Kriterien zur Überprüfung der Anwendbarkeit von Studienergebnissen

Ulrich Grouven, Lars Beckmann, Ralf Bender, Stefan Lange
IQWiG

IQWiG im Dialog, 21.06.2013

Übersicht

- Einführung / Problemstellung
- Teil A:
  - Empfohlenes Vorgehen der *Agency for Healthcare Research and Quality* (AHRQ) / *GRADE Working Group*
  - Biometrische Ansätze
- Teil B:

  Ein methodischer Vorschlag zum konkreten Vorgehen bei der Ableitung von Nutzenbelegen
- Schlussfolgerungen

Problemstellung

Ergebnisse Studienpopulation (SP)

Übertragbarkeit

Zielpopulation (ZP)

Auch:
Anwendbarkeit,
externe Validität,
Verallgemeinerbarkeit,
Directness,
Relevance

Betrachtete Fälle

(1) ZP und SP weichen bzgl. relevanter Patienten-charakteristika voneinander ab (unterschiedliche Verteilung relevanter Kovariablen)

→ **Teil A**

(2) Spezielle Datensituation:
ZP ist eine Teilmenge der SP

→ **Teil B**

Teil A

Wie wird Übertragbarkeit bewertet?

Zitate

„…there is currently no consensus about how to assess the external validity of study results.“ (Dekkers et al. 2010)

*„The **state of the art** currently is simply to compare the covariates one by one and to make qualitative statements regarding the similarity of subjects in a trial and some target population…” (Stuart et al. 2011)*

Was macht AHRQ? *(Atkins et al. 2011)*

Definition **„applicability“:**

“…the extent to which the effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to the population of interest under “real-world” conditions.“

Ziel

„… describe a systematic but practical approach for considering applicability in the process of reviewing, reporting, and synthesizing evidence from eligible studies.“

Vorgehen AHRQ (I)

Vergleich von Studien- und Zielpopulation anhand des **PICOS**-Schemas (**P**atient, **I**ntervention, **C**omparator, **O**utcome, **S**etting)

Generelle Punkte

1. Anwendbarkeit sollte getrennt für unterschiedliche Outcomes bewertet werden.
2. Anwendbarkeit ist kontext-abhängig (Verwendung eines allgemeingültigen Bewertungsschemas (Check-Liste) nicht empfohlen).
3. Mögliche Faktoren, die Übertragbarkeit beeinflussen, sollten separat von der Qualität der verfügbaren Evidenz diskutiert werden.

Vorgehen AHRQ (II)

Spezifische Schritte

1. Bestimmung der **wichtigsten Faktoren**, die die Übertragbarkeit beeinflussen können (relevante Effektmodifikatoren, Basis-Risiko).
2. Erstellung von **Evidenztabellen** mit Studieninformationen mit den in Schritt 1 identifizierten Faktoren (besondere Kennzeichnung von *Effectiveness Trials*).
3. Beschreibung und Bewertung der **Abweichungen** zwischen Studien- und Zielpopulation und deren potenzielle Auswirkungen auf die Übertragbarkeit.
4. Erstellung einer zusammenfassenden **Übersichtstabelle** mit Studien-übergreifenden Informationen (gemäß PICOS-Schema) und **Bewertung der Gesamt-Evidenz**

Was macht GRADE? *(Guyatt et al. 2011)*

Applicability als ein möglicher Aspekt von ***indirectness*** behandelt (neben Verwendung von Surrogatendpunkten und Durchführung indirekter Vergleiche)

Grundsätzlich ähnliches Vorgehen wie bei AHRQ:

→ Inhaltliche Bewertung von *Applicability* gemäß **PICOS**-Schema

Biometrische Ansätze

Grundlegende Idee

Adjustierung der Studienergebnisse für Patientencharakteristika (Kovariablen-Struktur) in der ZP
(→ „Umrechnung“ des Effektes von SP auf ZP)

Auswahl der Kovariablen

- Einfluss auf den Behandlungseffekt
- unterschiedliche Ausprägung zwischen SP und ZP

Mögliche Ansätze

- *Direkte Standardisierung*
  (Post-Stratification, Kalibrierung)
- *Modell-basierte Standardisierung*
  (Regressionsmodelle)

