Meta-Analysen re-visited

Methodische Aspekte der Evidenzsynthese

Doris Böhm

11. GESENT Kongress, 11. Dezember 2015

Inhalt

- Einführung
- Grundlagen - Effektmaße
- Statistische Modelle für Meta-Analysen / Heterogenität
- Beispiele und Ausblick
- Diskussion

Meta-Analyse – Einführung

Notwendigkeit der Zusammenfassung der zunehmenden wachsenden Zahl an Veröffentlichungen zu klinischen Studien.

G. Glass: „…*literature is growing on dozens of topics in education is growing at an astounding rate…*“

Skepsis:

- Garbage in, garbage out
- Mixing apples and oranges

Chance:

- Umfassende und systematische Übersicht
- Erhöhung der statistischen Power
- Untersuchung der Unterschiedlichkeit der Studienergebnisse
- Verbesserung von Publikations-Standards

Meta-Analysen - Einführung

Meta-Analysen – Einführung

Cochrane Collaboration

Statistisches Verfahren, um die Ergebnisse mehrerer Studien, die die gleiche Frage bearbeiten, quantitativ zu einem Gesamtergebnis zusammenzufassen und dadurch die Aussagekraft (Genauigkeit der Effektschätzer) gegenüber Einzelstudien zu erhöhen.

Meta-Analysen werden mit zunehmender Häufigkeit in systematischen Reviews eingesetzt. Allerdings beruht nicht jede Meta-Analyse auf einem systematischen Review.

Meta-Analyse

- Definition einer Fragestellung
- Literatursuche / Selektion von Studien
- Datenbeschaffung / Datenextraktion
- Beurteilung der Qualität der Einzelstudien
- Analyse / Interpretation
- Präsentation der Resultate

Meta-Analysen – Einführung

Standards für die Berichterstattung

- **QUOROM**: Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Lancet, 354, 1896-1900.

- **PRISMA**: Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Journal of Clinical Epidemiology, 62, 1006-1012.

- Ressing, Blettner, Klug (2009): Teil 6 der Serie zur Bewertung wissenschaftlicher Publikationen: Systematische Übersichtsarbeiten und Metaanalysen. Dtsch Arztebl Int 2009; 106(27): 456–6

Effektmaße

	Merkmal 1	Merkmal 2	Zeilensumme
Stichprobe 1	a	b	a + b
Stichprobe 2	c	d	c + d
Spaltensumme	a + c	b + d	n = a+b+c+d

Gemeinsames „Effektmaß“

Approximativ Normalverteilt

Effektschätzer für **binäre Zielgrößen** (Ansprechrate, Heilung)

- Risiko Differenz (RD)
- Relatives Risiko (RR)
- Odds Ratio (OR)

Effektschätzer für **stetige Zielgrößen** (GFR, FEV_1, Scores)

- Mittelwertdifferenz (MWD)
- Standardisierte Mittelwertdifferenz (SMWD, z.B. Hedges‘ g)

Effektschätzer für **Time-to-event Daten** (Gesamtüberleben, PFS)

- Hazard Ratio (HR)

Effektmaße

Exkurs: Effektmaße für ein binäres Zielkriterium:

	Therapieabbruch		
	ja	nein	
Therapie 1	40	60	100
Therapie 2	50	50	100
	90	110	200

