Maßzahlen zur Heterogenität in Metaanalysen – kritisch diskutiert

Gerta Rücker

Institut für Medizinische Biometrie und Medizinische Informatik, Universitätsklinikum Freiburg
Deutsches Cochrane Zentrum, Freiburg

ruecker@imbi.uni-freiburg.de

IQWiG im Dialog, Köln, Freitag, den 17. Juni 2011

Outline

What is measured? – Sources of heterogeneity

How to measure? – Measures of heterogeneity

Why measuring heterogeneity at all?

What next?

Sources of heterogeneity in meta-analysis

Julian Higgins (Higgins, 2008, Title of a commentary):
"Heterogeneity in meta-analysis should be expected and appropriately quantified"

- **Clinical heterogeneity** in patient baseline characteristics, not necessarily reflected in the effect measure
- **Heterogeneity from study-related sources**, e.g. design-related heterogeneity
- **Small-study effects** - more about this below!
- **'Statistical heterogeneity'**, quantified on the effect measurement scale
    - term often used for a **treatment-study interaction** that may or may not be clinically relevant
    - Only this is what we are measuring when using popular measures such as Q or I^2

Fixed and random effects model

- **Fixed effect model** (x_i observed treatment effect in study i)

$$x_i = \mu + \sigma_i \, \epsilon_i, \quad \epsilon_i \sim N(0, 1)$$

  μ fixed global mean
  σ_i^2 within-study sampling variance, ϵ_i random error

- **Random effects model** (DerSimonian and Laird, 1986; Fleiss, 1993)

$$x_i = \mu + \sqrt{\sigma_i^2 + \tau^2} \, \epsilon_i, \quad \epsilon_i \sim N(0, 1)$$

  True study means vary randomly around a fixed global mean
  τ^2 between-study (heterogeneity) variance

Fixed and random effects model

- Pooled effect estimate: Weighted mean of the study estimates

$$\hat{x} = \frac{\sum w_i x_i}{\sum w_i}$$

- Inverse variance weights w_i, w_i^* and variance estimators v_F, v_R:

	Weights	Variance of pooled estimate
Fixed effect model	$w_I = \frac{1}{\hat{\sigma}_i^2}$	$v_F = \frac{1}{\sum w_i}$
Random effects model	$w_i^* = \frac{1}{\hat{\tau}^2 + \hat{\sigma}_i^2}$	$v_R = \frac{1}{\sum w_i^*}$

- $v_R \geq v_F$
- Large heterogeneity (large τ^2) $\Rightarrow$ Random effects model weights tend to be more similar $\Rightarrow$ Smaller studies get higher weights

Extended random effects model

- **Extended random effects model**
  Take account of possible small-study effects by allowing the effect to depend on the standard error:

  $$x_i = \mu + \sqrt{\sigma_i^2 + \tau^2}\;(\alpha + \epsilon_i), \quad \epsilon_i \sim N(0,1),$$

  where α is the bias introduced by small-study effects ('publication bias')
- α interpreted as the expected shift in the standardised treatment effect estimate for 'small' studies (infinite standard error):

  $$\mathrm{E}\left(\frac{x_i - \mu}{\sigma_i}\right) \to \alpha, \quad \sigma_i \to \infty$$

Measures of heterogeneity in meta-analysis: Cochran's Q

- Notation
  - k number of trials in a meta-analysis
  - Trial i ($i = 1, \ldots, k$): Treatment effect estimate x_i with SE s_i
  - $w_i = 1/s_i^2$ inverse variance weights
- **Cochran's Q:** Weighted sum of squared distances of the study means from the fixed effect estimate (Cochran, 1954)

$$Q = \sum_{i=1}^{k} w_i \left(x_i - \frac{\sum w_j x_j}{\sum w_j} \right)^2$$

- Under homogeneity χ^2-distributed with $k - 1$ degrees of freedom
- Exact distribution under heterogeneity derived by Biggerstaff and Jackson (2008)

Measures of heterogeneity in meta-analysis: Generalised Q

- **Generalised Q:** Weighted sum of squared distances of the study means from the random effects model estimate (DerSimonian and Kacker, 2007; Viechtbauer, 2007; Bowden et al., 2011)

$$Q = \sum_{i=1}^{k} w_i^* \left(x_i - \frac{\sum w_j^* x_j}{\sum w_j^*} \right)^2$$

- Under homogeneity χ^2-distributed with $k-1$ degrees of freedom
- Reiteration leads to an alternative estimator for τ^2 (Paule and Mandel, 1982)

Measures of heterogeneity in meta-analysis: τ^2

- **Between-study variance** τ^2, e.g., moment-based estimate (DerSimonian and Laird, 1986):

$$\hat{\tau}^2_{DL} = \max\left\{0, \frac{Q-(k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}\right\}$$

