No solution yet – few studies and heterogeneity

Theodor Framke, Martina Kottas, Andrea Smith, Kristina Weber, Armin Koch

19.09.2017

Hannover Medical School

Introduction

Meta-analysis is a powerful tool to summarize individual studies

Various different methods

- Frequentist approaches and Bayesian approaches
- Fixed and Random effects approaches

Often it is thought that meta-analysis is

- used to summarize ***many*** studies
- only ***explanatory secondary*** research

Meta-analyses are not only explanatory!

- Health Technology Assessment (HTA)
- Drug approval (e.g. in rare diseases)

Introduction

In general, problems arise with

- Heterogeneity between studies
- Too few studies (rare diseases!)
  - → Asymptotic properties of methods
  - → Probably increased heterogeneity

http://media.propertycasualty360.com/propertycasualty360/article/2011/11/07/applesorganges1172011-crop-600x338.jpg

Overview

Andrea Smith: Meta-Analysis with 2 studies

Theodor Framke: Heterogeneity in Meta-Analysis with few studies

Martina Kottas: Triggers of heterogeneity – alternative detection rules

Kristina Weber: Extrapolation – adult to pediatric population

Comparison of methods for MA

Simulation study to analyze frequentist meta-analysis methods for k=2,6

- Fixed effect (inverse variance, (FE))
- Random effects (inverse variance, DerSimonian and Laird, (DL))
- Hartung and Knapp (DL-estimator, (HK))

Extension

- Mantel-Haenszel method (MH), default in RevMan
- Should perform better then inverse variance method when data are sparse
- Some attention has been drawn to alternatives to estimate between-study variance for RE meta-analysis

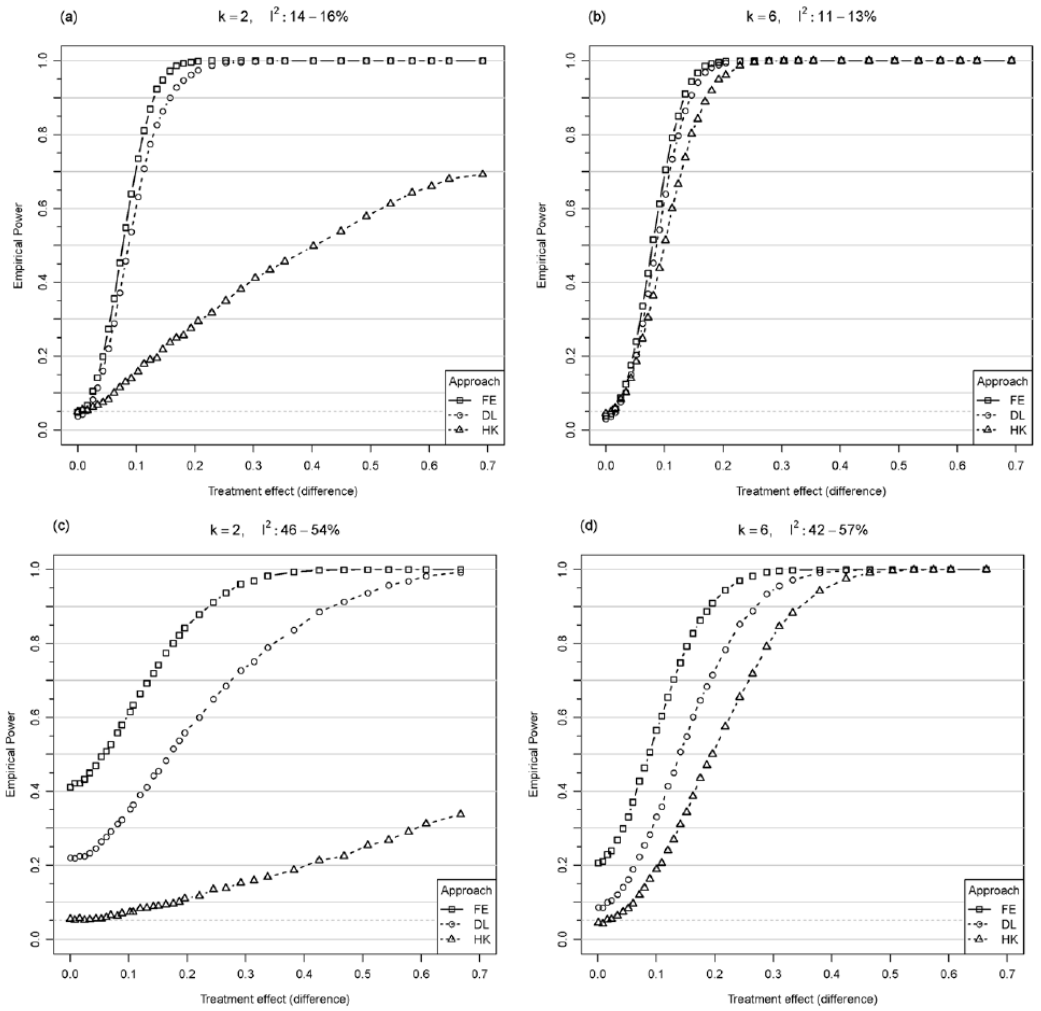
  → E.g. Veroniki et al. (2016) lists 16 different estimators for the between-study variance

