



**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
- Heterogenitätsintervall

## 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

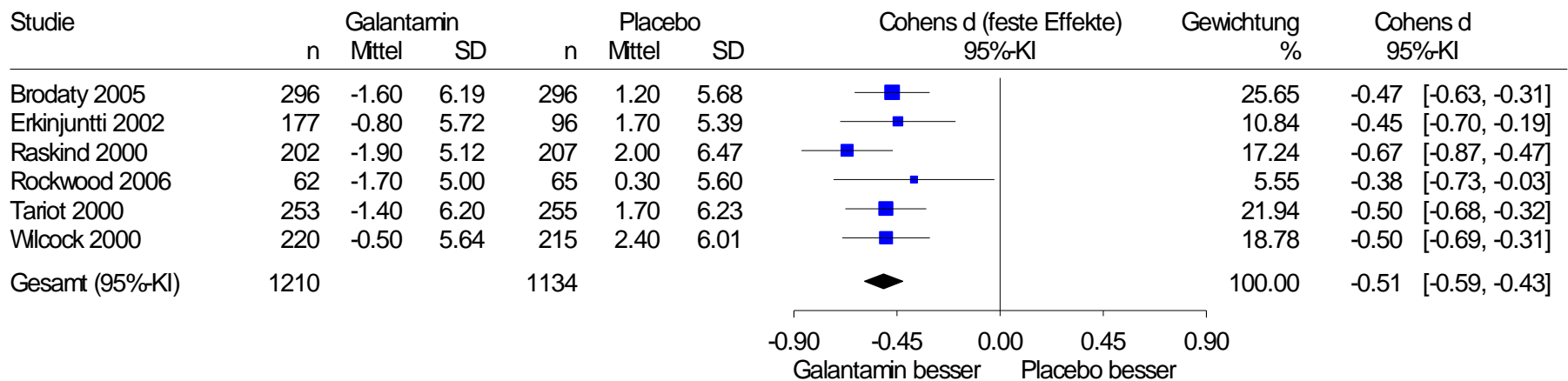
- übliche statistische Maßzahlen: Cochran's Q,  $I^2$

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

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

## Beispiel: ChE-Hemmer bei Alzheimer-Demenz

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



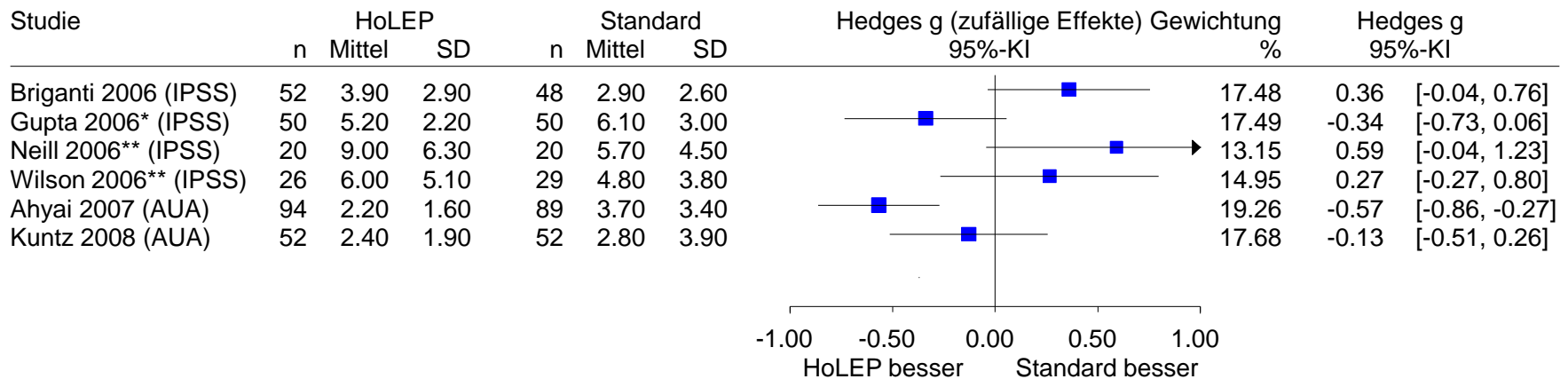
Heterogenität:  $Q=3.45$ ,  $df=5$  ( $p=0.630$ ),  $I^2=0\%$

Gesamteffekt: Z Score= $12.04$  ( $p=0.000$ )

Angaben zur Heterogenität

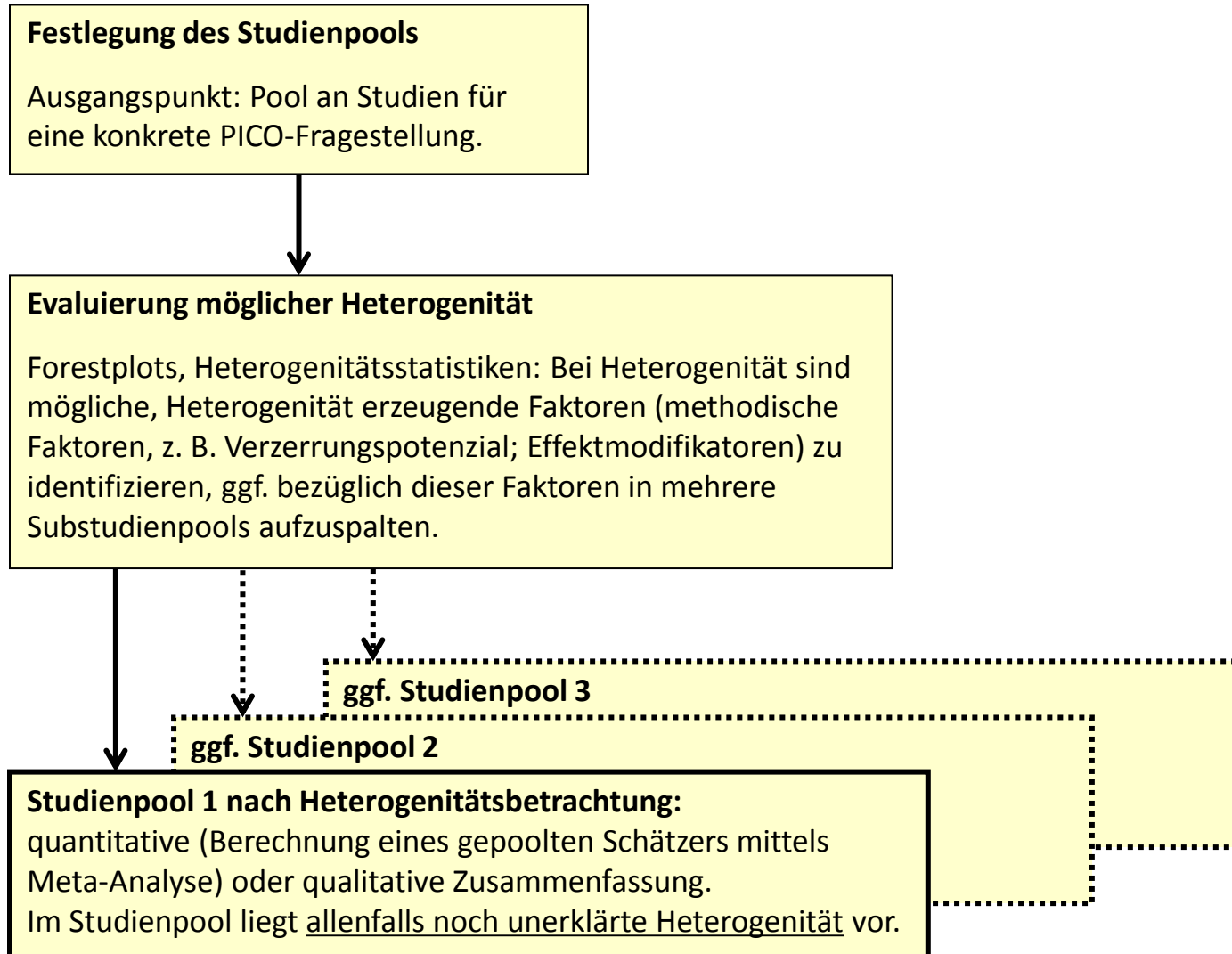
# Beispiel: Nichtmedikamentöse lokale Verfahren der Benignen Prostatahyperplasie

