Epidemiology and Psychiatric Sciences (2012), **21**, 27–29. © Cambridge University Press 2011
doi:10.1017/S2045796011000643

This Section of *Epidemiology and Psychiatric Sciences* regularly appears in each issue of the Journal to cover methodological aspects related to the design, conduct, reporting and interpretation of clinical and epidemiological studies. The aim of these Editorials is to help developing a more critical attitude towards research findings published in international literature, promoting original research projects with higher methodological standards, and implementing the most relevant results of research in every-day clinical practice.

Corrado Barbui, *Section Editor* and Michele Tansella, *Editor* EPS

Heterogeneity: the issue of apples, oranges and fruit pie

M. Purgato[1,2]* and C. E. Adams[2]

[1] *Section of Psychiatry and Clinical Psychology, Department of Public Health and Community Medicine, University of Verona, Verona, Italy*
[2] *Division of Psychiatry, University of Nottingham, Nottingham, UK*

Heterogeneity refers to any kind of variation among studies contributing to the same outcome in a systematic review. There are three broad types of heterogeneity: clinical heterogeneity, methodological heterogeneity and statistical heterogeneity. In this paper, we describe these three types of heterogeneity and the main statistical approaches to measure heterogeneity.

Received 19 August 2011; Revised 19 August 2011; Accepted 23 August 2011

Key words: Heterogeneity, methods, meta-analysis, systematic review.

Systematic reviews and meta-analyses should represent convincing and reliable evidence relevant to many aspects of medicine and health care, providing summary estimates of the effects of treatments (Egger & Davey Smith, 1997). When results of the studies included for a given outcome show identical effects, we describe them as homogeneous. When there is variation in the findings of different trials contributing to the same outcome, there is some greater or lesser degree of heterogeneity. Homogeneity is not necessarily more desirable than heterogeneity. Both should be considered. Certainly, perceptions of diversity or its lack may influence meta-analysts on what data to combine, what data to avoid combining, what methods to use to combine and how to interpret the results they eventually get (Ioannidis, 2008).

Types of heterogeneity

There are three broad types of heterogeneity: ***clinical heterogeneity*** refers to variability in the participants, types of intervention and outcomes. For example, if a set of trials focusing on the effects of drug X for condition Y all conducted in different age groups, this is clinical heterogeneity (Fletcher, 2007). ***Methodological heterogeneity*** refers to variability in study design, conduct and risk of bias. If there is no clinical heterogeneity but randomisation in some trials is clearly open to being tampered with when in the others it is not, this is methodological heterogeneity. Finally, ***statistical heterogeneity*** exists when the observed intervention effects being more different from each other than one would expect due to random error (chance) alone (Higgins & Green, 2009). Statistical heterogeneity is the manifestation of clinical or methodological heterogeneity in the results of the trials.

Statistical heterogeneity can often be spotted by simple visual inspection of forest plots (Fig. 1). These plots show each individual study estimate (represented by a square) and confidence intervals (represented as

* Address for correspondence: Dr. M. Purgato, Section of Psychiatry and Clinical Psychology, Department of Public Health and Community Medicine, University of Verona, Piazzale L.A. Scuro 10, 37134 Verona, Italy.
(Email: marianna.purgato@univr.it)

https://doi.org/10.1017/S2045796011000643 Published online by Cambridge University Press

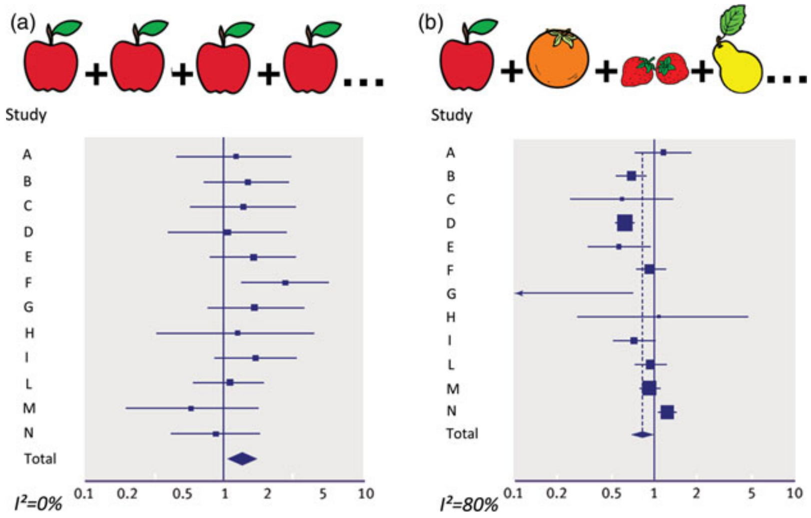

Fig. 1. Forest plots showing low heterogeneity level (a) and high heterogeneity level (b). (A colour version of this figure is available online at http://journals.cambridge.org/eps)

lines extending from the square to the upper and lower limits). The size of the square represents the weighting of each study in the overall synthesis. When confidence intervals overlap, heterogeneity is lower (Fig. 1a) than when confidence intervals do not overlap (Fig. 1b). Viewing data on graphs helps to have an immediate understanding but formal inferences should not depend on visual impressions alone (Ioannidis, 2008). Statistical approaches may help.