  → Paule-Mandel (PM) seems promising (e.g. Novianti et al. 2014; Langan et al. 2015; Langan et al. 2016)

Results

Alternative hypothesis for treatment effect and no heterogeneity

Alternative hypothesis for treatment effect and heterogeneity

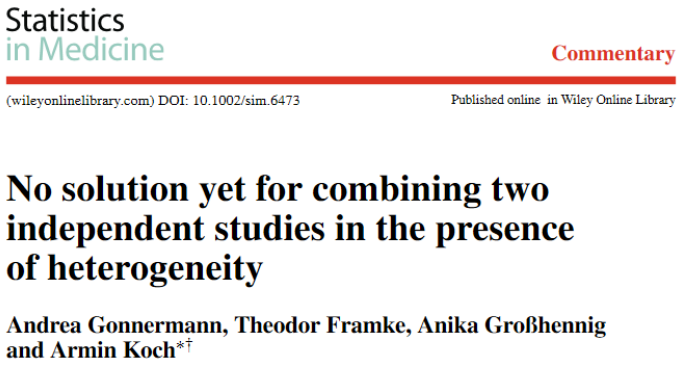


Conclusion

Result: *There is still no solution…*

- Type I error increases when heterogeneous studies are summarized
- HK-approach is relatively safe, but lacks power for k<5
- Comparable problems arise in stratified studies as methods are analogues → Type I error in one study with heterogeneous subgroups increases as well → careful assessment of subgroups required
- Summarizing only homogeneous studies with FE-approach?
- Conclusions do not change w.r.t. PM, MH

→ Detecting heterogeneity is not easy!


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*†

Triggers/warning signals

If treatment effect is significant (based on superiority trial):

I. Q-rule: p-value of Cochran‘s Q ≤ 0.15

II. Regulator‘s rule: $OR_i < \exp\left(\frac{\log(OR)}{2}\right)$ or $OR_i > \exp\left(2 \cdot \log(OR)\right)$

III. Epidemiologist’s rule: $OR_i < \exp\left(\frac{\log(OR)}{4}\right)$ or $OR_i > \exp\left(4 \cdot \log(OR)\right)$

IV. CI-rule: the point estimate of a subgroup is not included in the confidence interval of the overall treatment effect

Simulation Results

- ***Q-rule…***

+ Type I error as specified
+ Performance well in balanced SGs
- in unbalanced SGs power loss

- ***Regulator's rule…***

+ In balanced SGs small Type I error and power similar to Q-rule
- In unbalanced SGs high Type I error