$$RD = \quad 0.4 - 0.5 = -0.1$$

$$RR = \quad \frac{0.40}{0.50} = 0.8$$

$$OR = \quad \frac{2:3}{1:1} = 0.667$$

Relative Effektmaße

Annahme: Approximativ Normalverteilt

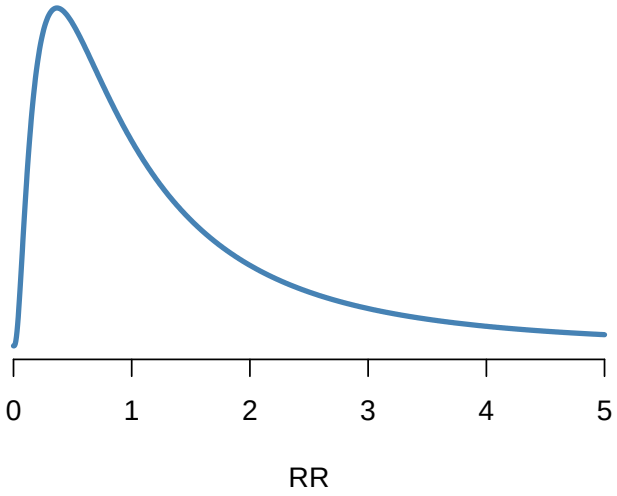
$RR \in [0;\infty]$

RR=1 ⇔ kein Unterschied zwischen den Behandlungsgruppen

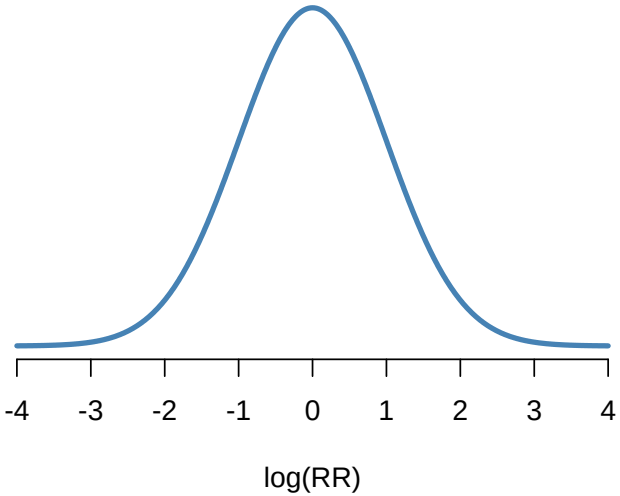
Für RR=1 liegt in [0;1] genauso viel Dichte wie in [1;∞]

⇒ RR ist approximativ lognormalverteilt, d. h. Dichte von log(RR) normalverteilt.

Dichte von RR



Dichte von log(RR)

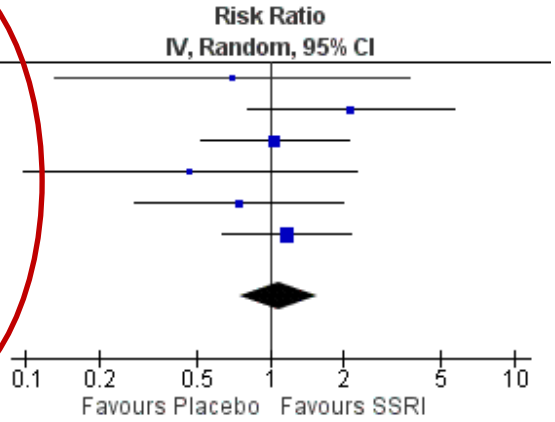


Das Prinzip

- Berechnung des Gesamtschätzers als gewichtetes Mittel der Schätzer aus den Einzelstudien
- Gewicht = Inverse der Varianz $\hat{\theta} = \frac{\sum w_i y_i}{\sum w_i}$ $w_i = \frac{1}{v_i}$

Study or Subgroup	SSRI Events	SSRI Total	Placebo Events	Placebo Total	Weight	Risk Ratio IV, Random, 95% CI
Wermuth 1998	2	18	3	19	4.7%	0.70 [0.13, 3.73]
Devos 2008	8	15	4	16	13.9%	2.13 [0.81, 5.64]
Fregni 2004	9	21	9	22	26.5%	1.05 [0.52, 2.12]
Menza 2008	2	18	4	17	5.4%	0.47 [0.10, 2.25]
Leentjes 2003	3	6	4	6	13.7%	0.75 [0.28, 2.00]
Antonini 2006	10	16	8	15	35.7%	1.17 [0.64, 2.15]
Total (95% CI)		**94**		**95**	**100.0%**	**1.08 [0.75, 1.55]**
Total events	34		32			

Heterogeneity: Tau² = 0.00; Chi² = 3.82, df = 5 (P = 0.58); I² = 0%
Test for overall effect: Z = 0.42 (P = 0.67)



Quelle: Skapinakis et.al BMC Neurology 2010: Response

Equal versus Random Effects

… oder doch ein paar Formeln

Fixed Effect Model (FEM) – *oder richtiger*: Equal Effect Model:

- Man geht von der Existenz eines "wahren" Behandlungseffektes θ aus, der in allen Studien derselbe ist.
- Unterschiede zwischen den Ergebnissen der Einzelstudien sind nur auf Zufallsfehler zurückzuführen

$$\boldsymbol{y_i = \theta + \varepsilon_i} \qquad \boldsymbol{\varepsilon_i \sim N(0, v_i)}$$

Random Effects Model (REM)

