

Heterogenität in Metaanalysen berichten und visualisieren

Dr. Waldemar Siemens
Institut für Evidenz in der Medizin, Universitätsklinikum Freiburg

Lunch & Learn
21.01.2026

Gefördert durch:



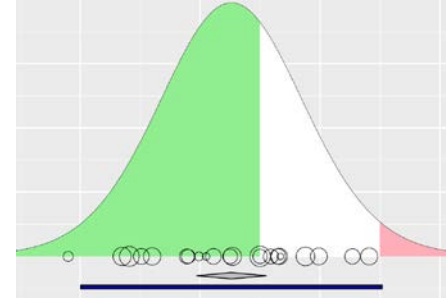
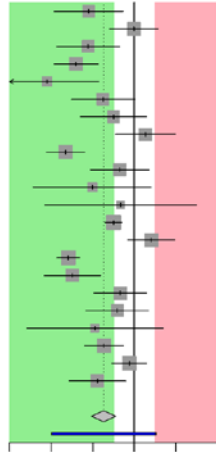
Bundesministerium
für Gesundheit

aufgrund eines Beschlusses
des Deutschen Bundestages

**Zuverlässige Evidenz.
Informierte Entscheidungen.
Bessere Gesundheit.**



Heterogenität in Metaanalysen berichten und visualisieren



Interessenskonflikt

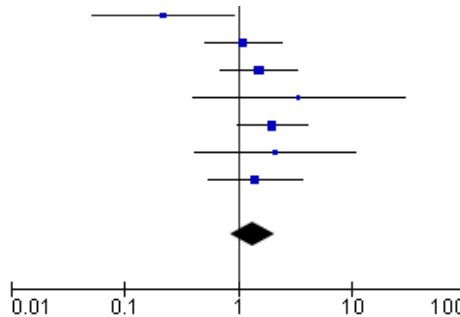
- Wissenschaftlicher Mitarbeiter und Statistiker am Institut für Evidenz in der Medizin (für die Cochrane Deutschland Stiftung), Universitätsklinikum Freiburg
- Sprecher des Fachbereichs Methodik des Netzwerks Evidenzbasierte Medizin
- Außerordentliches Mitglied der Arzneimittelkommission der Deutschen Ärzteschaft

Übersicht

1. **Was ist Heterogenität?**
2. 95% Prädiktionsintervall
3. Erwartete Anteile von Studien mit relevantem Nutzen oder Schaden
4. Zusammenfassung

Was ist Heterogenität?

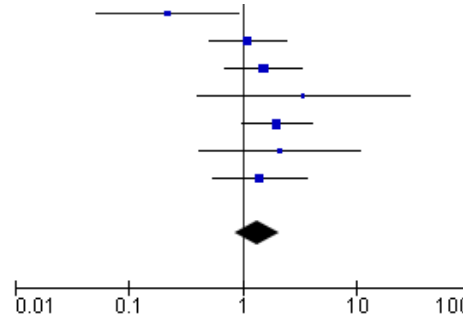
- Primärstudie: Variation der Werte zwischen Personen?
- Metaanalyse: Variation der Effekte zwischen Studien?



Heterogenität in Metaanalysen

In Metaanalysen kann man verschiedene **Arten von Heterogenität** unterscheiden, zum Beispiel

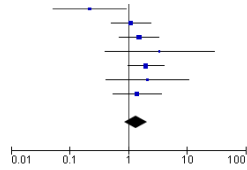
- klinische Heterogenität
 - methodische Heterogenität
 - zufällige Schwankungen
- } Messbar als (statistische) Heterogenität



Einflussfaktoren auf den Behandlungseffekt vs. eigene Fragestellung



- Population: Darf ich Kinder und Erwachsene in der Metaanalyse zusammenfassen?
- Interventionstyp: Darf ich pharmakologische und nicht-pharmakologische Interventionen poolen?
- Vergleichsgruppe: Darf ich aktive und passive Vergleichsgruppen in der selben Metaanalyse als Vergleich wählen?
- Outcome: Darf ich krankheitsspezifische und -unspezifische Lebensqualitätsfragebögen poolen?
- Studiendesign: Darf ich randomisierte und nicht-randomisierte Studiendesigns in der Metaanalyse kombinieren?
- Bias: Darf ich Studien mit niedrigem und sehr hohem Risiko für Bias in der Metaanalyse zusammenfassen?
- ...



