FACHBEREICH MEDIZIN FRANKFURT
GOETHE-UNIVERSITÄT

# Von „Real-world Data“ zu „Real-world Evidence“ – Methodische Grundvoraussetzungen

XV. Expertengespräch am 6. November 2020

im Gedenken an Herrn Dr. Holger Maria Rohde

Annette Lehn, lehn@med.uni-frankfurt.de

# Inhalt

- Motivationshistorie
- Motivationsbeispiel
- Klassische Auswertung
- Alternative Auswertung
- Propensity-Score-Methode
- Ausblick

# Motivationshistorie

## Goldstandard – Randomisierte kontrollierte klinische Studien:

- Randomisierung:
  - ⇒ Strukturgleichheit innerhalb der Behandlungsgruppen, d.h. gleichmäßige Verteilung (Balanciertheit) aller bekannten und unbekannten „Störgrößen“.
  - ⇒ Behandlungsgruppen sind vergleichbar bzgl. möglicher Einflussgrößen (z.B. Alter und Geschlecht).
- Die Wahrscheinlichkeit der Gruppenzugehörigkeit zu den beiden Behandlungen ist $\frac{1}{2}$.

⇒ Ein direkter Vergleich der Gruppen und damit eine kausale Aussage zum Behandlungseffekt möglich.

⇒ Der Goldstandard spiegelt eher Idealbedingungen wider.

# Motivationshistorie

Es gibt auch Kritik an randomisierten kontrollierten klinischen Studien:

Fehlende „externe Validität“ – Verallgemeinerbarkeit (Windeler 2008):

- Schlussfolgerungen aus den Ergebnissen sind für die spezielle Studienpopulation valide und reliabel ⇒ „interne Validität“, aber die Ergebnisse lassen sich in einigen Fällen schlecht auf die Realität übertragen.

Beispiel:

- Studienpatienten mittleren Alters haben einen großen untersuchten Behandlungseffekt.
  Ist dann auch derselbe Effekt unter derselben Behandlung in einer älteren Person außerhalb der Studie zu erwarten?

⇒ Dieser Kritik können sich allerdings auch nicht-randomisierte klinische Studien nicht entziehen.

# Motivationshistorie

Beispiel zur Überprüfung der Wirksamkeit von Penicillin bei bakteriellen Infektionen:

- Eine offensichtlich wirksame Behandlung wird gegen Placebo verglichen, obwohl es dabei sicher keine unbekannten „Störgrößen“ gibt, die einen nennenswerten Einfluss auf den Behandlungseffekt haben könnten.
- ⇒ Die Durchführung solcher randomisierter kontrollierter Studien werden als „unnötig“ erachtet (Black 1996).

Beispiel zur Auswahl von Zielgrößen:

- Häufige Verwendung von Surrogatendpunkten, die eine Aussage zum Effekt in einer kurzen Studienlaufzeit liefern.
  Aber Unterschiede aus patientenrelevanten Endpunkten zeigen sich erst nach längerer Untersuchungszeit und werden daher selten herangezogen.

# Motivationshistorie

Beispiel zu ethischen Aspekten:

- Keine Ethikkommission würde eine zufällige Einteilung von Patienten zur Behandlung auf einer Intensivstation im Vergleich zur Behandlung auf einer Normalstation zulassen (Black 1996).

- Patienten, die während des Screenings von der Studienpopulation ausgeschlossen werden, „Screening Failures“, tendieren eher zu einer schlechteren Prognose, als die letztendlich eingeschlossenen Patienten.
  - ⇒ Limitierte Übertragbarkeit der final gewonnenen Studienergebnisse auf die reale Population (McKee 1999).

# Motivationshistorie

## Nicht-randomisierte kontrollierte Studien:

- Die Zuteilung zur Behandlungsgruppe erfolgt unter Berücksichtigung der Einflussgrößen:
  - ⇒ Ungleiche Verteilung (Imbalance) aller bekannten und unbekannten „Störgrößen“.
  - ⇒ Behandlungsgruppen sind nicht vergleichbar bzgl. möglicher Einflussgrößen (z.B. Alter und Geschlecht).
- Die Wahrscheinlichkeit der Gruppenzugehörigkeit zu den beiden Behandlungen ist bedingt durch die unabhängigen Kovariablen.