- Vorliegende Studien als Zufallsstichprobe aus einer Grundgesamtheit von Studien
- Es gibt nicht einen einzigen wahren Behandlungseffekt, sondern Verteilung von Effekten
- Einbeziehung der Variabilität zwischen den Studien

$$\boldsymbol{y_i = \theta_i + \varepsilon_i = \mu + u_i + \varepsilon_i} \qquad \boldsymbol{u_i \sim N(0, \tau^2)}$$

Gesamtschätzer

- Berechnung des Gesamtschätzers als gewichtetes Mittel der Schätzer aus den Einzelstudien
- Gewicht = Inverse der Varianz

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$$

mit Varianz

$$\mathrm{var}(\hat{\theta}) = \frac{1}{\sum_{i=1}^{k} w_i}$$

k = Anzahl der Studien

w_i = Gewichtungsfaktor der i-ten Studie

Heterogenität

Q-Test basierend auf Cochran‘s Q-Statistik

- Teststatistik für Homogenitätstest $H_0: \ \theta_1 = \theta_2 = \ ... = \ \theta_k$
- $Q = \sum_{i=1}^{k} w_i(y_i - \bar{y})^2$ unter der Nullhypothese Chi-quadrat verteilt mit k-1 Freiheitsgraden
- Große Werte für Q sprechen gegen die Annahme homogener Effekte

Heterogenität

Maß für Heterogenität: Higgins I^2

$$I^2 = 100\% \times \frac{Q - (k-1)}{Q}$$

Interpretation

0% - 40%	→ geringe Heterogenität
30% - 60%	→ moderate Heterogenität
50% - 90%	→ bedeutende Heterogenität
75% - 100%	→ erhebliche Heterogenität

Schätzer für τ^2

- **DerSimonian und Laird Schätzer (in RevMan implementiert)**
- Hedges Schätzer
- Hunter-Schmidt Schätzer
- Sidik-Jonkman Schätzer
- Maximum likelihood Schätzer
- Restricted maximum likelihood Schätzer
- Empirical Bayes / Paule-Mandel Schätzer
- …

DerSimonian-Laird Schätzer für τ^2

$$\hat{\tau}^2 = \frac{Q-(k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}} \qquad w_i = \frac{1}{v_i}$$

- Falls der Schätzer negativ ist, dann wird er auf Null gesetzt
- Der Schätzer ist z.B. in RevMan implementiert
- Es wurde gezeigt, dass der DerSimonian und Laird Ansatz zu Überschreitungen des Signifikanzniveaus führt.

Adjustierung der Schätzung von τ^2

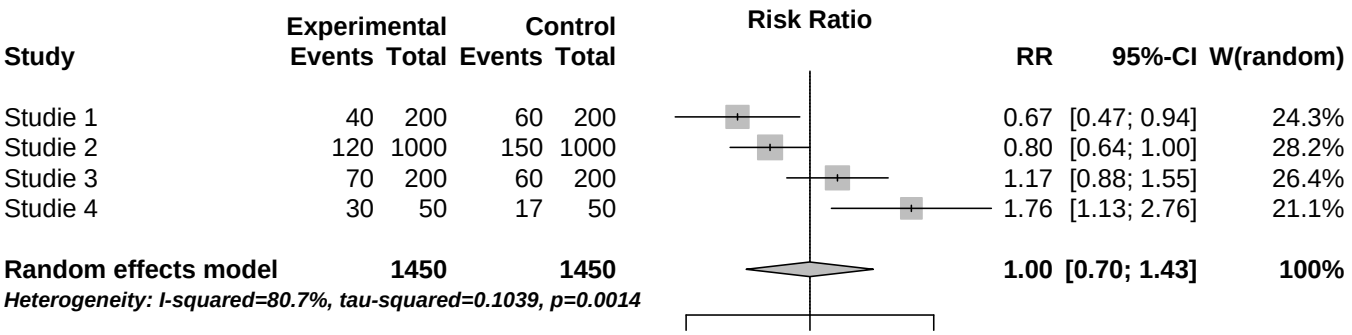
- RE-Modell: Annahme zufälliger Abweichungen zwischen den wahren Effekten der Einzelstudien $\theta_i \sim N(\mu, \tau^2)$
  - diese Variabilität wird durch τ^2 ausgedrückt (Heterogenität)
  - Aber wahres τ^2 nicht bekannt, weshalb α-Fehler und KI für die Schätzung von μ im Allgemeinen zu klein.

