Statistics
in Medicine

Commentary

(wileyonlinelibrary.com) DOI: 10.1002/sim.6473 Published online in Wiley Online Library

No solution yet for combining two independent studies in the presence of heterogeneity

Andrea Gonnermann, Theodor Framke, Anika Großhennig and Armin Koch*†

Meta-analysis plays an important role in the analysis and interpretation of clinical trials in medicine and of trials in the social sciences but is of importance in other fields (e.g., particle physics [1]) as well. In 2001, Hartung and Knapp [2,3] introduced a new approach to test for a nonzero treatment effect in a meta-analysis of k studies. Hartung and Knapp [2, 3] suggest to use the random effects estimate according to DerSimonian and Laird [4] and propose a variance estimator q so that the test statistics for the treatment effect is t distributed with $k - 1$ degrees of freedom. In their paper on dichotomous endpoints, results of a simulation study with 6 and 12 studies illustrate for risk differences, log relative risks and log odds ratios, the excellent properties regarding control of the type I error, and the achieved power [2]. They investigate different sample sizes in each study, and different amounts of heterogeneity between studies and compare their new approach (Hartung and Knapp approach (HK)) with the fixed effects approach (FE) and the classical random effects approach by DerSimonian and Laird (DL). It can be clearly seen that, with increasing heterogeneity, the FE as well as the DL does not control the type I error rate, while the HK keeps the type I error rate in nearly every situation and in every scale.

Advantages and disadvantages of the two standard approaches and respective test statistics have been extensively discussed (e.g., [5–7]). While it is well known that the FE is too liberal in the presence of heterogeneity, the DL is often thought to be rather conservative because heterogeneity is incorporated into the standard error of the estimate for the treatment effect and this should lead to larger confidence intervals and smaller test statistics for the treatment effect ([8] chapter 9.4.4.3). This was disproved among others by Ziegler and Victor [7], who observed in situations with increasing heterogeneity severe inflation of the type I error for the DerSimonian and Laird test statistic. Notably, the asymptotic properties of this approach will be valid, if both the number of studies and the number of patients per study are large enough ([8] chapter 9.54, [9, 10]). Although power issues of meta-analysis tests have received some interest, comparisons between the approaches and the situation with two studies were not the main interest [11, 12]. Borenstein *et al.* ([10], pp. 363/364) recommend the random effects approach in general for meta-analysis and do not recommend meta-analyses of small numbers of studies.

However, meta-analyses of few and of even only two trials are of importance. In drug licensing in many instances, two successful phase III clinical trials have to be submitted as pivotal evidence for drug licensing [13], and summarizing the findings of these studies is required according to the International Conference on Harmonisation guidelines E9 and M4E ([14, 15]). It is stated that 'An overall summary and synthesis of the evidence on safety and efficacy from all the reported clinical trials is required for a marketing application [...]. This may be accompanied, when appropriate, by a statistical combination

Institute for Biostatistics, Hannover Medical School, Hannover, Germany
**Correspondence to: Armin Koch, Institute for Biostatistics, Hannover Medical School, Hannover, Germany.*
†*E-mail: Koch.Armin@mh-hannover.de*

of results' ([14], p. 31). For the summary, 'The use of meta-analytic techniques to combine these estimates is often a useful addition, because it allows a more precise overall estimate of the size of the treatment effects to be generated, and provides a complete and concise summary of the results of the trials' ([14], p. 32). While in standard drug development, this summary will include usually more than two studies; in rare diseases for the same intervention, barely ever more than two studies are available because of the limited number of patients. Likewise, decision making in the context of health technology assessment is based on systematic reviews and meta-analyses. Often in practice, only two studies are considered homogeneous enough from clinical grounds to be included into a meta-analysis and then form the basis for decision making about reimbursement [16].

Despite the fact that meta-analysis is non-experimental observational (secondary) research [17] and *p*-values should be interpreted with caution, meta-analyses of randomized clinical trials are termed highest-level information in evidence-based medicine and are the recommended basis for decision making [18]. As statistical significance plays an important role in the assessment of the meta-analysis, it is mandatory to understand the statistical properties of the relevant methodology also in a situation, where only two clinical trials are included into a meta-analysis. We found Cochrane reviews including meta-analyses with two studies only, which are considered for evidence-based decision making even in the presence of a large amount of heterogeneity ($I^2 \approx 75\%$) [19–21]

Table I. Overview of the empirical type I error and power.

