Gibt es einen Unterschied zwischen randomisierten und nicht-randomisierten Studien in ähnlichen Patientengruppen? - Evidenz aus einer „Meta-Propensity Score-Analyse“ in der Herzchirurgie

Kuß O[1], Legler T[1], Börgermann J[2]

[1]Institut für Medizinische Epidemiologie, Biometrie und Informatik, Martin-Luther Universität Halle-Wittenberg, Halle (Saale)

[2]Herz- und Diabeteszentrum Nordrhein-Westfalen, Universitätsklinik der Ruhr-Universität Bochum, Bad Oeynhausen

Inhalt

- Exkurs: Propensity Score-Analyse

- Einleitung
- Material und Methoden
- Ergebnisse
- Diskussion

Propensity Score-Analyse I

- **Spezielle Methode zur Auswertung nicht-randomisierter Therapie-Studien**

- "Erfinder": Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. 1983;70:41-55.

- Hinweise, dass PS-Analysen herkömmlichen stat. Methoden (Regressionsanalyse) überlegen sind (Cook et al., 1989; Robinson et al, 1991; Cepeda et al. 2003; Martens et al., 2008)

Propensity Score-Analyse II

- **Definition:** Der Propensity Score ist die (bedingte) Wahrscheinlichkeit, behandelt zu werden (bedingt auf die gemessenen Kovariablen)

- **Zwei-Schritte:**

  1. Schritt: Logistische Regression mit Behandlung (!) als Zielgröße. Schätze PS.

  2. Schritt: Schätze Behandlungseffekt „unter Berücksichtigung“* des PS.

  *:
  1. Matching
  2. Regressionsadjustierung
  3. Stratifikation
  4. IPW

Propensity Score-Analyse III

- **Matching:** Matche Patienten mit gleichem PS, von den einer behandelt wurde, der andere nicht.

  Wenn beide Patienten die gleiche Behandlungsw‘t haben, aber nur einer von beiden behandelt wurde, dann ist es egal, welcher behandelt wurde!

- RCTs sind auch PS-Analysen (zumindest ein Spezialfall): PS bekannt und PS=0.5

Propensity Score-Analyse IV

TABLE 1. Preoperative data

	Off pump (n = 597)	On pump (n = 597)	*P* value
Age (y)	62.7 ± 9.2	62.4 ± 8.6	.552
≥75 y	52 (8.7%)	37 (6.2%)	.098
Female sex	84 (14.1%)	82 (13.7%)	.867
Diabetes	133 (22.3%)	150 (25.1%)	.247
Preoperative AMI	275 (46.1%)	296 (49.6%)	.224
ECV	135 (22.6%)	142 (23.8%)	.631
Redo	10 (1.7%)	20 (3.4%)	.064
Unstable angina	182 (30.5%)	159 (26.6%)	.141
Urgency	113 (18.9%)	112 (18.8%)	.941
LM	98 (16.4%)	118 (19.8%)	.133
2-Vessel disease	359 (60.1%)	371 (62.1%)	.543
3-Vessel disease	238 (39.9%)	226 (37.9%)	.543
EF (%)	58.8 ± 12.2	57.8 ± 18.8	.345
≤35%	24 (4.0%)	27 (4.5%)	.668
Logistic EuroSCORE	4.3%	4.1%	.458

AMI, Acute myocardial infarction; *ECV*, extracardiac vasculopathy; *LM*, left main; *EF*, ejection fraction.

Calafiore AM et al, J Thorac Cardiov Sur 2005;130:340-5.

Propensity Score-Analyse V

TABLE 1. Preoperative data

	Off pump (n = 597)	On pump (n = 597)	*P* value
Age (y)	62.7 ± 9.2	62.4 ± 8.6	.552
3-vessel disease	236 (39.5%)	226 (37.9%)	.543
EF (%)	58.8 ± 12.2	57.8 ± 18.8	.345
≤35%	24 (4.0%)	27 (4.5%)	.668
Logistic EuroSCORE	4.3%	4.1%	.458

AMI, Acute myocardial infarction; *ECV*, extracardiac vasculopathy; *LM*, left main; *EF*, ejection fraction.

ACHTUNG:
Balanciertheit nur für bekannte, gemessene Kovariablen, die im PS-Modell waren!

Calafiore AM et al, J Thorac Cardiov Sur 2005;130:340-5.

Propensity Score-Analyse VI

TABLE 1. Preoperative data

	Off pump (n = 597)	On pump (n = 597)	*P* value
Age (y)	62.7 ± 9.2	62.4 ± 8.6	.552
≥75 y	52 (8.7%)	37 (6.2%)	.098
Female sex	84 (14.1%)	82 (13.7%)	.867
Diabetes	133 (22.3%)	150 (25.1%)	.247
Preoperative AMI	275 (46.1%)	296 (49.6%)	.224
ECV	135 (22.6%)	142 (23.8%)	.631
Redo	10 (1.7%)	20 (3.4%)	.064
Unstable angina	182 (30.5%)	159 (26.6%)	.141
Urgency	113 (18.9%)	112 (18.8%)	.941
LM	98 (16.4%)	118 (19.8%)	.133
2-Vessel disease	359 (60.1%)	371 (62.1%)	.543
3-Vessel disease	238 (39.9%)	226 (37.9%)	.543
EF (%)	58.8 ± 12.2	57.8 ± 18.8	.345
≤35%	24 (4.0%)	27 (4.5%)	.668
Logistic EuroSCORE	4.3%	4.1%	.458
Unknown Killer-Gene	**249 (50%)**	**0 (0%)**	**<0.0001**
Unmeasured Killer-Virus	**478 (80%)**	**120(20%)**	**<0.0001**

Calafiore AM et al, J Thorac Cardiov Sur 2005;130:340-5.