- Knapp & Hartung (2003)-Methode adjustiert die Schätzung der Varianz von μ (und berücksichtigt damit die Unsicherheit der Schätzung von τ^2)
  - Schätzung der Varianz von μ durch $Var(\hat{\mu}) = \frac{s^2}{\sum w_i}$ mit $s^2 = \frac{1}{k-1}\sum wi(yi - \hat{\mu})^2$ und KI für $\hat{\mu}$ mit Quantil der t-Verteilung mit k-1 Freiheitsgraden (k: Studienanzahl)
  - Vergleiche: im gewöhnlichen RE-Modell wird die Varianz von μ durch $Var(\hat{\mu}) = \frac{1}{\sum w_i}$ geschätzt und KI für $\hat{\mu}$ mit Quantil der Standardnormalverteilung berechnet.

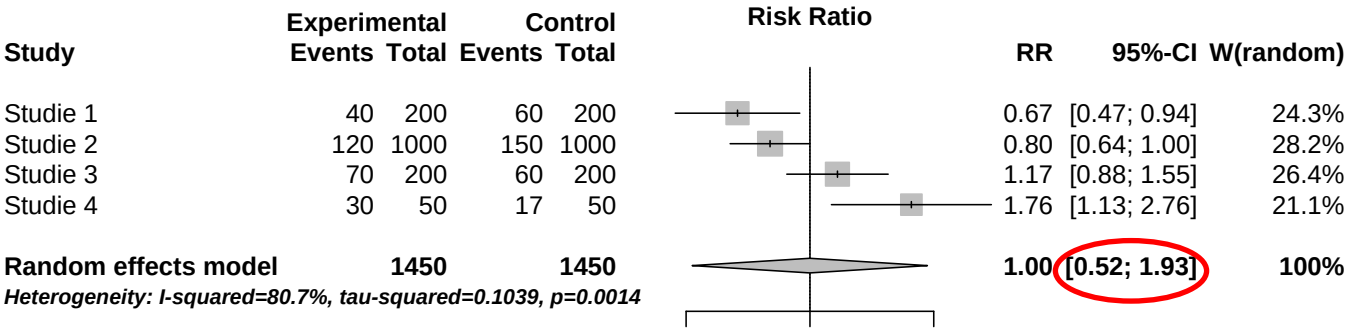
- **Fazit:** Die K&H-Methode erzielt einen verbesserten Schätzer der Varianz von μ wodurch Konfidenzintervall für den Gesamtschätzer des relativen Risiko (RR) breiter wird. Bei kleiner Studienanzahl kann KI wesentlich breiter werden → siehe nächste Folie.

Beispiel

- RE-Modell



- RE-Modell mit Knapp & Hartung Adjustierung für $Var(\hat{\mu})$

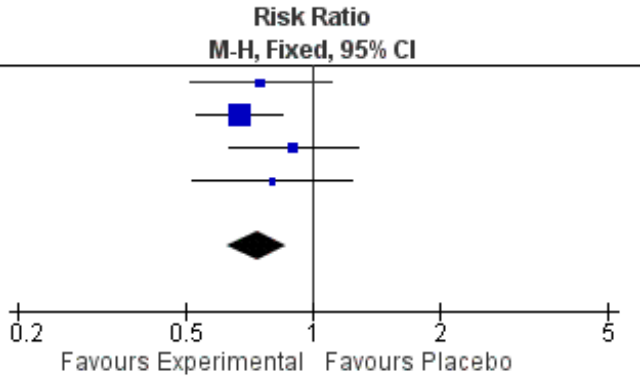


Heterogenität

Study or Subgroup	Experimental Events	Experimental Total	Placebo Events	Placebo Total	Weight	Risk Ratio M-H, Fixed, 95% CI
Studie 1	30	100	40	100	15.1%	0.75 [0.51, 1.10]
Studie 2	100	1000	150	1000	56.6%	0.67 [0.53, 0.85]
Studie 3	45	200	50	200	18.9%	0.90 [0.63, 1.28]
Studie 4	20	50	25	50	9.4%	0.80 [0.52, 1.24]
Total (95% CI)		**1350**		**1350**	**100.0%**	**0.74 [0.62, 0.87]**
Total events	195		265			