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



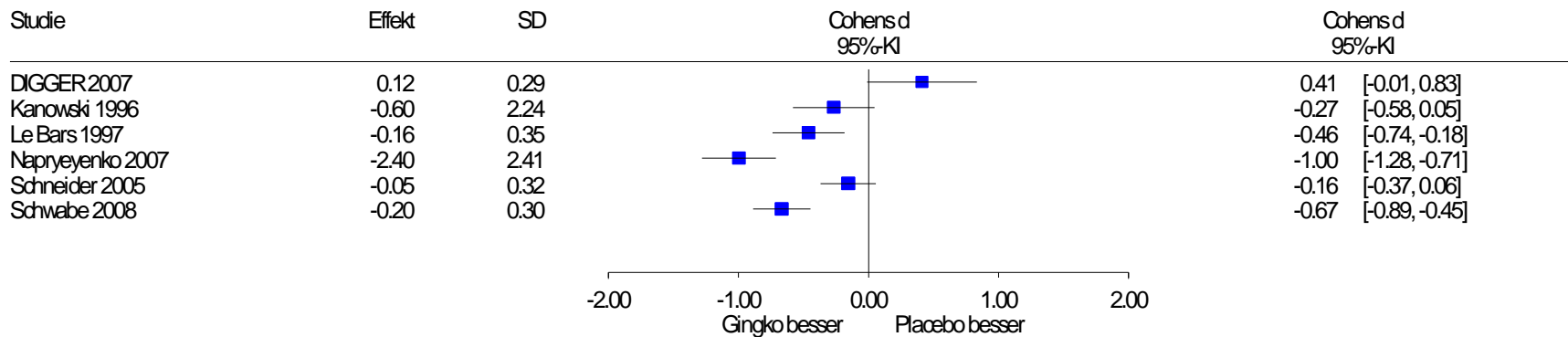
Heterogenität:  $Q=22.43$ ,  $df=5$  ( $p=0.000$ ),  $I^2=77.7\%$   
Gesamteffekt: Z Score=-0.06 ( $p=0.951$ ),  $\tau^2=0.154$

# Identifizierung zusammenfassbarer Studienpools



## Beispiel: Ginkgo Biloba

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

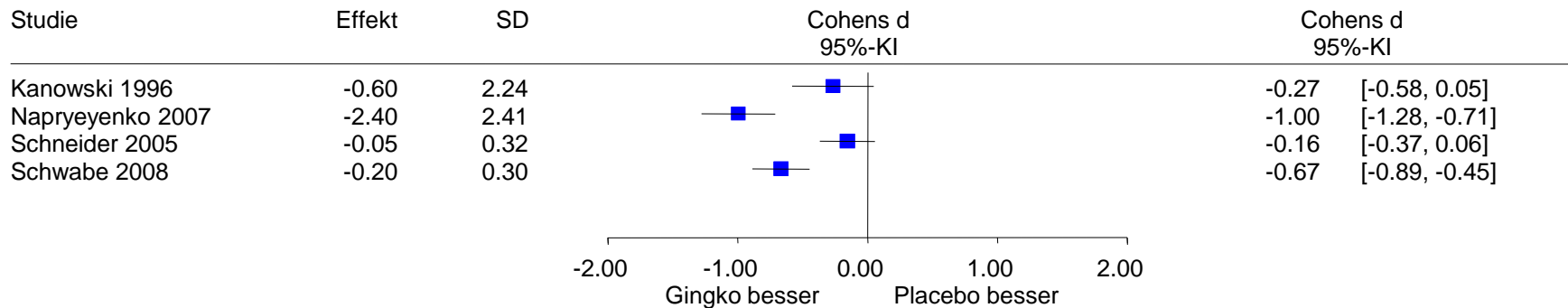


Heterogenität:  $Q=42.37$ ,  $df=5$  ( $p=0.000$ ),  $I^2=88.2\%$ ,  $\tau^2=0.144$