Direkte Standardisierung

- Stratifizierte Effektschätzung mit Gewichten gemäß Kovariablen-Struktur in der ZP
- *Notwendige Daten*
  - Individuelle Patientendaten (IPD) für Effektschätzung in der SP
  - Verteilung der verwendeten Kovariablen über die Strata in der ZP
- *Nachteil*

  nicht einsetzbar bei vielen kategoriellen oder stetigen Kovariablen

Modell-basierte Standardisierung

Z.B. Verwendung von **Propensity Scores**

- Berechnung von Maßzahlen zur Quantifizierung des Unterschiedes zwischen SP und ZP (Stuart et al. 2011)
- Standardisierte Effektschätzung für ZP
  (Erweiterung der direkten Standardisierung,
  z.B. Verwendung stetiger Kovariablen möglich)
  (Cole & Stewart 2010)
- *Notwendige Daten*
  IPD in SP und ZP

Exkurs Propensity Scores (PS)

Ziel

Adjustierung von Effektschätzern bzgl. Confounding durch relevante Kovariablen (Ausgangspunkt: Vergleich 2 Behandlungsgruppen)

Definition

(Bedingte) Wahrscheinlichkeit zur Interventionsgruppe zu gehören (Z = 1) bei gegebenen Kovariablen-Werten (X), d.h. p = P(Z=1 | X)

Berechnung

I.d.R. mittels logistischer Regressionsmodellen: $logit(p) = a + bX$

(a, b Modellparameter)

→ Patienten mit identischen Werten des PS haben vergleichbare Kovariablen-Struktur

→ Anwendung von PS (durch Matching, Stratifizierung oder Adjustierung) führt zu Strukturgleichheit der Behandlungsgruppen (aber nur für die berücksichtigten Kovariablen!)

Vorgehen (I) (Cole & Stewart 2010)

Hier: Propensity-Score $P(S_i = 1 \mid Z_i)$ definiert als *bedingte Wahrscheinlichkeit für Selektion in die SP*

(1) Effektschätzung (Hazard Ratio (HR)) mittels Cox Proportional Hazards Modell für die SP

(2) Berechnung eines Gewichtungsfaktors W_i (inverse probability-of – selection weight) mittels Propensity Scores für jeden Datensatz:

$$W_i = \begin{cases} \dfrac{P(S_i = 1)}{P(S_i = 1 \mid \mathbf{Z}_i)}, & \mathrm{S}_i = 1, \\ 0, & \mathrm{S}_i = 0, \end{cases}$$

i = 1,..,n (Größe der ZP)

mit

S_i Indikator für Mitglied der SP

Z_i Vektor mit relevanten Kovariablen in der ZP

Vorgehen (II) (Cole & Stewart 2010)

(3) Adjustierte Effektschätzung (HR^*) für die ZP

Modifizierte Partial Likelihood-Funktion:

$$L(\gamma) = \prod_{i=1}^{n} \left[\frac{\exp(\gamma X_i) \times W_i}{\sum_{k=1}^{n} R_k(t_i) \times \exp(\gamma X_k) \times W_k} \right]^{Y_i},$$

Y_i = Indikator für Ereignis

X_i = Behandlungseffekt

i = 1,..,n

Individuelles Gewicht

→ Modifiziertes Hazard-Ratio HR^* für ZP

Allgemeine Anmerkungen (Teil A)

- *Interne* Validität ist Voraussetzung für *externe* Validität (CONSORT 2010, Dekkers et al. 2010, Windeler 2008)
- *Externe* Validität ist nicht absolut, sondern hängt explizit von der betrachteten Zielpopulation ab; *externe* Validität ist – im Gegensatz zur *internen* Validität – kein *Studienkriterium*, sondern ein *Situationskriterium* (CONSORT 2010, Windeler 2008)
- Die entscheidende Frage ist nicht, ob sich die *Patienten* zwischen SP und ZP unterscheiden, sondern der *Behandlungseffekt* (Windeler 2008)