Heterogeneity: Chi² = 2.07, df = 3 (P = 0.56); I² = 0%

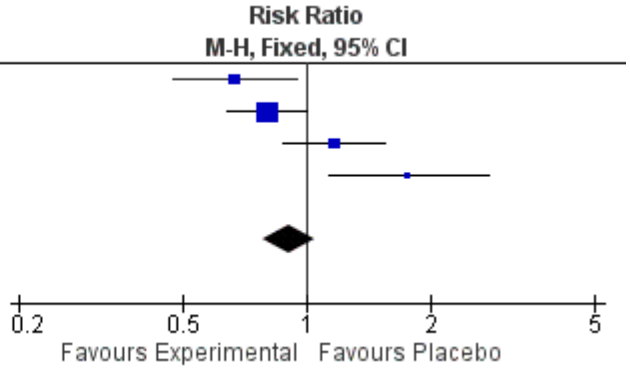
Test for overall effect: Z = 3.66 (P = 0.0003)



Study or Subgroup	Experimental Events	Experimental Total	Placebo Events	Placebo Total	Weight	Risk Ratio M-H, Fixed, 95% CI
Studie 1	40	200	60	200	20.9%	0.67 [0.47, 0.94]
Studie 2	120	1000	150	1000	52.3%	0.80 [0.64, 1.00]
Studie 3	70	200	60	200	20.9%	1.17 [0.88, 1.55]
Studie 4	30	50	17	50	5.9%	1.76 [1.13, 2.76]
Total (95% CI)		**1450**		**1450**	**100.0%**	**0.91 [0.78, 1.05]**
Total events	260		287			

Heterogeneity: Chi² = 15.74, df = 3 (P = 0.001); I² = 81%

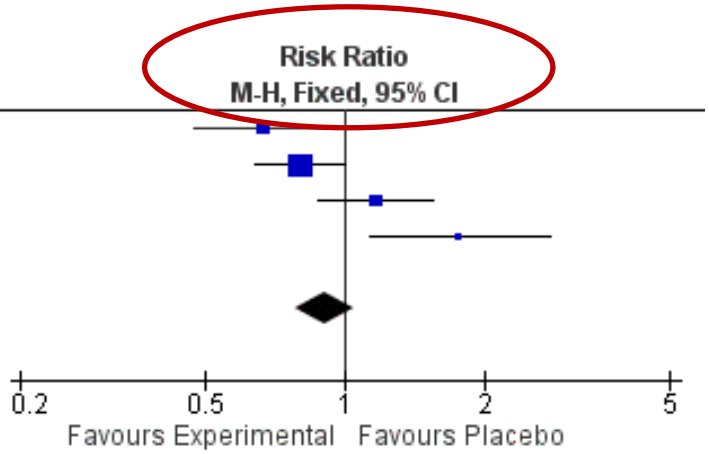
Test for overall effect: Z = 1.31 (P = 0.19)



Heterogenität

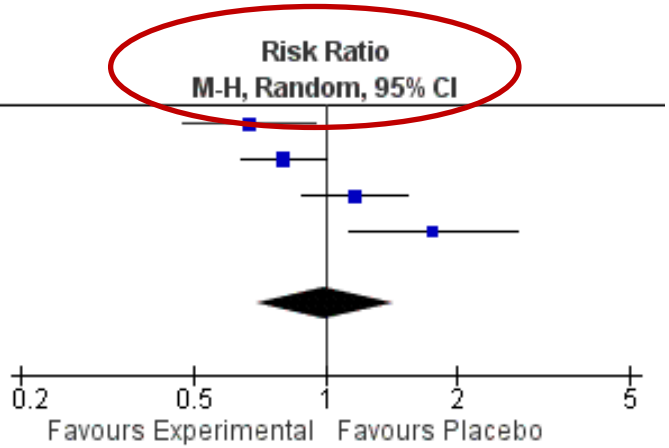
Study or Subgroup	Experimental Events	Total	Placebo Events	Total	Weight	Risk Ratio M-H, Fixed, 95% CI
Studie 1	40	200	60	200	20.9%	0.67 [0.47, 0.94]
Studie 2	120	1000	150	1000	52.3%	0.80 [0.64, 1.00]
Studie 3	70	200	60	200	20.9%	1.17 [0.88, 1.55]
Studie 4	30	50	17	50	5.9%	1.76 [1.13, 2.76]
Total (95% CI)		**1450**		**1450**	**100.0%**	**0.91 [0.78, 1.05]**
Total events	260		287			

Heterogeneity: Chi² = 15.74, df = 3 (P = 0.001); I² = 81%
Test for overall effect: Z = 1.31 (P = 0.19)