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

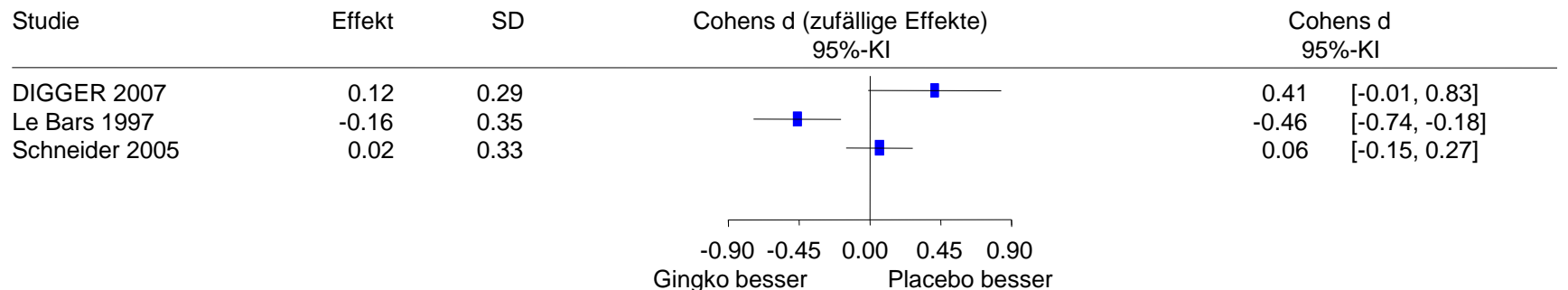


## hohe Dosierung



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

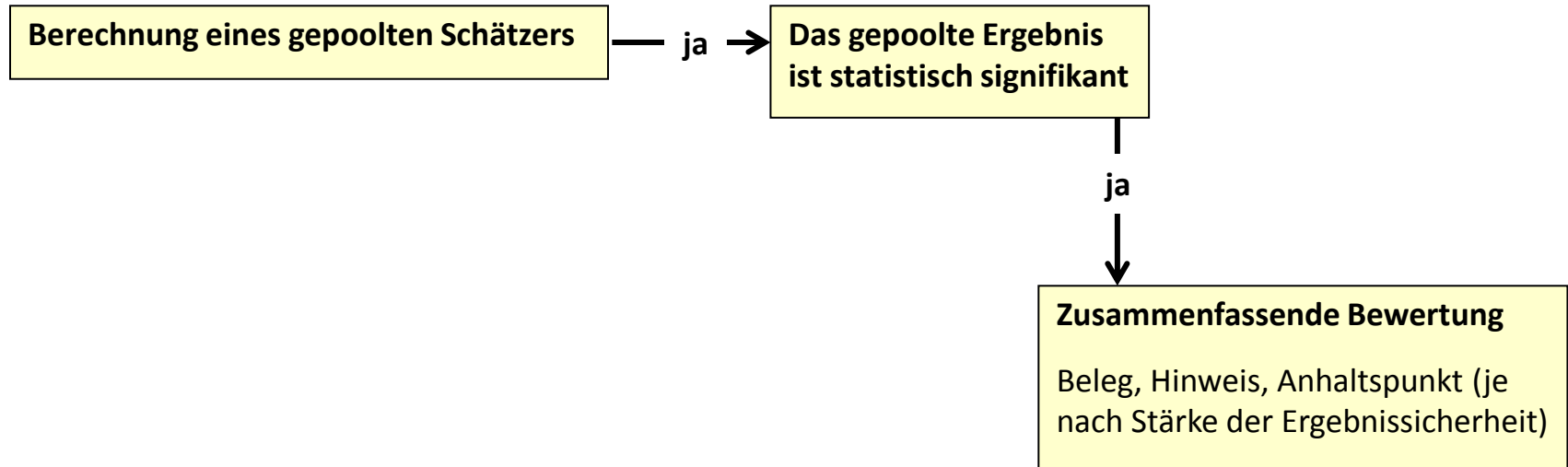
## niedrige Dosierung



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

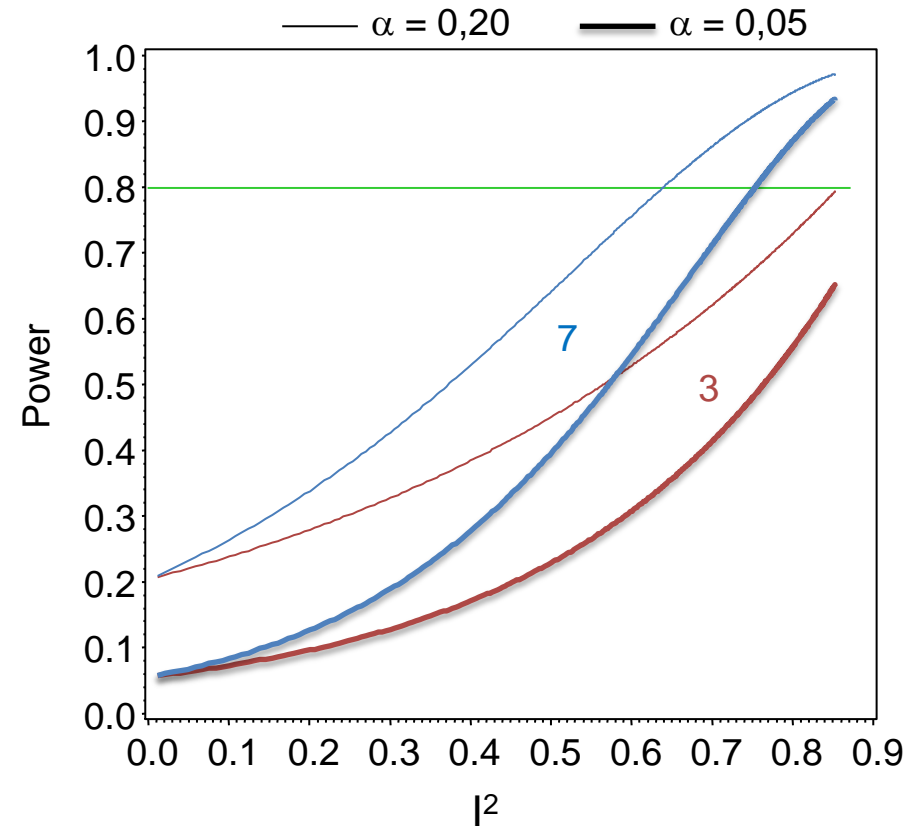
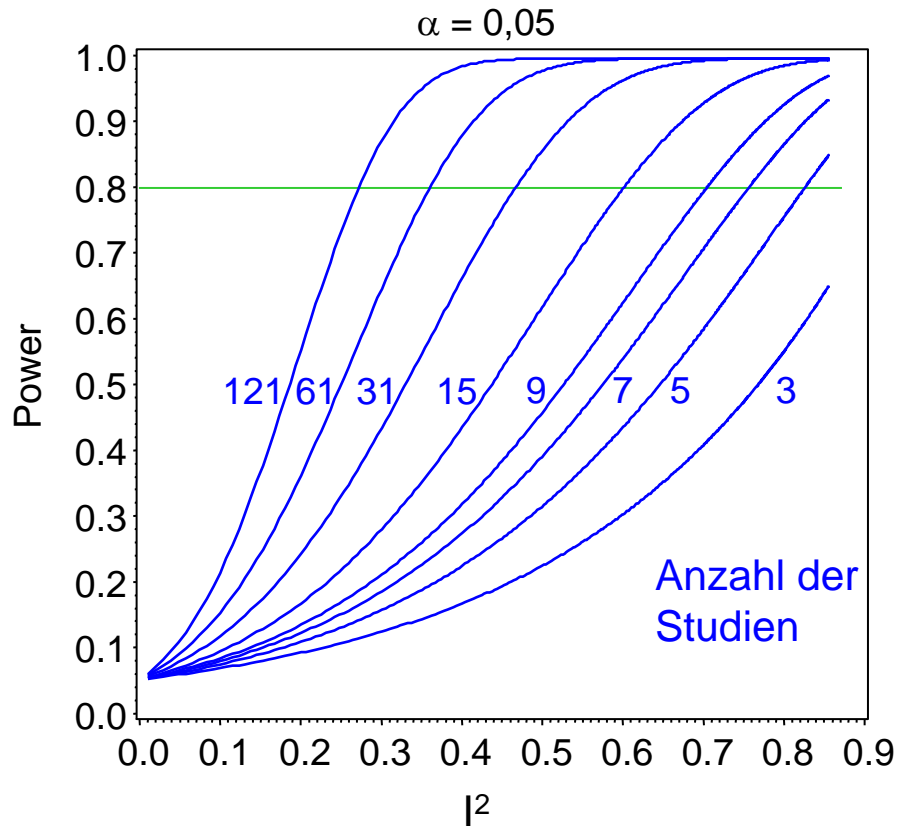


## Regelmäßige Formulierung in Berichtsplänen

- „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.“

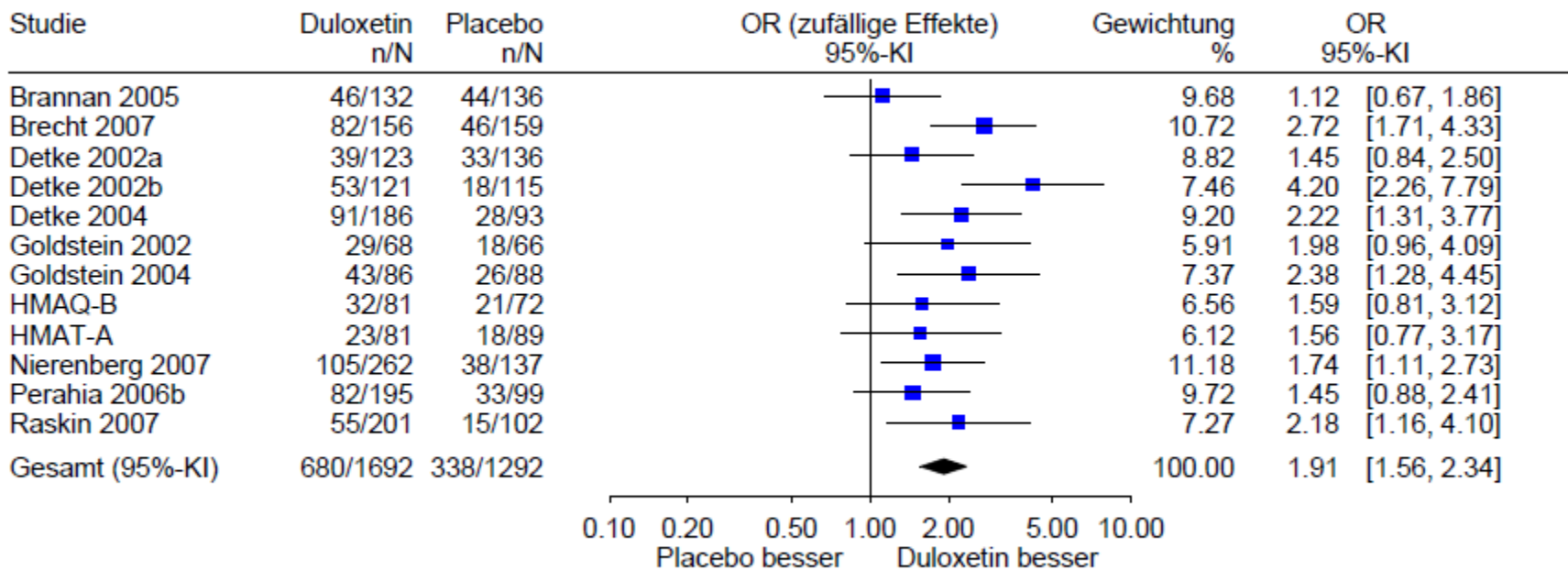
Grund: Anzahl der Studien in Meta-Analysen regelmäßig klein (2 - 5)  
→ sehr geringe Power des Heterogenitätstests