Klassifikation: Klinische und/oder methodische vs. statistische Heterogenität

Klinische und/oder methodische Heterogenität

UND

statistische Heterogenität

Übersicht

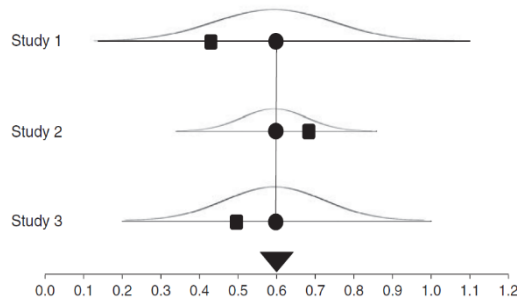
1. Was ist Heterogenität?
- 2. 95% Prädiktionsintervall**
3. Erwartete Anteile von Studien mit relevantem Nutzen oder Schaden
4. Zusammenfassung

Common-effect vs random-effects model: Annahmen

Common-effect model:

- Allen Studien liegt der gleiche „wahre“ Therapieeffekt zugrunde
- Unterschiede in den geschätzten Therapieeffekten sind auf die Zufallsstreuung zurückzuführen

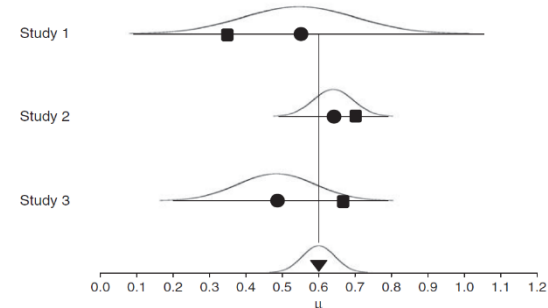
$$\text{Gewicht z.B. } w_i = \frac{1}{SE(x_i)^2} \text{ (Intrastudienvarianz)}$$



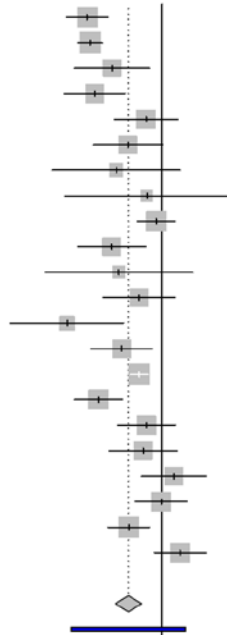
Random-effects model:

- Unterschiedliche Effekte für jede Studie (Normalverteilungsannahme)
- Kleinere Studien bekommen mehr Gewicht als beim common-effect model
- Weitere 95% KIs als beim common-effect model

$$\text{Gewicht } w_i = \frac{1}{SE(x_i)^2 + \tau^2} \text{ (Intrastudienvarianz + Interstudienvarianz } \tau^2 \text{)}$$



95% Prädiktionsintervall im random-effects model



95% Wahrscheinlichkeitsbereich für (wahren)
zukünftigen Effekt einer vergleichbaren Studie!

$$PI = \hat{\mu}_{RE} \pm t_{K-1,0.975} \sqrt{\left\{ \hat{\tau}^2 + SE(\hat{\mu}_{RE})^2 \right\}}$$

In the random-effects model, the pooled estimate is represented as $\hat{\mu}_{RE}$. Here, $t_{K-1,0.975}$ denotes the 97.5% quantile of the t-distribution with $K-1$ degrees of freedom resulting in a 95% PI, where K is the number of studies. $\hat{\tau}^2$ represents the estimated between-study variance, and $SE(\hat{\mu}_{RE})^2$ is the estimated variance of the pooled estimate in the random-effects meta-analysis.