Study or Subgroup	Experimental Events	Total	Placebo Events	Total	Weight	Risk Ratio M-H, Random, 95% CI
Studie 1	40	200	60	200	24.3%	0.67 [0.47, 0.94]
Studie 2	120	1000	150	1000	28.2%	0.80 [0.64, 1.00]
Studie 3	70	200	60	200	26.4%	1.17 [0.88, 1.55]
Studie 4	30	50	17	50	21.2%	1.76 [1.13, 2.76]
Total (95% CI)		**1450**		**1450**	**100.0%**	**1.00 [0.70, 1.43]**
Total events	260		287			

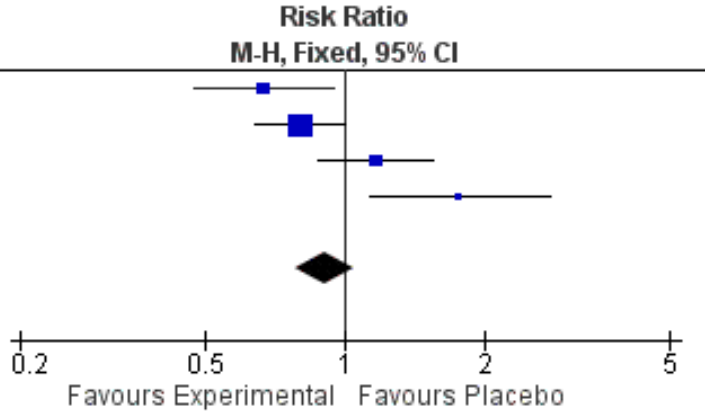
Heterogeneity: Tau² = 0.11; Chi² = 15.74, df = 3 (P = 0.001); I² = 81%
Test for overall effect: Z = 0.00 (P = 1.00)



Heterogenität

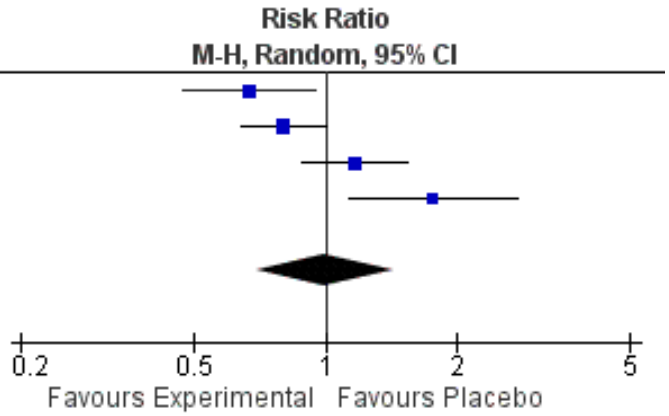
Study or Subgroup	Experimental Events	Experimental Total	Placebo Events	Placebo Total	Weight	Risk Ratio M-H, Fixed, 95% CI
Studie 1	40	200	60	200	20.9%	0.67 [0.47, 0.94]
Studie 2	120	1000	150	1000	52.3%	0.80 [0.64, 1.00]
Studie 3	70	200	60	200	20.9%	1.17 [0.88, 1.55]
Studie 4	30	50	17	50	5.9%	1.76 [1.13, 2.76]
Total (95% CI)		**1450**		**1450**	**100.0%**	**0.91 [0.78, 1.05]**
Total events	260		287			

Heterogeneity: Chi² = 15.74, df = 3 (P = 0.001); I² = 81%
Test for overall effect: Z = 1.31 (P = 0.19)



Study or Subgroup	Experimental Events	Experimental Total	Placebo Events	Placebo Total	Weight	Risk Ratio M-H, Random, 95% CI
Studie 1	40	200	60	200	24.3%	0.67 [0.47, 0.94]
Studie 2	120	1000	150	1000	28.2%	0.80 [0.64, 1.00]
Studie 3	70	200	60	200	26.4%	1.17 [0.88, 1.55]
Studie 4	30	50	17	50	21.2%	1.76 [1.13, 2.76]
Total (95% CI)		**1450**		**1450**	**100.0%**	**1.00 [0.70, 1.43]**
Total events	260		287			

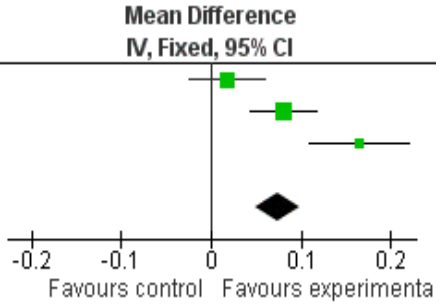
Heterogeneity: Tau² = 0.11; Chi² = 15.74, df = 3 (P = 0.001); I² = 81%
Test for overall effect: Z = 0.00 (P = 1.00)