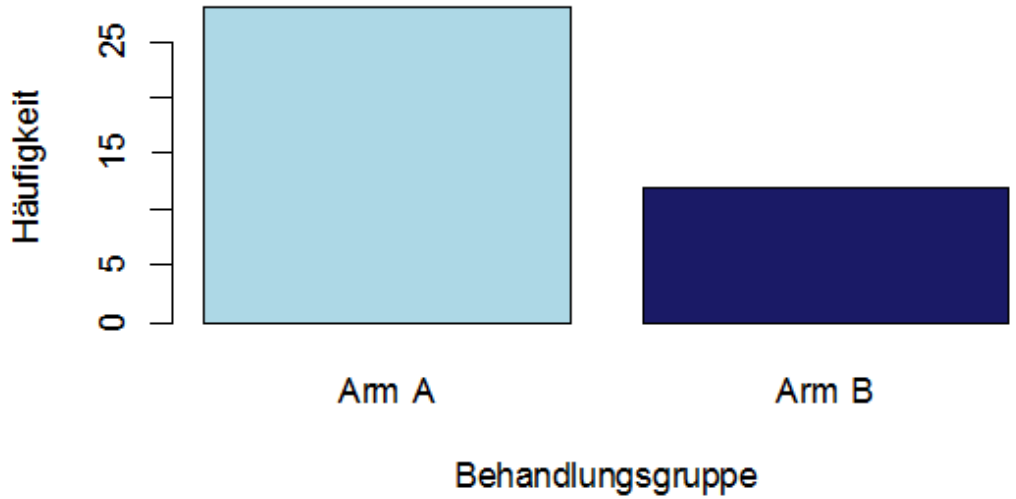
⇒ Direkter Vergleich und damit eine kausale Aussage zum Behandlungseffekt nicht möglich. Vorher Adjustierung der Daten.

⇒ Nicht-randomisierte Studien spiegeln eher Realbedingungen wider.

# Motivationsbeispiel

## Klassische Situation – Studienpopulation, z.B. n=40 Patienten:

- ➢ Zweigruppenvergleich, z.B. Behandlungen A versus B.
- ⇒ Patienten in A können sich von den Patienten in B bezüglich ihrer Anzahl unterscheiden.



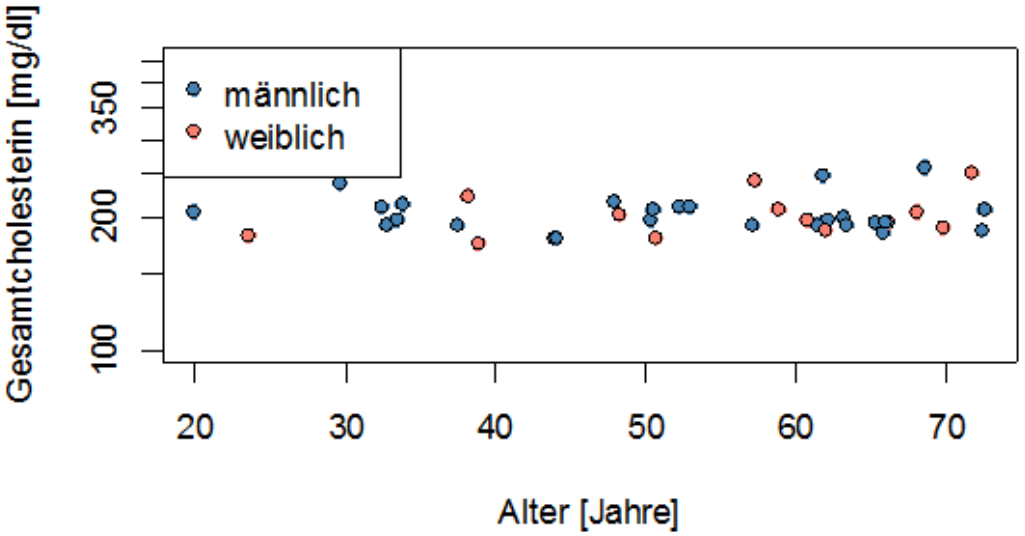
# Motivationsbeispiel

## Klassische Situation – Studienpopulation, z.B. n=40 Patienten:

- Zweigruppenvergleich, z.B. Behandlungen A versus B.
- Hauptzielparameter (quantitativ), z.B. Gesamtcholesterin im Blut, mit Einflussfaktoren Alter und Geschlecht.

⇒ Patienten in A können sich von denen in B auch in den Patienten-merkmalen unterscheiden.

