Workshop der Arbeitsgruppe Therapeutische Forschung

Bias-Kontrolle und P-Wert-Kalibrierung am Beispiel von Negativ-Kontrollen

27. September 2021

Dr. Tobias Bluhmki
Senior Manager Biostatistics – Real World Evidence

Bristol Myers Squibb™

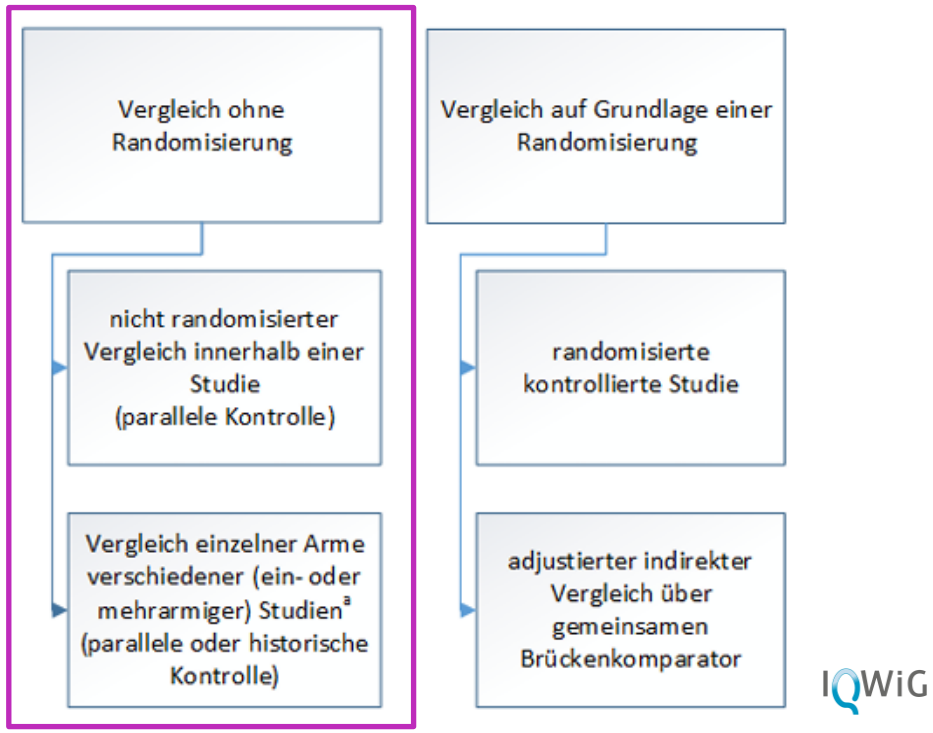
gmds 2021 TMF

Disclaimer

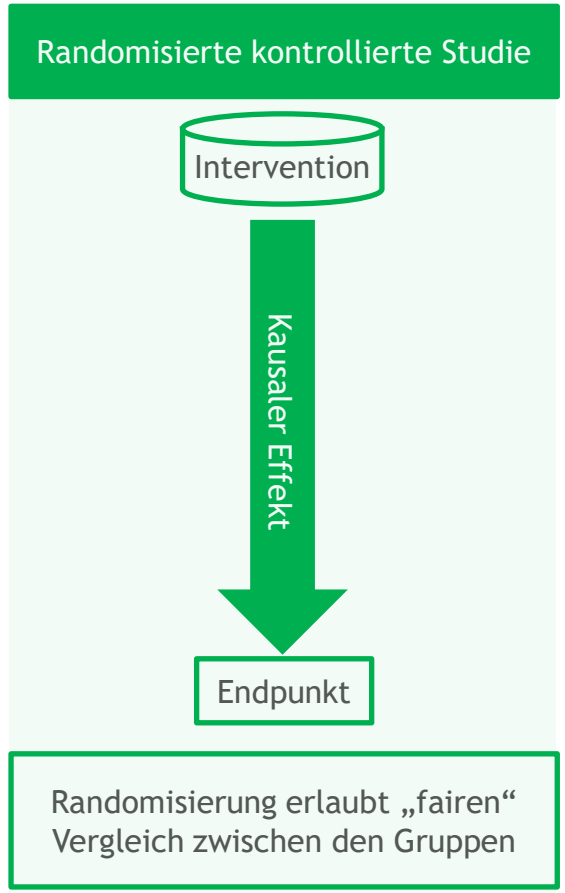
Die in dieser Präsentation gezeigten Inhalte und zum Ausdruck gebrachte Auffassung sind die Meinung des Vortragenden und nicht unbedingt die von Bristol Myers Squibb. Bristol Myers Squibb garantiert nicht die Richtigkeit und Verlässlichkeit der hier gezeigten Informationen.