Beispiel



Cochrane Database of Systematic Reviews

Exercise-based cardiac rehabilitation for adults with heart failure (Review)

Molloy C, Long L, Mordi IR, Bridges C, Sagar VA, Davies EJ, Coats AJS, Dalal H, Rees K, Singh SJ, Taylor RS

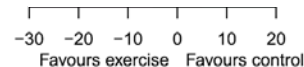
Beispiel: Statistische Heterogenität

Study

Belardinelli 1999
McKelvie 2002
Koukouvou 2004
Austin 2005
Passino 2006
Dracup 2007
Nilsson 2008
Jolly 2009
Andryukhin 2010
Davidson 2010
Gary 2010 (1)
Gary 2010 (2)
Yeh 2011
Witham 2012
Antonicelli 2016
Chen 2018
Dalal 2018
Du 2018
Lang 2018
Peng 2018
Jaarsma 2020
Wang 2021

Random effects model

Heterogeneity:



MLWHF: Minnesota Living with Heart Failure

Molloy et al., 2023; Borenstein, 2024
Gonzalez-Saenz de Tejada M et al. 2019

- **MLWHF** Fragebogen: Bereich 0 bis 105; große Werte: schlechte Lebensqualität (MCID: ± 5)
- **p-Wert** des χ^2 Tests für Heterogenität: $p < 0,0001$
 H_1 : Wahrer Effekt ist nicht gleich in allen Studien
- **τ^2 : Zwischenstudienvarianz** der Studieneffektschätzer: 34,3 Punkte im Quadrat
- **τ : Standardabweichung** der Studieneffektschätzer: 5,9 Punkte
- **I^2** : Prozentzahl der Variabilität aufgrund von wahrer Heterogenität statt zufälliger Unterschiede der Studieneffekte (Verhältnis: Varianz wahrer Effekte zur Gesamtvarianz): **82%**: substantiell/beträchtlich

95% Prädiktionsintervall:

95% Wahrscheinlichkeitsbereich eines zukünftigen Effekts einer vergleichbaren Studie: **-19,98 bis 5,21** Punkte; schneidet MCID=5 auf beiden Seiten.

Übersicht

1. Was ist Heterogenität?
2. 95% Prädiktionsintervall
- 3. Erwartete Anteile von Studien mit relevantem Nutzen oder Schaden**
4. Zusammenfassung

Erwartete Anteile von Studien mit relevantem Nutzen oder Schaden

Siemens et al. *BMC Medical Research Methodology* (2025) 25:275
<https://doi.org/10.1186/s12874-025-02733-9>