- Many alternative proposals for estimating τ^2, such as the ML or REML estimator (Knapp et al., 2006; Viechtbauer, 2007; DerSimonian and Kacker, 2007, and further refs)
- As τ is measured on the same scale as the effect, it can be directly used to quantify variability:
  - If studies with odds ratios of 0.8, 1 and 1.25 seem too heterogeneous to be pooled, this corresponds to a threshold of $\tau_0^2 = 0.05$

Measures of heterogeneity in meta-analysis: H^2 and R^2 (Higgins and Thompson, 2002)

- H^2 describes the inflation of the observed Q compared to what we would expect in the absence of heterogeneity:

$$H^2 = \frac{Q}{k-1}$$

- R^2 describes the quadratic inflation of the random effects confidence interval compared to that from the fixed effect model:

$$R^2 = \frac{v_R}{v_F}$$

Measures of heterogeneity in meta-analysis: I^2

- **I-squared** I^2 (Higgins and Thompson, 2002; Higgins et al., 2003)

$$I^2 = \max\left\{0, \frac{Q-(k-1)}{Q}\right\}$$

- I^2 is the proportion of variation in point estimates that is due to heterogeneity rather than within-study errors:

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}$$

given a so-called 'typical' within-study variance $\hat{\sigma}^2 = \frac{\sum w_i(k-1)}{(\sum w_i)^2 - \sum w_i^2}$

- I^2 increases with increasing precision/study size (Rücker et al., 2008)
- I^2 tends to 100% if sampling error approximates zero
- I^2 inapplicable as a measure of heterogeneity independent of the precision of the trials

Measures of heterogeneity in meta-analysis: D^2

- **Diversity** D^2 (Wetterslev et al., 2009)

$$D^2 = \frac{v_R - v_F}{v_R}$$

- Relative variance reduction when the model is changed from a random effects to a fixed effect model
- Like I^2, D^2 interpreted as a proportion:

$$D^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_D^2}$$

  where $\hat{\sigma}_D^2 = \frac{\hat{\tau}^2 v_F}{v_R - v_F}$ represents sampling error
- $D^2 \geq I^2$ for all meta-analyses

Measures of heterogeneity in meta-analysis: G^2

- **Adjusted for small-study effects:** G^2 (Rücker et al., 2010a)
- Based on the extended random effects model
- G^2 estimated by

$$G^2 = 1 - R^2_{reg} = \frac{\text{Residual sum of Squares}}{\text{Total Sum of Squares}}$$

  from regressing standardised shrunk treatment effects x_i'/s_i on $1/s_i$
- G^2 interpreted as the proportion of variation in the treatment effect that is not explained by a fixed effect model **that allows for small study effects**

Properties of measures of heterogeneity in meta-analysis

Measure	type	range	systematically increasing with number of studies[1]	systematically increasing with size of studies
τ^2	model parameter, τ interpretable on effect scale	$[0, \infty)$	no	no
Q	test statistic	$[0, \infty)$	yes	yes
Gen. Q	test statistic	$[0, \infty)$	yes	yes
H^2	test statistic	$[0, \infty)$	no	yes
R^2	test statistic	$[1, \infty)$	no	yes
I^2	test statistic	$[0, 1)$	no	yes
D^2	test statistic	$[0, 1)$	no	yes
G^2	adjusts for small-study effects	$[0, 1)$	no	no

[1] in meta-analysis

Relations between measures of heterogeneity (simplified)

Determine: from	H^2	I^2	R^2	D^2
$\hat\tau^2, \hat\sigma^2$ or $\hat\sigma^2_D$	$H^2 = \frac{\hat\tau^2+\hat\sigma^2}{\hat\sigma^2}$	$I^2 = \frac{\hat\tau^2}{\hat\tau^2+\hat\sigma^2}$	$R^2 = \frac{\hat\tau^2+\hat\sigma^2_D}{\hat\sigma^2_D}$	$D^2 = \frac{\hat\tau^2}{\hat\tau^2+\hat\sigma^2_D}$
v_F, v_R			$R^2 = \frac{v_R}{v_F}$	$D^2 = \frac{v_R-v_F}{v_R}$
Q	$H^2 = \frac{Q}{k-1}$	$I^2 = \frac{Q-(k-1)}{Q}$		
H^2		$I^2 = \frac{H^2-1}{H^2}$		
I^2	$H^2 = \frac{1}{1-I^2}$			
R^2				$D^2 = \frac{R^2-1}{R^2}$
D^2			$R^2 = \frac{1}{1-D^2}$	

$R^2 \geq H^2$, similar to H^2; $D^2 \geq I^2$, similar to I^2

G^2 cannot be directly derived from any of these

Common misinterpretation of I^2

Note: I^2 is not a population parameter, but a simple transformation of the test statistic Q!

- Misinterpretation of I^2 is common (Higgins, 2008; Rücker et al., 2008)
- Example I: Patsopoulos et al. (2008) present an algorithm that excludes studies from a meta-analysis aiming to achieve I^2 below a desired pre-set threshold
- Example II: Borm et al. (2009): *'The evidence provided by a single trial is less reliable than its statistical analysis suggests'*
  - Assuming a fixed 'true' I^2, the authors argue that P-values of single trials should be adjusted for heterogeneity
  - Observing larger I^2 values for large trials, they call for 'many small trials' instead of large trials
  - This is a misinterpretation of the role of I^2 (Rücker et al., 2009)
- The same considerations hold for D^2

Why measuring heterogeneity at all?