⇒ Patientenmerkmale können neben der Behandlung auch einen Einfluss auf die Zielvariable ausüben.



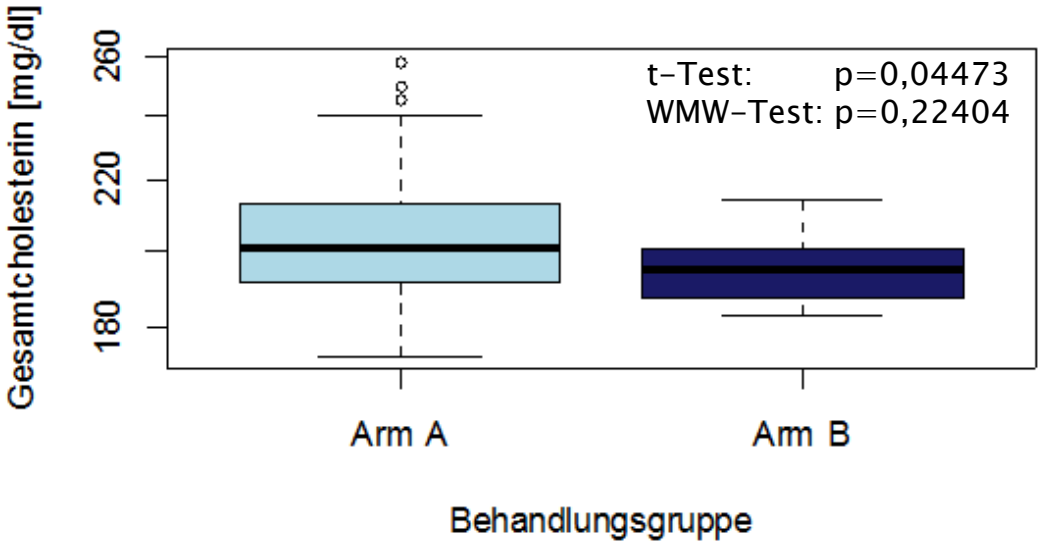
# Motivationsbeispiel

## Klassische Situation – Studienpopulation, z.B. n=40 Patienten:

- Zweigruppenvergleich, z.B. Behandlungen A versus B.
- Hauptzielparameter (quantitativ), z.B. Gesamtcholesterin im Blut, mit Einflussfaktoren Alter und Geschlecht.
- Zweiseitige Hypothese:
  $\boldsymbol{H_0}$: $\boldsymbol{\mu_A = \mu_B}$ -> kein Lage-Unterschied zwischen den Gruppen
  $\boldsymbol{H_1}$: $\boldsymbol{\mu_A \neq \mu_B}$ -> Lage-Unterschied zwischen den Gruppen

Testverfahren:

- t-Test (parametrisch)
- WMW-Test (nichtparametrisch, verteilungsfreies Rang-testverfahren)



# Motivationsbeispiel

Vorher ohne Gewichte,
mittleres Alter pro Gruppe:

| Gruppe | Mean | Std. Error |
|---|---|---|
| A | 49,31 | 2,555 |
| B | 61,62 | 3,903 |

p-Wert: 0,0120

Gruppen sind bzgl.
Alter unterschiedlich!

Nachher mit Gewichten,
mittleres Alter pro Gruppe:

| Gruppe | Mean | Std. Error |
|---|---|---|
| A | 61,78 | 3,104 |
| B | 61,62 | 2,331 |

p-Wert: 0,9681

Gruppen sind bzgl.
Alter nicht unterschiedlich!

=> Balancierung der Patientenmerkmale!