		Empirical type I error for $p_C = p_T = 0.2$			Empirical power for $p_T = 0.2$ and $p_T = 0.3$		
k	Het (mean I^2)	FE	DL	HK	FE	DL	HK
2	0.15	0.0466	0.0382	0.0481	0.7171	0.6074	0.1487
3	0.15	0.0459	0.0352	0.0477	0.7142	0.6169	0.2976
4	0.14	0.0417	0.0311	0.0473	0.6965	0.6146	0.3999
5	0.13	0.0391	0.0308	0.0473	0.7008	0.6267	0.4720
6	0.12	0.0373	0.0306	0.0447	0.6808	0.6147	0.5015
2	0.25	0.1313	0.0895	0.0469	0.6501	0.4980	0.1137
3	0.25	0.1016	0.0684	0.0525	0.6537	0.5030	0.2115
4	0.25	0.0861	0.0613	0.0467	0.6552	0.5113	0.2762
5	0.25	0.0900	0.0614	0.0467	0.6463	0.5030	0.3148
6	0.25	0.0835	0.0574	0.0423	0.6302	0.4948	0.3395
2	0.50	0.4142	0.2184	0.0489	0.5998	0.3447	0.0706
3	0.50	0.2884	0.1367	0.0493	0.5892	0.3362	0.1131
4	0.50	0.2391	0.1104	0.0467	0.5814	0.3377	0.1476
5	0.50	0.2231	0.0956	0.0443	0.5535	0.3089	0.1611
6	0.50	0.2077	0.0864	0.0421	0.5541	0.3171	0.1767
2	0.75	0.7306	0.2866	0.0639	0.7455	0.3050	0.0567
3	0.75	0.5384	0.1786	0.0509	0.6097	0.2307	0.0695
4	0.75	0.4664	0.1385	0.0501	0.5673	0.2082	0.0747
5	0.75	0.4303	0.1223	0.0466	0.5473	0.1982	0.0853
6	0.75	0.4023	0.1114	0.0468	0.5263	0.1936	0.0900

k, number of studies; Het, heterogeneity; FE, fixed effects approach; DL, DerSimonian and Laird approach; HK, Hartung and Knapp approach; p_c, event rate in control group; p_T, event rate in treatment group.

Note: In a random effects model for log odds ratios, normally distributed logit (p_T) and logit (p_C) were simulated. These values have been back transformed to p_T and p_C to generate binomially distributed number of successes for a given sample size per treatment arm. Median response rates are reported because of skewed distribution after back transformation of the logits generated in the first step of the simulation. The total sample size of the meta-analyses is 480 patients, with balanced treatment arms and $480/k$ patients per study to investigate the impact of the number of studies in the meta-analysis. Mean I^2 from the simulations is reported to describe the degree of heterogeneity.

We repeated the simulation study for dichotomous endpoints of Hartung and Knapp [2] with programs written in R 3.1.0 [22] to compare the statistical properties of the FE, the DL, and the HK for testing the overall treatment effect θ (H_0: $\theta = 0$) in a situation with two to six clinical trials. We considered scenarios under the null and alternative hypothesis for the treatment effect with and without underlying heterogeneity. We present the findings for the odds ratio with $p_C = 0.2$ and did vary probability of success in the treatment group p_T to investigate the type I error and the power characteristics. The total sample size per meta-analysis was kept constant in the different scenarios ($n = 480$) and n/k number of patients per study to clearly demonstrate the effect of the number of included studies on power and type I error of the various approaches. Likewise, we attempted to avoid problems with zero cell counts or extremely low event rates that may impact on type I error and power as well. I^2 was used to describe heterogeneity because thresholds have been published (low: $I^2 = 25\%$, moderate: $I^2 = 50\%$, and high: $I^2 = 75\%$) [23] for the quantification of the degree of heterogeneity with this measure. We termed $I^2 \leqslant 15\%$ negligible, and this refers to simulations assuming no heterogeneity (i.e., the fixed effects model).

Table I summarizes the results of our simulation study. The well-known anticonservative behavior of the FE and the DL in the presence of even low heterogeneity is visible for small numbers of studies in the meta-analysis. Particularly for the FE, the increase in the type I error is pronounced. With more than four studies even in situations with substantial heterogeneity, the HK perfectly controls the type I error. There is almost no impact on the power of the test in situations with no or low heterogeneity, and overall, it seems as if the only price to be paid for an increased heterogeneity is a reduced power of the test.

