Heterogeneity and sources of bias

Olaf Dekkers/Sønderborg/2016

Heterogeneity: to pool or not to pool?

Heterogeneity: What do we mean?

- Heterogeneity
- Diversity
- Statistical heterogeneity
- Variation between studies
- Bias
- Study quality

What do we mean?

- Sources of between study variation
  - Design elements
  - Patient characteristics
  - Treatments
  - Effect measures/outcomes
  - Effect estimates

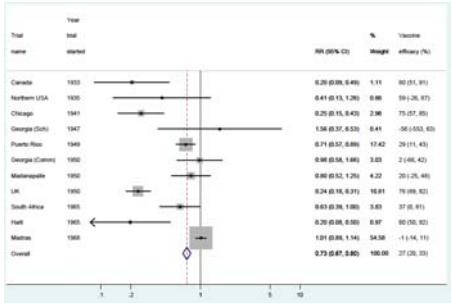
What do we mean?

- Sources of between study variation
  - Design elements
  - Patient characteristics
  - Treatments
  - Effect measures/outcomes
  - Effect estimates → Statistical judgement
    Q-statistics
    I-squared

Q-statistics

- Test for overall homogeneity $Q=\Sigma w_i (\theta_i-\theta)^2$
  - Chi-square with df = n-1
- The power of the test is low for few studies
  - Fails to detect heterogeneity often
- The power is too high for meta-analysis with much studies

Q-statistics

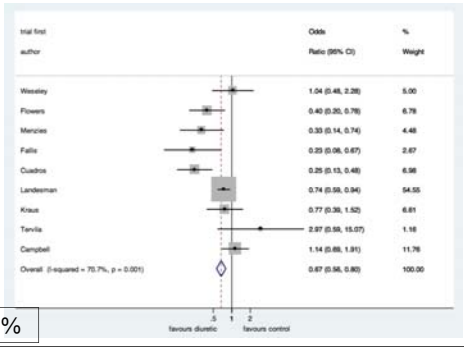


P-value Q-test: < 0.001

I^2

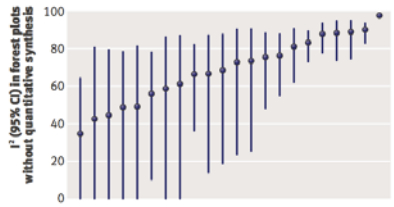
- Tries to separate
  - Variability due to differences in true effects (heterogeneity)
  - Random variation
- I^2 is formally not a test
- It describes the proportion variability between trials not due to chance
- Main disdvanatge: depends on sample size of trials

I^2



I^2 71%

I^2 of studies with high heterogeneity



I^2 point estimates and 95% CI for forest plots with at least 4 studies and no quantitative synthesis because of perceived high statistical heterogeneity

What do we mean?

- Sources of between study variation
  - Design elements
  - Patient characteristics
  - Treatments
  - Effect measures/outcomes
  - Effect estimates → Statistical judgement *Q-statistics* *I-squared*

To pool or not to pool?

Table 1 | Reasons for not showing summary estimates in forest plots from systematic peviews in Cochrane database 2005 issue 4

Reason	No (%) of systematic reviews (n=135)*
Statistical heterogeneity too high	32 (24)
Different interventions compared	41 (30)
Different metrics or outcomes evaluated	26 (19)
Different metric of same outcome	7
Different outcome	20
Different study designs	21 (16)
Non-randomised studies	3
Other design issues	18
Different study participants, settings	21 (16)
Data with many counts per participant	5 (4)
Data too limited	11 (8)
Clinical heterogeneity (not otherwise specified)	5 (4)
Synthesis considered inappropriate (not specified why)	3 (2)
Non-normality of data	1 (1)
No reason given	10 (7)
Artefact†	3 (2)
Quantitative synthesis given in text	7 (5)

Example I

BMJ RESEARCH

Sexual abstinence only programmes to prevent HIV infection in high income countries: systematic review

Kristen Underhill, research officer,[1] Paul Montgomery, lecturer,[1] Don Operario, lecturer[1]

Underhill, BMJ 2007, 335;248

Example I

As a result of data unavailability, lack of intention to treat analyses, and heterogeneity in programme and trial designs, we determined that a statistical meta-analysis would be inappropriate. Instead we present individual trial results using RevMan and provide a narrative synthesis.