Gold-Standard: Randomisierte kontrollierte Studie

- „fairer“ Vergleich (Strukturgleichheit) erlaubt kausale Aussage „…**wegen** Intervention…“
- Beobachtete Assoziation gleichzusetzen mit wahrem kausalen Effekt



IQWiG Rapid Report A19-43 (2020): *Konzept zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach §35a SGB V*



Unverzerrte Effektschätzung

Adaptiert nach [1]

Adjustierung bei nicht-randomisierten Vergleichen

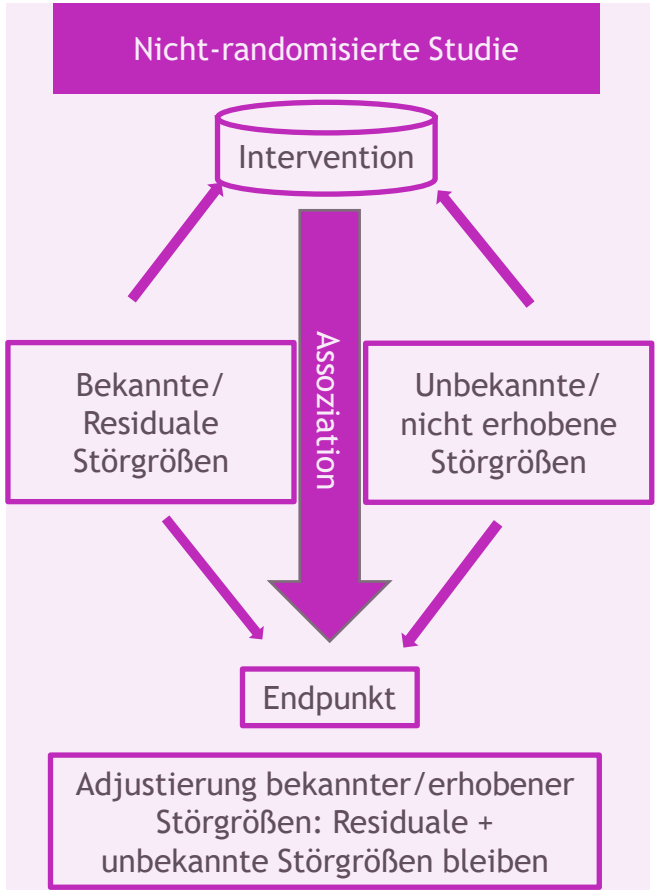
Zufälliger Fehler | Systematische Verzerrung

Trennung des kausalen Effekts einer Intervention auf einen Endpunkt von Assoziationen hervorgerufen durch andere Mechanismen

1. Informationsbias (systematische Messfehler, Misklassifikation)
2. Selektionsbias
3. Confounding

Confounder (Störgröße): Risikofaktor für den zu interessierenden Endpunkt, der mit der zu interessierenden Intervention assoziiert ist und nicht in der Kausalkette zwischen der Intervention und dem Endpunkt steht

→ Kein „fairer“, kausal zu interpretierender Vergleich
→ Adjustierung (z.B. mittels Propensity Score)



Verzerrte Effektschätzung

Adaptiert nach [1]

Confounding in der Nutzenbewertung

Annahme Kausalitätsaussagen (nach Adjustierung für bekannte/erhobene Störgrößen):
no unmeasured confounding