Propensity Score-Analyse VII

- Cook EF, Goldman L. Performance of Tests of Significance Based on Stratification by A Multivariate Confounder Score Or by A Propensity Score. Journal of Clinical Epidemiology 1989;42(4):317-24.
- Robinson LD, Jewell NP. Some Surprising Results About Covariate Adjustment in Logistic-Regression Models. International Statistical Review 1991 August;59(2):227-40.
- Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. American Journal of Epidemiology 2003 August 1;158(3):280-7.
- Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. International Journal of Epidemiology 2008 Oct;37(5):1142-7.

Gibt es einen Unterschied zwischen randomisierten und nicht-randomisierten Studien in ähnlichen Patientengruppen? - Evidenz aus einer „Meta-Propensity Score-Analyse“ in der Herzchirurgie

Kuß O[1], Legler T[1], Börgermann J[2]

[1]Institut für Medizinische Epidemiologie, Biometrie und Informatik, Martin-Luther Universität Halle-Wittenberg, Halle (Saale)
[2]Klinik für Herz- und Thoraxchirurgie, Universitätsklinikum Jena

Einleitung I: RCTs und Non-RCTs

- Effekte von therapeutischen Interventionen sollten, wenn möglich, in randomisierten kontrollierten Studien (RCTs) geprüft werden.

- RCTs haben manchmal eine geringe **externe** Validität (Rothwell, 2005).

- **Konsequenz:** Mangelnde **interne** Validität aller systematischen Vergleiche von randomisierten und nicht-randomisierten Studien!

Einleitung II: RCTs und Non-RCTs

Konkret:

Wenn **RCTs** in **hoch-selektionierten** Populationen durchgeführt werden,
Non-RCTs dagegen in **unselektionierten** Populationen,
dann sind Unterschiede nicht notwendigerweise auf die **fehlende Randomisierung** zurückzuführen.

Sie könnten auch durch die **unterschiedlichen Patientenpopulationen** zustande kommen!

Einleitung III: „Meta-Randomisierung“?

- Idealerweise sollte eine „meta-randomisierte“ Studie durchgeführt werden:
  Studiengruppen, die eine Studie zu einer bestimmten klinischen Frage durchzuführen bereit sind, werden zufällig ausgewählt („meta-randomisiert“), eine randomisierte oder eine nicht-randomisierte Studie durchzuführen.

  → Strukturgleichheit bzgl. aller „Meta-Confounder“
  → Kausaler Effekt der Randomisierung ist intern valide messbar.

- Technisch möglich? Ethisch akzeptabel?

Einleitung IV: „Meta-Propensity Score“!

- Unsere Lösung: Gematchte „Meta-Propensity Score-Analyse”

  1. Matche RCTs und Non-RCTs für wichtige „Meta-confounder“ (zusammengefasst durch einen „Meta-Propensity Score“)
  2. Vergleiche Effektschätzer in der “meta-gematchten” Stichprobe

  Einschränkung: In der Gruppe der Non-RCTs werden nur Propensity-Score Analysen betrachtet.

Einleitung V: Klinische Fragestellung

- Vergleich zweier OP-Techniken (mit und ohne Einsatz der Herz-Lungen-Maschine, On- und Off-pump) in der Bypasschirurgie

- " … one of the most hotly debated and polarizing issues in cardiac surgery …" (Sellke et al., 2005).

- Public health-Relevanz: In Deutschland wurden im Jahr 2007 49.788 (isolierte) Bypass-OPs durchgeführt, 10,1% davon „Off-pump" (Gummert et al., 2008).

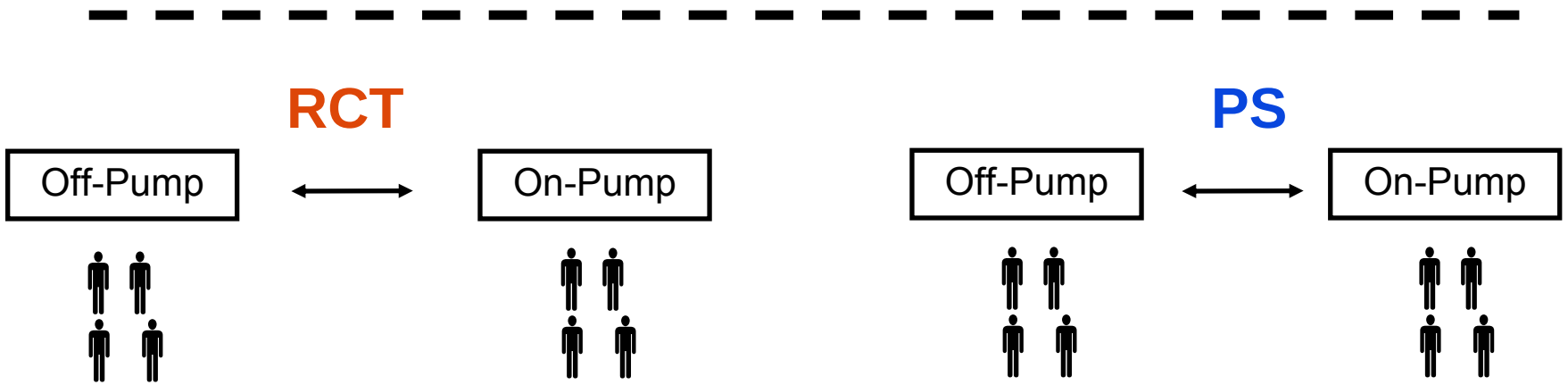
Einleitung VI: „Meta-Propensity Score“ – Versuch einer Erklärung