Example I

Trial	Control	No analysed	Odds ratio (95% CI)	Odds ratio (95% CI)
Diagnosis of sexually transmitted infection				
Kirby peer led[w6]	Usual care	1545		2.06 (0.67 to 6.32)
Kirby adult led[w6]	Usual care	2313		2.73 (1.05 to 7.14)
Kirby[w6]	Usual care	3761		0.77 (0.29 to 2.09)
Kirby[w6]	Usual care	372		0.31 (0.03 to 3.03)
Trenholm[w8]	Usual care	277		0.99 (0.28 to 3.46)
Trenholm[w8]	Usual care	277		1.46 (0.48 to 4.49)
Trenholm[w8]	Usual care	163		1.73 (0.35 to 8.64)
Trenholm[w8]	Usual care	323		0.83 (0.28 to 2.42)

Example I

Results The search identified 13 trials enrolling about 15 940 US youths. All outcomes were self reported. Compared with various controls, no programme affected incidence of unprotected vaginal sex, number of partners, condom use, or sexual initiation. One trial observed adverse effects at short term follow-up (sexually transmitted infections, frequency of sex) and long term follow-up (sexually transmitted infections, pregnancy) compared with usual care, but findings were offset by trials with non-significant results. Another trial observed a protective effect on incidence of vaginal sex compared with usual care, but this was limited to short term follow-up and countered by trials with non-significant findings. Heterogeneity prevented meta-analysis.

Conclusion Programmes that exclusively encourage abstinence from sex do not seem to affect the risk of HIV infection in high income countries, as measured by self reported biological and behavioural outcomes.

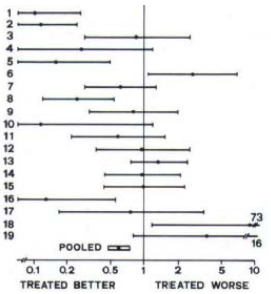
To pool or not to pool

- Clinical heterogeneity
  - Reconsider study eligibility?
  - Is pooling results defendable?
- Don't rely on I^2 for ultimate verdict
- Outcome heterogeneity
  - Ways to deal with
- Statistical heterogeneity
  - There are methods to account for statistical heterogeneity
    - Random effect models/Prediction intervals
    - Restriction/sensitivity analysis
    - Meta-regression

Explaining heterogeneity

"Heterogeneity should be the starting point for further examination" M.Egger

Heterogeneity in meta-analysis



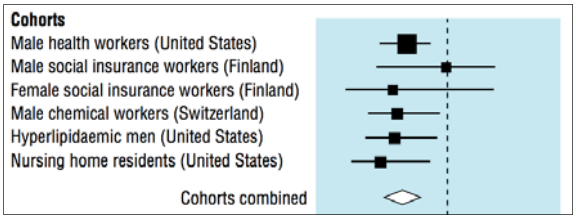
Explanations

- Differences in design elements/risk of bias
  - Adequate concealment
  - Blinding
  - Loss to follow-up
- Differences in clinical characteristics
  - Age
  - Co-medications
  - Differences in outcome

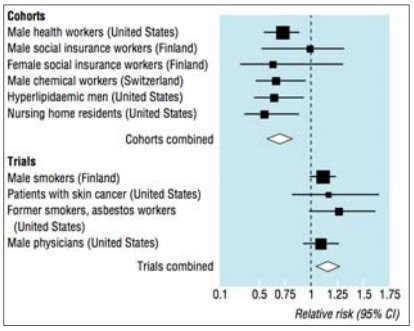
How to deal with heterogeneity

- Heterogeneity can (should!?) be the starting point for further investigation
- Explanation of heterogeneity is an important goal
  - Sensitivity analysis
  - Meta-regression

Assessment of heterogeneity



Assessment of heterogeneity



Heterogeneity

- Clinical characteristics and study characteritics can cause heterogeneity
- Design elements, clinical characteritics (at study level) and risk of bias used to explore heterogeneity
- Absence of heterogeneity does not mean absence of bias

Risk of bias assessment


+

= ??

Risk of bias

I. Bias vs risk of bias
II. Quality vs risk of bias
III. Risk of bias vs reporting
IV. Scales and scores
V. Risk of bias: empirical evidence

i. Bias vs Risk of bias

Risk of bias?

Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial

*Philip D Home, Stuart J Pocock, Henning Beck-Nielsen, Paula S Curtis, Ramon Gomis, Markolf Hanefeld, Nigel P Jones, Michel Komajda, John J V McMurray, for the RECORD Study Team**

Risk of bias?