- ***Epidemiologist's rule…***

+ Small type I error
- No more than 55% power

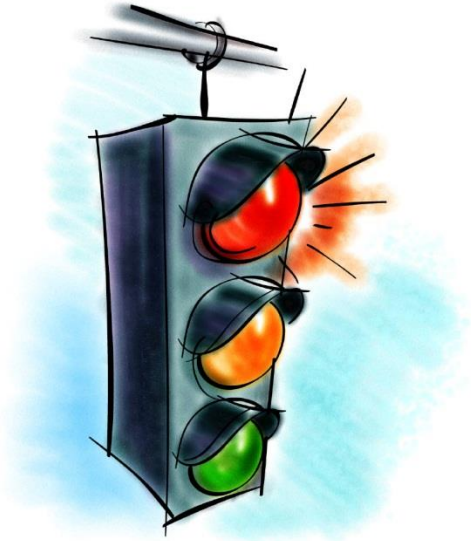
- ***CI-rule…***

+ in balanced SGs small type I error
- In unbalanced SGs high type I error

Situation	SG1: SG2	Q-rule (I)	R-rule (II)	Epi-rule (III)	CI-rule (IV)
H0	50:50	0.1534	0.0686	0.0164	0.0530
H1 True II		0.5483	0.5150	0.2702	0.3693
H1 True III		0.7582	0.7419	0.4900	0.6400
H0	70:30	0.1484	0.1832	0.0608	0.2079
H1 True II		0.3449	0.5201	0.3356	0.4367
H1 True III		0.5177	0.6939	0.5143	0.6197
H0	90:10	0.1440	0.4585	0.1808	0.5308
H1 True II		0.2031	0.5952	0.4011	0.6102
H1 True III		0.3026	0.6983	0.5526	0.7058

Conclusion

- There is no perfect rule until now
- Signals should be properly understood
  (“yellow traffic light“)
- Cannot be perfect, some false positives need to be accepted



Leitthema

Bundesgesundheitsbl 2015 · 58:274–282
DOI 10.1007/s00103-014-2105-2
Online publiziert: 8. Januar 2015

A. Gonnermann · M. Kottas · Armin Koch
Institut für Biometrie, Medizinische Hochschule Hannover, Hannover, Deutschland

Biometrische Entscheidungsunterstützung in Zulassung und Nutzenbewertung am Beispiel der Implikationen von heterogenen Ergebnissen in Untergruppen der Studienpopulation

Auch um dem Vorwurf zu begegnen, dass in einer randomisierten klinischen Studie neue Arzneimittel in einem „Windkanal- die möglichst größte Population zu identifizieren, für die die vorgenannten Kriterien erfüllt sind. Es ist jedoch quasi der diendaten begründet wird, auffällig. Viele Unterschiede in den Entscheidungsstrategien bestehen dabei nur scheinbar, weil

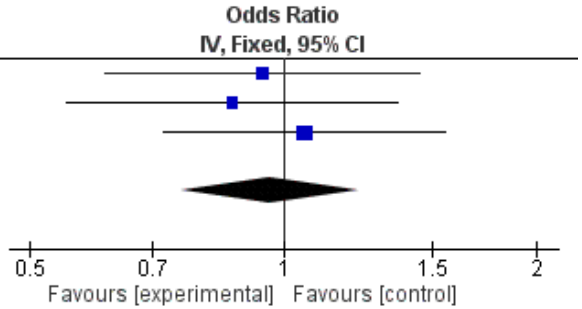
Pediatric extrapolation

Adult studies in de novo kidney transplants with EVR NIM(log(OR)): 0.54)

Study or Subgroup	Experimental Events	Total	Control Events	Total	Weight	Odds Ratio IV, Fixed, 95% CI	Year
Vitko 2004	58	194	61	196	31.7%	0.94 [0.61, 1.45]	2004
Lorber 2005	48	193	54	196	28.8%	0.87 [0.55, 1.37]	2005
Tedesco 2010	70	277	67	277	39.5%	1.06 [0.72, 1.56]	2010
Total (95% CI)		**664**		**669**	**100.0%**	**0.97 [0.76, 1.23]**	
Total events	176		182				