„Meta“-Ebene

Inhaltlich-Klinische Ebene

Einleitung VI: „Meta-Propensity Score“ – Versuch einer Erklärung

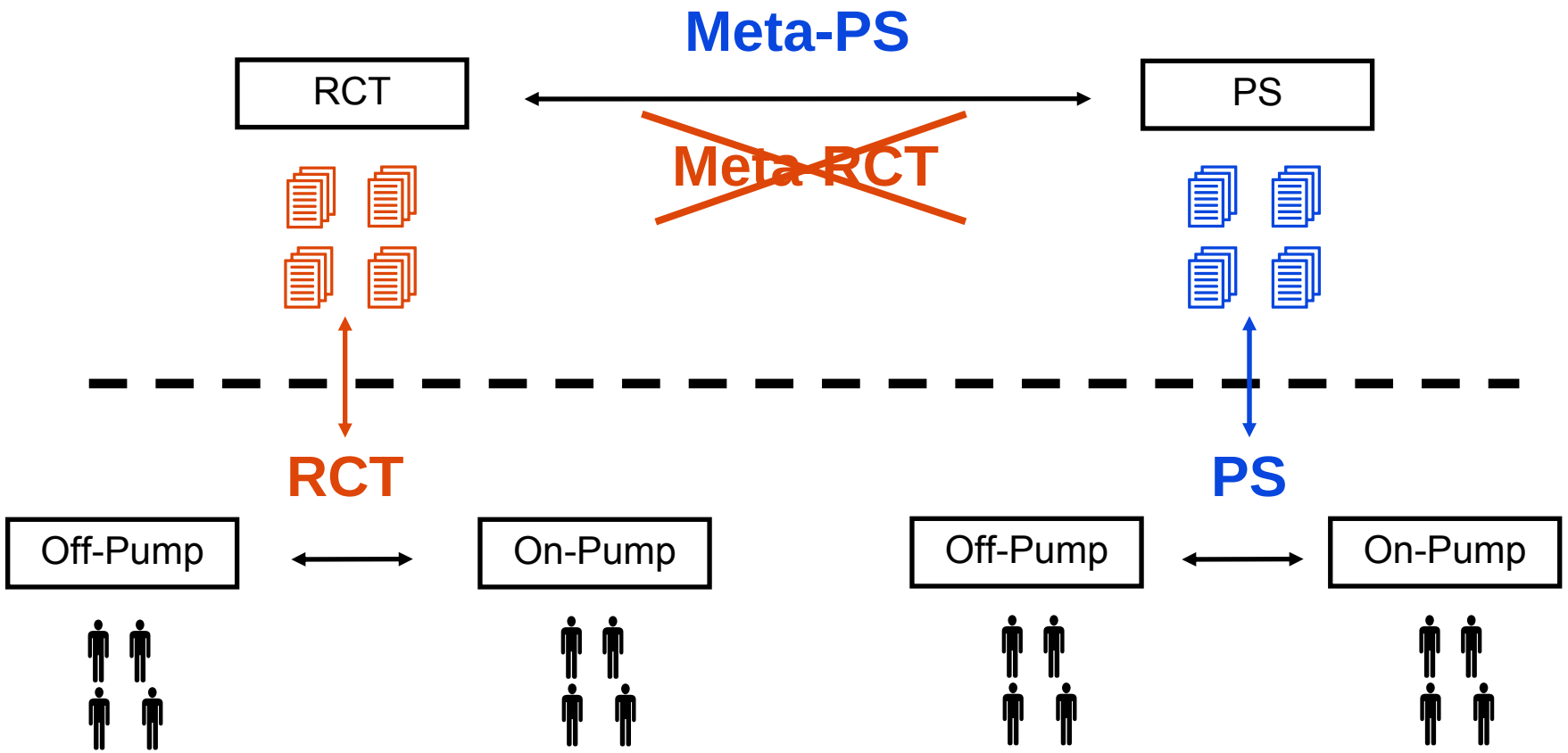
„Meta“-Ebene



Inhaltlich-Klinische Ebene

Einleitung VI: „Meta-Propensity Score“ – Versuch einer Erklärung

„Meta“-Ebene



Inhaltlich-Klinische Ebene

Material und Methoden I: Studien

- Systematische Suche nach allen RCTs (alle RCTs aus den 3 größten und aktuellsten MAs von RCT + eigene MEDLINE-Suche)
- Systematische Suche nach allen PS-Analysen (Kuss et al., 2009(?))
- Einschlusskriterien für Studien:
  - Information zu Studien-Population und anderen “Meta-Confoundern”
  - Information zu mindestens einer von 10 dichotomen klinischen In-Hospital Zielgrößen (Postoperativer Tod, Schlaganfall, Herzinfarkt, Nierenversagen, …)