EuroIntervention

Favourable effect of statin therapy on early survival at the time of percutaneous coronary intervention for ST-elevation myocardial infarction and shock

Bias vs risk of bias

- We do (often) not know whether the results are biased
- But: we can assess the risk of bias

ii. Study quality vs Risk of bias

Study quality?

ORIGINAL ARTICLE

Evaluation of gastric emptying rate in patients with fibromyalgia: a case control study

Serpil Erdogan · Gulcan Gurer · Hamdi Afsun · Yuksel Kucukzeybek

Materials and methods

Fifteen FM (ages 17–68 years) and 15 age-matched healthy women volunteers (ages 20–65 years) as a control group were enrolled in this study. The classification and evalua-

Quality vs risk of bias

- Quality is the best the authors have been able to do
- Low study quality ≠ high risk of bias
  - Good quality but still high risk of bias
    - Unblinded study of surgical intervention
  - Low quality but no risk of bias
    - Lacking sample size calculation

III. Reporting vs risk of bias

Report vs conduct

Methodological approach	RCTs using but not reporting the methodological approach, % (*n*/*N*)[a]	95% confidence intervals, %
Concealment of randomization	96 (52/54)	87–100
Blinding of		
Participants	20 (5/25)	7–41
Health care providers	65 (41/63)	52–77
Data collectors[b]	65 (53/82)	53–75
Outcome assessors[c]	79 (64/81)	69–87
Data analysts	50 (47/94)	40–60

[a] *n*/*N* = the number of authors who informed us they had used the methodological approach despite not reporting it in their RCT publication

Reporting vs risk of bias

- We are actually judging reporting
- Reporting not always good proxy for conduct
- 'Solution': make a category 'not reported'

IV: Use of quality assessment scales

Quality scores

ORIGINAL CONTRIBUTION

The Hazards of Scoring the Quality of Clinical Trials for Meta-analysis

Juni JAMA 1999

JAMA paper

- Different quality scores (n=25) applied to one meta-analysis
- Based on quality score studies were divided into high and low quality
- Summary estimate by quality
  - Standard assumption: better quality results in more valid estimates

Quality scales

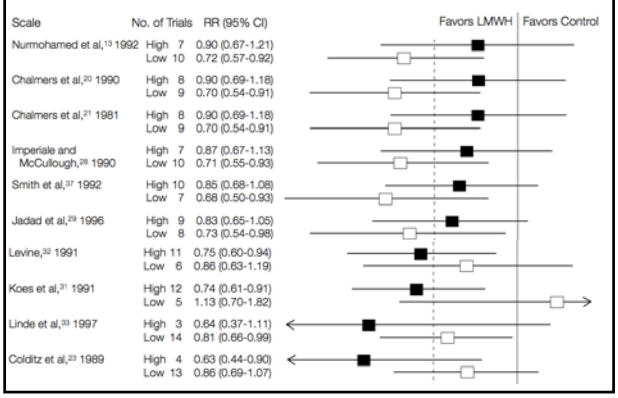
Table 1. Characteristics of 25 Scales for Quality Assessment of Clinical Trials

Scale	No. of Items	Weight Given to Methodological Key Domains, %*		
		Randomization	Blinding	Withdrawals
Andrew,[17] 1984	11	9.1	9.1	9.1
Beckerman et al,[18] 1992	24	4.0	12.0	16.0
Brown,[19] 1991	6	14.3	4.8	0
Chalmers et al,[20] 1990	3	33.3	33.3	33.3
Chalmers et al,[21] 1981	30	13.0	26.0	7.0
Cho and Bero,[22] 1994	24	14.3	8.2	8.2
Colditz et al,[23] 1989	7	28.6	0	14.3
Detsky et al,[24] 1992	14	20.0	6.7	0
Evans and Pollock,[25] 1985	33	3.0	4.0	11.0
Goodman et al,[26] 1994	34	2.9	2.9	5.9
Gøtzsche,[27] 1989	16	6.3	12.5	12.5
Imperiale and McCullough,[28] 1990	5	0	0	0
Jadad et al,[29] 1996	3	40.0	40.0	20.0

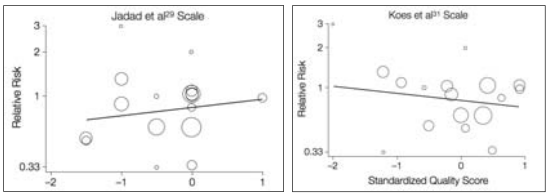
Table 2. Median Scores From 25 Quality Assessment Scales for 17 Trials Comparing Heparins for Thromboprophylaxis in General Surgery, and Thresholds for Definition of High Quality*