Heterogeneity: $Chi^2 = 0.44$, $df = 2$ ($P = 0.80$); $I^2 = 0\%$

Test for overall effect: $Z = 0.28$ ($P = 0.78$)



Aim: extrapolation to the paediatric population with one study

Investigation of two different scenarios:

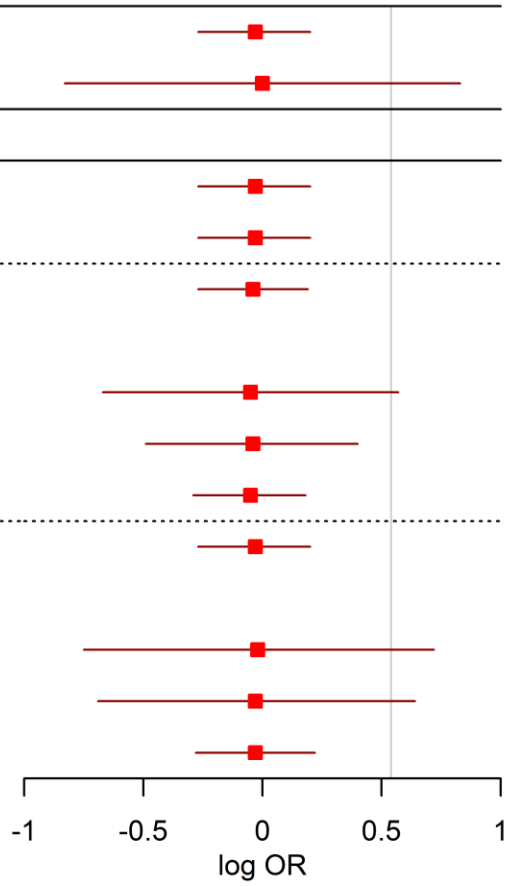
study	EVR events/treated	MPA events/treated	log (OR) 95% CI P-value
Scenario 1	16/53 30.2%	16/53 30.2%	0.00 (-0.83; 0.83) 1.00
Scenario 2	22/53 41.5%	16/53 30.2%	0.50 (-0.31; 1.30) 0.33

Approaches to a summary evaluation of individual sources of information

- **Frequentist Meta-Analysis**
  - Joint analysis of existing and new trial (eventually looking into heterogeneity) in a fixed (FEM) or a random (REM) effects model
- **Bayesian Meta-Analysis**
  - Joint analysis of existing and new trial in a FEM or a REM (Smith et al., 1995)
- **Bayesian meta-analytic predictive approach**
  - Analysis of new trial „in light of“ the already existing trial in a FEM or a REM (Viele et al., 2014 and Spiegelhalter et al., 2004)

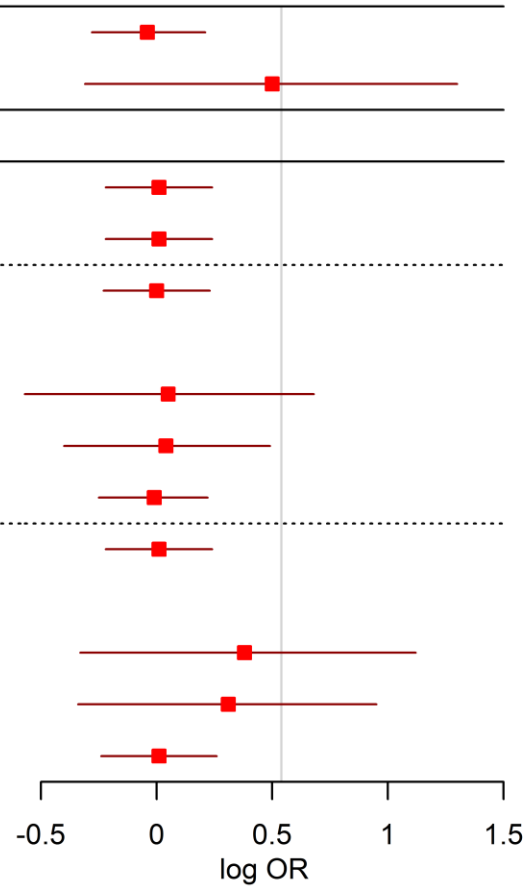
Results with Scenario 1 (assumed homogeneity)