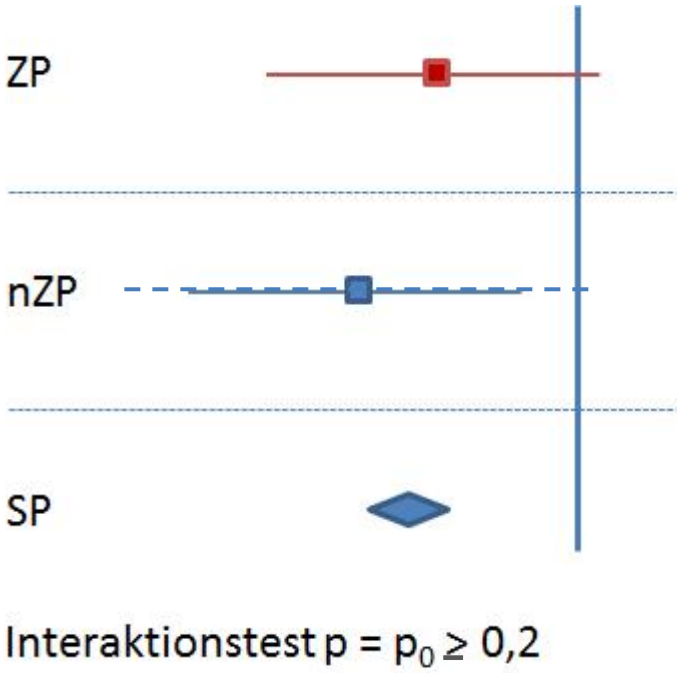
Zusammenfassung (Teil A)

- Das empfohlene Vorgehen von AHRQ und GRADE bei der Bewertung der *externen* Validität basiert auf *inhaltlichen* Aspekten und beinhaltet in besonderem Maße eine *subjektive* Komponente
- Die vorgestellten biometrischen Ansätze sind aufgrund hoher Unsicherheiten (ungemessenes Confounding, Modellannahmen, Extrapolation) und in der Praxis i.d.R. nicht verfügbarer Daten (IPD für SP und ZP) nicht unmittelbar einsetzbar für die Nutzenbewertung des IQWiG

Teil B

Motivation: Konkrete Datensituation

→ SP unterteilt sich in ZP und Nicht-ZP (nZP) (z.B. aufgrund der Zulassung, Festlegung der zweckmäßigen Vergleichstherapie)



→ kein signifikanter Effekt in ZP

→ (kein) signifikanter Effekt in nZP

→ signifikanter Effekt in SP

→ Bei Betrachtung von ZP: **kein Zusatznutzen**

Frage: Können die Ergebnisse der gesamten SP für Aussagen über die ZP herangezogen werden?

Häufiges Vorgehen

- Berechnung eines Interaktionstests zwischen ZP und nZP
- Aus Nichtsignifikanz ($p \geq 0{,}2$) wird Gleichheit von ZP und nZP gefolgert
- Somit können die Ergebnisse von SP zur Ableitung einer Nutzenaussage für ZP herangezogen werden

Problem

Es handelt sich um eine *Äquivalenzfragestellung*. Aus einem nicht signifikanten Heterogenitätstest kann nicht auf die Äquivalenz der Populationen geschlossen werden *(„absence of evidence is not evidence of absence“)*

Frage

Unter welchen Umständen ist es gerechtfertigt, die Ergebnisse der SP für Aussagen über die ZP heranzuziehen?

Vorschlag für methodisches Vorgehen

Ziel

Ausschluss von qualitativer Interaktion (d.h. in ZP liegt kein oder gegenläufiger Effekt vor)

Frage

Wie wahrscheinlich ist das beobachtete Ergebnis, wenn in Wahrheit kein Effekt in der ZP vorliegt?

Vorgehen

Simulation basierend auf beobachteten Ergebnissen und unter der Annahme, dass kein Effekt in ZP vorliegt:

→ betrachte Anteil an Simulationen mit beobachtetem (und extremerem Ergebnis) („empirischer“ p-Wert p_{emp})

Resultat

Falls p_{emp} klein, dann Heranziehung der Ergebnisse der gesamten SP zur Ableitung von Nutzenaussagen