Scale	Median Score (Range), %	Threshold for High Quality, %†
Poynard,[30] 1988	38.5 (15.4-76.9)	50.0
Chalmers et al,[21] 1981	39.8 (8.6-76.8)	NA
Spitzer et al,[36] 1990	48.1 (25.9-78.8)	NA
Beckerman et al,[18] 1992	50.0 (25.0-75.0)	52.0
Linde et al,[31] 1997	50.0 (14.3-92.9)	71.4
Chalmers et al,[20] 1990	55.6 (11.1-88.9)	66.7
Cho and Bero,[22] 1994	55.6 (37.8-75.6)	NA
Detsky et al,[24] 1992	56.7 (23.3-89.3)	NA
Colditz et al,[23] 1989	57.1 (14.3-85.7)	NA
Gøtzsche,[27] 1989	57.1 (7.1-71.4)	NA
Smith et al,[37] 1992	57.1 (25.7-85.7)	50.0
Jonas et al, 1993‡	58.3 (33.3-88.9)	76.0
Imperiale and McCullough,[28] 1990	60.0 (20.0-100)	80.0
Jadad et al,[29] 1996	60.0 (0-100)	60.0
Kees et al,[21] 1991	60.0 (20.0-78.6)	50.0
Reisch et al,[38] 1989	62.5 (37.5-87.5)	NA
Onghena and Van Houdenhove,[34] 1992	62.9 (34.3-100)	NA
Evans and Pollock,[25] 1985	63.8 (32.5-88.2)	NA
Levine,[32] 1991	64.4 (26.8-79.5)	60.0
Goodman et al,[26] 1994	67.7 (31.0-83.2)	60.0
Kleijnen et al,[33] 1991	70.0 (30.0-98.0)	55.0
Nurmohamed et al,[12] 1992	75.0 (25.0-100)	87.5
Andrew,[17] 1984	77.3 (45.5-90.9)	72.7
Brown,[19] 1991	81.0 (52.4-95.2)	81.0
[illegible]	[illegible]	[illegible]

Figure 1. Results From Sensitivity Analyses Dividing Trials in High- and Low-Quality Strata, Using 25 Different Quality Assessment Scales



Quality and effect sizes



Quality as a weight factor in pooling?

- Assumtpion: High quality studies provide better estimates
- Use of scales/aggregate scores should be discouraged
  - Choice of scales is arbitrarily
- Preferably: use risk of bias assessment to explore heterogeneity ***per item***:
  - Restriction
  - Sensitivity analysis
  - Meta-regression

V: Risk of bias: empirical evidence

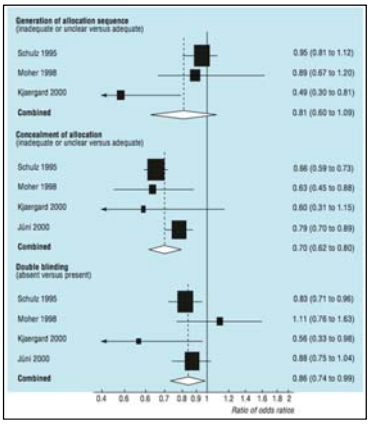
Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial

Philip D Home, Stuart J Pocock, Hemming Beck-Nielsen, Paula S Curtis, Ramon Gomis, Markolf Hanefeld, Nigel P Jones, Michel Komajda, John J V McMurray, for the RECORD Study Team*

Table. Summary of Reviewed Case-Report Forms in RECORD Trial

	No. (%) of Case-Report Forms Reviewed		
	Rosiglitazone (n = 278)	Control (n = 271)	Total (N = 549)
With problems	45 (16.2)	25 (9.2)	70 (12.8)
Favoring rosiglitazone	44 (15.8)	13 (4.8)	57 (10.4)
Favoring control	1 (0.4)	12 (4.4)	13 (2.4)

Abbreviation: RECORD, Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycemia in Diabetes.