Study		log OR	
adult MA		-0.03	
Scenario 1		0	
Analysis method	**Prior**	**log OR**	**Heterogeneity**
F FE MA		-0.03	q=0.44, $\hat{\tau}^2$= 0.00
F RE MA		-0.03	
B FE MA		-0.04	
B RE MA			
prior: E(τ^2) = 0.33		-0.05	$\hat{\tau}^2$= 0.31
prior: E(τ^2) = 0.14		-0.04	$\hat{\tau}^2$= 0.14
prior: E(τ^2) = 0.001		-0.05	$\hat{\tau}^2$= 0.001
B FE MAP	adult	-0.03	
B RE MAP			
prior: E(τ^2) = 0.33	adult	-0.02	$\hat{\tau}^2$= 0.42
prior: E(τ^2) = 0.14	adult	-0.03	$\hat{\tau}^2$= 0.16
prior: E(τ^2) = 0.001	adult	-0.03	$\hat{\tau}^2$= 0.001



Results with Scenario 2 (log OR = 0.50, at the margin)

Study		log OR	
adult MA		-0.04	
Scenario 2		0.5	
Analysis method	**Prior**	**log OR**	**Heterogeneity**
F FE MA		0.01	q=0.44, $\hat{\tau}^2$= 0.00
F RE MA		0.01	
B FE MA		0	
B RE MA			
prior: E(τ^2) = 0.33		0.05	$\hat{\tau}^2$= 0.32
prior: E(τ^2) = 0.14		0.04	$\hat{\tau}^2$= 0.15
prior: E(τ^2) = 0.001		-0.01	$\hat{\tau}^2$= 0.001
B FE MAP	adult	0.01	
B RE MAP			
prior: E(τ^2) = 0.33	adult	0.38	$\hat{\tau}^2$= 0.43
prior: E(τ^2) = 0.14	adult	0.31	$\hat{\tau}^2$= 0.16
prior: E(τ^2) = 0.001	adult	0.01	$\hat{\tau}^2$= 0.001



Assessment of the exemplary analyses

Many approaches …

- If meta-analysis is used as a tool to arrive at an overall conclusion, no difference between a frequentist approach or a Bayesian approach can be detected: actually summary estimates will always be dominated by adult data.
- Using the predictive approach might allow that the pediatric data stand against the adult data (in case a prior is chosen that will allow for heterogeneity), however then even in case of homogeneity nothing can be concluded with the current sample-size.
- If heterogeneity is restricted, the impact of the adult data is increased (similar to frequentist MA).
- Precise pre-specification of the assumptions is required / recommended.
- Such considerations could be used to determine sample-size for a pediatric trial.

Discussion

How to **summarize**?

→ Homogeneity: Fixed effect approach preferred

→ Hartung-Knapp REM good T1E control. Lacks power for small k in homogeneous situation. Heterogeneity: interpretation still problematic

→ Extension to MH, PM: also no solution yet

→ No optimal rule available for detection of heterogeneity (false positives vs. overlooking heterogeneity?)

Extrapolation: What can be done?

→ Avoiding “overweight” in the MA-approach (e.g with content-wise selection of adult patients, only use data from young adults to weigh in for the assessment of adolescent pediatric patients)

→ Be precise about the weight of the prior information

→ Change of emphasis from “Does it work?” towards “Is there evidence for differential effects?”