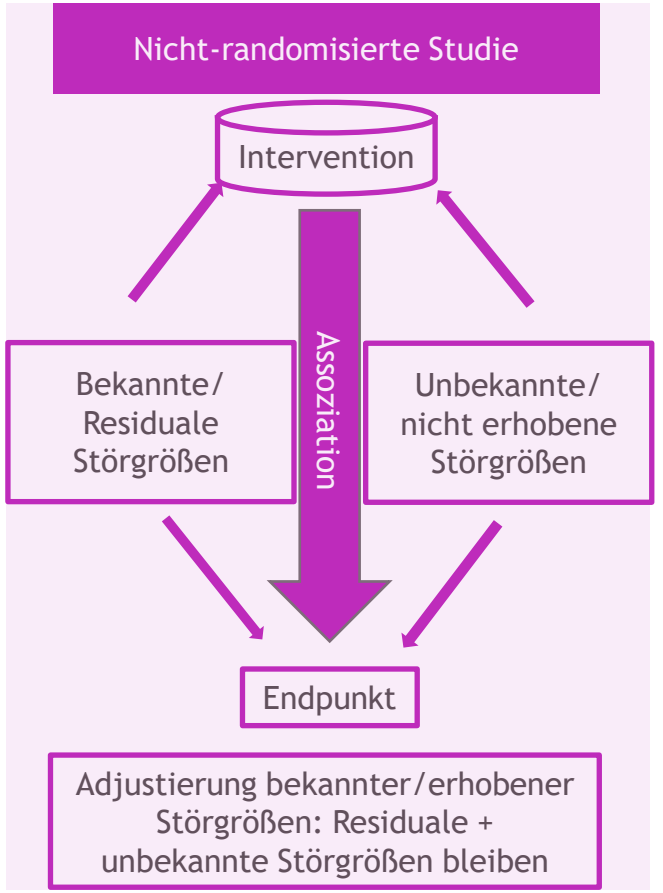
IQWiG Rapid Report A19-43 (2020):
„[…] aufgrund potenziell unbekannter Confounder aus den in der Studie beobachteten Effekten erst ab einer bestimmten Effektstärke eine Aussage zum Nutzen oder Schaden einer Intervention abgeleitet werden."
→ „Schwellenwert" für Grenze Konfidenzintervall (z.B. für relatives Risiko von 2-5)

Wichtig für die frühe Nutzenbewertung:

1. Sensitivität der Ergebnisse, z.B. durch
   - Plausibilitätschecks
   - Confounding functions [2]
   - Rosenbaum bounds [3]
   - E-values (siehe nächster Vortrag)
   - U.v.m
2. Bias-Detektion (Überprüfung der Adjustierung)
3. Bias-Kontrolle/-Korrektur
4. **Quantifizierung einer ausreichenden Effektstärke?**

} Negativkontrollen

Viele methodische Ansätze sind verheißungsvoll und bereits länger in der Literatur beschrieben, jedoch gibt es noch keine Beispiele aus der frühen Nutzenbewertung



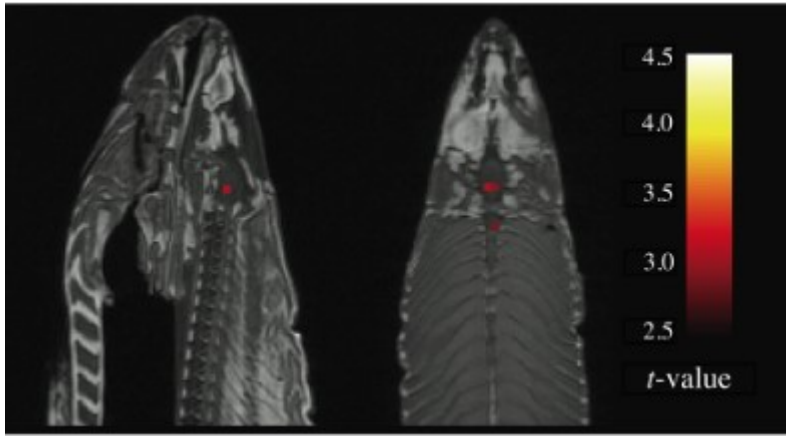
Verzerrte Effektschätzung

Adaptiert nach [1]