Beispiele

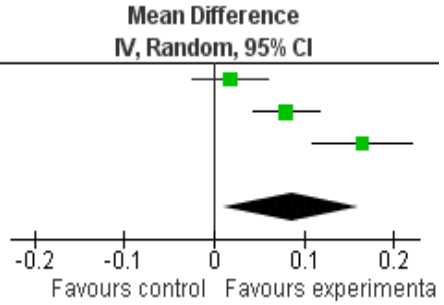
	Experimental			Control				Mean Difference
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Fixed, 95% CI
Study 1	0.979	0.409	730	0.961	0.417	766	36.0%	0.02 [-0.02, 0.06]
Study 2	1.15	0.378	745	1.07	0.367	745	44.1%	0.08 [0.04, 0.12]
Study 3	1.3	0.405	400	1.136	0.406	400	20.0%	0.16 [0.11, 0.22]
Total (95% CI)			**1875**			**1911**	**100.0%**	**0.07 [0.05, 0.10]**

Heterogeneity: Chi² = 16.82, df = 2 (P = 0.0002); I² = 88%
Test for overall effect: Z = 5.81 (P < 0.00001)



	Experimental			Control				Mean Difference
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random, 95% CI
Study 1	0.979	0.409	730	0.961	0.417	766	34.0%	0.02 [-0.02, 0.06]
Study 2	1.15	0.378	745	1.07	0.367	745	34.7%	0.08 [0.04, 0.12]
Study 3	1.3	0.405	400	1.136	0.406	400	31.3%	0.16 [0.11, 0.22]
Total (95% CI)			**1875**			**1911**	**100.0%**	**0.09 [0.01, 0.16]**

Heterogeneity: Tau² = 0.00; Chi² = 16.82, df = 2 (P = 0.0002); I² = 88%
Test for overall effect: Z = 2.23 (P = 0.03)



Zwischenfazit

- Der zusammenfassende Effektschätzer wird als gewichtetes Mittel der Effektschätzer der Einzelstudien berechnet

- Als Gewicht wird die Inverse der Varianz des Effektschätzers gewählt

$$FEM: \quad w_i = \frac{1}{v_i} \qquad REM: \quad w_i = \frac{1}{v_i + \widehat{\tau^2}}$$

- Die Heterogenität der Studien sollte beurteilt werden und sollte ggfs. bei der Modellwahl (FEM versus REM) berücksichtigt werden
  - Q-Test
  - Higgins I^2

Zwischenfazit

Equal Effects

- Geht man von gleichen Effekten in den Einzelstudien aus, dann ist die Variabilität zwischen Studien nur auf Zufallsfehler zurückzuführen