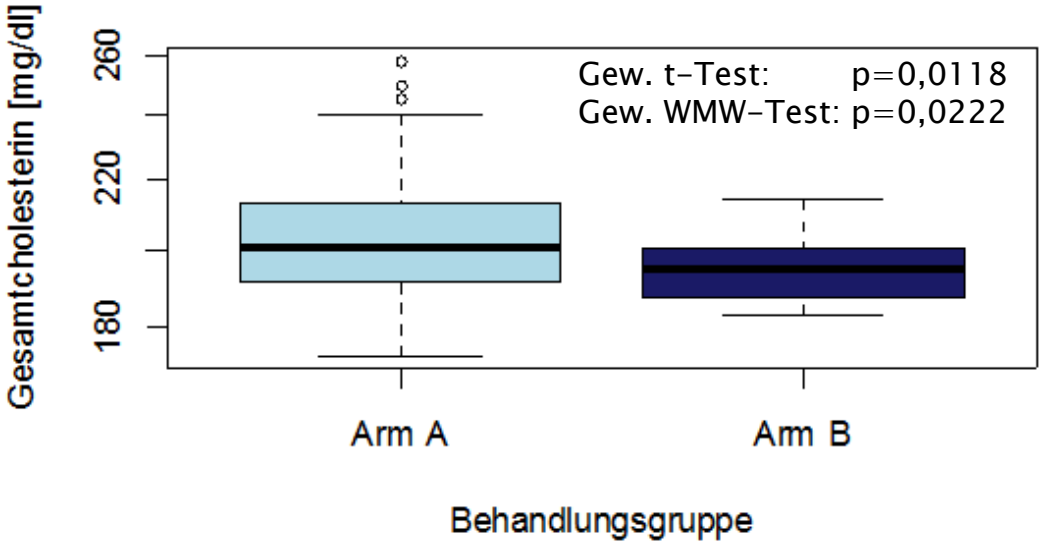
# Motivationsbeispiel

## Klassische Situation – Studienpopulation, z.B. n=40 Patienten:

- Zweigruppenvergleich, z.B. Behandlungen A versus B.
- Hauptzielparameter (quantitativ), z.B. Gesamtcholesterin im Blut, mit Einflussfaktoren Alter und Geschlecht.
- Zweiseitige Hypothese:
  $\boldsymbol{H_0}$: $\boldsymbol{\mu_A = \mu_B}$ -> kein Lage-Unterschied zwischen den Gruppen
  $\boldsymbol{H_1}$: $\boldsymbol{\mu_A \neq \mu_B}$ -> Lage-Unterschied zwischen den Gruppen

Testverfahren:

- Gewichteter t-Test (parametrisch)
- Gewichteter WMW-Test (nichtparametrisch)



# Klassische Auswertung

## Multiples Regressionsmodell:

- Herkömmliche Methode zum Vergleich der Gruppen trotz möglicher Imbalance.
- Unterschiede zwischen den Gruppen werden direkt anhand einer Zielgröße und unter Berücksichtigung der Patientenmerkmale (Alter, Geschlecht etc.) untersucht.

⇒ adjustiertes 1-Schritt-Verfahren

## Nachteile:

- Die Modelle schätzen immer einen Therapieeffekt, auch wenn die Gruppen extrem unterschiedlich sind ⇒ nicht sinnvoll.
- Die Anzahl der berücksichtigten Kovariablen ist beschränkt ⇒ sonst Problem des „Overfitting“.

# Propensity-Score-Methode

## Propensity-Score-Methode als alternatives 2-Schritt-Verfahren:

Der Propensity Score (e) für ein Individuum (i aus n) ist definiert als die bedingte Wahrscheinlichkeit (P) der Zugehörigkeit zur Behandlung B (Z=1) unter der Bedingung der unabhängigen Kovariablen (X):

$$e(X) = P(Z = 1|X)$$

1. Schritt: Berechnung von „Zuteilungswahrscheinlichkeiten" (PS) zur Behandlung B (Z=1) durch eine logistische Regression (unabhängig von der Zielgröße).