Material und Methoden II: Studien

- Strukturierte Datenerfassung (pilot-gestestetes CRF, zwei verblindete Reviewer [TL,OK], evtl. Konsensfindung mit drittem Reviewer [JB])

- Extrahierte Daten:
  - Allgemeine Angaben zur Studie (Studiendauer, Anzahl Zentren, Anzahl Patienten, Land, …)
  - Studienpopulation (Risikofaktoren)
  - Zielgrößen (Abs. Häufigkeiten oder Effektschätzer)

Material und Methoden III: „Meta-PS-Analysis“

- Einschlusskriterium für „Meta-Confounder“: Information in mindestens 2/3 aller RCTs und PS-Analysen
- Vereinfachende Annahme: Mittelwert = Median
- Falls notwendig: Transformation von kategorialen „Meta-Confoundern“ in stetige unter Annahme von Gleichverteilung in den Kategorien
- Multiple Imputation von fehlenden Werten im „Meta-PS-Modell“ (SAS® PROC MI, 1000 Datensätze)
- „Meta-PS-Modell“ als logistisches Modell mit stetigen „Meta-Confoundern“ und deren polynomiale . Transformationen bis zur dritten Ordnung, keine Interaktionsterme, Optimalitätskriterium: c-Statistik (= 89,6%)

Material und Methoden IV: „Meta-PS-Analysis“

- „Meta-Matching“ mit medianem (über die 1000 imputierten Datensätze) „Meta-PS“ und einem optimalen Matching-Algorithmus mit einer variablen Anzahl (höchstens 1:4) von Kontrollen (Soledad Cepeda et al., 2006)

- **Wichtig:** Festlegung des „Meta-PS-Modells“ **vor Auswertung** und **unabhängig von den Zielgrößen**

Material und Methoden V: Auswertung

- Unterschiede zwischen RCTs und PS-Analysen werden im „meta-gematchten“ Datensatz als Differenzen von „Meta-Odds ratios“ (mit 95%-KI, multivariate Delta-Methode) angegeben.
- „Rekonstruktion“ von Vierfeldertafeln in den PS-Analysen mit der Di Pietrantonj-Methode (2006, Umrechnung von Effektschätzern in zugrundeliegende Vierfeldertafeln).
- **Statistisches Modell:** Logistische Regression mit zufälligen Effekten
  Für einzelne Zielgrößen: Zwei zufällige Effekte (Studie, Matching-Stratum)
  Für Gesamteffekt: Zusätzlicher zufälliger Effekt Zielgröße
- Parameterschätzung durch PQL (SAS® PROC GLIMMIX)

Ergebnisse I: Studien

- Gefundene und eingeschlossene Studien: 69 Studien, davon 28 PS-Analysen und 51 RCTs
- 7 „Meta-Confounder“ mit Information in mindestens 2/3 aller RCTs und PS-Analysen
- Nach „Meta-Matching“: 39 Studien, davon 10 PS-Analysen (25.552 Patienten) und 29 RCTs (2.723 Patienten)
- Insgesamt 186 Effektschätzer aus den 39 Studien für die klinischen Zielgrößen:
  Postoperativer Tod: 38, Schlaganfall: 28, Herzinfarkt: 27, Nierenversagen: 12, …

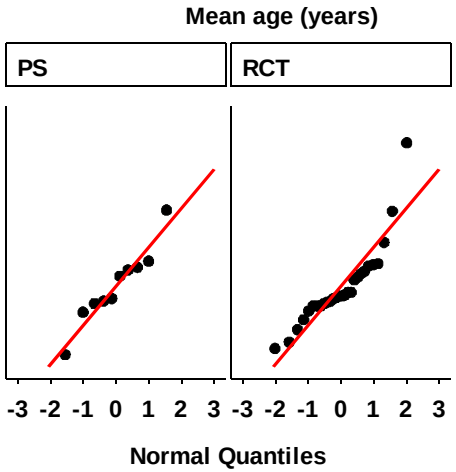
Ergebnisse II: Studien *vor* dem „Meta-Matching“

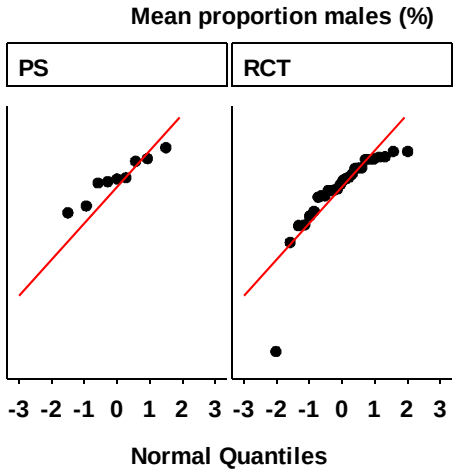
Meta-Confounder	PS-Analysen (N=28)	RCTs (N=51)	p-Wert	Standard. Diff. (%)
Studienregion			0.007	
Europa	17 (61%)	36 (71%)		
Nordamerika	10 (36%)	5 (10%)		
Sonstige	1 (3%)	10 (19%)		
Anzahl Zentren				
1	18 (65%)	47 (92%)	0.006	
>1	9 (32%)	3 (6%)		
Fehlend	1 (3%)	1 (2%)		
Mittl. Alter (Jahre)	65.8	63.1	0.002	75.1
Mittl. Anteil Männer (%)	72.1	77.1	0.138	-37.0
Mittl. Anteil prä-op. Herzinf.(%)	44.5	41.6	0.480	21.0
Mittl. LVEF (%)	58.8	62.7	0.033	-55.9
Mittl. Anteil Diabetiker (%)	26.2	24.4	0.595	13.9