Empirical evidence for risk of bias

- Is there evidence bias indeed has an effect on the outcome?
  - Extensive literature for randomized studies
  - Almost no literature for observational studies

Risk of bias

- We can assess risk of bias, not bias (sometimes we can)
- Study quality has no direct translation in terms of risk of bias
- We are actually assessing study reporting
- Use of scales and scores should be discouraged
- Empirical evidence for risk of bias mainly for RCTs

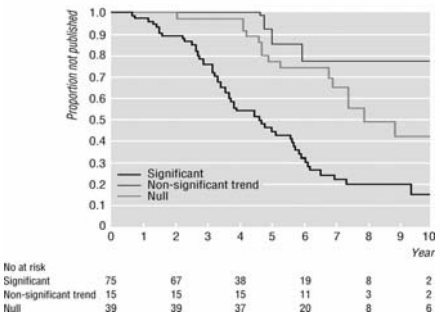
Publication bias

Publication bias

Studies with significant results are more likely to get published than studies without significant results, leading to publication bias

BMJ 1998;316:61-66

Publication bias



Publication bias

Unpublished (unavailable)

Available in principle (Chinese journals, meeting abstracts)

Easily available (Medline, Embase)

P-value higher P-value lower

Publication bias: empirical evidence

The NEW ENGLAND JOURNAL of MEDICINE

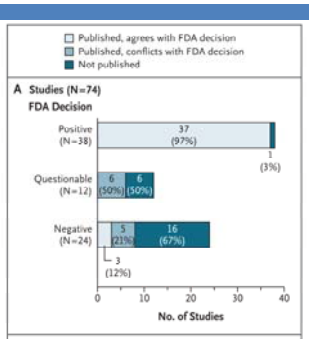
SPECIAL ARTICLE

Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy

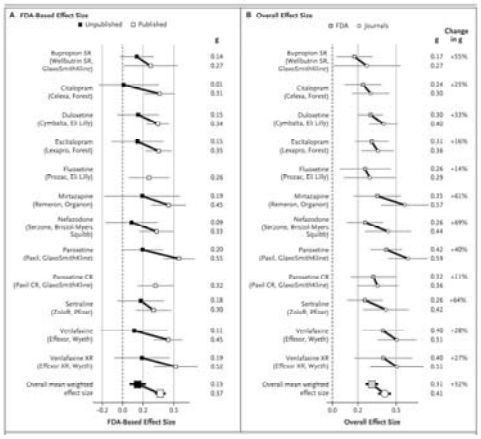
Erick H. Turner, M.D., Annette M. Matthews, M.D., Eftihia Linardatos, B.S., Robert A. Tell, L.C.S.W., and Robert Rosenthal, Ph.D.

N Engl J Med 2008;358:252-260

Publication bias



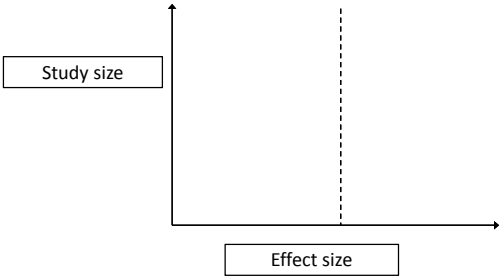
NEJM 2008;358:3



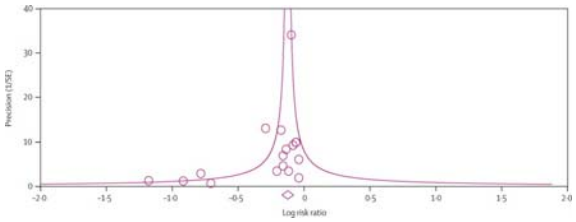
Funnel plot

- Visual way to detect (publication) bias
- The scatter should be symmetrical around overall effect
- Effect measure is plotted against a measure of precision

Funnel plot



Funnel plot (a)symmetry



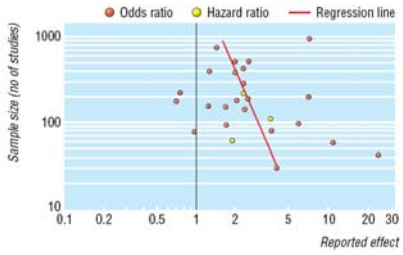
Sources of funnel plot asymmetry

- Publication bias
- True heterogeneity
- Chance

Use of symmetry tests

- Symmetry tests:
  - Power is low
  - Different tests can give different results
- Alternative approaches
  - Restriction to large trials
  - Cumulative meta-analysis

Publication bias in publication bias?



Funnel plot of 26 reports on publication bias, with reported effect as dependent variable

BMJ 2005;331:433-434