

Untersuchung von und Umgang mit Heterogenität in Nutzenbewertungen – ein Problemaufriss

Dr. Sandra Janatzek

Fachbereich Evidenzbasierte Medizin

Medizinischer Dienst des Spitzenverbandes Bund
der Krankenkassen (MDS)

Essen

Einordnung

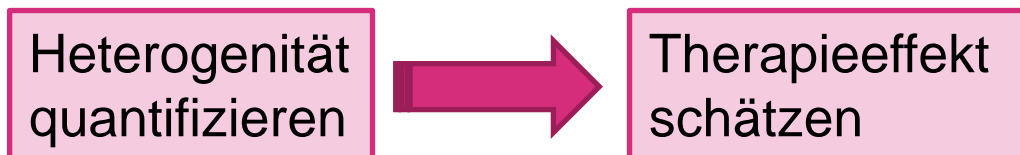
Kontext: Nutzenbewertung



Ziel: Aussagen zum Nutzen (und zum Schaden) ableiten

Quantifizierung + Untersuchung von Heterogenität „an sich“
ist dabei uninteressant.


**Interessant sind ausschliesslich die Auswirkungen auf
Nutzen-Aussagen !**



Fokus: Therapeutische Maßnahmen; direkte Vergleiche; publikationsbasierte Meta-Analysen

Vorgehen 1 („FEM bei Homogenität, REM bei Heterogenität“)

1) Sinnvolles **Aufsplitten** des Studienpools in mehrere Pools auf Basis methodischer und klinischer Unterschiede zwischen den Studien

- 
- unterschiedliche Kontrollintervention
 - unverblindet/verblindet
 - unterschiedliche Beobachtungsdauer
 - ...
- Unterschiede in der Prüftherapie (z.B. Dosierung, unterschiedliche Produkte)
 - Unterschiede in der Indikation (z.B. operable / inoperable Patienten)
 - ...

FEM = Modell mit festen Effekten (fixed effect model)

REM = Modell mit zufälligen Effekten (random effects model)

Unsere Erfahrung im MDS

- vorwiegend nicht-medikamentöse Maßnahmen
- intensive Bewertung der Einzelstudien steht im Vordergrund
- häufig nur wenige verwertbare Studien, die in mehrere Studienpools aufzusplitten sind → **sehr kleine Pools**, häufig nur 1 oder 2 Studien

Beispiel:

Asynchrone Balneophototherapie bei Psoriasis vulgaris

(IQWiG-Abschlussbericht
vom 21.12.2006, N04/04)

Therapievergleich		Studie(n)	Anzahl Studien (Anzahl RCTs)
Bade-PUVA	vs. orale PUVA	Collins 1992; Cooper 2000; Calzavara-Pinton 1994; Lowe 1986	4 (2 RCTs)
Bade-PUVA	vs. SB-UVB	Dawe 2003; Snellman 2004; Rosón 2005	3 (2 RCTs)
Bade-PUVA	vs. UVB	BP-BVDD-Studie	1 (1 RCT)
Bade-PUVA	vs. LW+UVB	BP-BVDD-Studie	1 (1 RCT)
Sole + SB-UVB	vs. SB-UVB	Dawe 2005; Léauté-Labrèze 2001	2 (2 RCTs)
Sole + UVB	vs. UVB	BP-BVDD-Studie	1 (1 RCT)
Sole + UVB	vs. LW+UVB	BP-BVDD-Studie	1 (1 RCT)
Sole + BB-UVB	vs. BB-UVB	Boer 1982	1 (0 RCTs)
Sole + BB-UVB	vs. LW+BB-UVB	Boer 1982	1 (0 RCTs)
Sole + SB-UVB	vs. Sole	Léauté-Labrèze 2001	1 (1 RCT)
Bade-PUVA	vs. Sole+UVB	BP-BVDD-Studie	1 (1 RCT)

Dann für jeden Studienpool separat:

- 2) Heterogenitätstest (Q-Test) zum Niveau 15% (10% ? 20% ?)
- 3) Wenn nicht signifikant, dann FEM-Analyse → MA-Schätzer
- 4) Wenn signifikant, dann nach möglichen Ursachen der Het. Suchen (Subgruppen-Analysen, Meta-Regressionen)
- 5) Falls keine Ursachen ident., dann REM-Analyse → MA-Schätzer
- 6) Falls Ursachen identifiziert werden, dann entsprechendes Aufsplitten des Studienpools (→ höchstens noch unerklärte Het.)
- 7) Für jeden Studienpool separat:
 - Heterogenitätstest (Q-Test) zum Niveau 15% (10% ? 20% ?)
 - Falls signifikant, dann REM-Analyse, sonst FEM-Analyse → MA-Schätzer

Bei welchem Ausmaß (unerklärter)
Heterogenität wird nicht gepoolt?