This is in strong contrast to the situation with only two studies. Again, the HK perfectly controls the prespecified type I error. However, even in a homogeneous situation, the power of the meta-analysis test was lower than 15% in situations where the power of the FE and the DL approximates 70% and 60%, respectively. In the presence of even low heterogeneity with the HK, there is not much chance to arrive

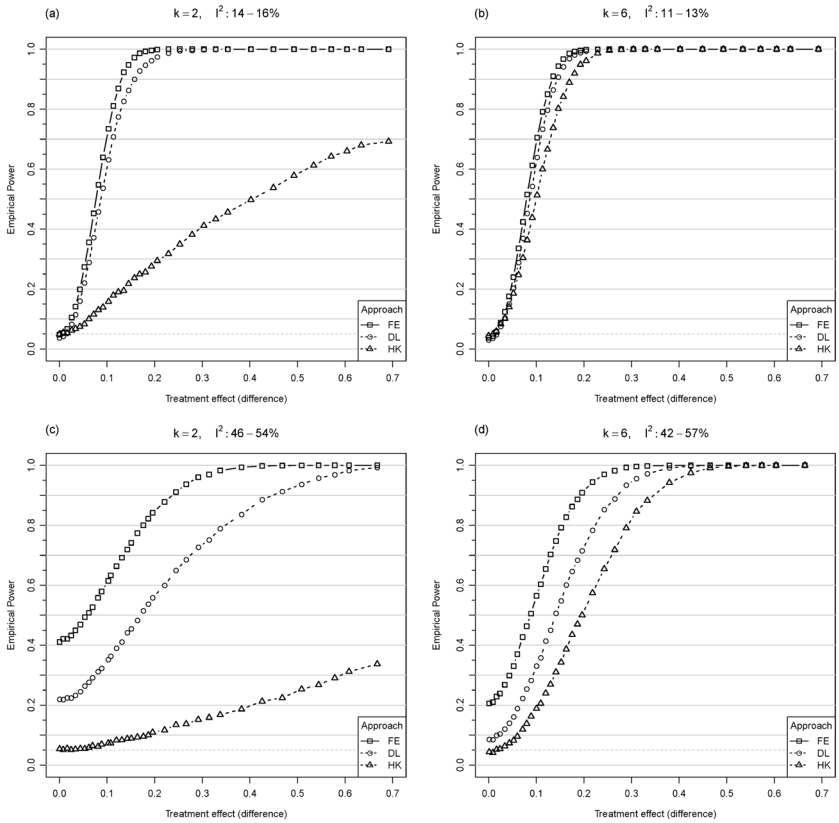

Figure 1. (a–d): Influence of heterogeneity in meta-analysis with two and six studies on empirical power. FE, fixed effects approach; DL, DerSimonian and Laird approach; HK, Hartung and Knapp approach. In the left column, simulation results with two studies are presented, whereas in the right column, situations with six studies are investigated. No heterogeneity is assumed in the top row, and in the bottom row, the impact of moderate heterogeneity is shown.

at a positive conclusion even with substantial treatment effects. Figure 1 summarizes the main finding of our simulation study with $k = 2$ and 6 studies impressively.

In the homogeneous situation with two studies, the DL and even better the FE can be used to efficiently base conclusions on a meta-analysis. In contrast, already with mild to moderate heterogeneity, both standard tests severely violate the prespecified type I error, and there is a high risk of false positive conclusion with the classical approaches.

This has major implications for decision making in drug licensing as well. We have noted previously that a meta-analysis can be confirmatory if a drug development program was designed to include a preplanned meta-analysis of the two pivotal trials [24]. As an example, thrombosis prophylaxis was discussed in the paper by Koch and Röhmel [24], where venous thromboembolism is accepted as primary endpoint in the pivotal trials. In case when both pivotal trials are successful, they can be combined to demonstrate a positive impact on, for example, mortality. This can be preplanned as a hierarchical testing procedure: first, both pivotal trials will be assessed individually before confirmatory conclusions will be based on the meta-analysis. As explained, neither the FE, nor the DL, nor the HK can be the methodology to be recommended for a priori planning in this sensitive area unless any indication for heterogeneity is taken as a trigger not to combine studies in a meta-analysis at all. It is our belief that not enough emphasis has been given to this finding in the original paper and the important role of heterogeneity is not acknowledged enough in the discussion of findings from meta-analyses, in general.

