IQWiG Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Institute for Quality and Efficiency in Health Care

IQWiG im Dialog 17.06.2011, Köln

Biometrische Verfahrensweise mit Heterogenität im IQWiG – aktueller Stand und Ausblick

Guido Skipka

Agenda

- Identifizierung und Quantifizierung von Heterogenität
- Identifizierung zusammenfassbarer Studienpools
- Zusammenfassung der Ergebnisse pro Studienpool
  - quantitativ
  - qualitativ
- Heterogene Ergebnisse zwischen Subgruppen
- Heterogenitätsinterval

Aussagen zur Beleglage zum Nutzen und Schaden

- Zusammenfassende Bewertung
  - Beleg: statistische Signifikanz mehrerer Studien hoher Ergebnissicherheit
  - Hinweis: statistische Signifikanz mehrerer Studien mäßiger Ergebnissicherheit bzw. einer Studie hoher Ergebnissicherheit
  - Anhaltspunkt (neu*): statistische Signifikanz mehrerer Studien geringer Ergebnissicherheit bzw. einer Studie mäßiger Ergebnissicherheit

- mehrere Studien: Konsistenz der Ergebnisse
  - qualitativ (gleichgerichtete Effekte, Heterogenität möglich)
  - quantitativ (homogene Effekte)

*: Allgemeine Methoden, Entwurf Version 4.0, 09.03.2011

Identifizierung und Quantifizierung von Heterogenität

- übliche statistische Maßzahlen: Cochran‘s Q, I^2

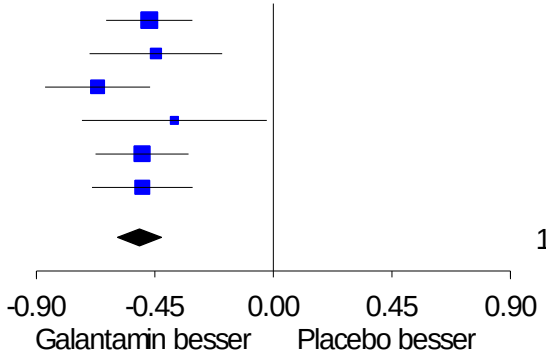
$$Q = \sum_i w_i(\theta_i - \theta_{FEM})^2$$

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

Beispiel: ChE-Hemmer bei Alzheimer-Demenz

Galantamin - kognitive Leistungsfähigkeit
Endpunkt: ADAS-cog/11 - Differenz zu Baseline
Distanzmaß: standardisierte Differenz der Mittelwerte

Studie	Galantamin			Placebo			Cohens d (feste Effekte) 95%-KI	Gewichtung %	Cohens d 95%-KI
	n	Mittel	SD	n	Mittel	SD			
Brodaty 2005	296	-1.60	6.19	296	1.20	5.68		25.65	-0.47 [-0.63, -0.31]
Erkinjuntti 2002	177	-0.80	5.72	96	1.70	5.39		10.84	-0.45 [-0.70, -0.19]
Raskind 2000	202	-1.90	5.12	207	2.00	6.47		17.24	-0.67 [-0.87, -0.47]
Rockwood 2006	62	-1.70	5.00	65	0.30	5.60		5.55	-0.38 [-0.73, -0.03]
Tariot 2000	253	-1.40	6.20	255	1.70	6.23		21.94	-0.50 [-0.68, -0.32]
Wilcock 2000	220	-0.50	5.64	215	2.40	6.01		18.78	-0.50 [-0.69, -0.31]
Gesamt (95%-KI)	1210			1134				100.00	-0.51 [-0.59, -0.43]



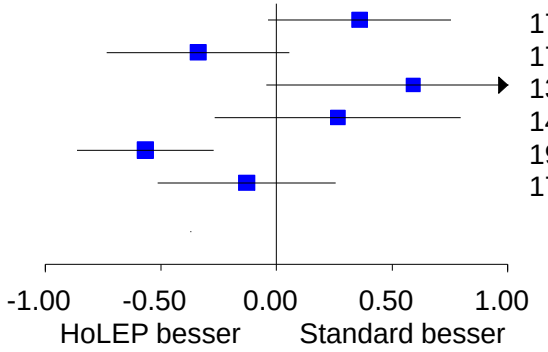
Heterogenität: Q=3.45, df=5 (p=0.630), I^2=0%
Gesamteffekt: Z Score=-12.04 (p=0.000)

Angaben zur Heterogenität

Beispiel: Nichtmedikamentöse lokale Verfahren der Benignen Prostatahyperplasie

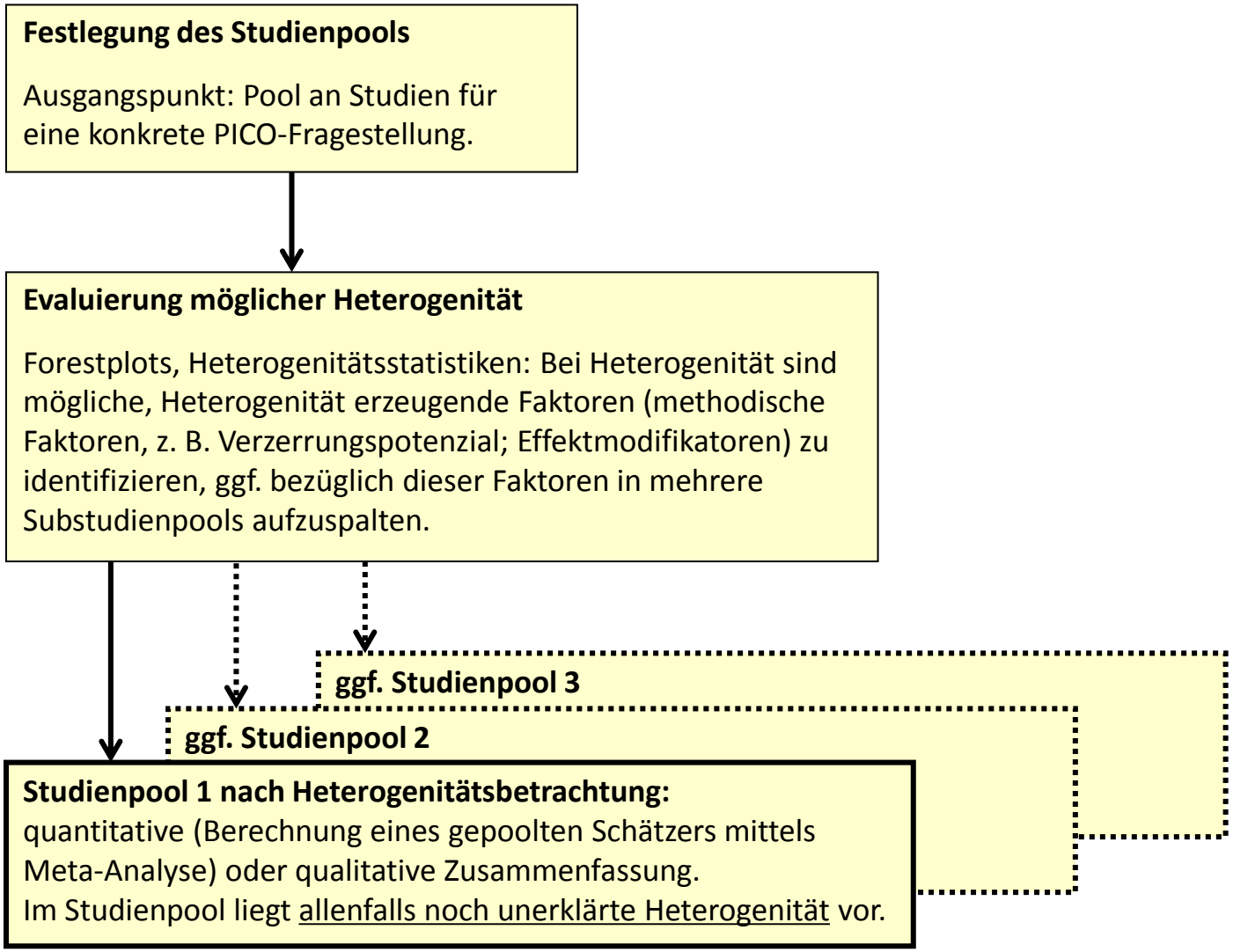
Meta-Analyse BPH: HoLEP vs. Standard
Symptom-Scores nach 6 Monaten
Distanzmaß: Standardisierte Differenz der Mittelwerte

Studie	HoLEP n	HoLEP Mittel	HoLEP SD	Standard n	Standard Mittel	Standard SD	Hedges g (zufällige Effekte) 95%-KI	Gewichtung %	Hedges g 95%-KI
Briganti 2006 (IPSS)	52	3.90	2.90	48	2.90	2.60		17.48	0.36 [-0.04, 0.76]
Gupta 2006* (IPSS)	50	5.20	2.20	50	6.10	3.00		17.49	-0.34 [-0.73, 0.06]
Neill 2006** (IPSS)	20	9.00	6.30	20	5.70	4.50		13.15	0.59 [-0.04, 1.23]
Wilson 2006** (IPSS)	26	6.00	5.10	29	4.80	3.80		14.95	0.27 [-0.27, 0.80]
Ahyai 2007 (AUA)	94	2.20	1.60	89	3.70	3.40		19.26	-0.57 [-0.86, -0.27]
Kuntz 2008 (AUA)	52	2.40	1.90	52	2.80	3.90		17.68	-0.13 [-0.51, 0.26]



