updated February 2008]. The Cochrane Collaboration 2008.

[6] Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *Br Med J* 2007;**335:**914–16.

[7] Moonesinghe R, Khoury MJ, Liu T, Ioannidis JP. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc Natl Acad Sci USA* 2008;**105:**617–22.

doi:10.1093/ije/dyn235
Advance Access publication 21 October 2008