Acknowledgements

This work has been funded partly by the FP7-HEALTH-2013-INNOVATION-1 project Advances in Small Trials Design for Regulatory Innovation and Excellence (ASTERIX) Grant Agreement No. 603160. We thank the editors and reviewers for useful comments to tailor the main messages.

References

1. Jackson D, Baker R. Meta-analysis inside and outside particle physics: convergence using the path of least resistance? *Research Synthesis Methods* 2013; **4**:125–126.
2. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* 2001; **20**(3):875–3889.
3. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine* 2001; **20**:1771–1782.
4. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188, available at 10.1016/0197-2456(86)90046-2.
5. Villar J, Mackey M. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. *Statistics in Medicine* 2001; **20**:3635–3647.
6. Thompson SG, Pocock S. Can meta-analyses be trusted? *Lancet* 1991; **338**:1127–1130.
7. Ziegler S. Victor. Gefahren der Standardmethoden für Meta-Analysen bei Vorliegen von Heterogenität. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* 1999; **30**:131–140.
8. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions,* 5.1.0 ed. The Cochrane Collaboration, 2011. [http://www.cochrane-handbook.org].
9. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Statistics in Medicine* 2001; **20**:825–840.
10. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd: Chichester, UK, 2009.
11. Roloff V, Higgins JPT, Sutton AJ. Planning future studies based on the conditional power of a meta-analysis. *Statistics in Medicine* 2013; **32**(1):11–24.
12. Hedges LV, Pigott TD. The power of statistical tests in meta-analysis. *Psychological Methods* 2001; **6**(3):203–217.
13. U.S. Department of Health and Human Services Food and Drug Administration (1998). Providing Clinical Evidence of Effectiveness for Human Drugs and Biological Products. *Guidance For Industry*, 1998. [Online: http://www.fda.gov/downloads/Drugs/.../Guidances/ucm078749.pdf] [Accessed on 18 December 2014].
14. International Conference on Harmonisation of Techniqual Requirements for Registration of Pharmaceuticals for Human Use. Statistical Principles for Clinical Trials E9. *ICH Harmonised Tripartite Guideline*, 1998. [Online: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf] [Accessed on 18 December 2014].
15. International Conference on Harmonisation of Techniqual Requirements for Registration of Pharmaceuticals for Human Use. The Common Technical Document For The Registration Of Pharmaceuticals For Human Use EFFICACY – M4E(R1). *ICH Harmonised Tripartite Guideline*, 2002. [Online: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/CTD/M4__R1__Efficacy/M4E__R1_.pdf] [Accessed on 18 December 2014].
16. Janatzek S. Untersuchung von und Umgang mit Heterogenität in Nutzenbewertungen - ein Problemaufriss. *IQWiG*, 2001. [Online: https://www.iqwig.de/download/IQWiG_im_Dialog_2011_Sandra_Janatzek.pdf] [Accessed on 18 December 2014].

17. Victor N. 'The challenge of meta-analysis': discussion. Indications and contra-indications for meta-analysis. *Journal of Clinical Epidemiology* 1995; **48**(1):5–8.
18. OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. Oxford Centre for Evidence-Based Medicine. *Oxford Centre for Evidence-Based Medicine*, 2011. [Online: http://www.cebm.net/index.aspx?o=5653] [Accessed on 18 December 2014].
19. Amato L, Davoli M, Minozzi S, Ferroni E, Ali R, Ferri M. Methadone at tapered doses for the management of opioid withdrawal (Review). *Cochrane Database of Systematic Reviews* 2005; **3**:Art. No.: CD003409. DOI: 10.1002/14651858
20. Poustie VJ, Wildgoose J. Dietary interventions for phenylketonuria (Review). *The Cochrane Database of Systematic Reviews* 2010; **1**:Art. No.: CD001304. DOI: 10.1002/14651858.
21. Zeng Y, Duan X, Xu J, Ni X. TPO receptor agonist for chronic idiopathic thrombocytopenic purpura. *Cochrane Database of Systematic Reviews* 2011; **7**:Art. No.: CD008235. DOI: 10.1002/14651858
22. R Development Core Team. *R: A Language and Environment for Statistical Computing, Reference Index Version 3.1.0*, R Foundation for Statistical Computing: Vienna, Austria, 2014. [http://www.R-project.org].
23. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *British Medical Journal* 2003; **327**:557–560.
24. Koch A, Röhmel J. Why are some meta-analyses more credible than others? *Drug Information Journal* 2001; **35**: 1019–1030.