Vorgehen 2 („generell REM“)

1) Sinnvolles **Aufsplitten** des Studienpools in mehrere Pools

Dann für jeden Studienpool separat:

2) Heterogenitätstest (Q-Test) zum Niveau 15% (10% ? 20% ?)

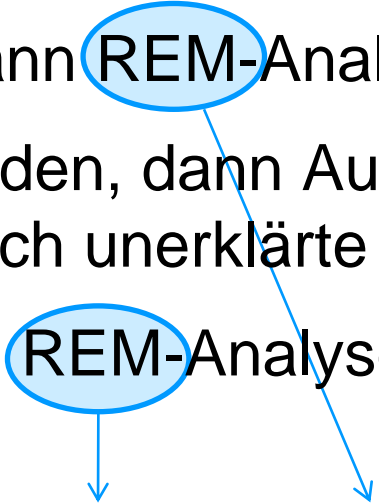
3) Wenn nicht signifikant, dann REM-Analyse → **MA-Schätzer**

4) Wenn signifikant, dann nach möglichen Ursachen der Het. Suchen (Subgruppen-Analysen, Meta-Regressionen)

5) Falls keine Ursachen ident., dann **REM-Analyse** → **MA-Schätzer**

6) Falls Ursachen identifiziert werden, dann Aufsplitten des Studienpools (→ höchstens noch unerklärte Het.)

7) Für jeden Studienpool separat: **REM-Analyse** → **MA-Schätzer**



Bei welchem Ausmaß (unerklärter)
Heterogenität wird nicht gepoolt?

Vorgehensweise in verschiedenen Institutionen

Aus den Methodenpapieren:

- **IQWiG:** Allgemeine Methoden 4.0
- **AHRQ:** Methods Guide for Effectiveness and Comparative Effectiveness Reviews, March 2011
- **NICE:** Guide to the methods of technology appraisal, June 2008. → Darin Verweis auf:

Centre for Reviews and Dissemination:
Systematic Reviews – CRD's guidance for undertaking reviews in health care, 2008

IQWiG

Modell

REM

Effektmaß

relatives Maß

Wie untersuchen, ob Heterogenität vorliegt?

- Heterogenitätstest zum Niveau 10-20%
- I^2

Wann + wie werden Ursachen der Heterogenität untersucht?

- Bei „großer Heterogenität“:
- Meta-Regressionen
 - Subgruppen-Analysen (?)

Immer poolen?

nein

Wann nicht?

Falls „Heterogenität zu groß“:
Höchstens dann poolen, wenn Einzelstudien deutliche + gleich gerichtete Effekte zeigen,
in Entscheidung fließen auch inhaltliche Gründe ein

Falls Ergebnisse eines Heterogenitäts- oder Interaktionstests bzgl. wichtiger Subgruppen signifikant zum Niveau 5%, dann kein Poolen aller Studien, sondern Subgruppen-Ergebnisse (→ getrennte Nutzaussagen)

AHRQ

Modell

REM

Effektmaß

Risikodifferenz (bei seltenen Outcomes: Relatives Risiko)

Wie untersuchen, ob Heterogenität vorliegt?

- visuelle Inspektion des Forest Plots und des kumulativen MA-Plots
- Heterogenitätstest zum Niveau 10%
- I^2 mit KI

Wann + wie werden Ursachen der Heterogenität untersucht?

keine Angabe zum „wann“

- Subgruppen-Analysen
- Meta-Regressionen
- Sensitivitätsanalysen
- Ausreißer-Elimination: Falls statistische Heterogenität durch 1 oder 2 Studien verursacht, können Sensitivitätsanalysen mit Ausschluss dieser Studien durchgeführt werden

Immer poolen?

nein

Wann nicht?

- Entscheidung **nicht** auf Basis des Heterogenitätstests
- Wenn **große** klinische + methodische Heterogenität und gleichzeitig **große** statistische Heterogenität

NICE

Modell

FEM **und** REM, um Robustheit zu prüfen

Effektmaß

keine Angabe

Wie untersuchen, ob Heterogenität vorliegt?

- visuelle Inspektion des Forest Plots
- Heterogenitätstest zum Niveau 10%
- I^2

Wann + wie werden Ursachen der Heterogenität untersucht?