Heterogenität: Q=22.43, df=5 (p=0.000), I^2=77.7%
Gesamteffekt: Z Score=-0.06 (p=0.951), tau^2=0.154

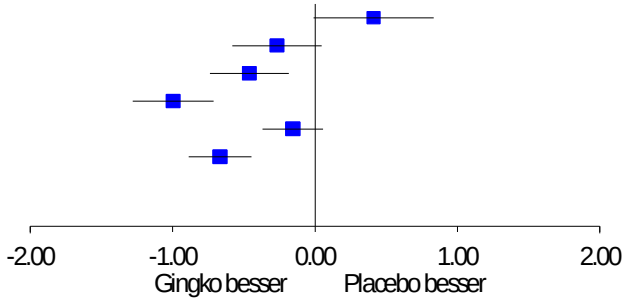
Identifizierung zusammenfassbarer Studienpools



Beispiel: Ginkgo Biloba

Gingko, Aktivitäten des täglichen Lebens
Endpunkt: NAB, GERRI - Gruppenunterschied zu Placebo
Distanzmaß: standardisierte Mittelwertdifferenz

Studie	Effekt	SD	Cohens d 95%-KI	
DIGGER 2007	0.12	0.29	0.41	[-0.01, 0.83]
Kanowski 1996	-0.60	2.24	-0.27	[-0.58, 0.05]
Le Bars 1997	-0.16	0.35	-0.46	[-0.74, -0.18]
Napryeyenko 2007	-2.40	2.41	-1.00	[-1.28, -0.71]
Schneider 2005	-0.05	0.32	-0.16	[-0.37, 0.06]
Schwabe 2008	-0.20	0.30	-0.67	[-0.89, -0.45]

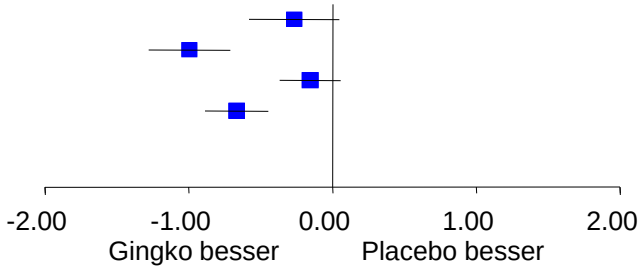


Heterogenität: Q=42.37, df=5 (p=0.000), I^2=88.2%, tau^2=0.144

Höhe der Dosierung als Heterogenität erzeugenden Faktor identifiziert
→ Meta-Analysen, getrennt nach Dosisstufen

hohe Dosierung

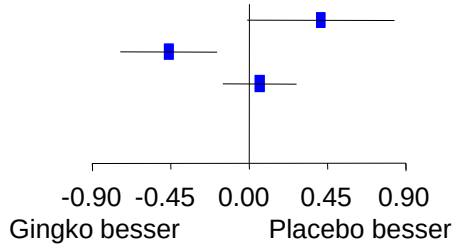
Studie	Effekt	SD	Cohens d 95%-KI
Kanowski 1996	-0.60	2.24	-0.27 [-0.58, 0.05]
Napryeyenko 2007	-2.40	2.41	-1.00 [-1.28, -0.71]
Schneider 2005	-0.05	0.32	-0.16 [-0.37, 0.06]
Schwabe 2008	-0.20	0.30	-0.67 [-0.89, -0.45]



Heterogenität: Q=26.11, df=3 (p=0.000), I^2=88.5%, tau^2=0.127

niedrige Dosierung

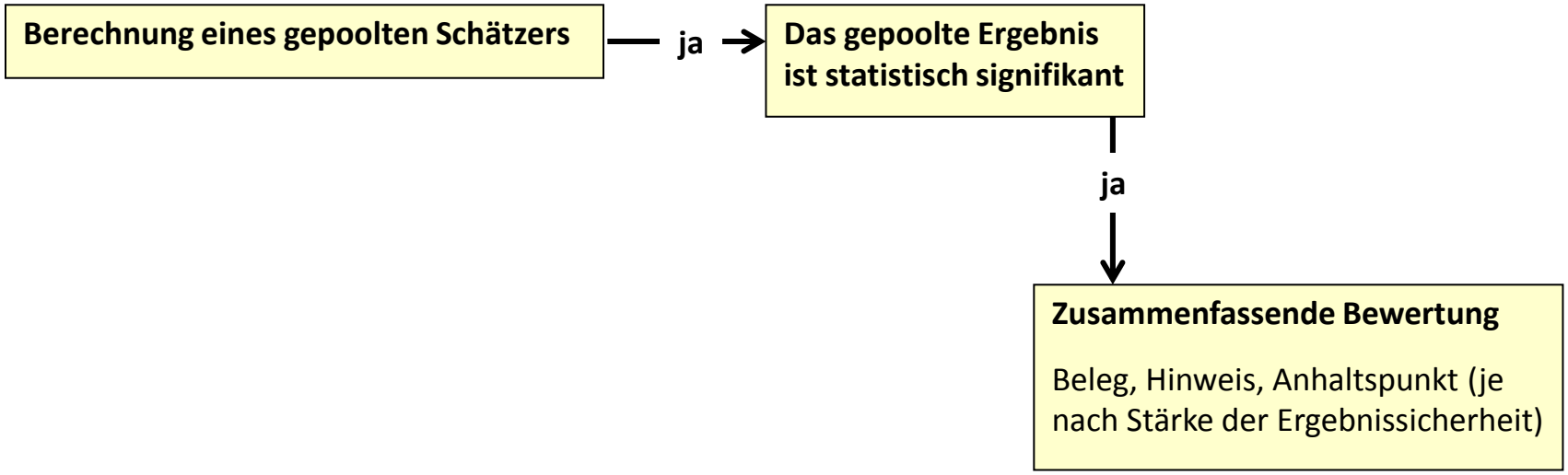
Studie	Effekt	SD	Cohens d (zufällige Effekte) 95%-KI	Cohens d 95%-KI
DIGGER 2007	0.12	0.29		0.41 [-0.01, 0.83]
Le Bars 1997	-0.16	0.35		-0.46 [-0.74, -0.18]
Schneider 2005	0.02	0.33		0.06 [-0.15, 0.27]



Heterogenität: Q=14.05, df=2 (p=0.001), I^2=85.8%, tau^2=0.132

keine weiteren Heterogenität erzeugenden Faktoren ersichtlich
→ unerklärte Heterogenität

Zusammenfassung der Ergebnisse pro Studienpool



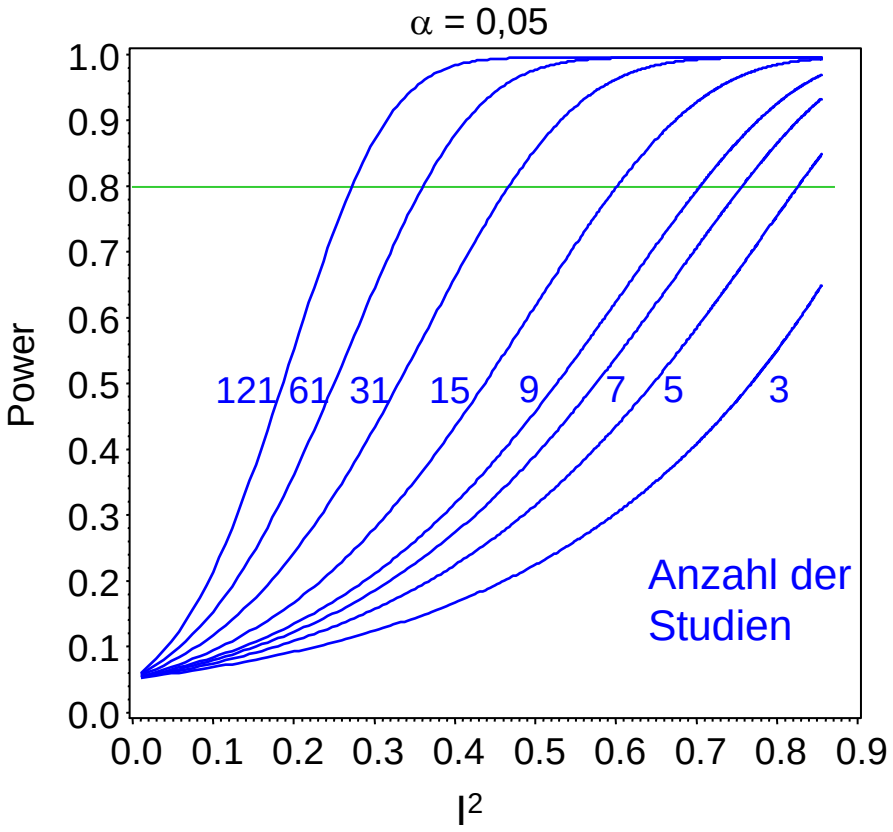
Regelhafte Formulierung in Berichtsplänen