BMC Medical Research
Methodology

RESEARCH

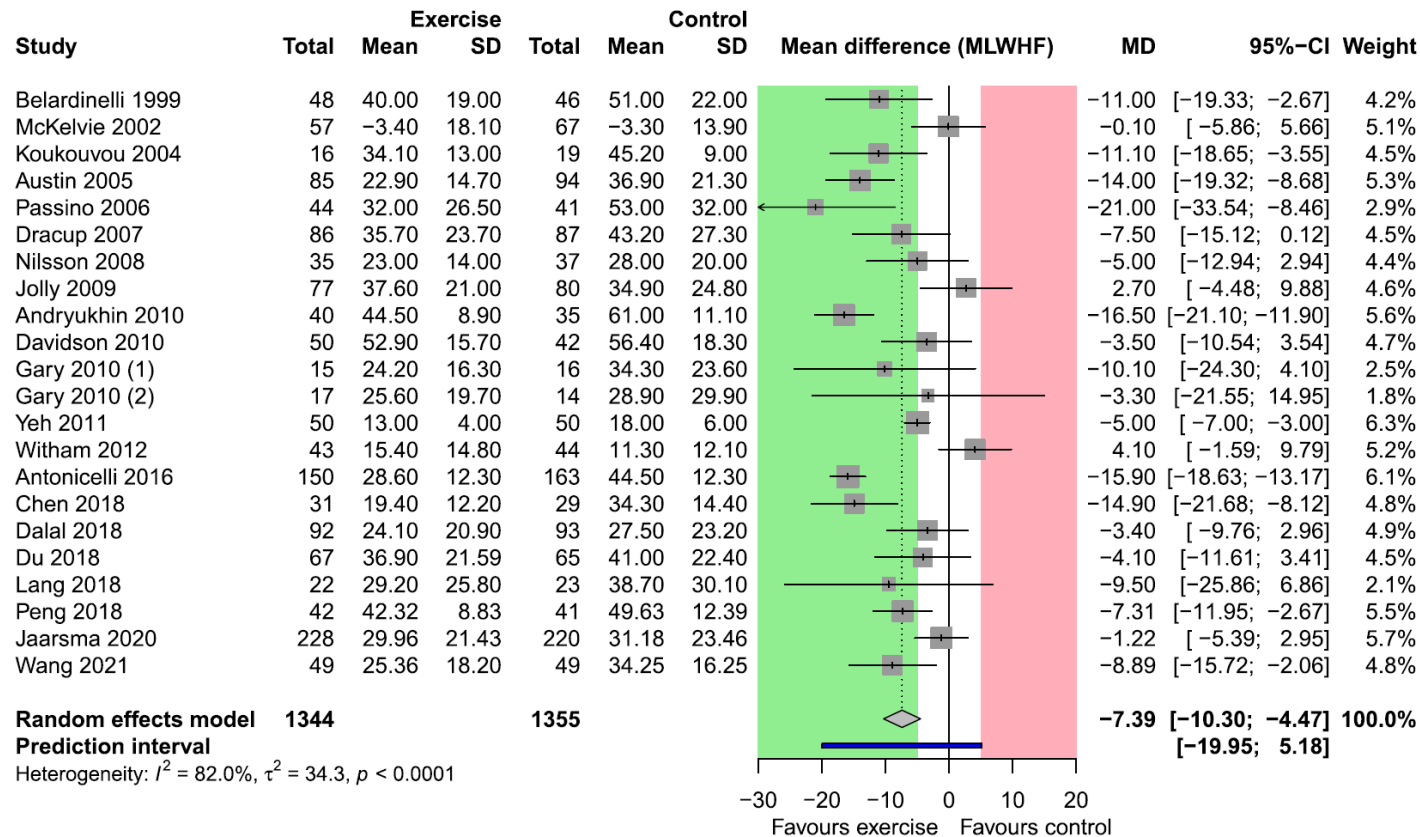
Open Access

Beyond prediction intervals in meta-analysis: reporting the expected proportion of comparable studies with clinically relevant benefit or harm



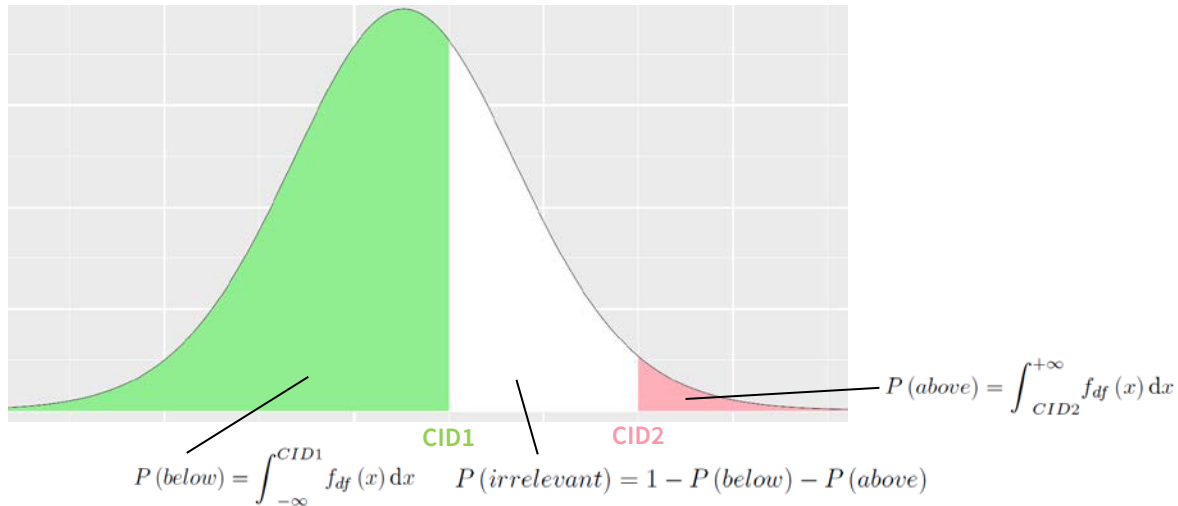
W. Siemens^{1*}, M. Borenstein², T. Evrenoglou³, J.J. Meerpohl^{1,4} and G. Schwarzer³