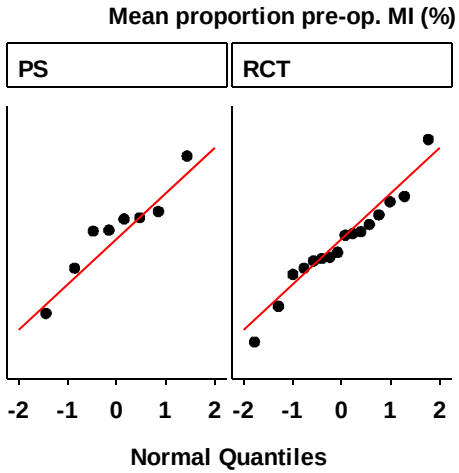
Ergebnisse III: Studien *nach* dem „Meta-Matching“

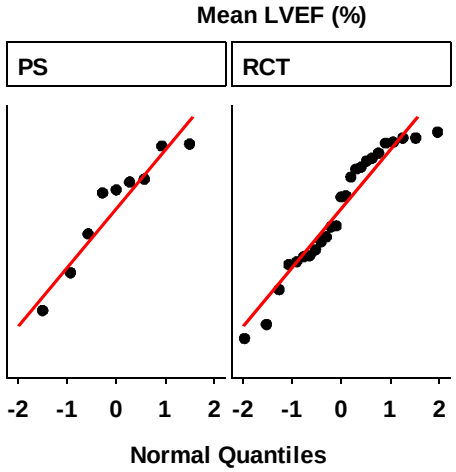
Meta-Confounder	PS-Analysen (N=10)	RCTs (N=29)	p-Wert	Standard. Diff. (%)
Studienregion			0.999	
Europa	8 (80%)	23 (80%)		
Nordamerika	1 (10%)	3 (10%)		
Sonstige	1 (10%)	3 (10%)		
Anzahl Zentren				
1	8 (80%)	25 (86%)	0.631	
>1	2 (20%)	3 (10%)		
Fehlend	0 (0%)	1 (4%)		
Mittl. Alter (Jahre)	64.1	63.9	0.916	3.9
Mittl. Anteil Männer (%)	80.5	76.9	0.431	30.5
Mittl. Anteil prä-op. Herzinf.(%)	44.0	39.9	0.530	27.6
Mittl. LVEF (%)	61.1	60.7	0.861	6.8
Mittl. Anteil Diabetiker (%)	24.8	25.2	0.925	-3.7

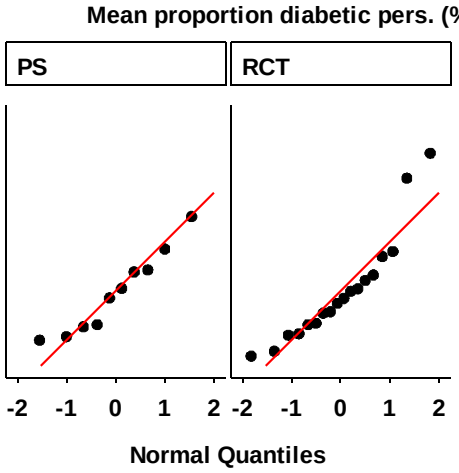
Ergebnisse IV: Verteilungen (Q-Q-Plots) der stetigen „Meta-Confounder“