## Power des Heterogenitätstests (appr. Formel nach Jackson 2006)



## Beispiel: SNRI bei Depression

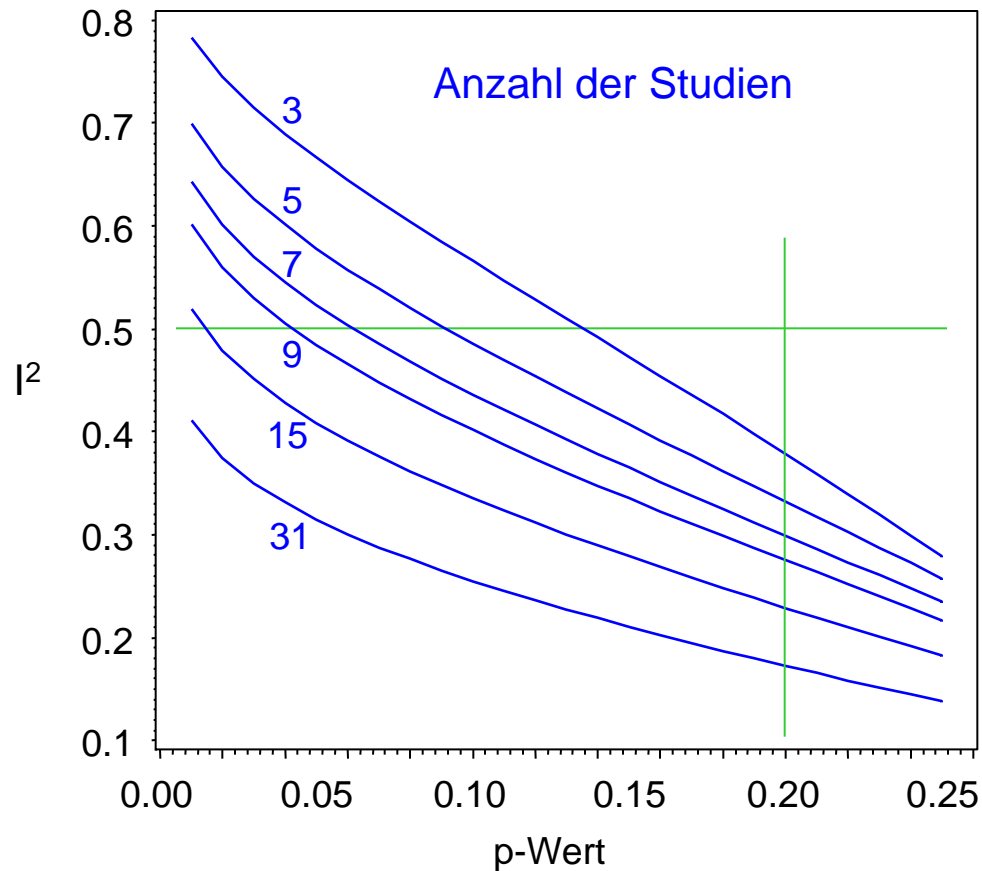
Remission  
Duloxetine vs. Placebo



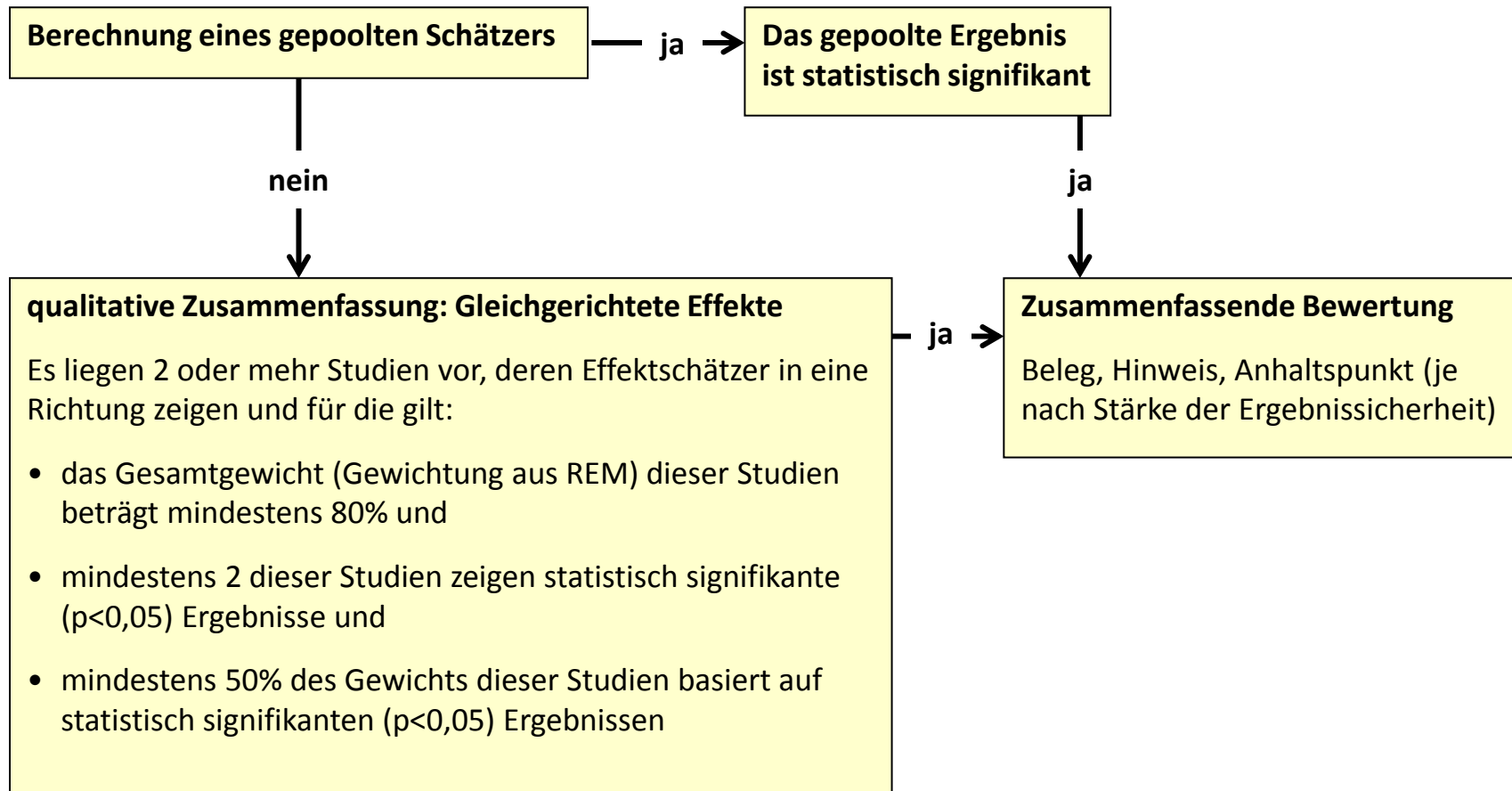
Heterogenität:  $Q=16.62$ ,  $df=11$  ( $p=0.120$ ),  $I^2=33.8\%$   
Gesamteffekt:  $Z$  Score= $6.31$  ( $p=0.000$ ),  $\tau^2=0.042$

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

## Zusammenhang zwischen $I^2$ und p-Wert



## Zusammenfassung der Ergebnisse pro Studienpool

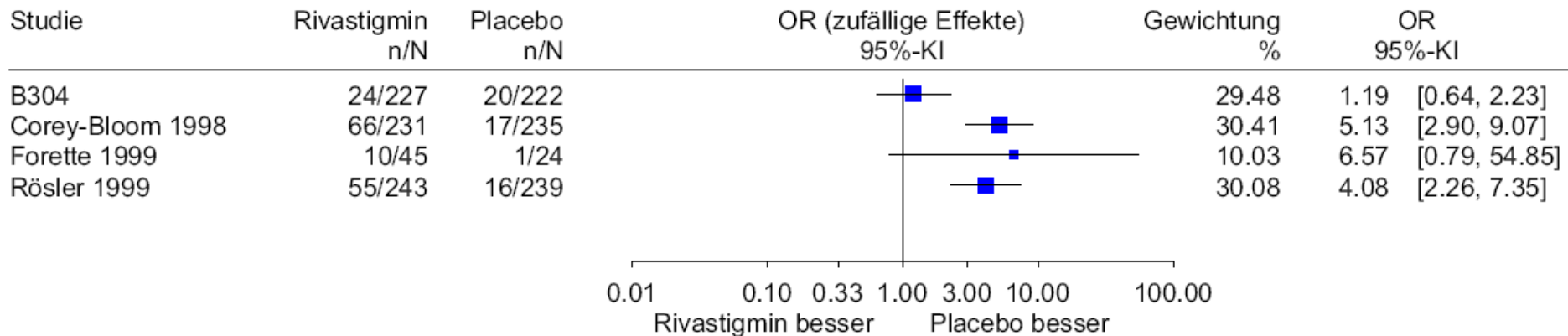


## Beispiel: ChE-Hemmer

Rivastigmin - Abbruch wegen UEs

Endpunkt: Studie aufgrund UE vorzeitig abgebrochen (ja, nein)

Distanzmaß: Odds Ratio der Anteile von Patienten mit Abbruch



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.

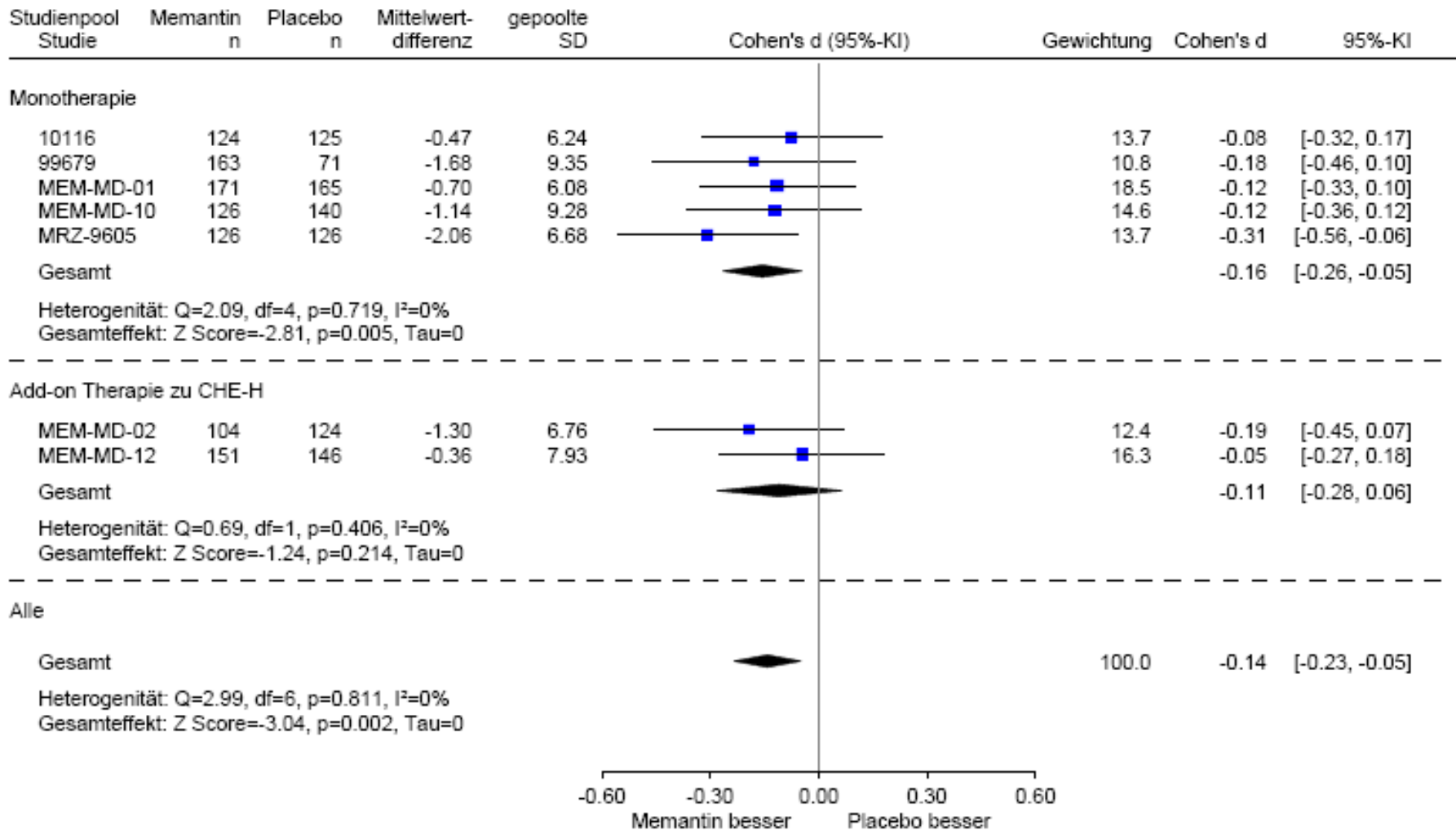
# Heterogene Ergebnisse zwischen Subgruppen