Bei **statistischer Heterogenität**:

- Subgruppen-Analysen
- Meta-Regressionen

Immer poolen?

nein

Wann nicht?

keine Angabe

Vorgehensweise in verschiedenen Institutionen

Übereinstimmung mit
internationaler Literatur

- Kein einheitliches Vorgehen bzgl. **Modellwahl (FEM / REM)**
- Kein einheitliches Vorgehen bzgl. Wahl des **Effektmaßes**
- Allen gemeinsam:
 - Unklarheit, **wann gepoolt wird** und wann nicht
 - Unklarheit, wann potentielle Ursachen der Heterogenität untersucht werden

Weitere offene Fragen:

- Heterogenitätstest (**Q-Test**) oder **I^2** oder ... ?
Welches Signifikanzniveau bzw. welcher Cut-off ?
- Ist der MA-Schätzer (mit KI) in der **Situation sehr weniger Studien** valide ?

FEM oder REM bzw. wann welches?

Schroll et al. (2011): Zufallsstichprobe von 60 Cochrane Reviews aus Cochrane Database of SR's 2008, Issue 1 (Reviews, die mindestens 1 MA enthalten, für das 1. Outcome im 1. Vergleich *alle* Studien enthalten und $I^2 > 50\%$)

BMC Medical Research Methodology 2011, 11:22

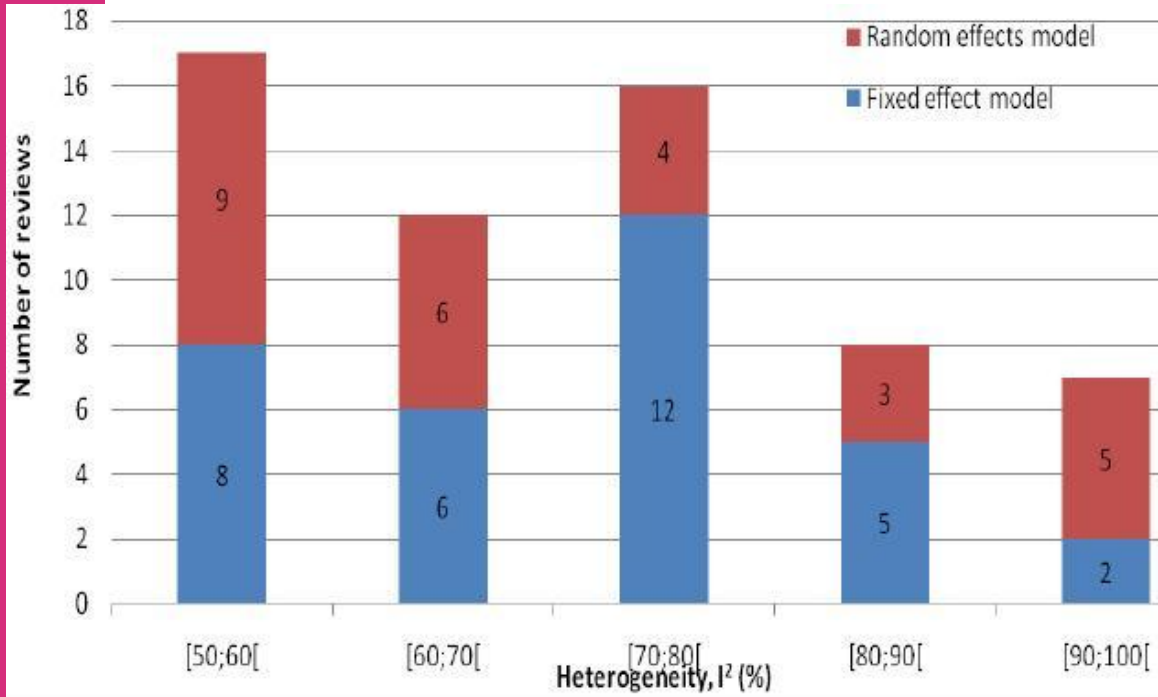


Table 1 Choice of model in relation to the P-value for the heterogeneity test

P	Newer reviews		Older reviews	
	Random	Fixed	Random	Fixed
< 0.0001	6	0	2	6
[0.0001;0.001[2	1	2	1
[0.001;0.01[1	2	0	3
[0.01;0.05[7	4	1	4
[0.05;0.1[6	2	0	8
> = 0.1	0	0	0	2
Total	22	9	5	24

Newer reviews are those updated after 1 June 2005 (n = 31). P = 0.007 for those reviews where the heterogeneity test yielded a P-value between 0.05 and 0.10.