Measures of statistical heterogeneity

A chi^2 test alone was used in the past to determine whether heterogeneity was present. It assesses whether observed differences in results are compatible with chance alone. However, the test is of low power, depending as it does on the number of *studies* included in the synthesis – not the numbers within those studies.

More recently the I^2 metric was proposed as a measure of the impact of heterogeneity. I^2 represents the approximate proportion of total variability in point estimates that can be attributed to heterogeneity rather than chance. It takes values from 0 to 100%: a value of 0% indicates no observed heterogeneity and larger values mean increasing heterogeneity. Cut-offs are used to claim that the degree of heterogeneity is large or not; for example, when the measure of the values of the point estimates are in the company of a statistically significant chi^2 test, an I^2 of 50% could be taken as a cut-off for large heterogeneity and 25% as a cut-off for modest heterogeneity (Higgins *et al.* 2003).

Finally, Tau (τ) is an estimate of the amount of heterogeneity itself, expressed as the standard deviation of the underlying treatment effects across studies. This is the only genuine measure of heterogeneity, but is often more difficult to interpret. For example, when the treatment effect is expressed as an odds ratio, Tau refers to a standard deviation in log odds ratios (Borenstein *et al.* 2009).

What not to do

It is important not to pursue homogeneity ruthlessly, either as a researcher or reader. Heterogeneity can contribute to a deeper understanding. Eysenck was right many years ago when he pithily criticised meta-analysis as ‘mega-silliness’ because of inadvisable adding of ‘apples and oranges’ (Eysenck, 1978), but his cutting remarks were too sweeping and not helpful. He passed his negative judgment on pioneering attempts at evidence synthesis that came out of educational psychology and the need to understand the effects of psychological approaches (Smith & Glass, 1977). In this way, Eysenck undermined syntheses of mental health evidence and set it back decades. It took that length of time for the rather more slow-witted experts in evidence synthesis to suggest that adding apples and oranges is not bad at all if fruit pie is the desired outcome (Deeks *et al.* 2011). In these decades, Eysenck did not seem to have moved on (Eysenck, 1995) and seemed not to realise that sophistication of evidence synthesis had. Thoughtless acceptance or dismissal of either homogeneity or heterogeneity is not sensible.

What to do

The degree of diversity should be considered, investigated and considered again. With statistically homogeneous results, discussion should centre on what this tells us about the finding when methods, health care systems, participants, interventions and outcomes are often diverse. Genetically identical white rats in perfect experimental conditions yield homogeneous results. Human beings should not.

Modest heterogeneity when all findings essentially agree in the direction of effect may often be welcome, with the diversity encouraging wider generalisability of the results. However, even modest heterogeneity could belie interesting differences in the degree of effect that might become apparent with more power. Heterogeneity generates hypotheses that could be tested in further analyses. For example, moderate heterogeneity could be apparent and when clinical differences in the studies are reconsidered and those involving only children are removed, the findings become more homogeneous and less statistically significant.

Even when heterogeneity is considerable, if there is no obvious reason for it, then leaving the heterogeneity in the synthesis can be justified, provided that the results are appropriately interpreted. Sometimes heterogeneity is a specious and unhelpful finding. For example, in a meta-analysis of very large trials, confidence intervals will be very narrow. Even very little differences in the findings of individual trials around which confidence intervals are negligible may generate statistically significant heterogeneity of no clinical meaning.

Finally, modest or considerable heterogeneity when studies show results in directly opposite directions is another matter. The most common reason for this is typing in the results incorrectly, but if that is not the case there is probably some methodological or clinical cause that can be investigated by taking in and out trials that for some [preferably pre-specified] reason group in different clusters to others. In this case it is inadvisable to leave the heterogeneity and continue to blindly synthesise the trials or accept the results of such a synthesis.

Acknowledgement

We are grateful to Professor Julian Higgins, MRC Biostatistics Unit, Cambridge, for checking the manuscript.

Conflicts of Interest

C.E.A. is Co-ordinating Editor of a Cochrane Group.

References

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009). *Introduction to Meta-Analysis*. John Wiley and Sons: New York.

Deeks JJ, Higgins JPT, Altman DG, Cochrane Statistical Methods Group (2011). Chapter 9: Analysing data and undertaking meta-analyses [9.5.1 What is heterogeneity?]. In *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0* (ed. J. P. T. Higgins and S. Green). The Cochrane Collaboration. Retrieved from http://www.cochrane-handbook.org (updated March 2011).

Egger M, Davey Smith G (1997). Meta-analysis: potentials and promise. *British Medical Journal* **315**, 1371–1374.

Eysenck HJ (1978). An exercise in mega-silliness. *American Psychologist* **33**, 517.

Eysenck HJ (1995). Meta-analysis squared—does it make sense? *American Psychologist* **50**, 110–111.

Fletcher J (2007). What is heterogeneity and is it important? *British Medical Journal* **334**, 94–96.

Higgins JPT, Green S (2009). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2. Retrieved from http://www.cochrane-handbook.org (updated September 2009).

Higgins JPT, Thompson SG, Deeks J, Altman DG (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* **327**, 557–560.

Ioannidis JP (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice* **14**, 951–957.

Smith ML, Glass GV (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist* **32**, 752–760.

https://doi.org/10.1017/S2045796011000643 Published online by Cambridge University Press