## Beispiel: Memantin

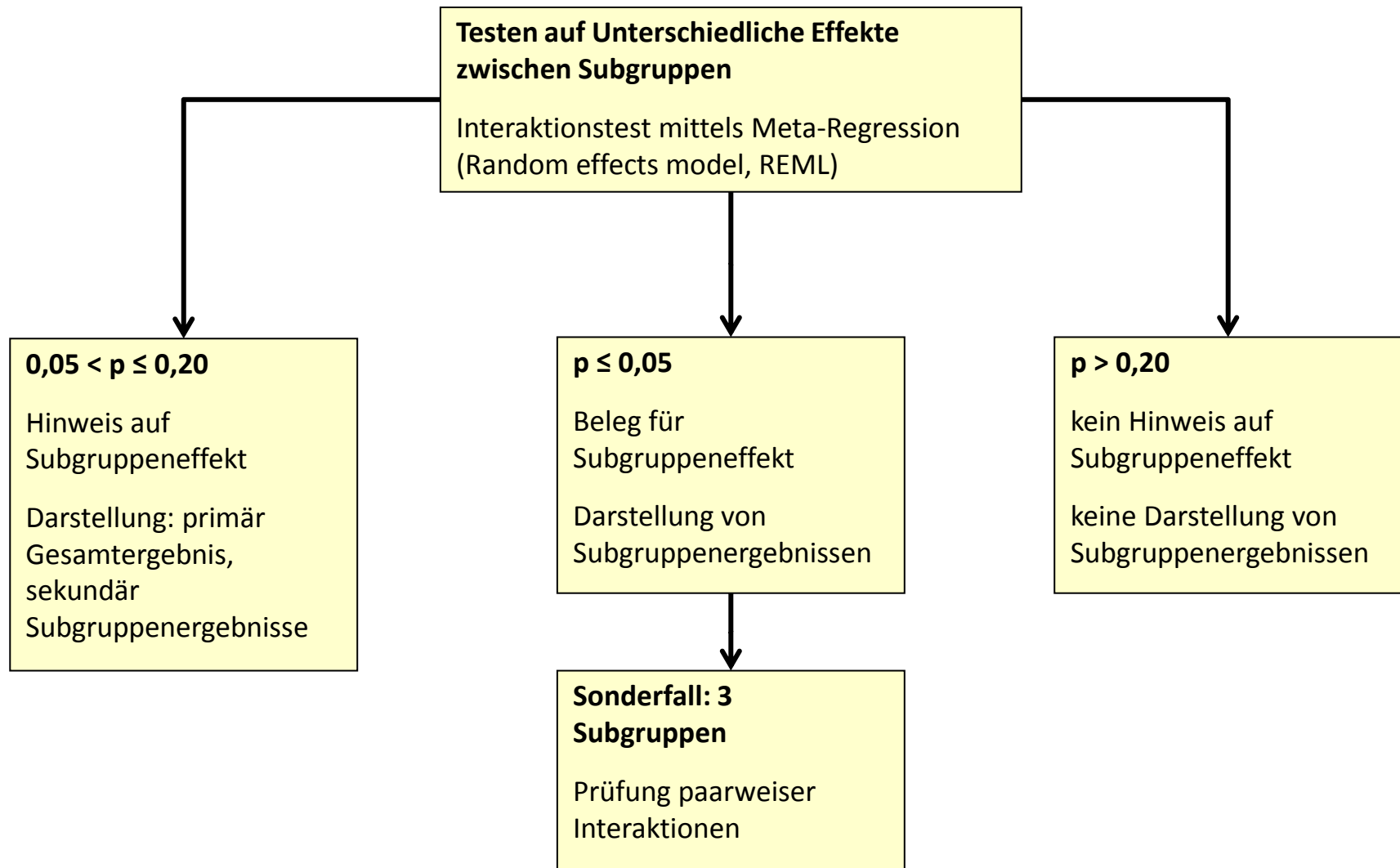
Memantin vs. Placebo

alltagspraktische Fähigkeiten: ADCS-ADL-23, ADCS-ADL-sev

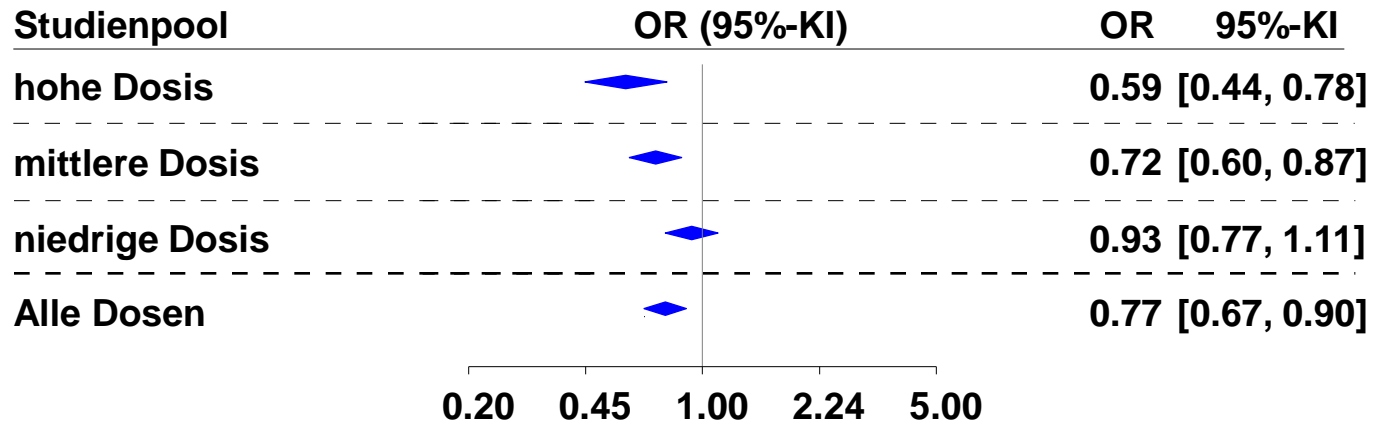
Modell mit zufälligen Effekten - DerSimonian und Laird



Heterogenität zwischen Subgruppen:  $Q=0.21$ ,  $df=1$ ,  $p=0.648$ ,  $I^2=0\%$



## fiktives Beispiel: Effektmodifikator → Dosierung

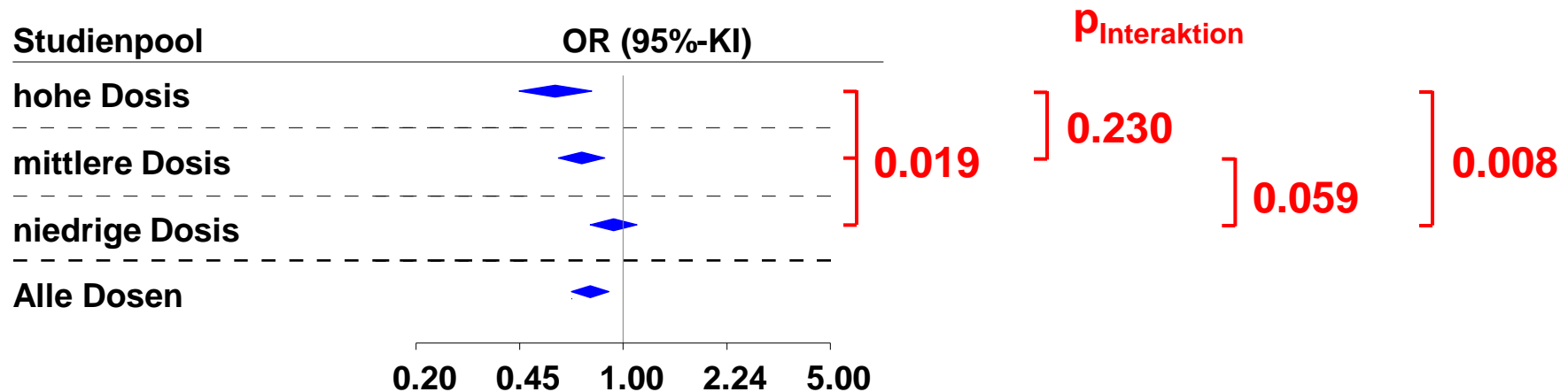


Heterogenität zwischen Studienpools:  $Q=7.88$ ,  $df=2$ ,  $p=0.019$ ,  $I^2=74.6\%$