- „Große“ Studien erhalten mehr Gewicht

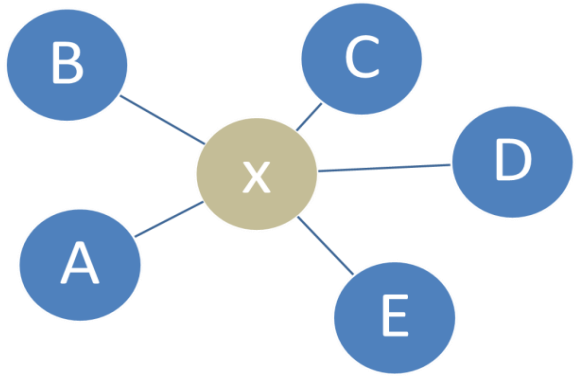
Zwischenfazit

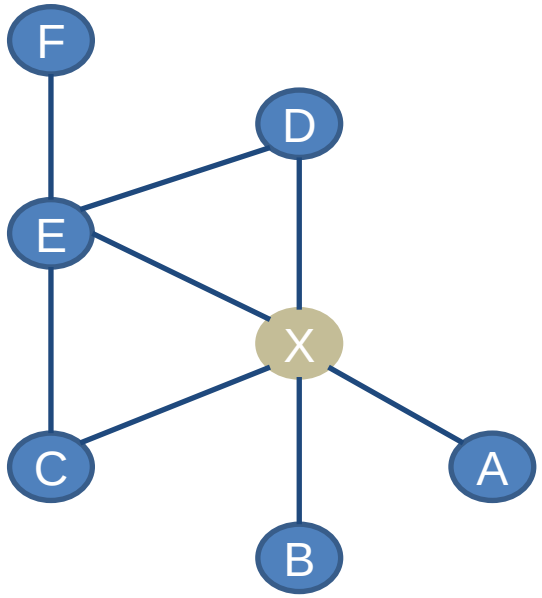
Random Effects

- Neben der Stichprobenvarianz ist auch eine Zwischen-Studien-Variabilität berücksichtigt
- Ist die Zwischen-Studien-Variabilität gleich Null dann ist der Schätzer des FEM gleich dem des REM
- Im Falle erheblicher Heterogenität werden kleinere Studien gegenüber dem FEM höher gewichtet
- Liefert breitere Konfidenzintervalle
- Der DerSimonian und Laird Ansatz kann zur Überschreitung des Fehlers erster Art führen (andere Schätzer für τ^2 sind hier zu präferieren)

Ausblick

- Meta-Analyse auf der Basis individueller Patientendaten der Einzelstudien (IPD Meta-Analyse)
- Meta-Regression
- Netzwerk Meta-Analysen / Mixed Treatment Comparison





Fazit

Meta-Analysen sind kein Ersatz für große, gut designte klinische Studien …

… umgekehrt jedoch, sind große klinische Studien auch kein Ersatz für eine umfassende Meta-Analyse

... turning complexity into clarity

Doris Böhm

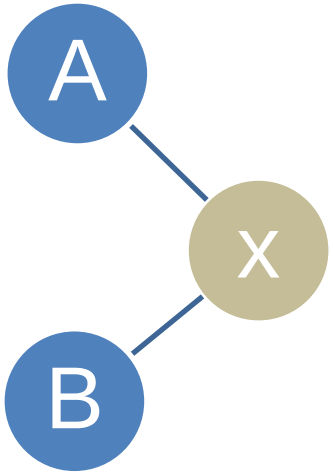
Rosa-Bavarese-Str. 5
D-80639 München

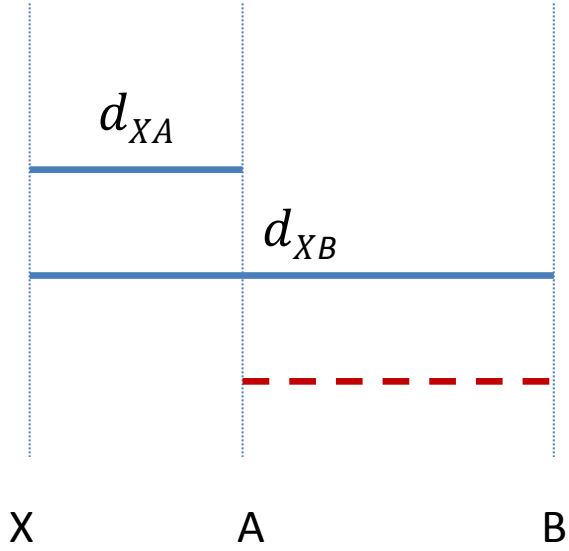
Phone +49 (0) 89-200 00 74-114
Mobile +49 (0) 157-71451035

Doris.Boehm@ams-europe.com
www.ams-europe.com

Backup

Indirekte Vergleiche





$$d_{AB} = d_{XB} - d_{XA}$$

direkter Vergleich

indirekter Vergleich

d_{XA} Effekt von Behandlung A relativ zu Behandlung X

Indirekter Vergleich

'Trial 1: Porsche versus Golf'

Porsche - Golf = 2s

'Trial 2: Volvo versus Golf'

Volvo - Golf = 8s

→ Indirect Comparison:

Volvo versus Porsche: 8-2=6s

Indirekter Vergleich – Grundlegende Annahmen

Ähnlichkeit (klinisch und methodisch)

- Vergleichbarkeit der Studien bezüglich Charakteristika der eingeschlossenen Patienten
- Art und Dauer der Behandlung (Dosis, Behandlungsdauer)
- Vergleichbarkeit bezüglich der Verteilung prognostischer Faktoren

Homogenität

- Vergleichbarkeit der geschätzten Effekte innerhalb aller direkten Paarvergleiche

Konsistenz

- Kommen direkte und indirekte Vergleiche zu denselben Schlüssen bzgl. des Behandlungseffekts / Ausmaß des Effekts?