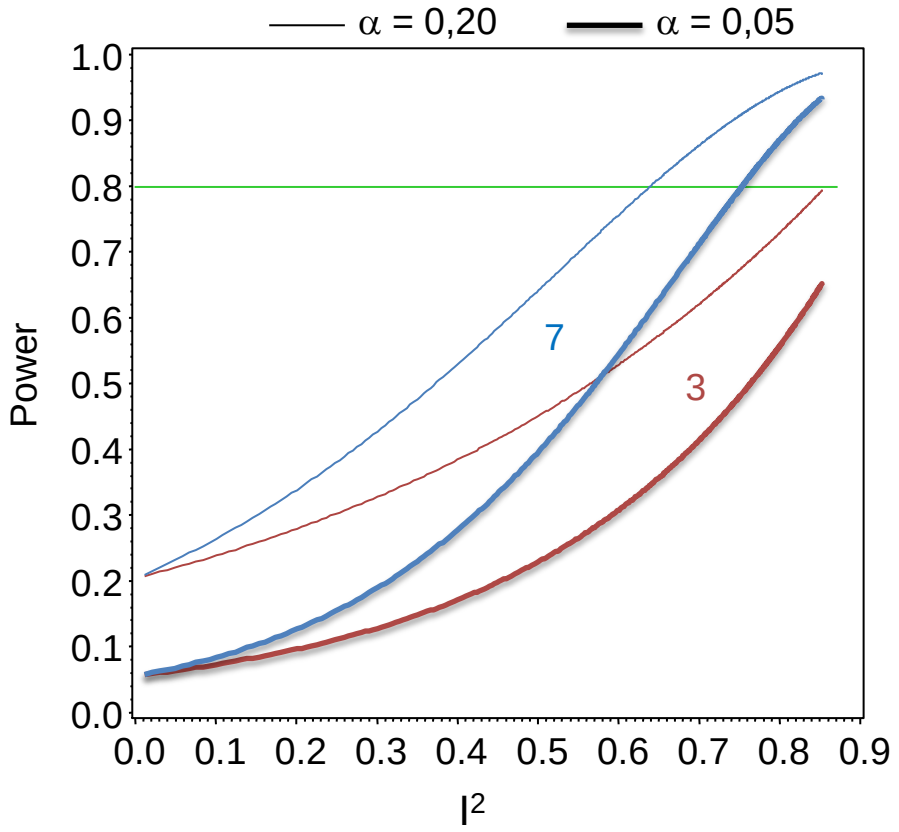
- „Die Meta-Analysen erfolgen in der Regel auf Basis von Modellen mit zufälligen Effekten [Der Simonian & Laird]. In begründeten Ausnahmefällen werden Modelle mit festen Effekten eingesetzt.

- „Die Einschätzung einer möglichen Heterogenität der Studienergebnisse erfolgt anhand des Maßes I^2 und des statistischen Tests auf Vorliegen von Heterogenität [Higgins et al., 2003]. Ist die Heterogenität der Studienergebnisse nicht bedeutsam ($p > 0,2$ für Heterogenitätstest), wird der gemeinsame (gepoolte) Effekt inklusive Konfidenzintervall dargestellt. Bei bedeutsamer Heterogenität werden die Ergebnisse nur in begründeten Ausnahmefällen gepoolt. Außerdem wird untersucht, welche Faktoren diese Heterogenität möglicherweise erklären könnten.“

  Grund: Anzahl der Studien in Meta-Analysen regelhaft klein (2 - 5)
  → sehr geringe Power des Heterogenitätstests

Power des Heterogenitätstests (appr. Formel nach Jackson 2006)

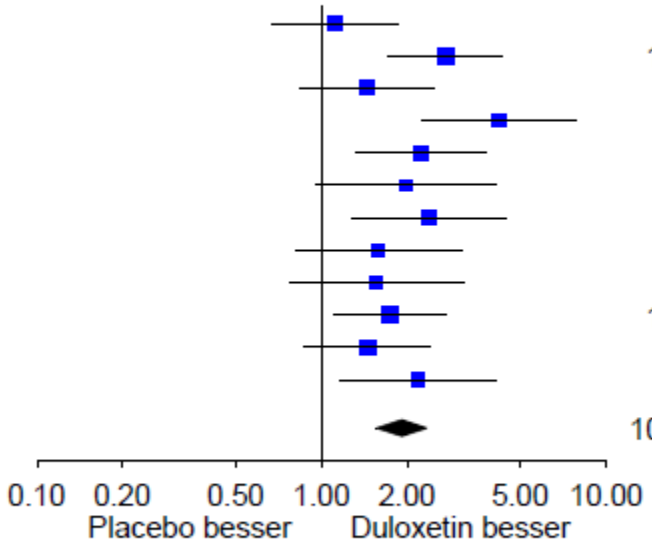




Beispiel: SNRI bei Depression

Remission
Duloxetin vs. Placebo

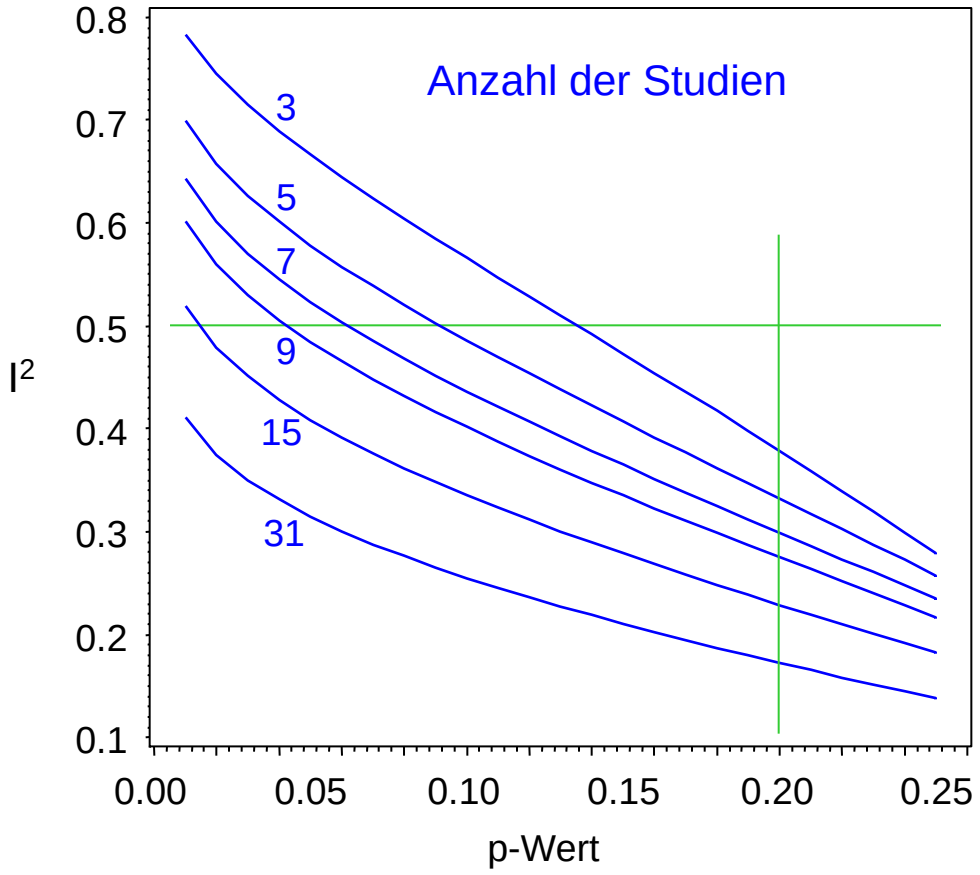
Studie	Duloxetin n/N	Placebo n/N	OR (zufällige Effekte) 95%-KI	Gewichtung %	OR 95%-KI
Brannan 2005	46/132	44/136		9.68	1.12 [0.67, 1.86]
Brecht 2007	82/156	46/159		10.72	2.72 [1.71, 4.33]
Detke 2002a	39/123	33/136		8.82	1.45 [0.84, 2.50]
Detke 2002b	53/121	18/115		7.46	4.20 [2.26, 7.79]
Detke 2004	91/186	28/93		9.20	2.22 [1.31, 3.77]
Goldstein 2002	29/68	18/66		5.91	1.98 [0.96, 4.09]
Goldstein 2004	43/86	26/88		7.37	2.38 [1.28, 4.45]
HMAQ-B	32/81	21/72		6.56	1.59 [0.81, 3.12]
HMAT-A	23/81	18/89		6.12	1.56 [0.77, 3.17]
Nierenberg 2007	105/262	38/137		11.18	1.74 [1.11, 2.73]
Perahia 2006b	82/195	33/99		9.72	1.45 [0.88, 2.41]
Raskin 2007	55/201	15/102		7.27	2.18 [1.16, 4.10]
Gesamt (95%-KI)	680/1692	338/1292		100.00	1.91 [1.56, 2.34]