John Copas (personal communication):
I'm cautious about ideas of "measuring" statistical heterogeneity, since these are just open to abuse, like having some magical threshold below which we can say that "heterogeneity can be ignored".

Alex Sutton (from an open peer review[2]):
My way of conducting meta-analysis is to estimate τ^2 (ideally with uncertainty), if it is non-zero then I use a random effect model, if it is 0 it reduces automatically to a fixed effect model. In a sense I avoid Q, I^2 or other statistics or hypothesis tests to decide model choice. Please clarify why we need Q, I^2, D^2 etc – is it to help decide on model choice or simply quantify the degree of heterogeneity or both?

[2] Bowden et al. (2011)

Conclusions and open questions

- **Random effects model**
  - may provide a valid estimate of the global mean and its confidence interval
  - does not explain heterogeneity
  - is susceptible to small-study effects (Rücker et al., 2010b)
- **Prediction interval**
  - indicates a range where future studies might be expected (Higgins et al., 2009)
- **Measures of heterogeneity**
  - only describe extent of treatment-study interaction (‘statistical heterogeneity’)
  - do not explain heterogeneity
  - do not describe other aspects of between-study heterogeneity

Adjusting for heterogeneity in meta-analysis: Metaregression

Subgroup analysis and metaregression may explain heterogeneity

Caveats:

- Covariates/subgroups should be pre-defined
- Risk of spurious findings (Higgins and Thompson, 2004)
- For aggregate data meta-analyses, covariates should be defined on study level factors due to the potential for ecological bias (Berlin et al., 2002)
  - Avoid: age mean, proportion of females
- For IPD (individual patient data), also patient-level covariates may be considered (Riley et al., 2010)
- Often no explanation can be found despite all efforts!

Why not pool nevertheless?

One may pool data despite considerable and unexplained heterogeneity if

- all studies are on the same side of the 0
- heterogeneity is not clinically relevant (look at τ)
- I^2 is large simply because studies are large

Next slides: References

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., and Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine*, 21:371–387.

Biggerstaff, B. J. and Jackson, D. (2008). The exact distribution of cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine*, 27(29):6093–6110.

Borm, G. F., Lemmers, O., Fransen, J., and Donders, R. (2009). The evidence provided by a single trial is less reliable than its statistical analysis suggests. *Journal of Clinical Epidemiology*, 62(7):711–715.

Bowden, J., Tierney, J. F., Copas, A. J., and Burdett, S. (2011). Quantifying, displaying and accounting for heterogeneity in the meta-analysis of rcts using standard and generalised Q statistics. *BMC Medical Research Methodology*, 11:41.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10:101–129.

DerSimonian, R. and Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, 28:105–114.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7:177–188.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2:121–145.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, 93:388–395.

Higgins, J. P. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, 37(5):1158–1160.

Higgins, J. P., Thompson, S. G., and Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society*, 172:137–159.

Higgins, J. P. T. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21:1539–1558.

Higgins, J. P. T. and Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine*, 23:1663–1682.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal*, 327:557–560.

Knapp, G., Biggerstaff, B. J., and Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal*, 48:271–285.

Patsopoulos, N. A., Evangelou, E., and Ioannidis, J. P. (2008). Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. *International Journal of Epidemiology*, 37(5):1148–1157.

Paule, R. and Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87(5):377–385.

Riley, R. D., Lambert, P., and Abo-Zaid, G. (2010). Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*, 340.

Rücker, G., Schwarzer, G., Carpenter, J., Binder, H., and Schumacher, M. (2010a). Treatment effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics*, 12(1):122–142. Doi:10.1136/jme.2008.024521.

Rücker, G., Schwarzer, G., Carpenter, J., and Schumacher, M. (2010b). 'Natural weighting' – is it natural? Letter to the editor. *Statistics in Medicine*, 29:2963–2966. DOI: 10.1002/sim.3957.

Rücker, G., Schwarzer, G., Carpenter, J. R., and Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, 8:79. doi:10.1186/1471-2288-8-79.

Rücker, G., Schwarzer, G., Carpenter, J. R., and Schumacher, M. (2009). Are large trials less reliable than small trials? Letter to the editor. *Journal of Clinical Epidemiology*, 62:886–889.

Sidik, K. and Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *JRSS Series C (Applied Statistics)*, 54(2):367–384.

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26:37–52.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48.

Wetterslev, J., Thorlund, K., Brok, J., and Gluud, C. (2009). Estimating required information size by quantifying diversity in random-effects model meta-analyses. *BMC Medical Research Methodology*, 9(1):86.

Appendix: I^2 (solid line) and P-values (dashed line) against n (Rücker et al., 2009)