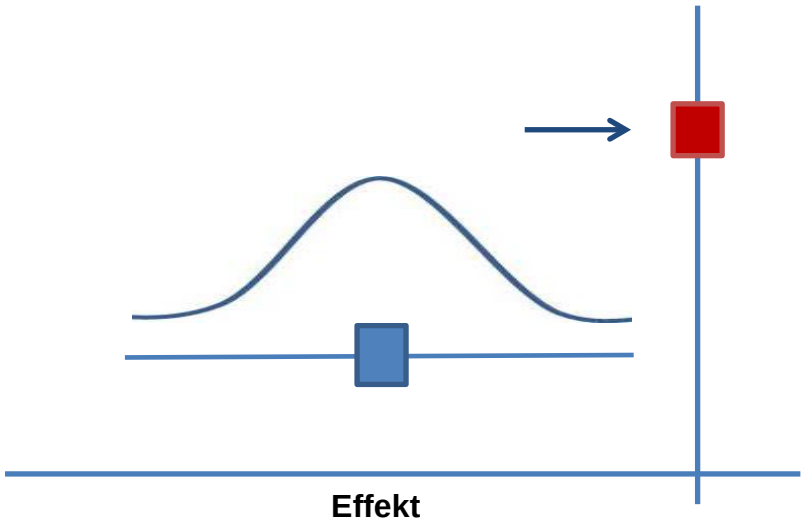
Simulation (I)

Hypothese: H_0: $RR_{ZP} \geq 1$ vs. H_1: $RR_{ZP} < 1$

Simulation (unter H_0) (basierend auf vorliegenden Daten → fix):

- Fallzahlen für Behandlungsgruppe und Kontrollgruppe (für ZP und nZP)
- Basisrisiko (aus SP): $R_{Basis} = n_{SP}^{K} / N_{SP}^{K}$
- p-Wert Interaktionstest ($p_{Interaktion}$)
- $RR_{nZP}\,[95\%\text{-}KI]$

Grafische Veranschaulichung



Simulation (II)

- Simuliere:
  - Ereigniszahlen $\tilde{n}^{B}_{ZP}, \tilde{n}^{K}_{ZP}, \tilde{n}^{K}_{nZP}$ aus Binomialverteilung $B(N^*, R_{Basis})$
    (* zugehörige Teilpopulation)
  - $\tilde{n}^{B}_{nZP}$ aus $B(N^{B}_{nZP}, \widetilde{RR}_{nZP} \times R_{Basis})$
  - $\widetilde{RR}_{nZP}$ aus $N\{log(RR_{nZP}), var(log(RR_{nZP}))\}$

  → Jeder Simulationsdurchgang liefert $\widetilde{RR}_{ZP}$ und p-Wert $\tilde{p}_{Interaktion}$ für Interaktionstest (aus Q-Statistik)

- Berechne empirischen p-Wert p_{emp} als Anteil der Simulations-durchgänge mit $\widetilde{RR}_{ZP} \leq RR_{ZP}$ und $\tilde{p}_{Interaktion} \geq p_{Interaktion}$

- Wenn $p_{emp} \leq 0{,}025$ → H_0 ablehnen

  ⇒ Beleg (Hinweis, Anhaltspunkt) für einen Zusatznutzen, Ausmaß: nicht quantifizierbar

Zusammenfassung / Schlussfolgerungen

- Die Frage der Übertragbarkeit ist komplex und vielschichtig
- Bei der Frage nach der Übertragbarkeit von Studien-ergebnissen auf eine relevante ZP spielt die *inhaltliche Bewertung* eine wesentliche Rolle
- Die vorgestellten biometrischen Ansätze sind mit zahlreichen Unsicherheiten behaftet und nicht unmittelbar relevant für die Nutzenbewertung
- Der vorgeschlagene Simulationsansatz (Teil B) ermöglicht die Ableitung einer Nutzenaussage für die relevante ZP in bestimmten Situationen

Literatur

1. Atkins D, Chang SM, Garthlehner G, et al. Assessing applicability when comparing medical interventions: AHRQ and the Effective Health Care Program. Journal of Clinical Epidemiology 2011; 64: 1198-1207.
2. Cole SR, Stuart EA. Generalizing evience from randomized clinical trials to target populations. American Journal of Epidemiology 2010; 172: 107-115.
3. Dekkers OM, von Elm E, Algra A, et al. How to assess the external validity of therapeutic trials: a conceptual approach. International Journal of Epidemiology 2010; 39: 89- 94.
4. Guyatt G, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence – indirectness. Journal of Clinical Epideimology 2011; 64: 1303-1310.
5. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 – Explanation and elaboration: updated guidelines for reporting parallel group randomsed trials. BMJ 2010; 340: c869.
6. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. Journal of The Royal Statistical Society A 2011; 174, Part 2: 369-386.
7. Windeler J. Externe Validität. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen 2008; 102: 253-260.