Negativkontrollen am Beispiel der „Dead Salmon Studie“ [4]

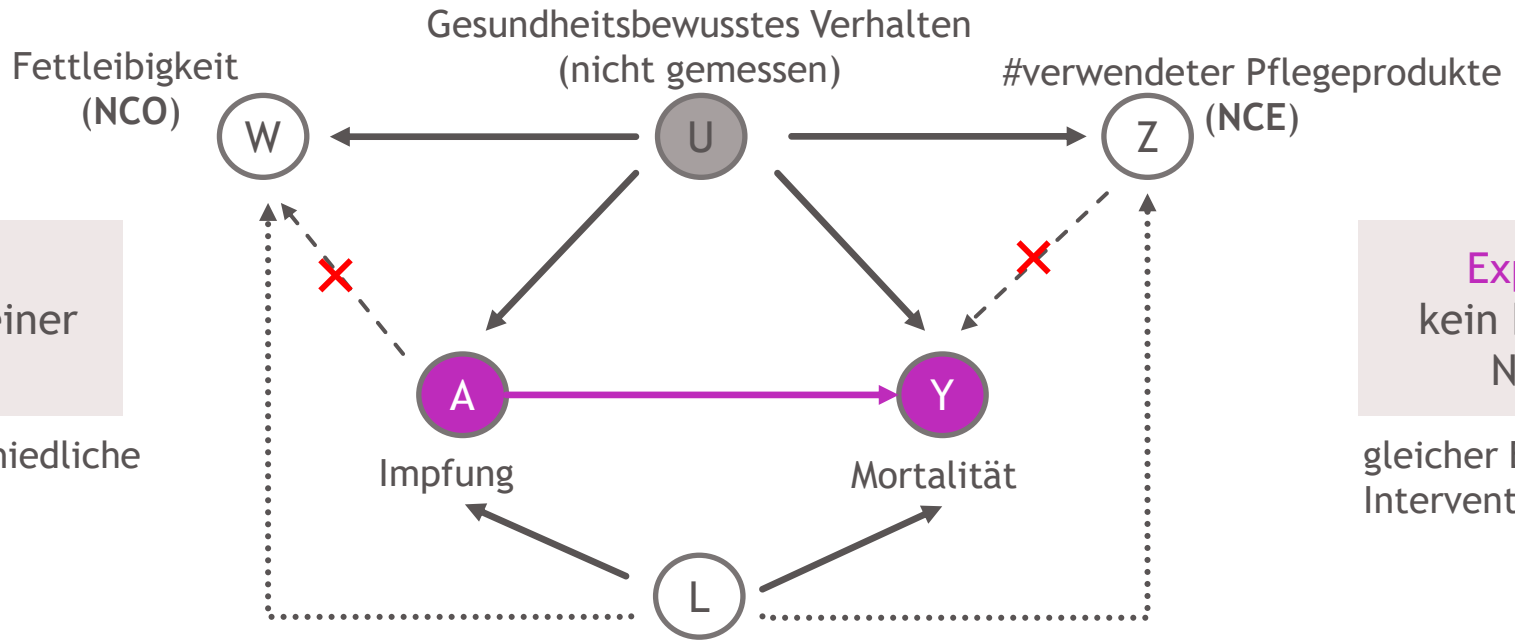
- **Intervention**: Bilder von Menschen in sozialen Situationen mit bestimmter Stimmung
- **Endpunkt**: Gehirnaktivität eines toten Lachses erhoben durch Magnetresonanztherapie

- Die Fähigkeit des toten Lachses menschliche Emotionen zu erkennen ist eine Negativkontrolle; es wird angenommen, dass die Nullhypothese (kein Effekt) wahr ist
- Die Studienergebnisse machen eine Aussage über die Qualität der Methode (und nicht des Subjekts!)
  - Scanner (kann Hintergrundrauschen generieren)
  - Statistische Analyse (kann das Ausmaß/die Art des Hintergrundrauschens nicht adäquat korrigieren)



Konzept der Negativkontrollen – kausaler Graph

Aktuelles Beispiel: Impfung (vereinfacht)