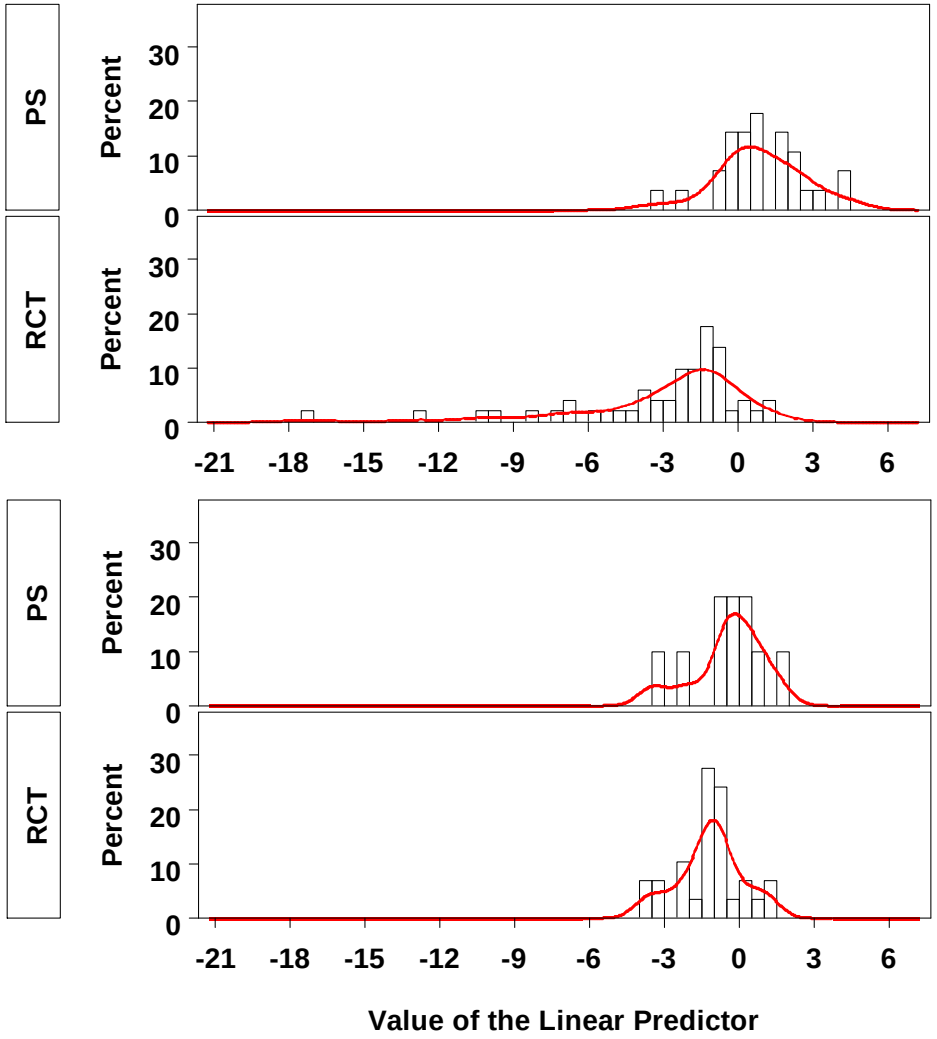




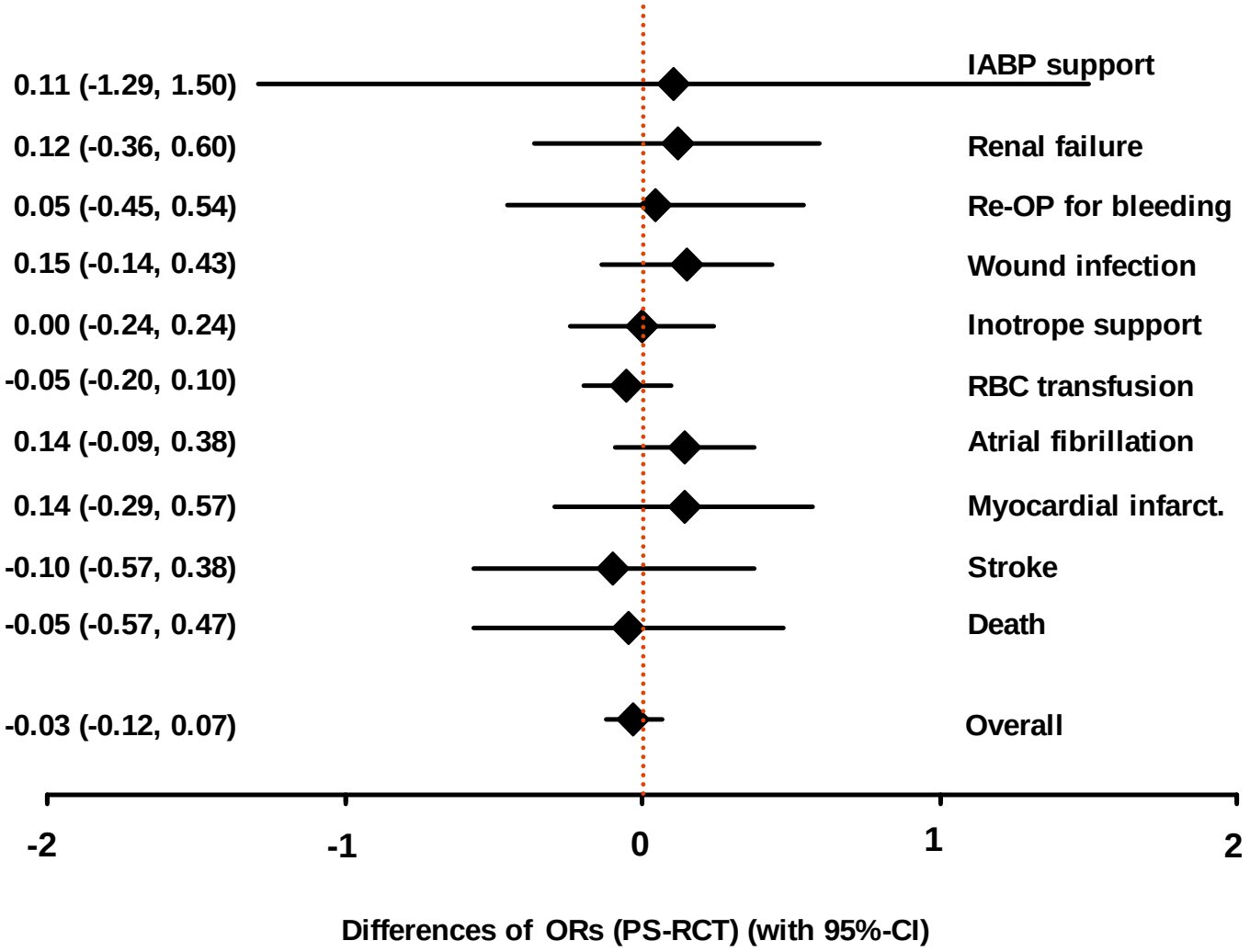
Ergebnisse V: Verteilung des linearen „Meta-Prädiktors“ vor und nach „Meta-Matching“