### Situation

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

## paarweise Testung auf Interaktion

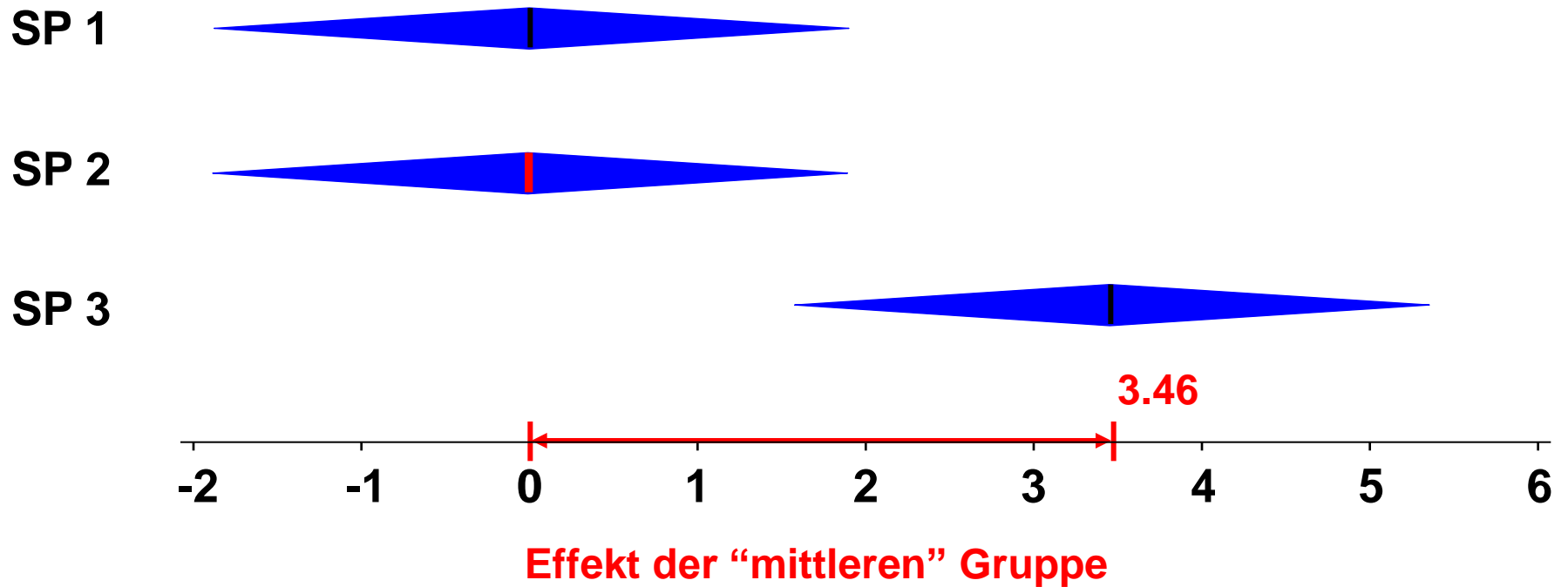


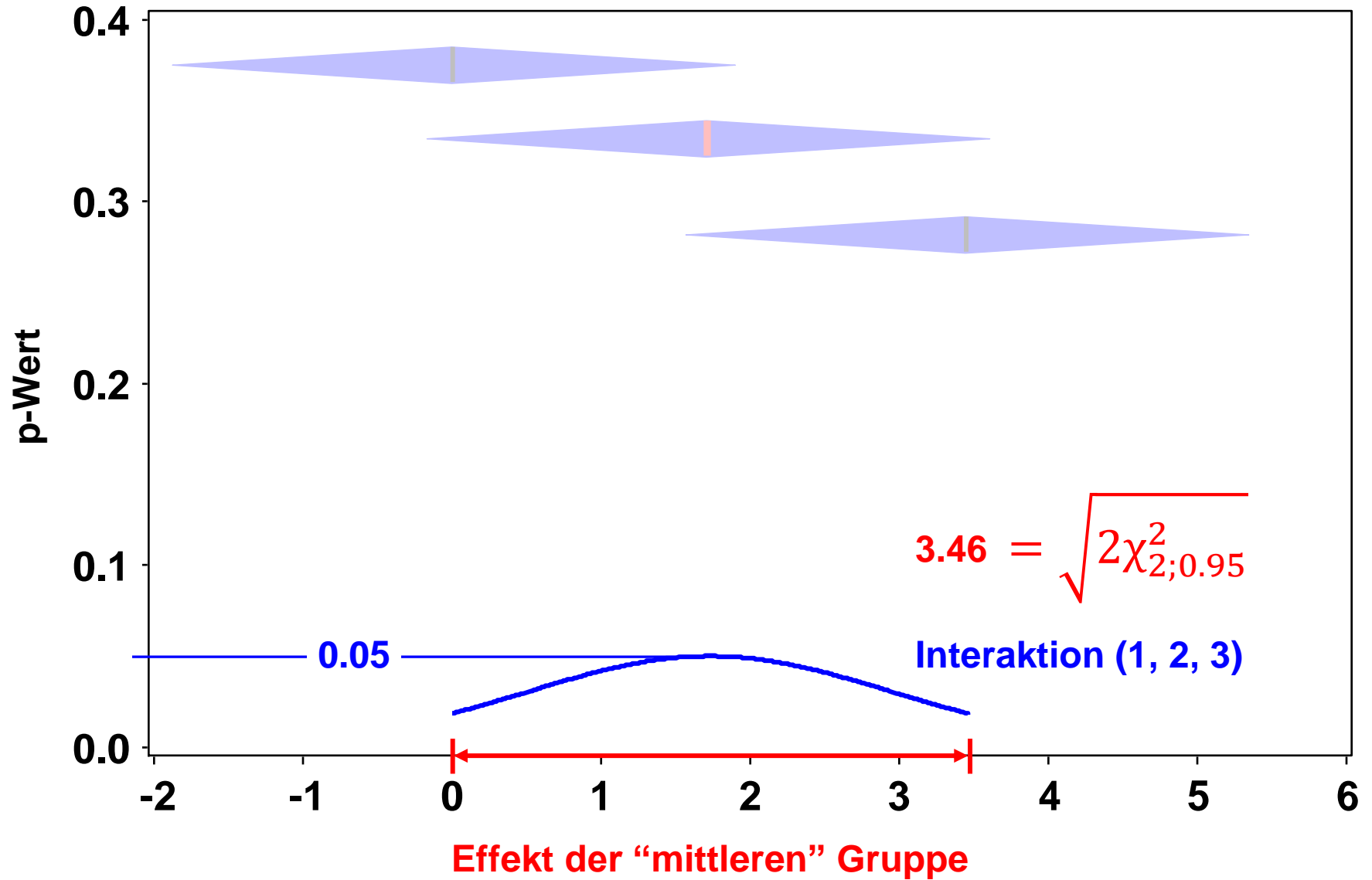
### häufiges Problem bei $\alpha = 0.05$ :

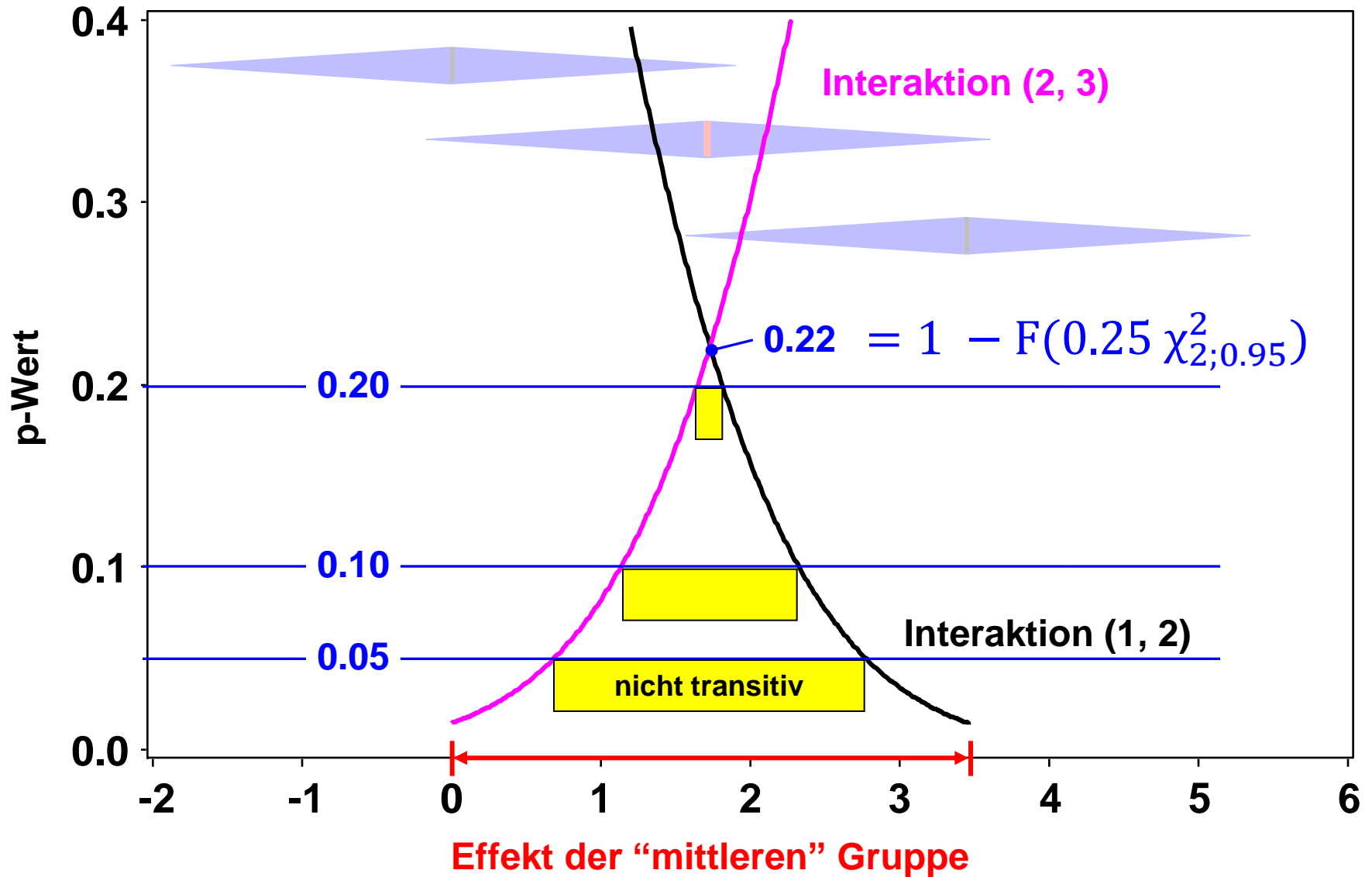
- Ein Unterschied zwischen hoher und niedriger Dosis ist belegt, nicht jedoch zwischen hoher und mittlerer sowie zwischen mittlerer und niedriger Dosis (nicht transitive Relation).
- Idee: Anhebung des Niveaus ( $\alpha > 0.05$ ) bei paarweisem Testen  
→ um wieviel ?  $\alpha = 0.10$  oder  $0.20$  oder  $0.30$  ?