Endpunktkontrolle:
kein bekannter Effekt einer Impfung auf NCO

gleiche Intervention, unterschiedliche Endpunkte

Expositionskontrolle:
kein bekannter Effekt von NCE auf Mortalität

gleicher Endpunkt, unterschiedliche Interventionen

Annahme Negativkontrollen: (approximative) U-Vergleichbarkeit [5]

Bias Detektion: Beobachtung eines Null-Effekts von A → W und/oder Z→Y (nach Adjustierung für L), dann A→Y Assoziation vermutlich nicht durch Störgrößen verzerrt

Bias-Kontrolle am Beispiel der P-Wert Kalibrierung


Research Article
Statistics in Medicine
Received 12 November 2012, Accepted 3 July 2013 Published online 30 July 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sim.5925

Interpreting observational studies: why empirical calibration is needed to correct p-values

Martijn J. Schuemie,[a,b*†] Patrick B. Ryan,[b,c] William DuMouchel,[b,d] Marc A. Suchard[b,e] and David Madigan[b,f]

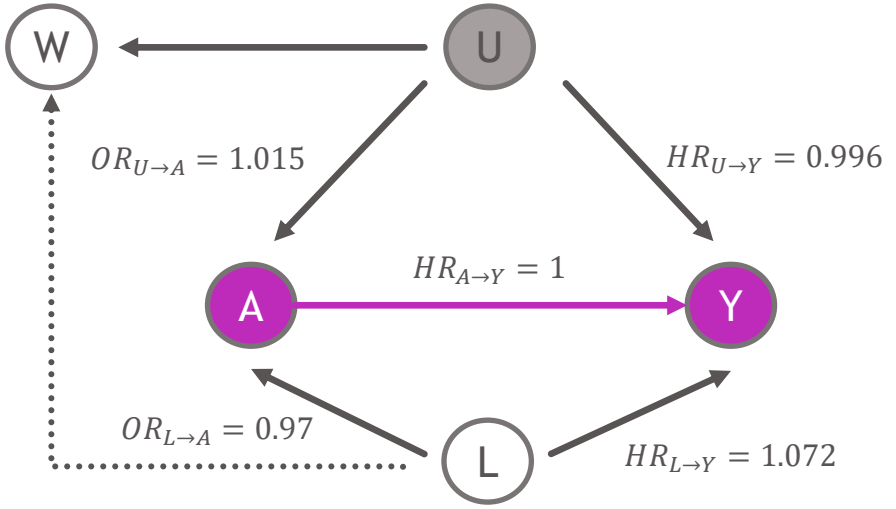
- Idee: Schätzung einer empirischen Null-Verteilung (für A→Y Effekt) unter Verwendung der geschätzten Effekte der Negativkontrollen
- Berücksichtigung von zufälligem und systematischen Fehler
- Ist der beobachtete Behandlungseffekt vor dem Hintergrund von Negativkontrollen auch noch signifikant?

Methodik zur Schätzung der empirischen Nullhypothese (Fokus: NCO) [6]

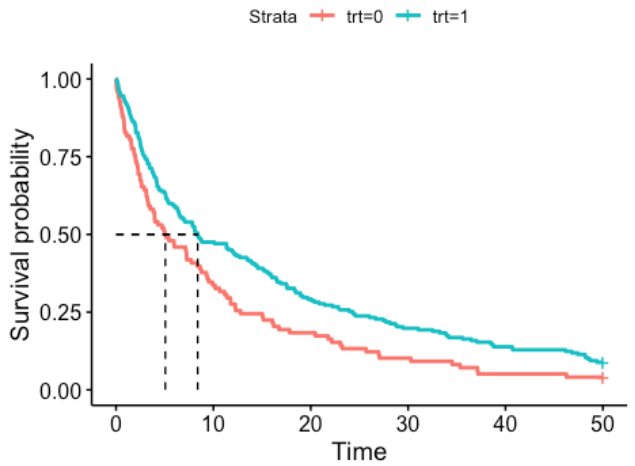
- y_i geschätzter log-Effekt (z.B. log HR, log RR) des i-ten Negativ-Interventions-Endpunkt Paares,
- τ_i dazugehöriger geschätzter Standardfehler
- θ_i wahre (aber unbekannte) Verzerrung des i-ten Negativ-Interventions-Endpunkt Paares
- σ^2 zufälliger Schätzfehler
- Analog zur Standard-p-Wert-Berechnung: $y_i \sim N(\theta_i, \tau_i^2)$
- Zusätzlich: $\theta_i \sim N(\mu, \sigma^2)$ → Null (Bias) Verteilung
- Schätzung von μ und σ^2 via Maximum-Likelihood
- Kalibrierter p-Wert für ein neues Interventions-Endpunkt Paar: $2 * \left(1 - \Phi\left(\frac{y_{n+1} - \hat{\mu}}{\sqrt{\hat{\sigma}^2 + \tau_{n+1}^2}}\right)\right)$