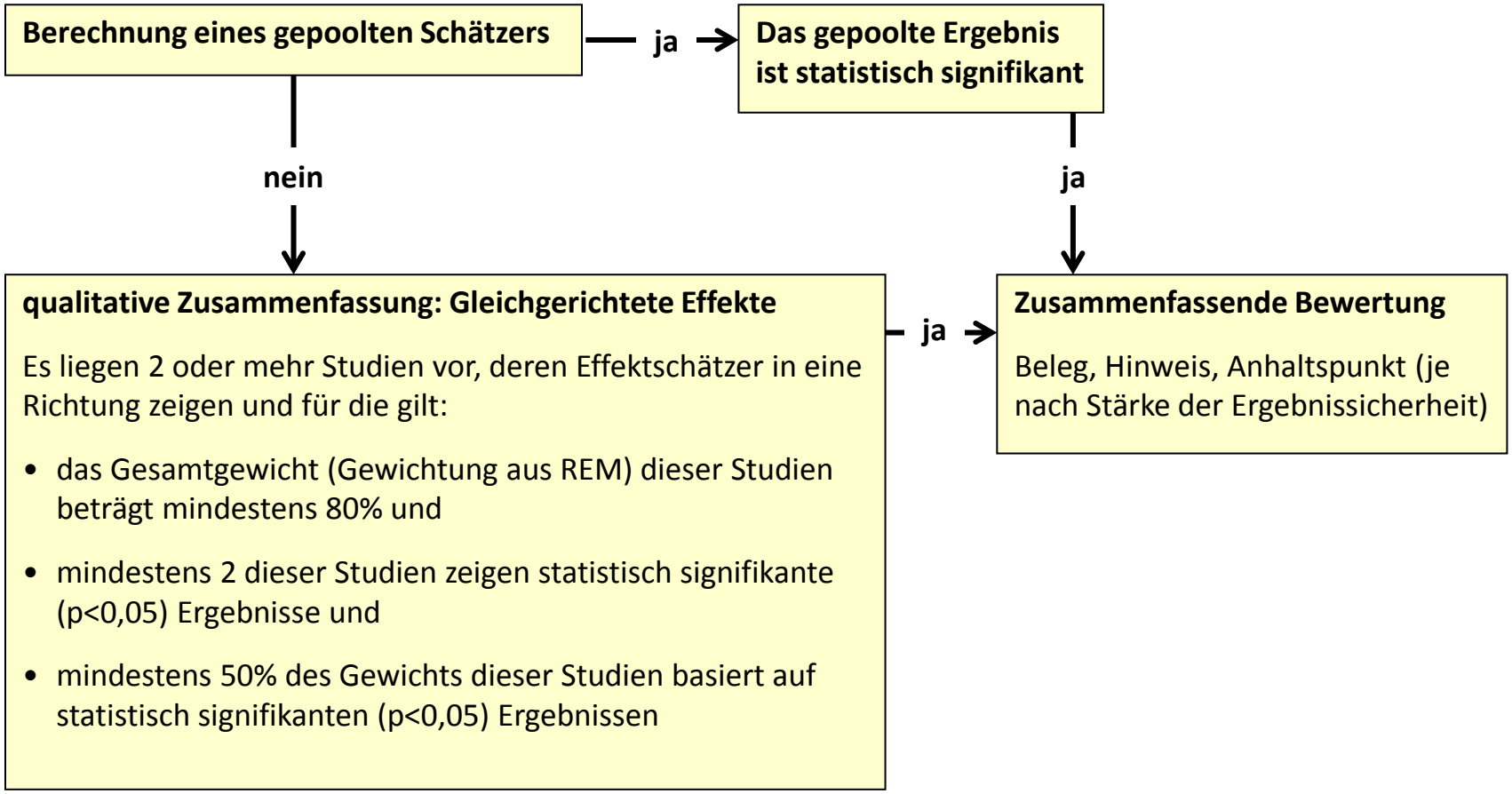
Heterogenität: Q=16.62, df=11 (p=0.120), I^2=33.8%
Gesamteffekt: Z Score=6.31 (p=0.000), tau^2=0.042

Viele Studien: Darstellung des gepoolten Schätzers, falls $I^2 < 50\%$

Zusammenhang zwischen I^2 und p-Wert



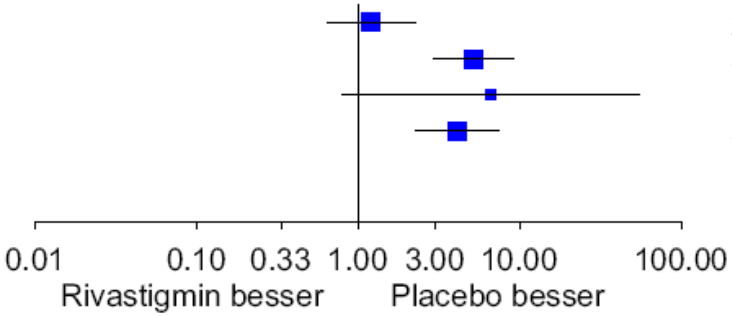
Zusammenfassung der Ergebnisse pro Studienpool



Beispiel: ChE-Hemmer

Rivastigmin - Abbruch wegen UEs
Endpunkt: Studie aufgrund UE vorzeitig abgebrochen (ja, nein)
Distanzmaß: Odds Ratio der Anteile von Patienten mit Abbruch

Studie	Rivastigmin n/N	Placebo n/N	OR (zufällige Effekte) 95%-KI	Gewichtung %	OR 95%-KI
B304	24/227	20/222		29.48	1.19 [0.64, 2.23]
Corey-Bloom 1998	66/231	17/235		30.41	5.13 [2.90, 9.07]
Forette 1999	10/45	1/24		10.03	6.57 [0.79, 54.85]
Rösler 1999	55/243	16/239		30.08	4.08 [2.26, 7.35]



Heterogenität: Q=13.31, df=3 (p=0.004), I^2=77.5%
Gesamteffekt: Z Score=2.88 (p=0.004), tau^2=0.451

Zwischenstand

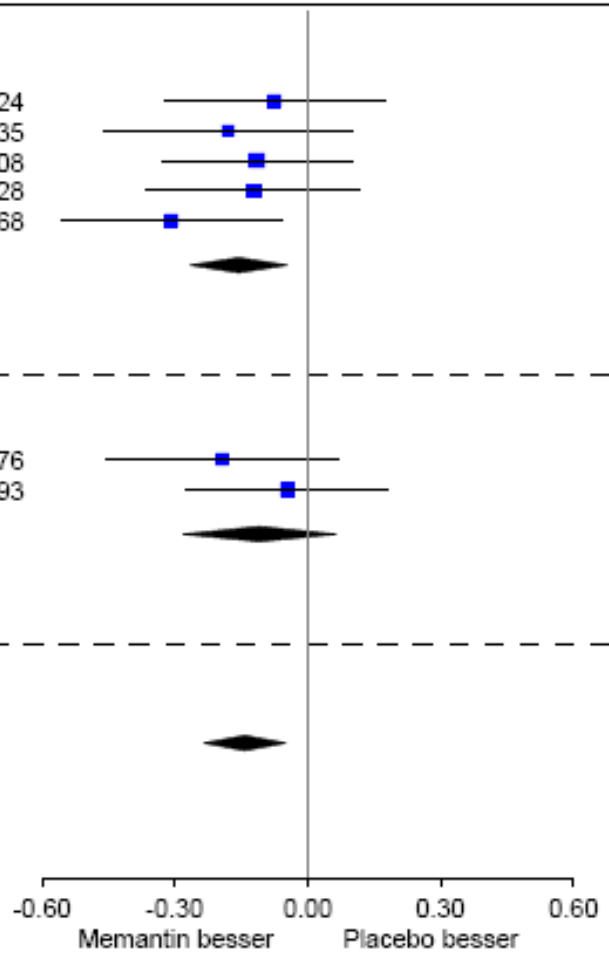
- Wann sind gepoolte Schätzer sinnvoll ?
  - „große“ Heterogenität ($p < 0,05$ oder $I^2 > 70\%$) → Poolen inadäquat
  - „kleine“ Heterogenität ($p > 0,20$ und $I^2 < 25\%$) → Poolen adäquat
  - „mittlere“ Heterogenität (sonst)
    - feste Grenzen für Heterogenitätsstatistiken?
      → einfach, transparent vs. zu pauschal
    - p-Wert oder I^2 oder ... ?
    - Grenzen in Abhängigkeit der Studienzahl oder ... wählen ?
    - Immer Poolen (aber Heterogenität berücksichtigen) ?

- Auch bei heterogenen Ergebnissen sind Aussagen zur Beleglage möglich.