## Annahmen

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

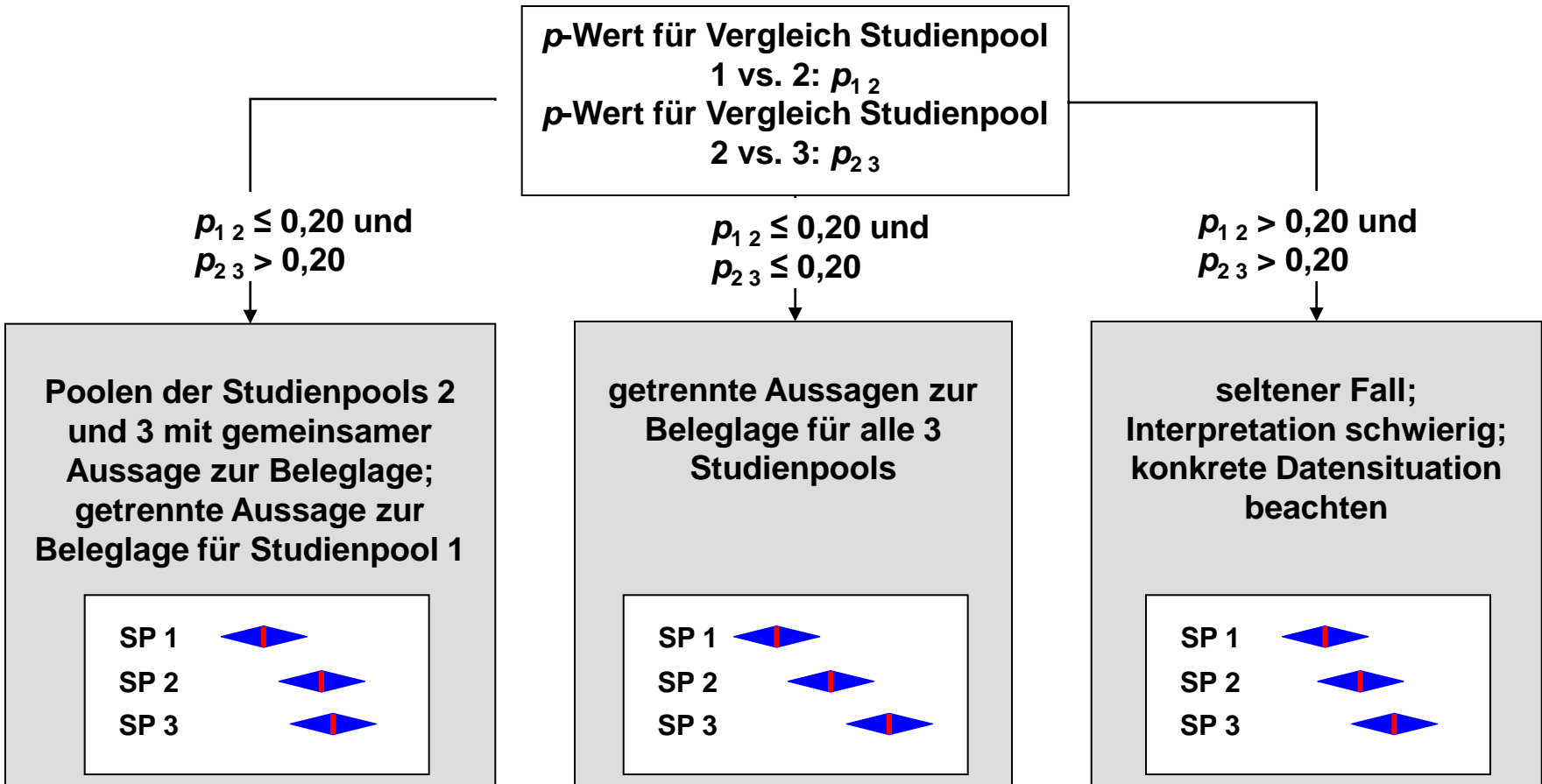








## Vorgehen bei 3 Subgruppenanalysen mit signifikanter Interaktion ( $\alpha = 0,05$ )



→ Skipka & Bender 2010

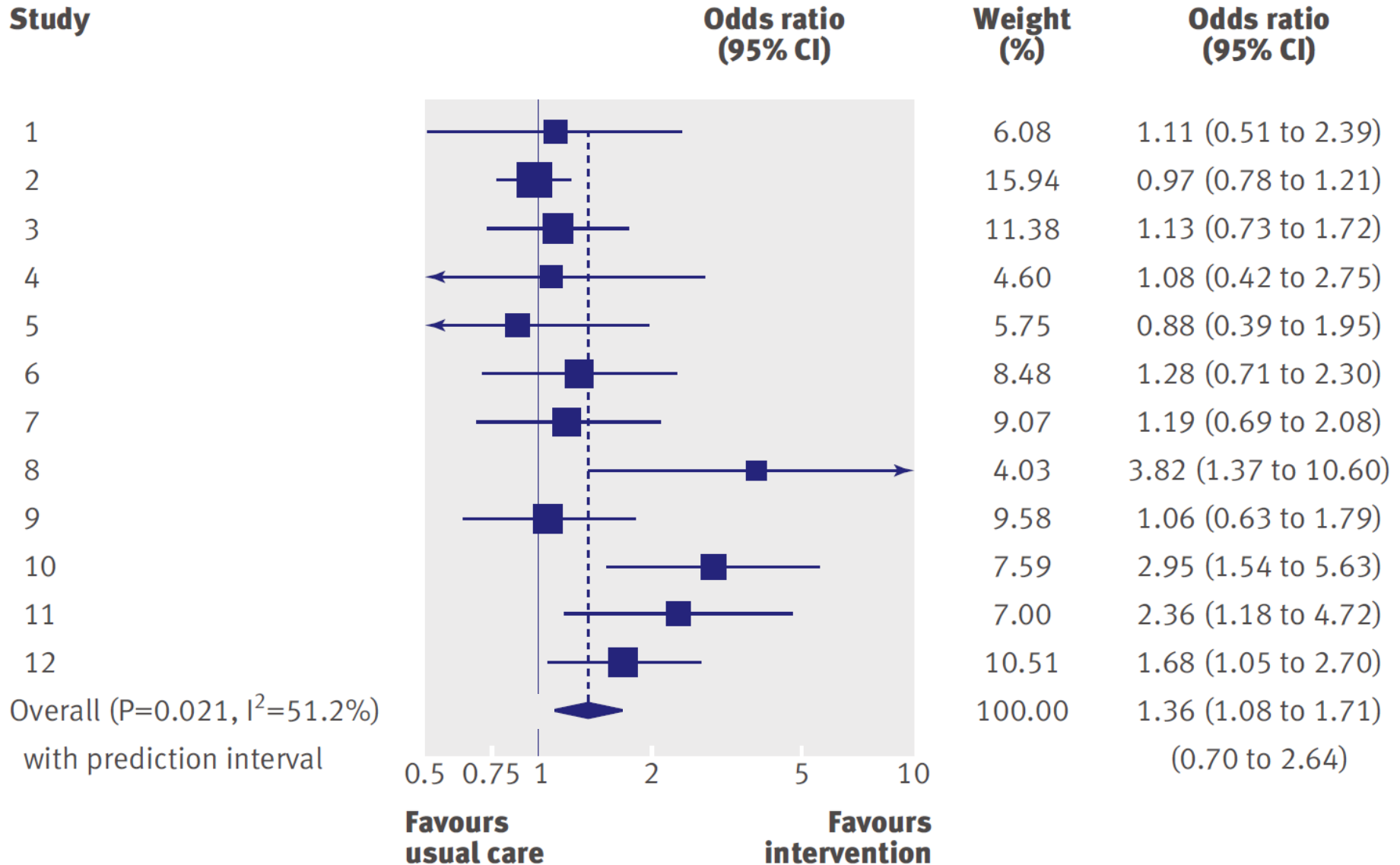
## Warum nicht (unter Annahme eines REM) immer poolen ?