P-Wert Kalibrierung: Simuliertes Datenbeispiel

- Ereigniszeitanalyse, z.B. Vergleich von Krebstherapien auf OS
- 300 Patienten
- Kein Behandlungseffekt von A auf Y (HR=1)
- $L \sim \mathrm{N}(60,13)$
- $U \sim \mathrm{N}(460,100)$
- Logit Modell für A
- Baseline Hazard für Y ist 0.009
- 50 NCOs W, variierende Baseline Hazards sowie Effekte $L \rightarrow W$ und $U \rightarrow W$
- Cox-Modell für W und Y [7], administrative Zensierung bei $t = 50$

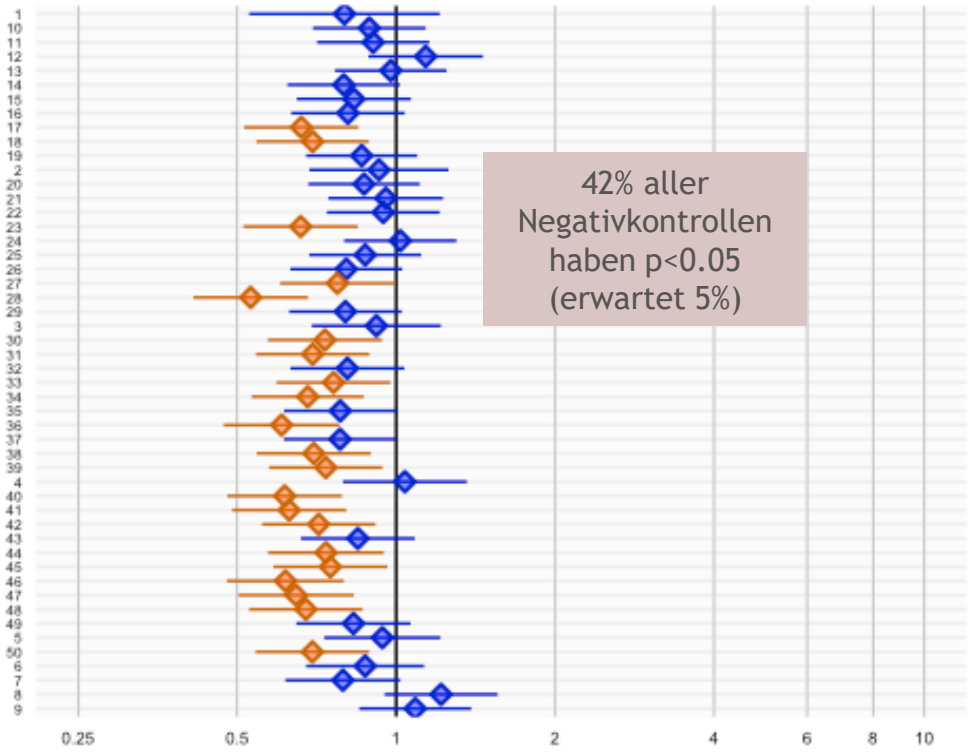


Ergebnis der nach L adjustierten Analyse: HR=0.76, 95% CI: 0.59- 0.97, p=0.03
→ Fälschlicherweise signifikanter Behandlungseffekt, da verzerrt durch nicht erhobene Störgröße U