Heterogene Ergebnisse zwischen Subgruppen
Beispiel: Memantin

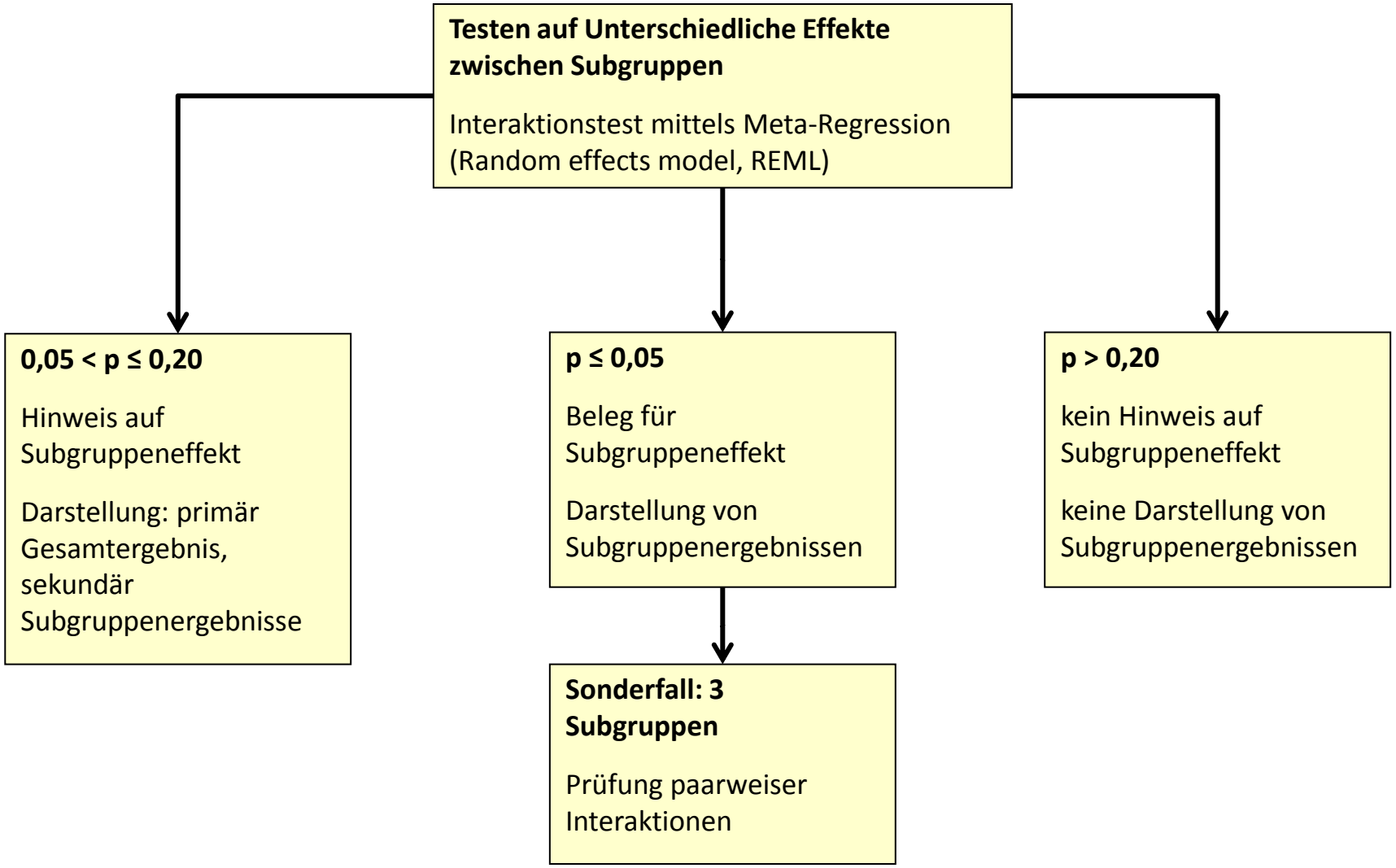
Memantin vs. Placebo
alltagspraktische Fähigkeiten: ADCS-ADL-23, ADCS-ADL-sev
Modell mit zufälligen Effekten - DerSimonian und Laird

Studienpool Studie	Memantin n	Placebo n	Mittelwert- differenz	gepoolte SD	Gewichtung	Cohen's d	95%-KI
Monotherapie							
10116	124	125	-0.47	6.24	13.7	-0.08	[-0.32, 0.17]
99679	163	71	-1.68	9.35	10.8	-0.18	[-0.46, 0.10]
MEM-MD-01	171	165	-0.70	6.08	18.5	-0.12	[-0.33, 0.10]
MEM-MD-10	126	140	-1.14	9.28	14.6	-0.12	[-0.36, 0.12]
MRZ-9605	126	126	-2.06	6.68	13.7	-0.31	[-0.56, -0.06]
Gesamt						-0.16	[-0.26, -0.05]
Heterogenität: Q=2.09, df=4, p=0.719, I²=0% Gesamteffekt: Z Score=-2.81, p=0.005, Tau=0							
Add-on Therapie zu CHE-H							
MEM-MD-02	104	124	-1.30	6.76	12.4	-0.19	[-0.45, 0.07]
MEM-MD-12	151	146	-0.36	7.93	16.3	-0.05	[-0.27, 0.18]
Gesamt						-0.11	[-0.28, 0.06]
Heterogenität: Q=0.69, df=1, p=0.406, I²=0% Gesamteffekt: Z Score=-1.24, p=0.214, Tau=0							
Alle							
Gesamt					100.0	-0.14	[-0.23, -0.05]
Heterogenität: Q=2.99, df=6, p=0.811, I²=0% Gesamteffekt: Z Score=-3.04, p=0.002, Tau=0							

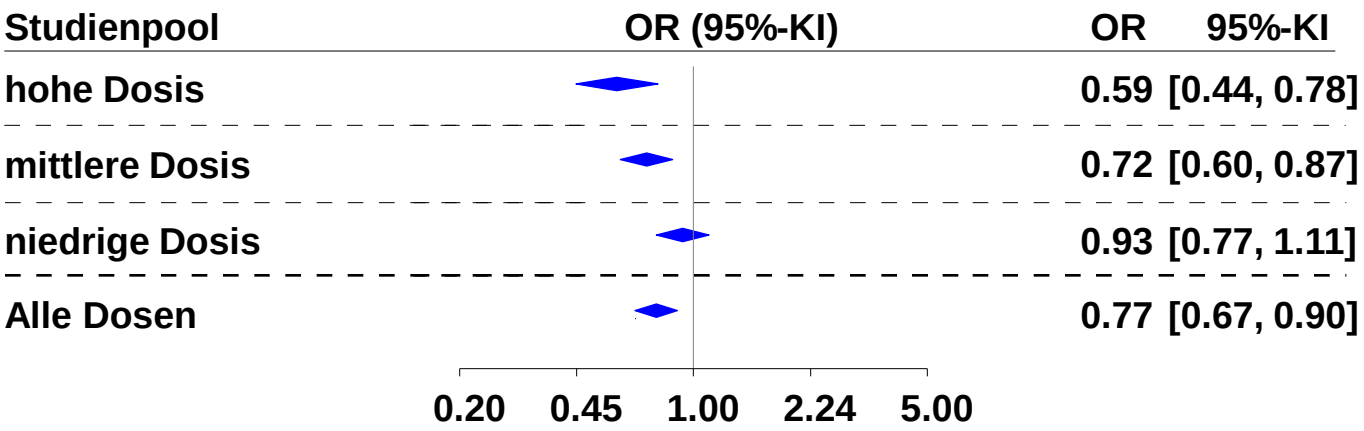


Heterogenität zwischen Subgruppen: Q=0.21, df=1, p=0.648, I²=0%

Vorgehen bei Subgruppenanalysen



fiktives Beispiel: Effektmodifikator → Dosierung

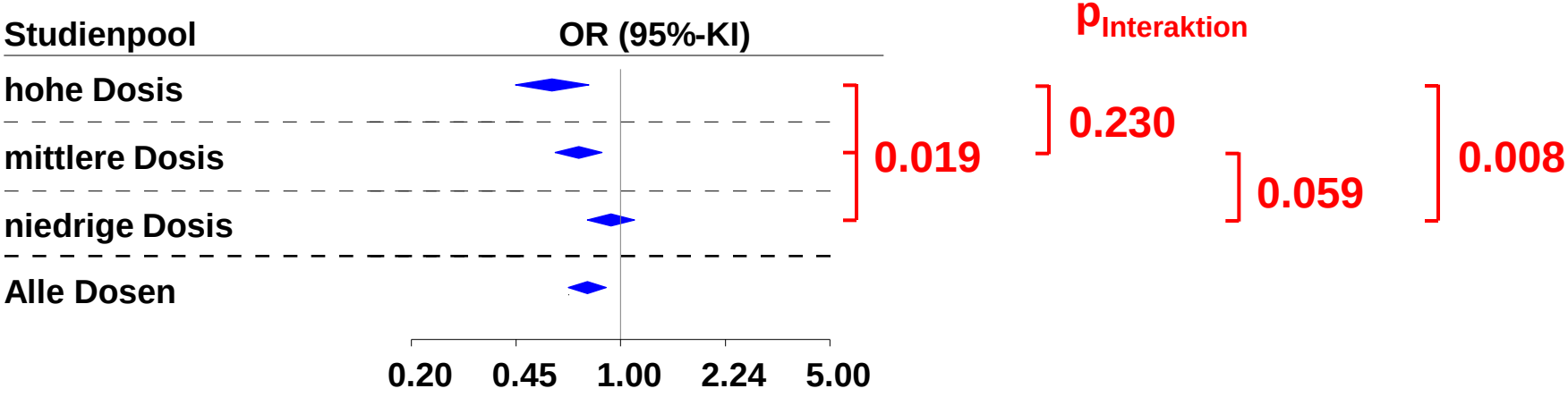


Heterogenität zwischen Studienpools: Q=7.88, df=2, p=0.019, I²=74.6%

Situation

- signifikanter Effekt (α = 0.05) bei gemeinsamer Betrachtung aller Dosierungen
- Dosierung ist signifikanter Effektmodifikator (α = 0.05)
- Frage: Können zumindest 2 der 3 Dosierungsgruppen zur Erhöhung der Präzision zusammengefasst werden?

paarweise Testung auf Interaktion



häufiges Problem bei α = 0.05:

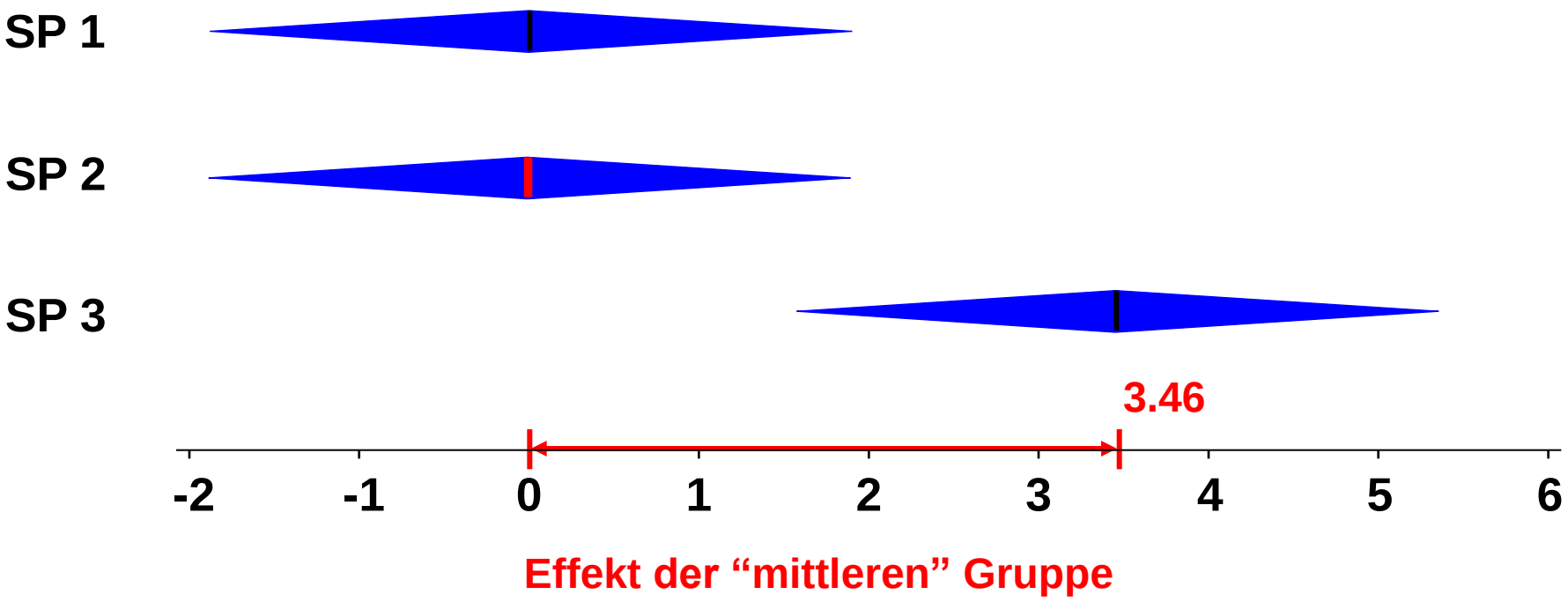
- Ein Unterschied zwischen hoher und niedriger Dosis ist belegt, nicht jedoch zwischen hoher und mittlerer sowie zwischen mittlerer und niedriger Dosis (nicht transitive Relation).
- Idee: Anhebung des Niveaus (α > 0.05) bei paarweisem Testen

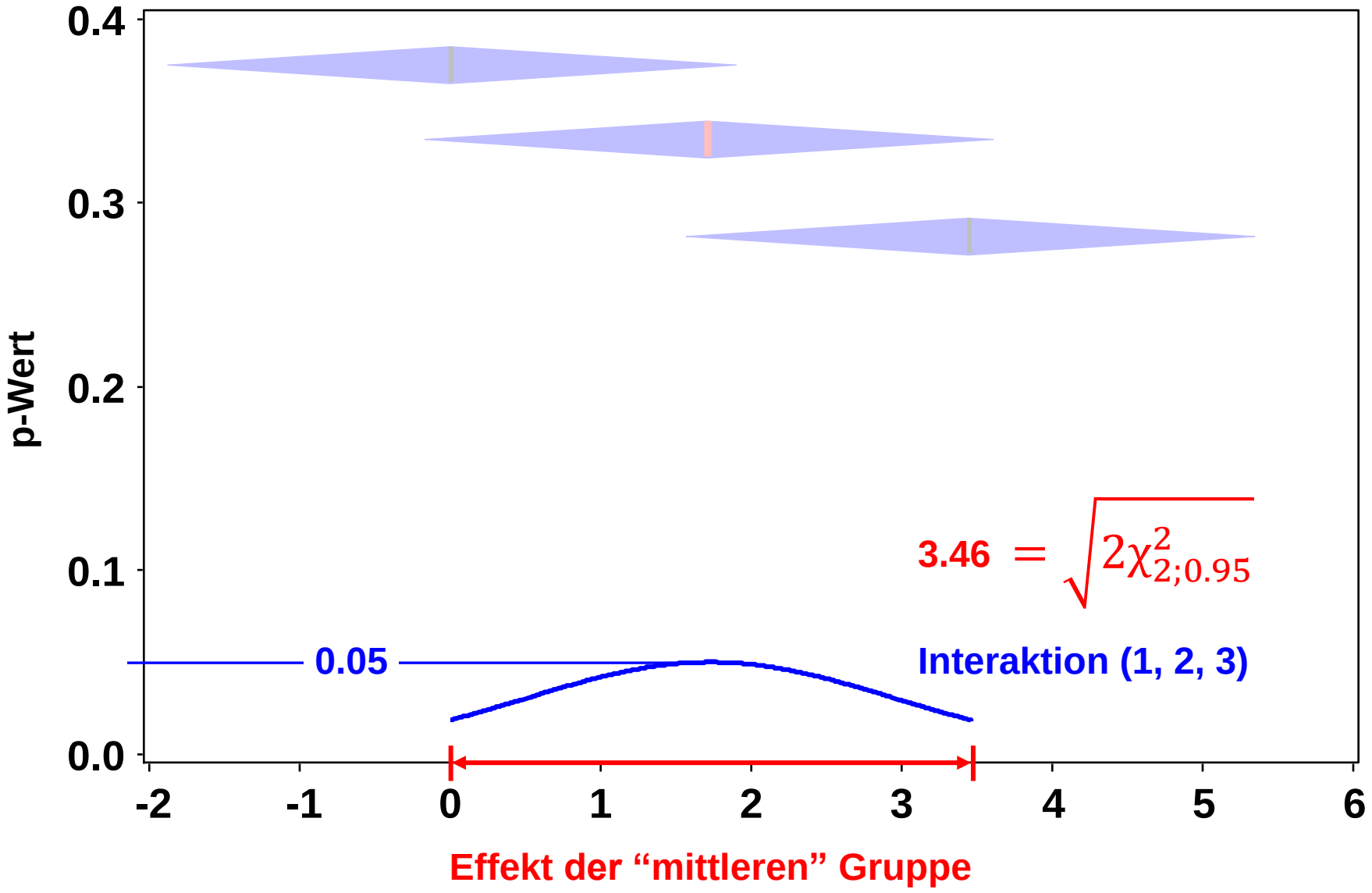
  → um wieviel ? α = 0.10 oder 0.20 oder 0.30 ?