- Modellwahl nach wie vor kontrovers diskutiert
- Tendenz, verstärkt generell REM einzusetzen

Welches Effektmaß ?

- Meist empfohlen: **relatives Maß**,
da i.R. homogenere Therapieeffekte als mit absolutem Maß
(z.B. Engels et al. 2000, *Statist. Med.* 19:1707-1728)
- Bei Vorliegen von Heterogenität:
Effektmaß wechseln und prüfen, ob dann Homogenität vorliegt
 - im Sinne einer Sensitivitätsanalyse etabliert
 - **Aber**: Dürfen wir daraus Nutzen-Aussage ableiten ?

Q-Test oder I^2 oder?

- kein Konsens

- Aufgabe von Q-Test, I^2 , ... :
 - Entscheidung, ob nach Ursachen von Heterogenität gesucht wird
 - Entscheidung für/gegen Poolen (bei unerklärter Het.)
 - Entscheidung über Modellwahl (bei unerklärter Het.)

Unter welchen Bedingungen wird nicht gepoolt?

▪ Entscheidung auf Basis des KI zu τ^2 :

Rücker et al. 2008 (*BMC Medical Research Methodology* 2008, 8:79),

Knapp et al. 2006 (*Biometrical Journal* 48 (2): 271-285)

→ Wenn obere KI-Grenze $> \tau^2_{\text{relevant}}$, dann nicht poolen

▪ Entscheidung auf Basis des Q-Tests:

zu welchem Niveau ?

▪ Entscheidung auf Basis von I^2 :

Zu welchem Cut-off ? 75% ?

Wird I^2 oder die obere KI-Grenze mit dem Cut-off verglichen ?

▪ Entscheidung auf Basis verschiedener Faktoren:

- I^2
- Q-Test
- Größe + Richtung der Therapieeffekte der Einzelstudien (→ Forest-Plot)
- klinische + methodische Unterschiede zwischen den Studien

→ Wie ?

(z.B. Higgins & Thompson 2002 (*Statist. Med.* 21: 1539-1558),
Virgili et al. 2009 (*Intern Emerg Med* 4: 423-427),
Cochrane Handbook)

MA mit sehr wenigen Studien

- Klassische Auswertung im REM hält Niveau nicht ein, auch nicht asymptotisch ($N \rightarrow \infty$)
- Niveau wird insbesondere bei **wenigen großen** Studien (mit größerer Varianz zwischen den Studien) überschritten, → umso **mehr**, je größer die Einzelstudien
- Asymptotischer Fehler 1. Art ($N \rightarrow \infty$), hier für Spezialfall gleicher Varianzen in den Einzelstudien:

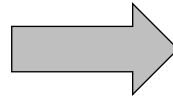
Anzahl Studien	Asympt. Fehler 1. Art
2	30%
3	19%
4	15%
5	12%

Ziegler, Koch, Victor 2001: Method Inform Med 40: 148-55

➤ Ist diese Niveauüberschreitung in Meta-Analysen akzeptabel?

Nein, denn Funktion von Meta-Analysen hat sich gewandelt:

Meta-Analysen
sind **explorativer**
Natur



Im Grunde leiten wir „**konfirmatorische Aussagen**“ ab, denn die Ergebnisse münden in Nutzen-Aussagen

➤ Es gibt Vorschläge, die das Niveau (approximativ) einhalten, z.B.

Hartung & Knapp 2001 (*Statist. Med.* 20: 3875-3889)

Hartung & Knapp 2001 (*Statist. Med.* 20: 1771-1782)

Biggerstaff & Tweedie 1997 (*Statist. Med.* 16: 753-768)

➤ Ihr Einsatz erscheint dringend notwendig ...

➤ ... es sei denn, die (relevanten) Niveauüberschreitungen treten nur in Situationen auf, in denen sowieso nicht gepoolt wird ?

Fazit

- Viele offene Fragen
- Es ist wesentlich, die Fragen „als Ganzes“ zu bearbeiten, im Kontext der Schätzung des Therapieeffektes
- Vermutung:
Einige Probleme lassen sich vermeiden, indem mehr Aufmerksamkeit darauf verwendet wird, auf Grundlage klinischer und methodischer Unterschiede zwischen den Studien zu entscheiden, ob/wie der **Studienpool aufgesplittet** und/oder reduziert werden sollte.
- These:
Ein **klarer Algorithmus** (ohne subjektive Entscheidungselemente) wird schwer begründbar sein. Wir sollten akzeptieren, dass Heterogenität inhaltlich (und mit statistischen Hilfsmitteln) hinterfragt werden muss.