vorher

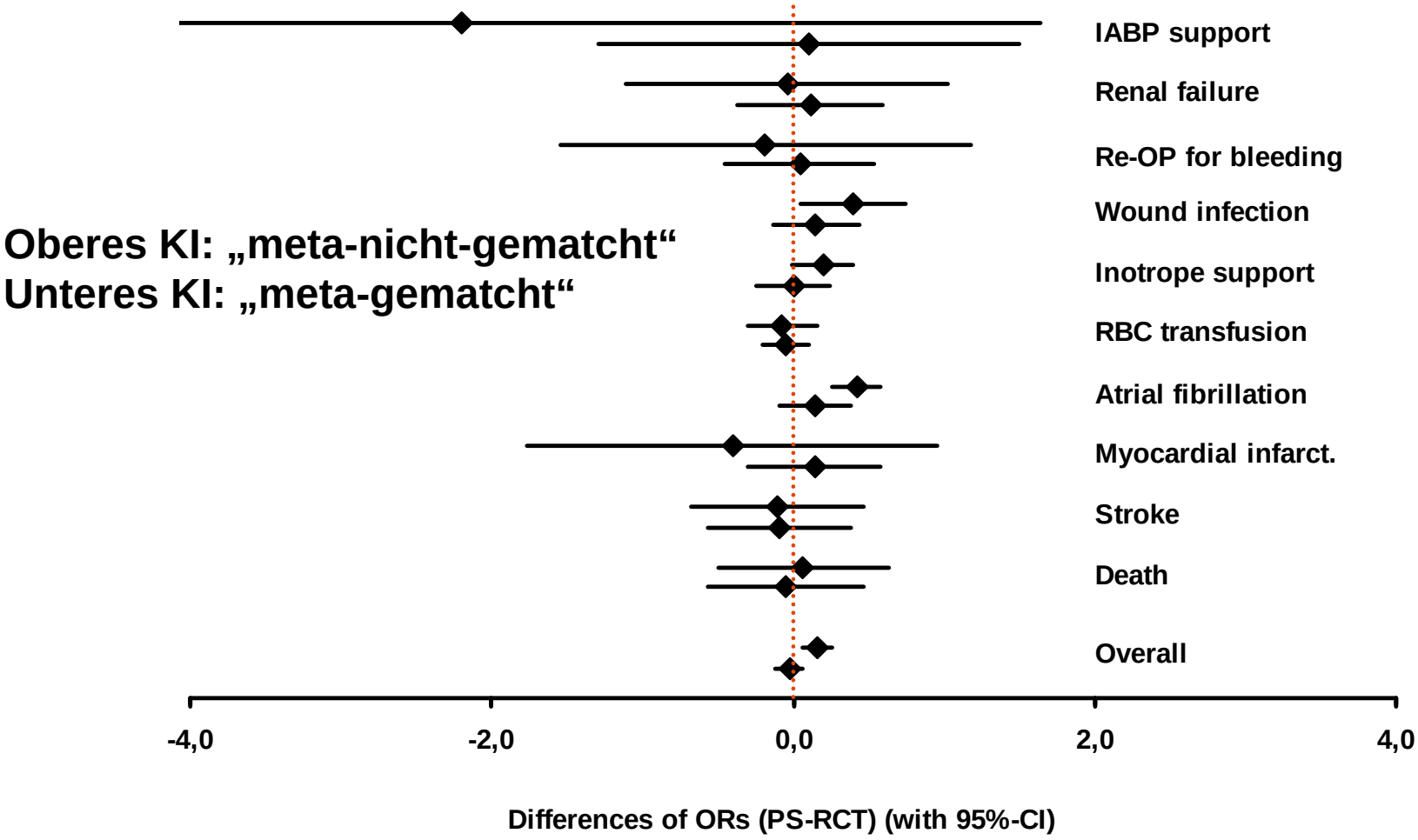
nachher



Ergebnisse VI: Differenzen der „Meta-ORs“(PS-RCT) im „meta-gematchten“ Datensatz



Ergebnisse VII: Differenzen der „Meta-ORs“(PS-RCT) im „meta-gematchten“ Datensatz und im „meta-nicht-gematchten“ Datensatz



Diskussion I

- In unserem Beispiel aus der Herzchirurgie waren die Unterschiede zwischen RCTs und PS-Analysen mit ähnlichen „meta-gematchten“Populationen marginal: (Globale Diff. in Meta-ORs [95%-CI]: -0.027 [-0.119, 0.066]) → **Kleiner Effekt der Randomisierung**
- Das ist nicht überraschend, sondern bestätigt vorliegende Evidenz (Noel et al., 1998; King et al., 2005; Vist et al., 2005; Hernan et al., 2008; Shadish et al., 2008; Furlan et al., 2008; Tannen et al., 2009).