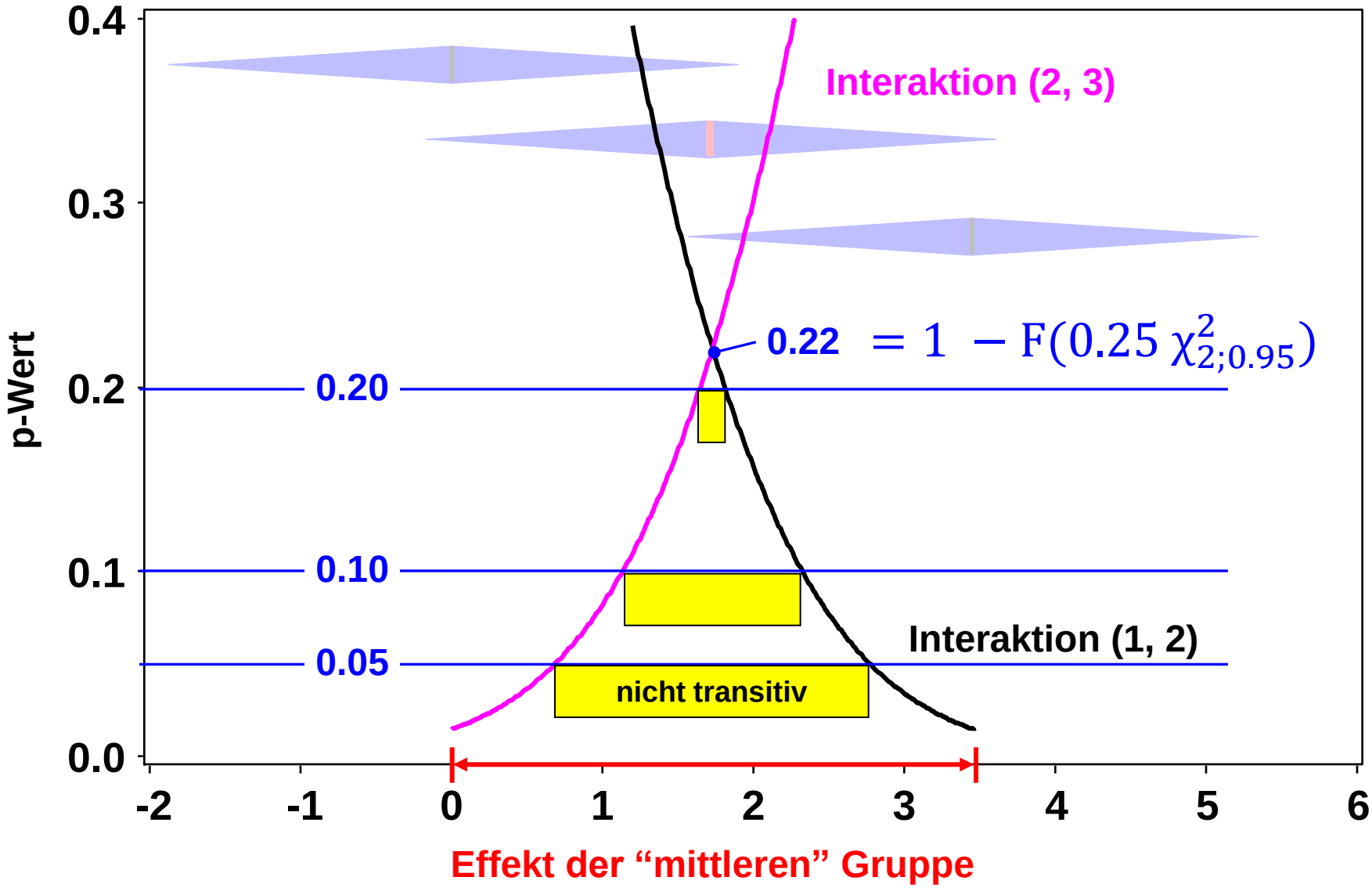
Szenario

Annahmen

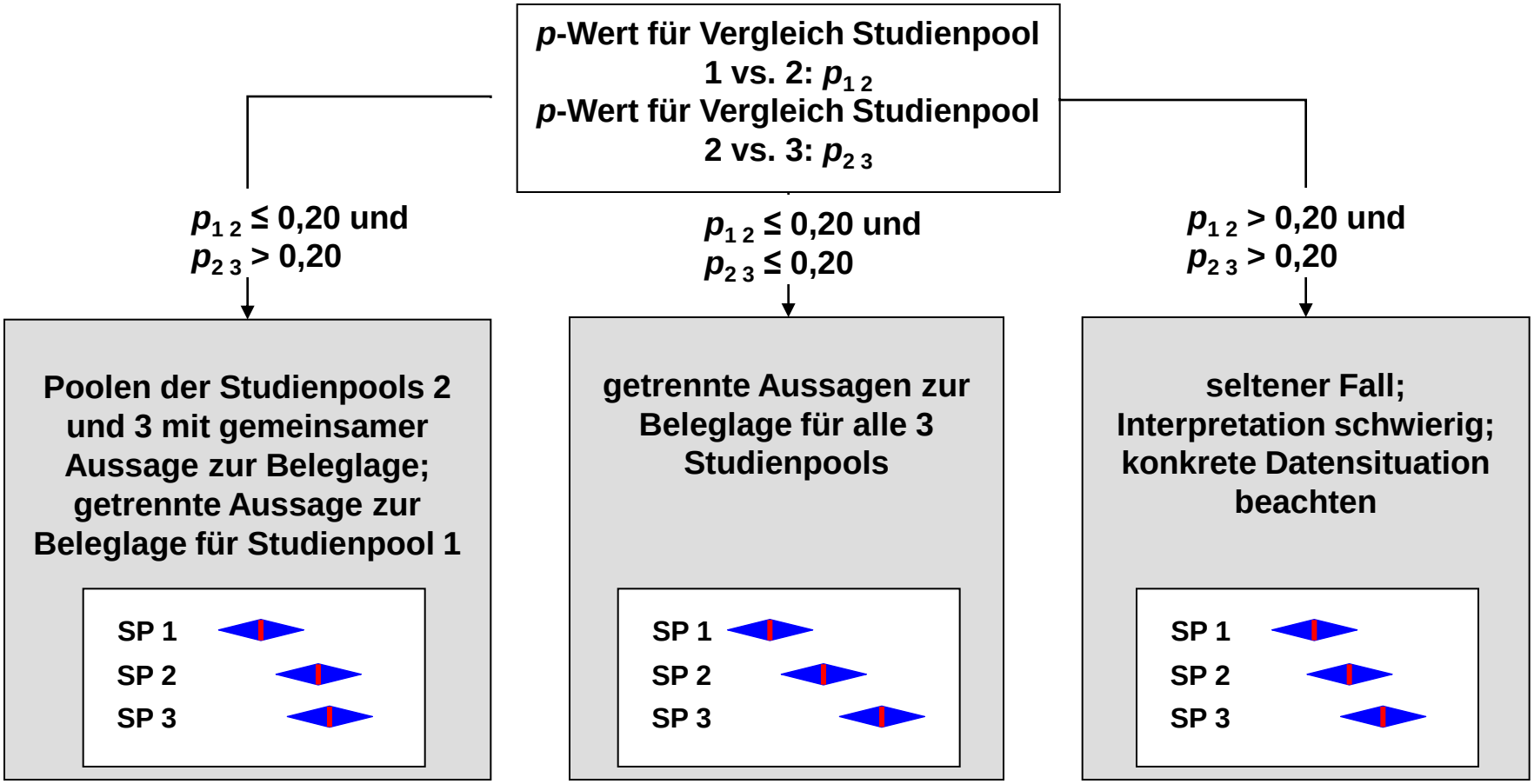
- Die Präzision der meta-analytischen Ergebnisse ist für alle 3 Studienpools (SP) gleich (SE = 1)
- Cochrans Q-Test




0.4
0.3
0.2
0.1
0.0
p-Wert
-2
-1
0
1
2
3
4
5
6
3.46 = $\sqrt{2\chi^2_{2;0.95}}$
0.05
Interaktion (1, 2, 3)
Effekt der “mittleren” Gruppe


0.4
0.3
0.2
0.1
0.0
p-Wert
Interaktion (2, 3)
0.22 = 1 − F(0.25 χ²2;0.95)
0.20
0.10
0.05
Interaktion (1, 2)
nicht transitiv
-2
-1
0
1
2
3
4
5
6
Effekt der “mittleren” Gruppe

Vorgehen bei 3 Subgruppenanalysen mit signifikanter Interaktion (α = 0,05)



→ Skipka & Bender 2010

Warum nicht (unter Annahme eines REM) immer poolen ?

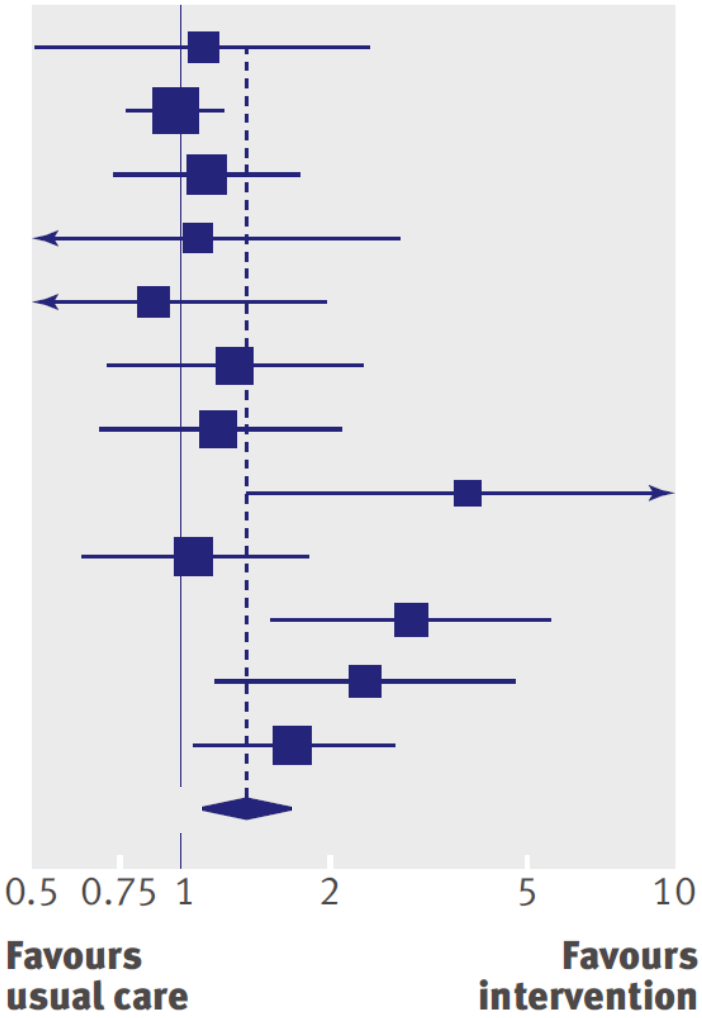
Interpretation of random effects meta-analyses

Richard D Riley,[1] Julian P T Higgins,[2] Jonathan J Deeks[1]

BMJ | 30 APRIL 2011 | VOLUME 342

How to calculate a prediction interval

$$\hat{\mu} - t_{k-2}\sqrt{\hat{\tau}^2 + \mathrm{SE}(\hat{\mu})^2}\ , \quad \hat{\mu} + t_{k-2}\sqrt{\hat{\tau}^2 + \mathrm{SE}(\hat{\mu})^2}$$