Forestplot MLWHF

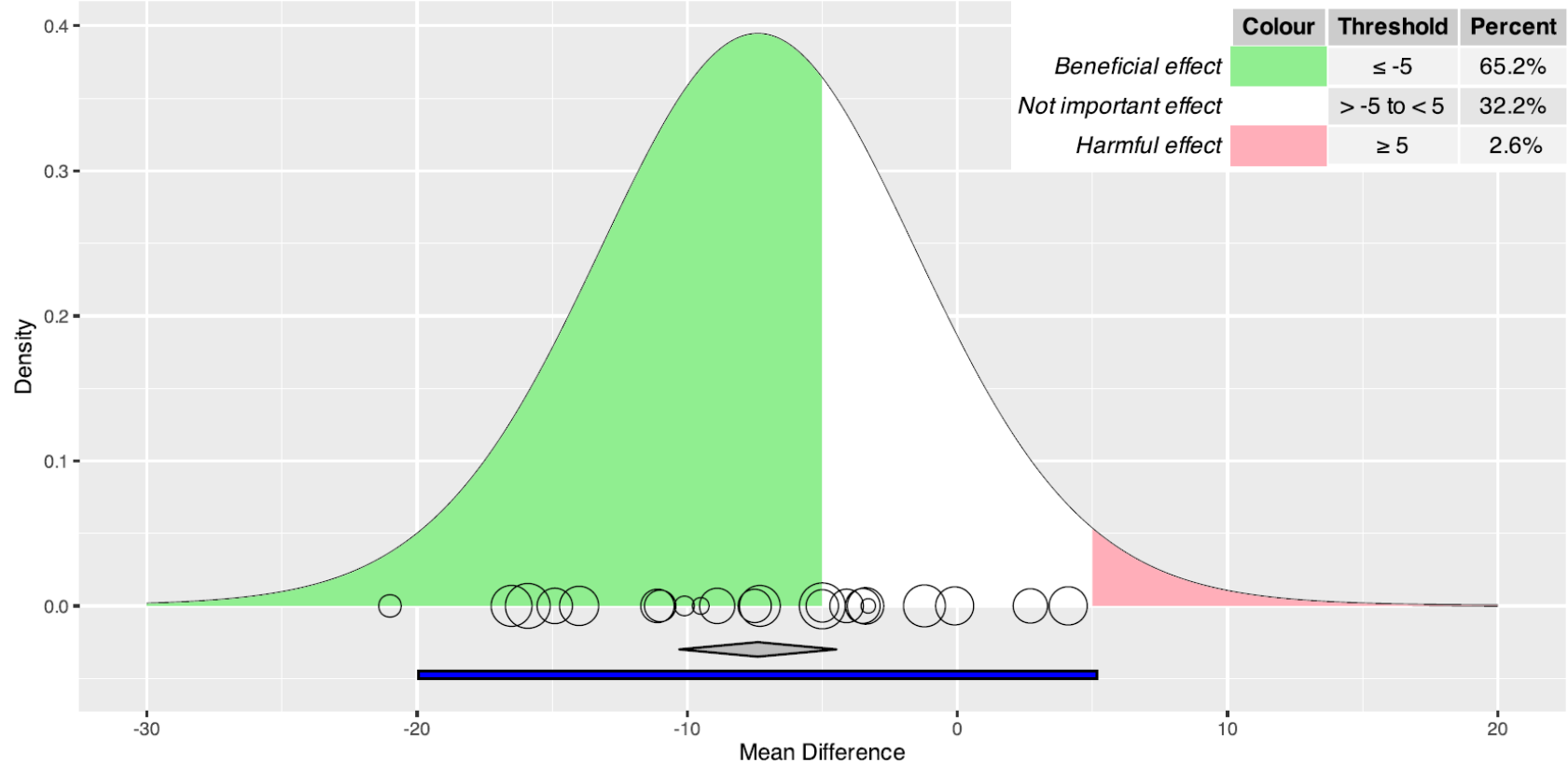


Erwartete Anteile von Studien mit relevantem Nutzen oder Schaden

- Erwartete Anteile vergleichbarer Studien unterhalb **CID1** und oberhalb **CID2**
- Integration der Fläche unter der t-Verteilung, die dem Prädiktionsintervall zugrunde liegt!