Diskussion I

- In unserem Beispiel aus der Herzchirurgie waren die Unterschiede zwischen RCTs und PS-Analysen mit ähnlichen „meta-gematchten“Populationen marginal: (Globale Diff. in Meta-ORs [95%-CI]: -0.027 [-0.119, 0.066]) → **Kleiner Effekt der Randomisierung**
- Das ist nicht überraschend, sondern bestätigt vorliegende Evidenz (Noel et al., 1998; King et al., 2005; Vist et al., 2005; Hernan et al., 2008; Shadish et al., 2008; Furlan et al., 2008; Tannen et al., 2009).

Furlan et al., 2008: “…homogeneity in terms of settings, population, interventions, and outcomes predicts the agreement between an NRS (non-randomised study) and an RCT of the same intervention …”.

Diskussion I

- In unserem Beispiel aus der Herzchirurgie waren die Unterschiede zwischen RCTs und PS-Analysen mit ähnlichen „meta-gematchten“Populationen marginal: (Globale Diff. in Meta-ORs [95%-CI]: -0.027 [-0.119, 0.066]) → **Kleiner Effekt der Randomisierung**
- Das ist nicht überraschend, sondern bestätigt vorliegende Evidenz (Noel et al., 1998; King et al., 2005; Vist et al., 2005; Hernan et al., 2008; Shadish et al., 2008; Furlan et al., 2008; Tannen et al., 2009).

King at al., 2005: “ …differences in outcome across the trials between randomized and preference groups were generally small, particularly in large trials and after accounting for baseline measures of outcome. Therefore, there was little evidence that preferences substantially interfere with the internal validity of randomized trials”

Diskussion II

- Weitere **Vorteile** unserer Studie, die für eine gute Vergleichbarkeit von RCTs und Non-RCTs sorgen:
  - Identisches Design der Non-RCTs (PS)
  - Identisches Intervention und Kontrolle
  - Identische Zielgrößen in RCTs und Non-RCTs
  - Identische Follow-Up-Länge
  - Valide klinische Outcomes
  - RCTs und Non-RCTs in ähnlichen Zeiträumen durchgeführt

Diskussion III

- **Gefahren/Nachteile:**
  - Publication bias?
  - Vereinfachende Annahmen zu simpel?
  - „Meta-Residual Confounding?“
    (Zur Erinnerung: Eigentlich brauchen wir eine „meta-randomisierte Studie“!)
  - Zu wenig (n=7) „Meta-Confounder“?

Diskussion IV

- **In der Zukunft:**
  Unsere Studie sollte unabhängig repliziert werden, am besten in einem anderen klinischen Bereich.

  Auch nach einer Replikation werden RCTs nicht obsolet werden!

  Allerdings: Die gegenwärtige Praxis, gut gemachte Non-RCTs (besser: PS-Analysen) aus systematischen Reviews für Therapieeffekte auszuschließen, sollte auf den Prüfstand!

Literatur

- Rothwell PM. External validity of randomised controlled trials: 'to whom do the results of this trial apply?' *Lancet 2005;* 365(9453):82-93.
- Sellke FW, DiMaio JM, Caplan LR et al. Comparing on-pump and off-pump coronary artery bypass grafting: numerous studies but few conclusions: a scientific statement from the American Heart Association council on cardiovascular surgery and anesthesia in collaboration with the interdisciplinary working group on quality of care and outcomes research. *Circulation*. 2005;111:2858-2864.
- Gummert JF, Funkat A, Beckmann A et al. Cardiac surgery in Germany during 2007: a report on behalf of the German Society for Thoracic and Cardiovascular Surgery. *Thorac Cardiovasc Surg* 2008;56(6):328-36.
- Kuss O, von Salviati B, Börgermann J. Off-pump versus On-Pump in Coronary Artery Bypass Grafting: A Systematic Review and Meta-Analysis of Propensity Score Analyses. In preparation.
- Soledad Cepeda M, Boston R, Farrar JT, Strom BL. Optimal matching with a variable number of controls vs. a fixed number of controls for a cohort study: trade-offs. *J Clin Epidemiol*. 2003;56(3):230-7.
- Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. *Statist Med* 2006; 25:2299–2322.
- Noel PH, Larme AC, Meyer J, Marsh G, Correa A, Pugh JA. Patient choice in diabetes education curriculum. Nutritional versus standard content for type 2 diabetes. *Diabetes Care* 1998;21(6):896-901.
- King M, Nazareth I, Lampe F et al. Impact of participant and physician intervention preferences on randomized trials: a systematic review. *JAMA* 2005;293(9):1089-99.
- Vist GE, Hagen KB, Devereaux PJ, Bryant D, Kristoffersen DT, Oxman AD. Systematic review to determine whether participation in a trial influences outcome. *BMJ* 2005;330(7501):1175.
- Hernan MA, Alonso A, Logan R et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 2008;19(6):766-79.
- Shadish WR, Clark MH, Steiner PM. Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *JASA* 2008;103(484):1334-43.

Literatur

- Furlan AD, Tomlinson G, Jadad AA, Bombardier C. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. J Clin Epidemiol 2008;61(3):209-31.
- Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. BMJ 2009;338:b81.