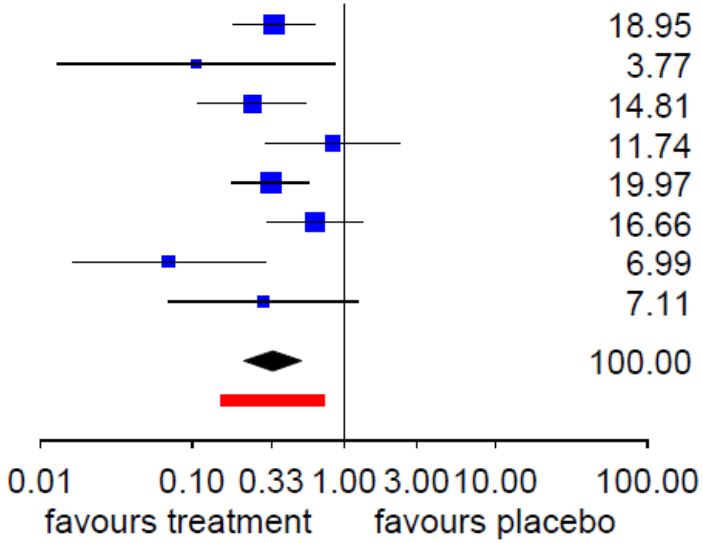

Study
Odds ratio (95% CI)
Weight (%)
Odds ratio (95% CI)
1 6.08 1.11 (0.51 to 2.39)
2 15.94 0.97 (0.78 to 1.21)
3 11.38 1.13 (0.73 to 1.72)
4 4.60 1.08 (0.42 to 2.75)
5 5.75 0.88 (0.39 to 1.95)
6 8.48 1.28 (0.71 to 2.30)
7 9.07 1.19 (0.69 to 2.08)
8 4.03 3.82 (1.37 to 10.60)
9 9.58 1.06 (0.63 to 1.79)
10 7.59 2.95 (1.54 to 5.63)
11 7.00 2.36 (1.18 to 4.72)
12 10.51 1.68 (1.05 to 2.70)
Overall (P=0.021, I²=51.2%) 100.00 1.36 (1.08 to 1.71)
with prediction interval (0.70 to 2.64)
0.5 0.75 1 2 5 10
Favours usual care
Favours intervention

Cochrane Colloquium 2006 (Poster)

Heterogenitätsinterval: $[\hat{\theta} - 1.96\hat{\tau};\ \hat{\theta} + 1.96\hat{\tau}]$

Outcome: cases of influenza
Random effects model (DerSimonian&Laird)

Study	Treatment n/N	Placebo n/N	OR 95% CI	Weight %	OR 95% CI
Oker-Blom 1970	16/141	41/152		18.95	0.35 [0.18, 0.65]
Muldoon 1976	1/53	8/52		3.77	0.11 [0.01, 0.88]
Monto 1979	8/136	28/139		14.81	0.25 [0.11, 0.57]
Kantor 1980	9/59	9/51		11.74	0.84 [0.31, 2.31]
Pettersson 1980	32/95	59/97		19.97	0.33 [0.18, 0.59]
Quarles 1981	15/107	20/99		16.66	0.64 [0.31, 1.34]
Dolin 1982	2/113	27/132		6.99	0.07 [0.02, 0.30]
Reuman 1989	3/317	5/159		7.11	0.29 [0.07, 1.25]
Total expectation (95% CI)	86/1021	197/881		100.00	0.34 [0.22, 0.53]
Total heterogeneity (95% CI)					[0.15, 0.74]



Heterogeneity: Q=12.44, df=7 (p=0.087), I^2=43.7%
Overall effect: Z Score=-4.84 (p=0.000), tau^2=0.160

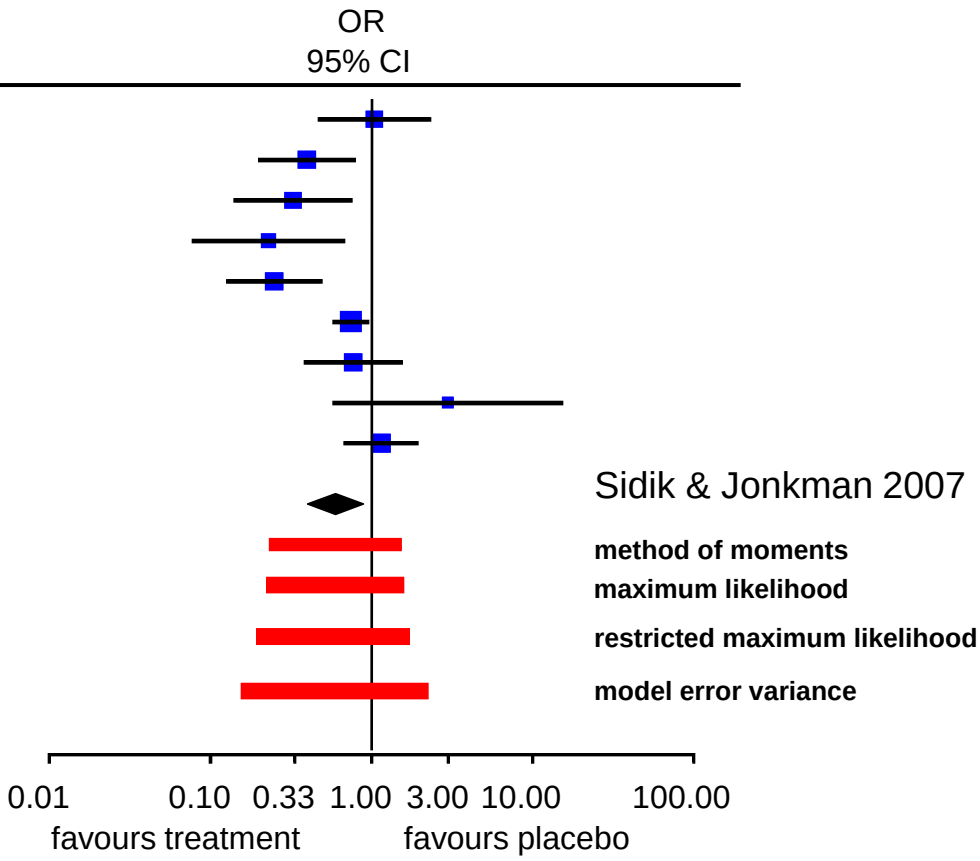
Schätzung der Varianz zwischen den Studien

Beispiel: Collins et al. 1985
(adaptiert von Viechtbauer 2007)

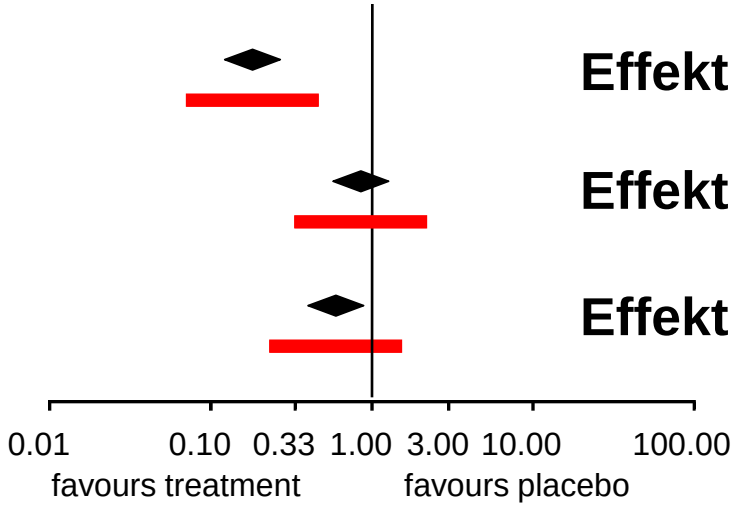
Outcome: cases of pre-eclampsia
Random effects model (DerSimonian&Laird)

Study	diuretics n/N	control n/N
1	14/131	14/136
2	21/385	17/134
3	14/57	24/48
4	6/38	18/40
5	12/1011	35/760
6	138/1370	175/1336
7	15/506	20/524
8	6/108	2/103
9	65/153	40/102
Total expectation (95% CI)	291/3759	345/3183
Total heterogeneity (95% CI)		

Heterogeneity: Q=27.27, df=8 (p=0.001), I^2=70.7%
Overall effect: Z Score=-2.54 (p=0.011), tau=0.48



Zusammenfassende Bewertung anhand des Heterogenitätsintervals



Effekt (variabler Stärke) liegt vor

Effekt liegt nicht vor

Effekt liegt nicht in allen Situationen vor

Zusammenfassung

- allgemein akzeptierte Verfahrensweise mit Heterogenität existiert (noch) nicht

- Umgang mit Heterogenität ist im IQWiG in weiten Teilen operationalisiert
  - festgelegte Schemata
  - fixe Kriterien zum regelhaften Vorgehen
  - Abweichungen begründet möglich

- größtes Problem: Güte der Heterogenitätsstatistiken bei wenigen Studien

Literatur

- Collins R, Yusuf S, Peto R. Overview of randomised trials of diuretics in pregnancy. *Br Med J (Clin Res Ed)* 1985;290(6461):17-23.
- Jackson D: The power of the standard test for the presence of heterogeneity in meta-analysis. *Stat Med* 2006;25(15):2688-2699.
- Riley RD, Higgins JP, Deeks JJ: Interpretation of random effects meta-analyses. Bmj 2011;342:964-967.
- Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med* 2007;26(9):1964-1981.
- Skipka G, Bender R: Intervention effects in the case of heterogeneity between three subgroups: assessment within the framework of systematic reviews. Methods Inf Med 2010;49(6):613-617.
- Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med* 2007;26(1):37-52.

Zum Glück gibt es Heterogenität !