$$\hat{Z} = \beta_0 + \beta_1 X_1 +, \ldots, +\beta_k X_k$$

mit $\beta_i$ Schätzer der Regressionskoeffizienten

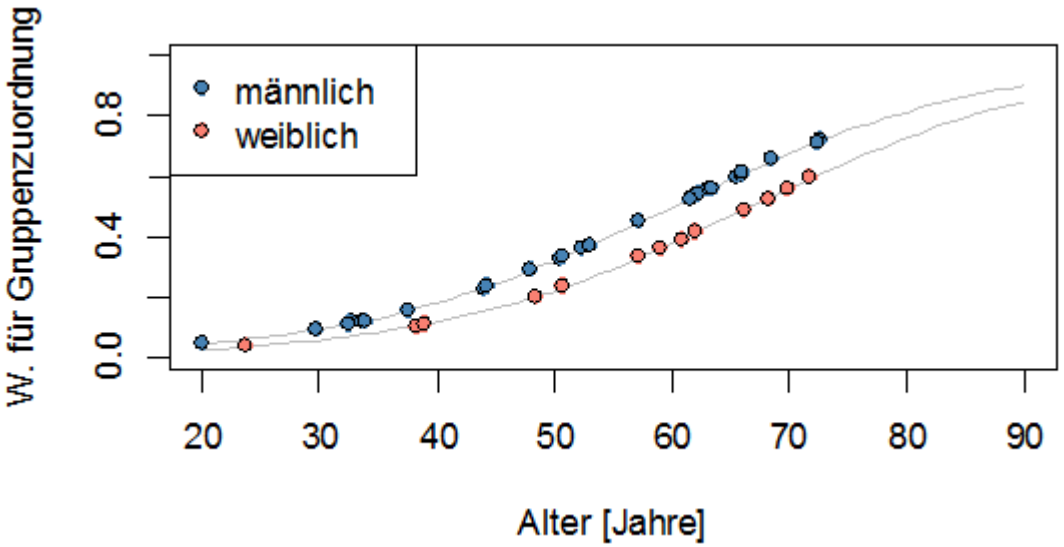
Propensity Scores werden als Odds dargestellt, $\frac{\hat{e}(X)}{1-\hat{e}(X)}$, und danach durch die Logit-Funktion transformiert.

$$\hat{e}(\mathbf{X}) = \frac{e^{\hat{Z}}}{1 + e^{\hat{Z}}}$$

Rosenbaum und Rubin 1983

# Propensity-Score-Methode

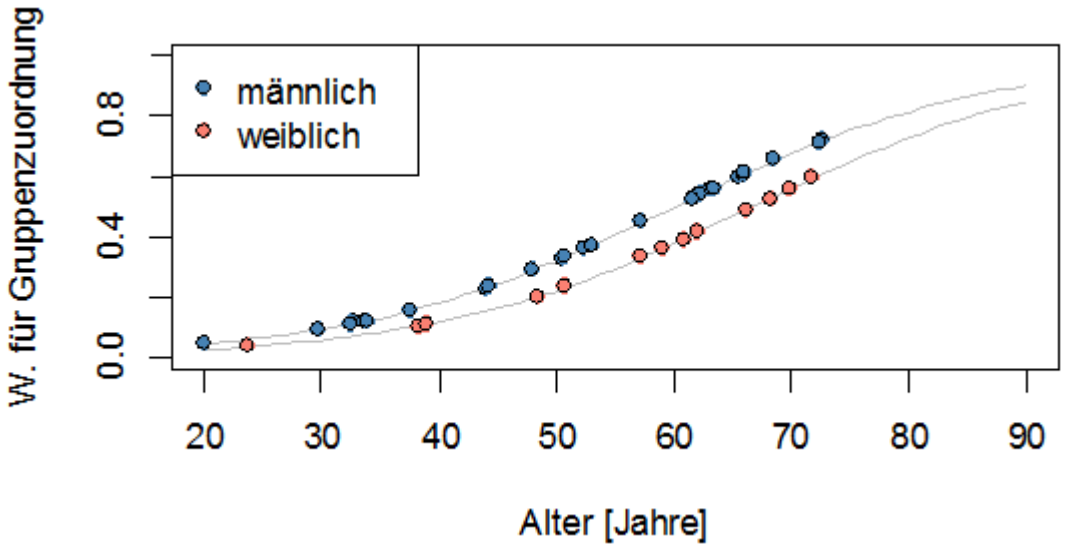
PS-Scores für Motivationsbeispiel:




Der Propensity Score, die „Zuteilungswahrscheinlichkeit“, für jeden Patienten, d.h. die vorhergesagte Wahrscheinlichkeit, dass der Patient der Behandlung B (Z=1) zugeordnet ist.
Cave: Zwei Patienten mit identischem PS sind nicht notwendigerweise identisch bzgl. ihrer Patientenmerkmale.

# Propensity-Score-Methode

PS-Scores für Motivationsbeispiel:




Die PS-Gewichte lassen sich als Odds der Wahrscheinlichkeiten der Behandlungszugehörigkeit zur Behandlung B (Z=1), PS, bestimmen.

Z=0: $\mathbf{w} = \frac{\mathbf{1}}{\mathbf{1-PS}}$ , wobei PS die Wahrscheinlichkeit für Z=1 ist.

Z=1: $\mathbf{w} = \frac{\mathbf{1}}{\mathbf{PS}}$ , wobei PS die Wahrscheinlichkeit für Z=1, d.h. w=1, ist.