Expected proportion plot



Wir erwarten, dass eine trainingsbasierte kardiologische Rehabilitation in etwa 65,2 % der Studien, die mit denen in der Metaanalyse vergleichbar sind, klinisch hilfreich ist ($MD \leq -5$) und in etwa 2,6 % der Studien klinisch schädlich ist ($MD \geq 5$). Wir gehen davon aus, dass etwa 32,2 % solcher Studien in den Bereich der trivialen Effekte fallen würden ($MD > -5$ bis < 5).

Übersicht

1. Was ist Heterogenität?
2. 95% Prädiktionsintervall
3. Erwartete Anteile von Studien mit relevantem Nutzen oder Schaden
- 4. Zusammenfassung**



Zusammenfassung

- Statistische Heterogenität = klinische & methodische Diversität + Zufallsschwankungen
- Alle statistischen Heterogenitätsmaße sind im Falle weniger Studien nicht präzise schätzbar.
- Statistische Heterogenität quantifizieren und im Kontext klinischer Relevanzgrenzen diskutieren!
 - ▶ Prüfung der klinischen und methodischen Eigenschaften der eingeschlossenen Studien
 - ▶ Visueller Check des Forest Plots: Spannweite und Überschreiten klinischer Relevanzgrenzen prüfen
 - ▶ Berechnung des 95% Prädiktionsintervalls (idR bei ≥ 5 Studien relativ verlässlich); ggf. auch I^2
 - „95% Wahrscheinlichkeitsbereich eines zukünftigen Effekts einer vergleichbaren Studie: -19,98 bis 5,21 Punkte“
 - ▶ Erwartete Anteile von Studien mit relevantem Nutzen oder Schaden
 - „Wir erwarten, dass eine trainingsbasierte kardiologische Rehabilitation in etwa 65,2 % der Studien, die mit denen in der Metaanalyse vergleichbar sind, klinisch hilfreich ist ($MD \leq -5$) und in etwa 2,6 % der Studien klinisch schädlich ist ($MD \geq 5$)“

Literatur

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis, 2nd Edition. Wiley 2021.

Borenstein M. Avoiding common mistakes in meta-analysis: Understanding the distinct roles of Q, I-squared, tau-squared, and the prediction interval in reporting heterogeneity. *Res Synth Methods*. 2024 Mar;15(2):354-368. doi: 10.1002/jrsm.1678. Epub 2023 Nov 8.

Gonzalez-Saenz de Tejada M, Bilbao A, Ansola L, Quirós R, García-Perez L, Navarro G, Escobar A. Responsiveness and minimal clinically important difference of the Minnesota living with heart failure questionnaire. *Health Qual Life Outcomes*. 2019 Feb 14;17(1):36. doi: 10.1186/s12955-019-1104-2. PMID: 30764842; PMCID: PMC6376687.

Deeks JJ, Higgins JP, Altman DG, McKenzie JE, Veroniki AA (editors). Chapter 10: Chapter 10: Analysing data and undertaking meta-analyses [last updated November 2024]. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editor(s). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.5. Cochrane, 2024. Available from cochrane.org/handbook.

Molloy CD, Long L, Mordi IR, Bridges C, Sagar VA, Davies EJ, Coats AJS, Dalal H, Rees K, Singh SJ, Taylor RS. Exercise-based cardiac rehabilitation for adults with heart failure - 2023 Cochrane systematic review and meta-analysis. *Eur J Heart Fail*. 2023 Dec;25(12):2263-2273. doi: 10.1002/ehf.3046. Epub 2023 Oct 31.

Siemens W, Borenstein M, Evrenoglou T, Meerpohl JJ, Schwarzer G. Beyond prediction intervals in meta-analysis: reporting the expected proportion of comparable studies with clinically relevant benefit or harm. *BMC Med Res Methodol*. 2025 Dec 7;25(1):275. doi: 10.1186/s12874-025-02733-9.