References

Gonnermann A, Kottas M, Koch A. (2015): Biometrische Entscheidungsunterstützung in Zulassung und Nutzenbewertung am Beispiel der Implikationen von heterogenen Ergebnissen in Untergruppen der Studienpopulation. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz, 58(3): 274-82, doi: 10.1007/s00103-014-2105-2.

Committee for Medicinal Products for Human Use (2014), Guideline on the investigation of subgroups in confirmatory clinical trials - DRAFT http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf [accessed 26.06.2014].

Wiksten, A., Rücker, G. & Schwarzer, G., 2016. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Statistics in Medicine*, 35(15), pp.2503–2515.

Li, Z., Chuang-Stein, C., Hoseyni, C. (2007), The Probability of Observing Negative Subgroup Results When the Treatment Effect Is Positive and Homogeneous Across All Subgroups, Drug Information Journal 2007 41:47.

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

Friede, T. et al. (2017). Meta-analysis of few small studies in orphan diseases. Research Synthesis Methods, 8(1), pp.79-91.

Veroniki, A.A. et al., 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods, 7(1), pp.55–79.

Langan, D., Higgins, J.P.T. & Simmonds, M., 2015. An empirical comparison of heterogeneity variance estimators in 12894 meta-analyses. Research Synthesis Methods, 6(2), pp.195–205.

Langan, D., Higgins, J.P.T. & Simmonds, M., 2016. Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. Research Synthesis Methods.

Smith, T.C., Spiegelhalter, D.J., Thomas, A. (1995). Bayesian Approaches to Random-Effects Meta-Analysis: A Comparative Study. Statistics in Medicine, 14: 2685-2699

David J. Spiegelhalter, Keith R. Abrams, and Jonathan P. Myles. Bayesian Approaches to Clinical Trials and Helath-Care Evaluation. Wiley, 2004.

Kert Viele, Scott Berry, Beat Neuenschwander, et al. (2014), Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical Statistics., 13(1), pp . 41-54.

The holy grail of heterogeneity assessment has not been found yet – but we will continue searching.

Thank you for your attention!

Results

Null hypothesis for treatment effect, **no heterogeneity**

k	N	Sample Size $Study_i$: $study_k$	MH	FE	DL	PM	HK	Q	Mean I^2
2	480	240:240	0,0501	0,0480	0,0355	0,0355	0,0496	0,1558	15,50
2	60	30:30	0,0355	0,0241	0,0229	0,0229	0,0523	0,1373	14,41
2	480	120:360	0,0470	0,0449	0,0349	0,0349	0,0495	0,1518	15,05
6	480	80:80:…:80	0,0489	0,0372	0,0308	0,0308	0,0422	0,1313	12,09
6	180	30:30:…:30	0,0428	0,0199	0,0182	0,0181	0,0375	0,0651	8,26
6	480	60:…:60:180	0,0495	0,0358	0,0292	0,0285	0,0413	0,1167	11,33

Effect size: Odds ratio. Baseline risk: 0.2

Results

Null hypothesis for treatment effect, but **heterogeneity**

k	N	Sample Size $Study_i$: $study_k$	MH	FE	DL	PM	HK	Q	Mean I^2
2	480	240:240	0,4320	0,4026	0,2096	0,2096	0,0485	0,5618	50,00
2	60	30:30	0,6007	0,4518	0,2199	0,2199	0,0475	0,5669	50,27
2	480	120:360	0,5327	0,5206	0,2408	0,2408	0,0573	0,5637	50,27
6	480	80:80:…:80	0,2557	0,2046	0,0892	0,0860	0,0411	0,7310	50,58
6	180	30:30:…:30	0,3051	0,1993	0,0941	0,0901	0,0488	0,7355	50,72
6	480	60:…:60:180	0,3295	0,2854	0,0983	0,0932	0,0517	0,7185	50,08

Effect size: Odds ratio. Baseline risk: 0.2