# Propensity-Score-Methode

## 2. Schritt: Prozeduren zur Sicherstellung der Balanciertheit der Patientenmerkmale durch die PS-Schätzer:

Auswahl an Prozeduren:

- PS-Matching (1:1 nearest neighbor matching)
- Inverse probability of treatment weighting (IPTW)
- Stratifizierung oder Subklassifikation
- Regressionsadjustierung für den PS

Dabei können mehr Kovariablen eingeschlossen werden als im herkömmlichen Regressionsmodell.

Für seltene Ereignisse ist die PS-Methode besonders überlegen.

⇒ Finaler Gruppenvergleich anhand der Zielgröße (jedes Skalentyps) durch ein geeignetes statistisches Testverfahren.

# Propensity-Score-Methode

Die folgenden Erläuterungen zu den PS-Prozeduren beziehen sich auf eine nicht-randomisierte klinischen Therapie-Studie.

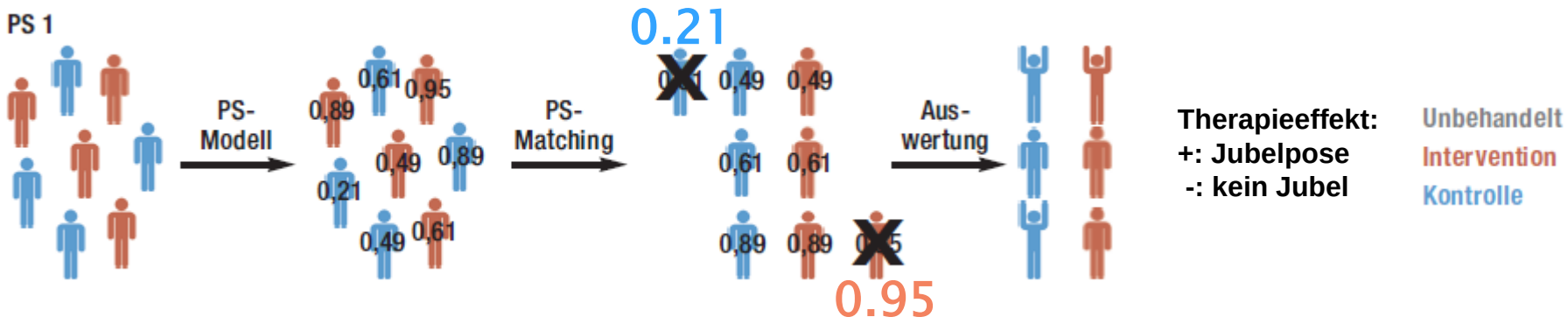
Die dazugehörigen Abbildungen sowie einige inhaltliche Aspekte dieses Vortrags sind folgender Publikation entnommen:

Propensity Score – eine alternative Methode zur Analyse von Therapieeffekten.
Oliver Kuss, Maria Blettner, and Jochen Börgermann.
Dtsch Ärztebl Int, 113:597-603, 2016.