### Interpretation of random effects meta-analyses

Richard D Riley,<sup>1</sup> Julian P T Higgins,<sup>2</sup> Jonathan J Deeks<sup>1</sup> **BMJ** | 30 APRIL 2011 | VOLUME 342

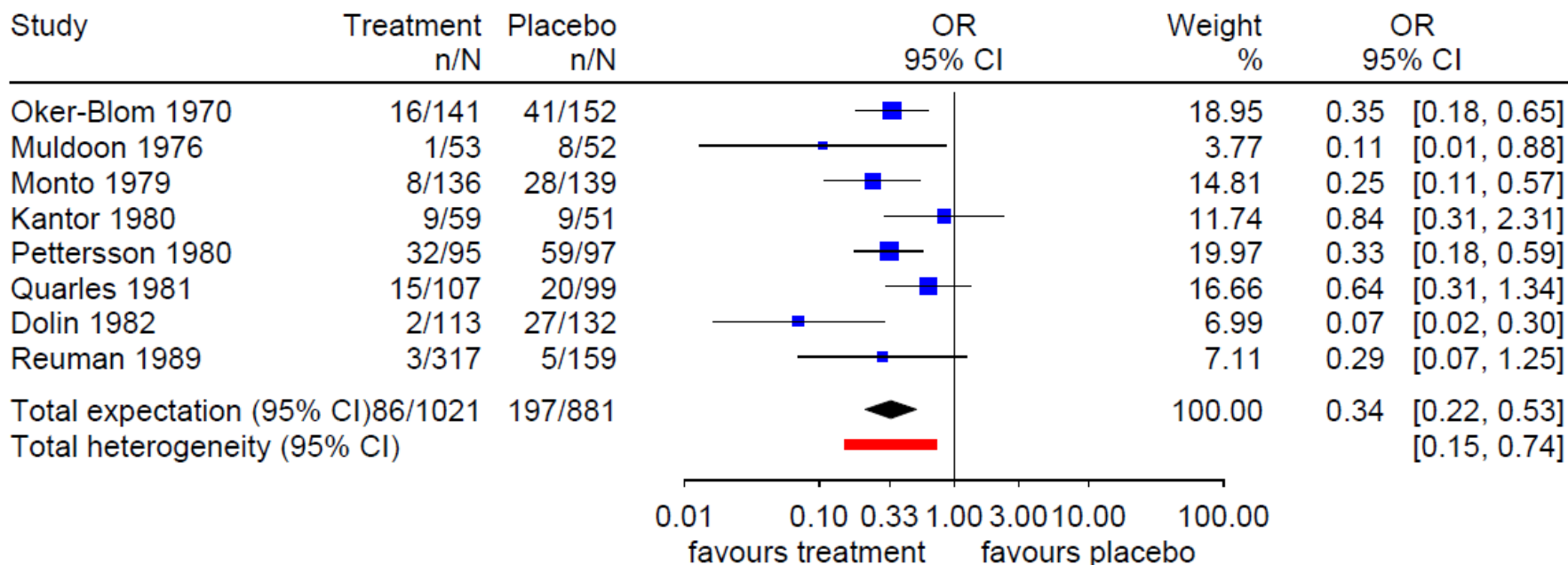
## How to calculate a prediction interval

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



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

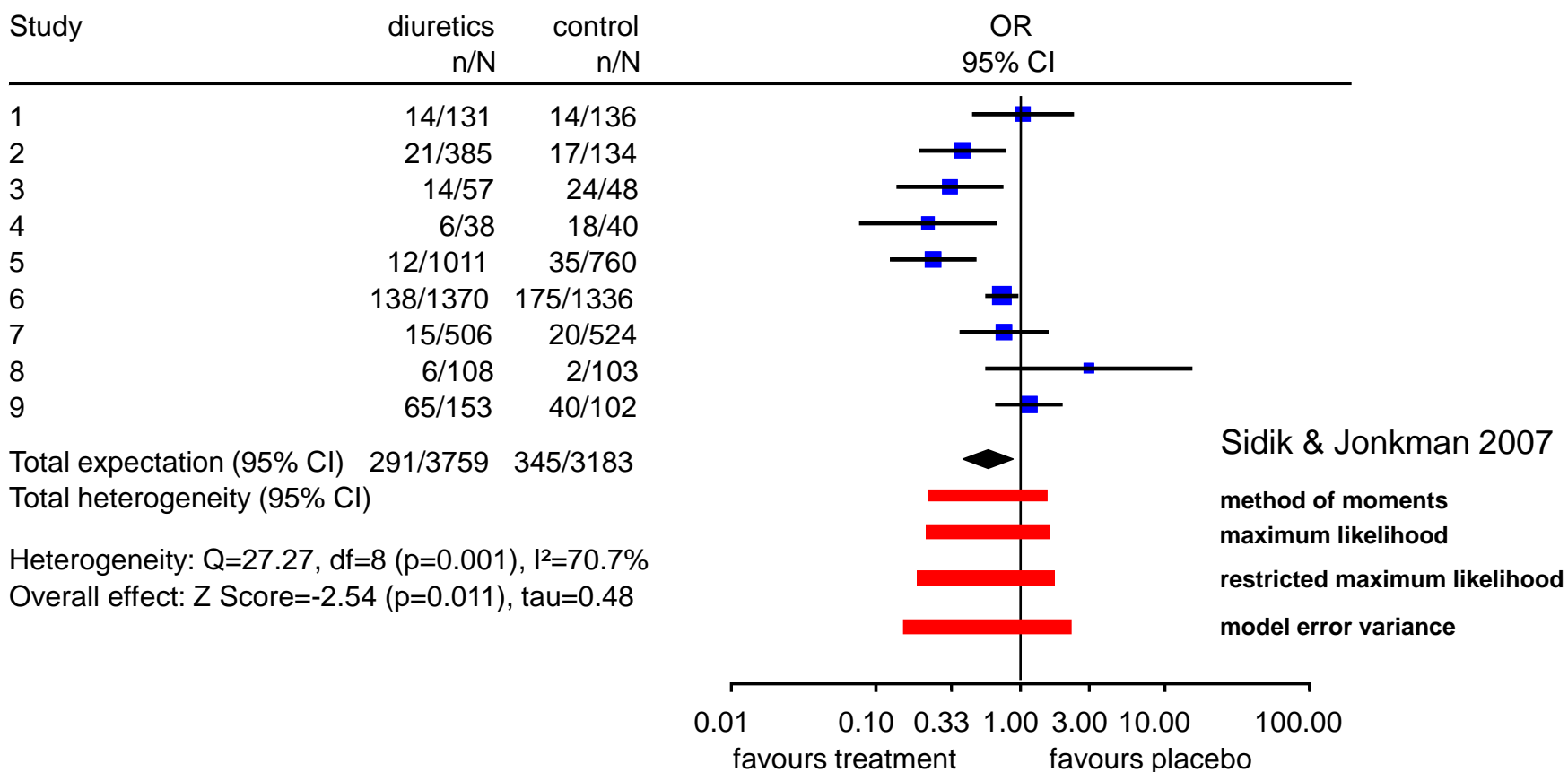
Outcome: cases of influenza  
Random effects model (DerSimonian&Laird)



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

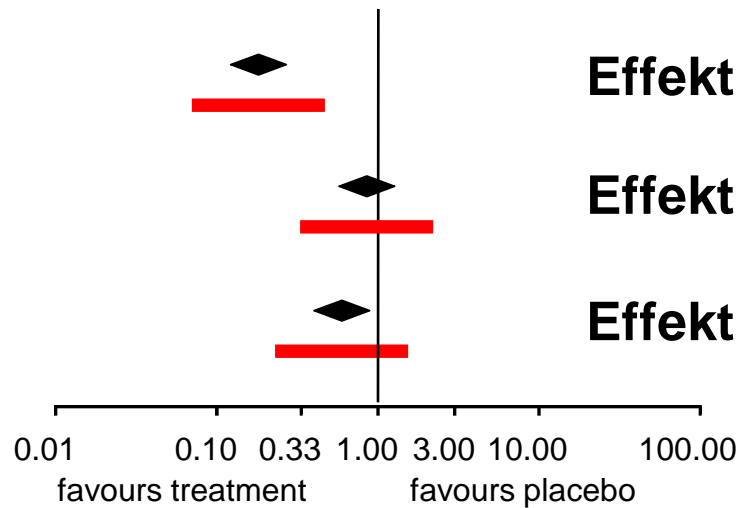
## Beispiel: Collins et al. 1985 (adaptiert von Viechtbauer 2007)

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



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

- 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 !**