**Vielen Dank für Ihre
Aufmerksamkeit!**

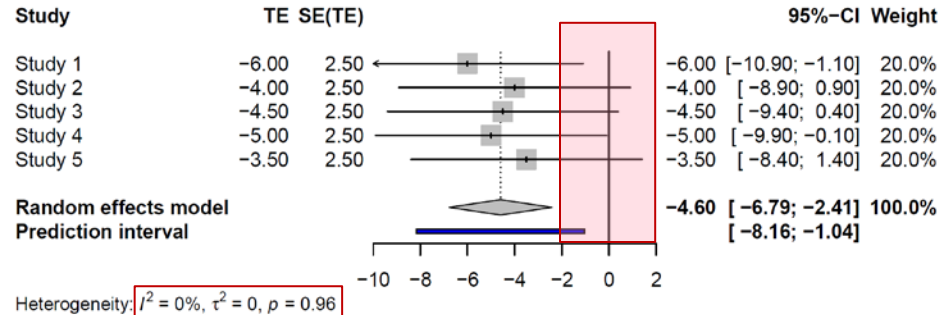
Zuverlässige Evidenz.
Informierte Entscheidungen.
Bessere Gesundheit.

Dr. Waldemar Siemens
waldemar.siemens@uniklinik-freiburg.de

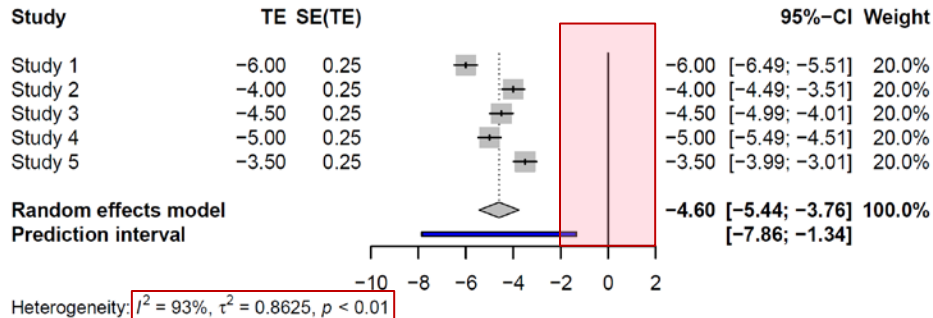


Inkonsistenz – Vorsicht mit I^2 und Chi^2 -Test

- Vergleichbare Effekte
- Große 95% KIs
- MCID: ± 2
- SE: 2,5



- Vergleichbare Effekte
- Präzise 95% KIs
- MCID: ± 2
- SE: 0,25



Using 95% confidence limits of $\hat{\tau}^2$ to calculate prediction interval and expected proportions of studies

$\hat{\tau}^2 * (95\% \text{ CI})$	34.3 (12.5 to 77.1)
95% prediction interval using lower limit	-15.03 to 0.29
95% prediction interval using point estimate	-19.95 to 5.18
95% prediction interval using upper limit	-26.21 to 11.31
Expected proportion of studies using lower limit	Beneficial [#] effect: 73.6% Not important [#] effect: 26.2% Harmful effect [#] : 0.1%
Expected proportion of studies using point estimate	Beneficial [#] effect: 65.2% Not important [#] effect: 32.2% Harmful effect [#] : 2.6%
Expected proportion of studies using upper limit	Beneficial [#] effect: 60.6% Not important [#] effect: 30.3% Harmful effect [#] : 9.1%

* DerSimonian-Laird estimator for $\hat{\tau}^2$ representing the estimated between-study variance; # beneficial effect

MD ≤ -5, not important effect MD > -5 to < 5, MD ≥ 5

MD: mean difference; NA: not applicable

Impact of other $\hat{\tau}^2$ estimates on prediction interval and expected proportions of studies

	DerSimonian-Laird estimator for $\hat{\tau}^2$	
95% prediction interval	-19.95 to 5.18	
Expected proportion of studies	Beneficial [#] effect:	65.2%
	Not important [#] effect:	32.2%
	Harmful effect [#] :	2.6%
	REML estimator for $\hat{\tau}^2$	
95% prediction interval	-19.45 to 4.69	
Expected proportion of studies	Beneficial [#] effect:	65.7%
	Not important [#] effect:	32.1%
	Harmful effect [#] :	2.2%
	Paule-Mandel estimator for $\hat{\tau}^2$	
95% prediction interval	-18.83 to 4.09	
Expected proportion of studies	Beneficial [#] effect:	66.4%
	Not important [#] effect:	31.8%
	Harmful effect [#] :	1.8%

$\hat{\tau}^2$ represents the estimated between-study variance

[#] beneficial effect MD ≤ -5, not important effect MD > -5 to < 5, MD ≥ 5

MD: mean difference; REML: Restricted maximum-likelihood estimator