# Propensity-Score-Methode

Deutsches Ärzteblatt | Jg. 113 | Heft 35–36 | 5. September 2016

## PS 1: PS-Matching (1:1, auch 1:n, nearest neighbor matching)



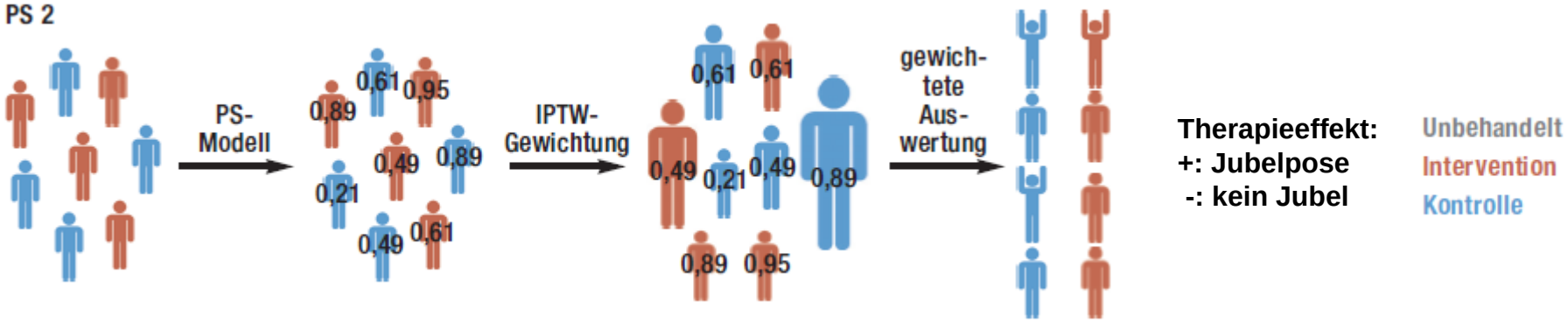

- Ein Interventions-Patient wird zu genau dem Kontroll-Patient gematcht, der den ähnlichsten PS-Schätzer aufweist ⇒ vergleichbare Gruppen.
- Schwellenwert c festlegen (Max. erlaubte Differenz zwischen zwei Patienten. Je kleiner c, desto höher Anforderungen an statistischen Zwilling).
- Balanciertheit der Merkmale z.B. mit standardisierten Differenzen überprüfbar.
- Nachteil: Patienten, für die kein Matching-Partner gefunden wurde, werden ausgeschlossen ⇒ Reduktion der Fallzahl, Verlust an statistischer Power, aber robust bei extremen Propensity Scores.

Auswertung auf reduzierter Population mit geeignetem statistischen Verfahren.

# Propensity-Score-Methode

Deutsches Ärzteblatt | Jg. 113 | Heft 35–36 | 5. September 2016

## PS 2: Inverse probability of treatment weighting (IPTW)




- Patienten mit höheren IPTW-Gewichten sind größer dargestellt.
- Pat. mit Intervention: PS-Gewicht=1/PS, Pat. mit Kontrolle: PS-Gewicht=1/(1-PS).
  - ⇒ Nicht robust bei extremen PS-Gewichten.
  - ⇒ Vergleichbare gewichtete Gruppen.
- Balanciertheit der Merkmale z.B. mit gewichteten standardisierten Differenzen überprüfbar.

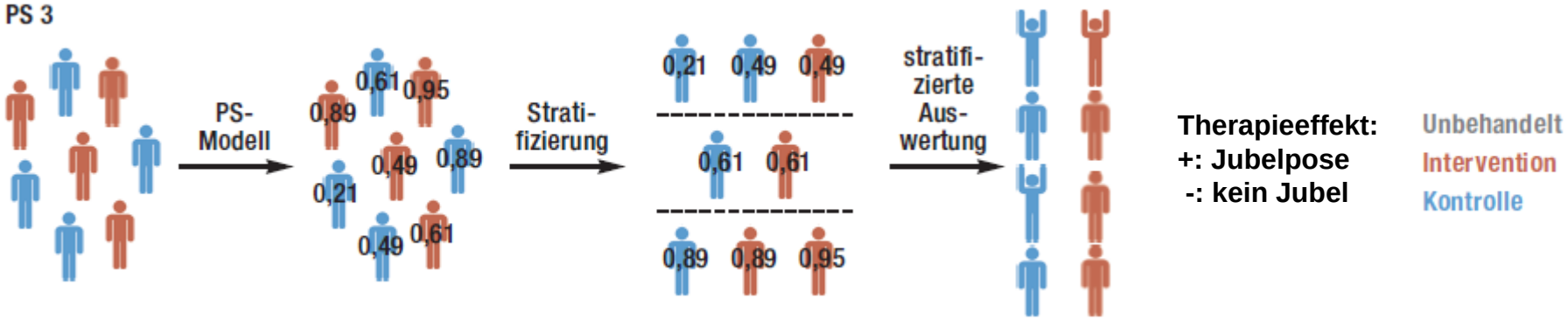
Auswertung auf Gesamtpopulation unter Berücksichtigung der PS-Gewichte.

⇒ Gewichtete gemischte lineare Modelle oder gewichtete nichtparametrische Rang-basierte Verfahren.

# Propensity-Score-Methode

Deutsches Ärzteblatt | Jg. 113 | Heft 35–36 | 5. September 2016

## PS 3: Stratifizierung oder Subklassifikation




- Geordnete PS werden gleichmäßig stratifiziert in Terzile (üblicher Quintile).
  - ⇒ Robust bei extremen Propensity Scores.
  - ⇒ Vergleichbare Gruppen innerhalb eines Stratums.
- Balanciertheit der Merkmale mit z.B. standardisierten Differenzen pro Stratum überprüfbar.