P-Wert Kalibrierung: Verteilung der NCOs

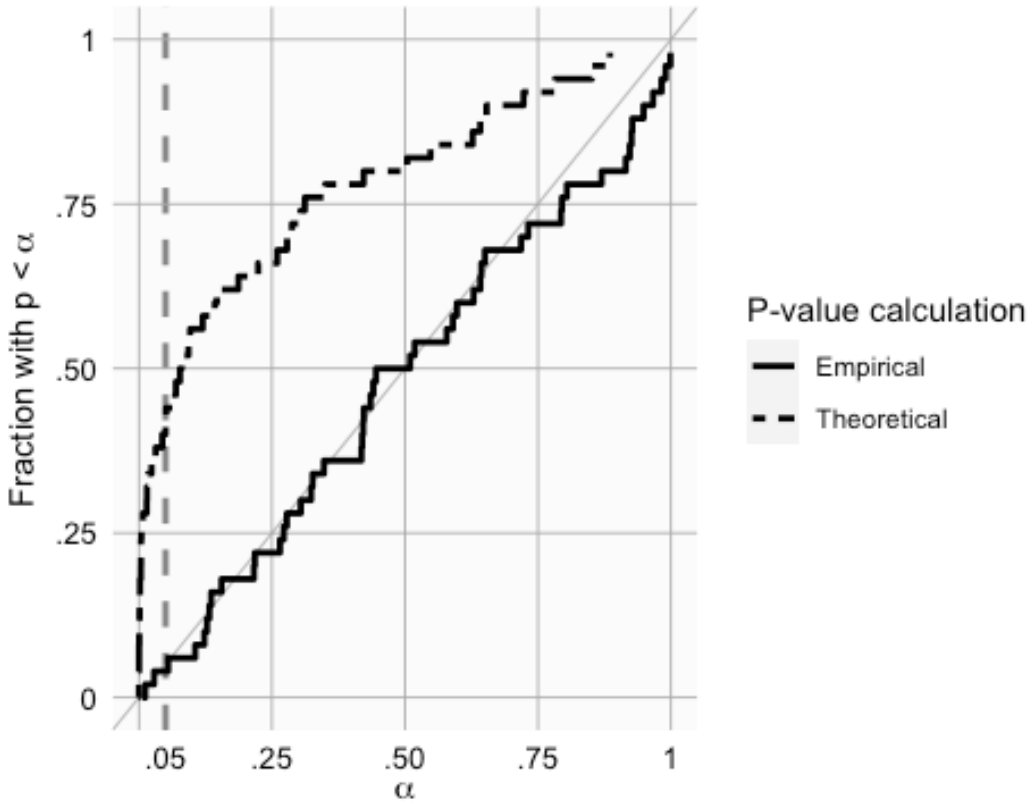
Forest Plot des (nach L adjustierten) Behandlungseffekts von A auf NCOs



42% aller Negativkontrollen haben p<0.05 (erwartet 5%)

Orange: Ablehnung von H_0 (HR=1)

Performance Check:
% Negativkontrollen mit p-Wert <α ist approx. α



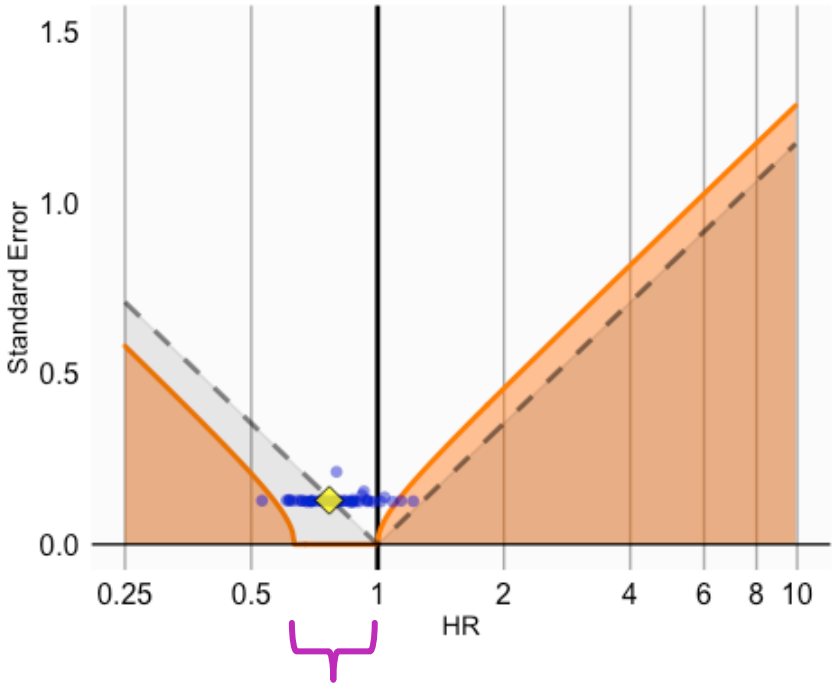
Empirische Null-(Bias-)Verteilung: $\theta_i \sim N(-0.23, 0.12)$

↓

Negative Verzerrung

P-Wert Kalibrierung: Simuliertes Datenbeispiel

- Blaue Punkte: HR (SE) der Negativkontrollen
- HR (SE) Schätzer unterhalb der grauen gestrichelten Linie besitzen traditionellen p-Wert<0.05
- HR (SE) Schätzer unterhalb der orangen Fläche besitzen einen kalibrierten p-Wert<0.05
- Unter dieser empirischen Nullverteilung kann die Nullhypothese für den Behandlungseffekt A→Y nicht abgelehnt werden: p=0.83>0.05



Es lässt sich ableiten, ab welcher Effektstärke - trotz möglichem Confounding - nach Kalibrierung noch signifikante Aussagen treffen lassen (mit ausreichender Präzision HR>0.2)

Diskussion

- Verwendung von Negativkontrollen um empirisch die zufällige und systematische Verzerrung in einer (nicht-randomisierten) Studie nach Adjustierung für bekannte (und erhobene) Störgrößen zu evaluieren (idealerweise: $\hat{\mu} = \hat{\sigma}^2 = 0$)
- Bias-Detektion & Bias-Korrektur (von p-Werten) → R-Paket auf CRAN verfügbar [6]
- Struktur und Herkunft der Verzerrung der Negativkontrollen nicht notwendigerweise identisch; Verzerrung sollte einer gemeinsamen (Normal-)Verteilung entstammen

  Ziel: Korrigieren des Typ I Fehlers (fälschliche Ablehnung von H_0) durch Kalibrierung, möglicherweise zu Kosten der Power [6,8,9-11]

- Ansatz lässt sich auf Kalibrierung von Konfidenzintervallen übertragen [12]
- Kalibrierung als supportive Analyse um ausreichenden Effektstärke zu quantifizieren?
- Annahmen: U-Vergleichbarkeit (vergleichbare Menge an Ursachen von A und W/Z), insb. Verzerrung der Negativkontrollen und zu interessierendem Effekt gleichermaßen normalverteilt
- Herausforderung: Identifikation & Verfügbarkeit von passenden & ausreichend vielen Negativkontrollen

Literaturverzeichnis

[1] Rush et al. (2018) Eur Heart J, 39(37): 3417-3438

[2] Kasza et al. (2017), Int J Epi 46(4):1303-1311

[3] Liu et al. (2013), Prev Sci 14(6): 570-580

[4] Bennet et al. (2009), J Serend Unexp Res 1:1–5

[5] Lipsitch et al. (2010), Epidemiology 21(3): 383-388

[6] Schuemie et al.(2014), Statist Med 33:209-218

[7] Bender et al. (2005), Statist Med 24(11):1713-1723

[8] Veronesi et al. (2020), Int J Epi 49(3):876-884

[9] Gruber & Tchetgen Tchetchen (2016), Statist Med, 35:3869-3882

[10] Franklin (2016), Statist Med, 35:3889-3894

[11] Schuemie et al.(2016), Statist Med 35:3883-3888

[12] Schuemie et al. (2018), PNAS, 115(11):2571-2577


Bristol Myers Squibb™