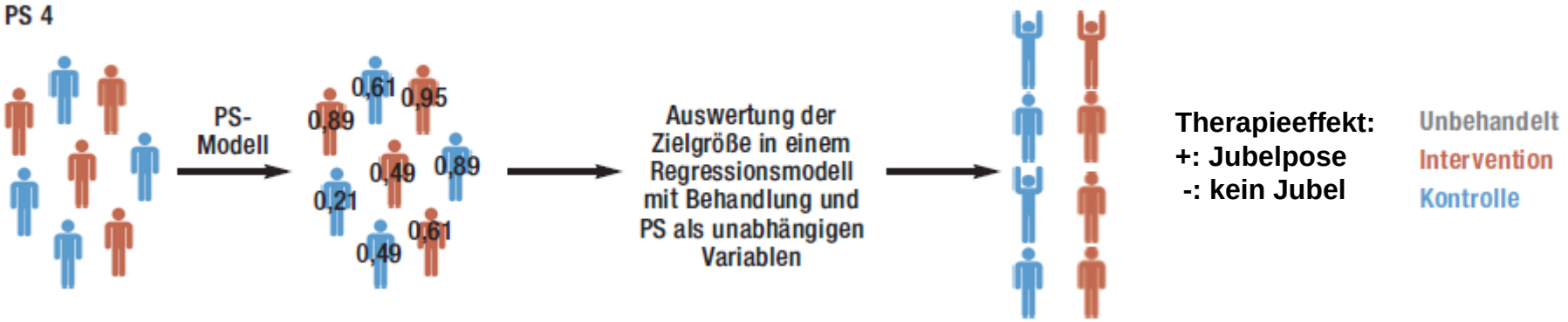
Auswertung des Therapieeffekts wird pro Stratum geschätzt.
Zusammenfassung des Effekts für Gesamtpopulation mit Hilfe von:

⇒ Metaanalyse-Methoden oder
für nichtparametrische Zielgrößen, z.B. stratifizierter van Elteren-Test

# Propensity-Score-Methode

Deutsches Ärzteblatt | Jg. 113 | Heft 35–36 | 5. September 2016

## PS 4: Regressionsadjustierung für den PS



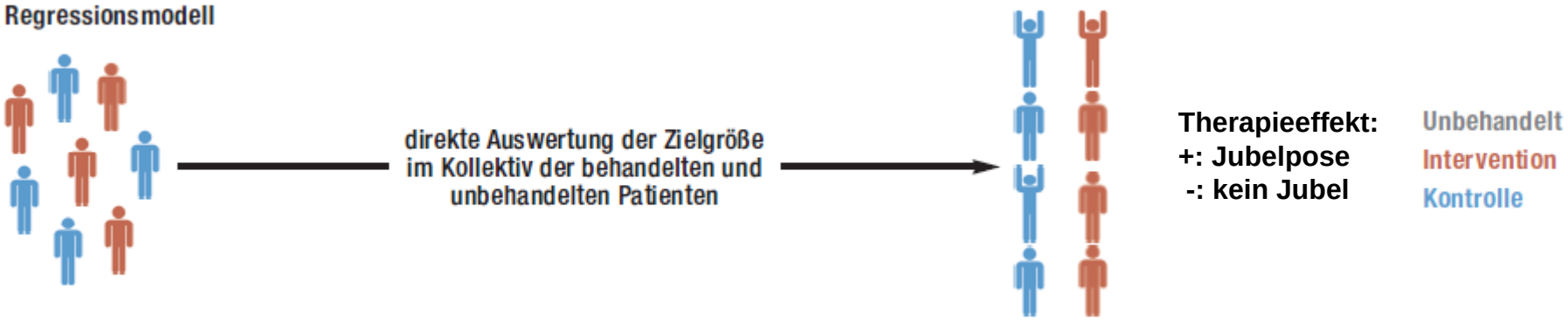

- Geschätzter Behandlungseffekt ist adjustiert für alle zur Bestimmung des PS einbezogenen Patientenmerkmale.
  - ⇒ Robust bei extremen Propensity Scores.
  - ⇒ Keine vergleichbaren Gruppen.

Auswertung der Zielgröße mit Hilfe von Regressionsmodellen unter Berücksichtigung der Behandlung und des PS-Schätzers als unabhängige Variablen.

# Propensity-Score-Methode

Deutsches Ärzteblatt | Jg. 113 | Heft 35–36 | 5. September 2016

## Herkömmliches Regressionsmodell



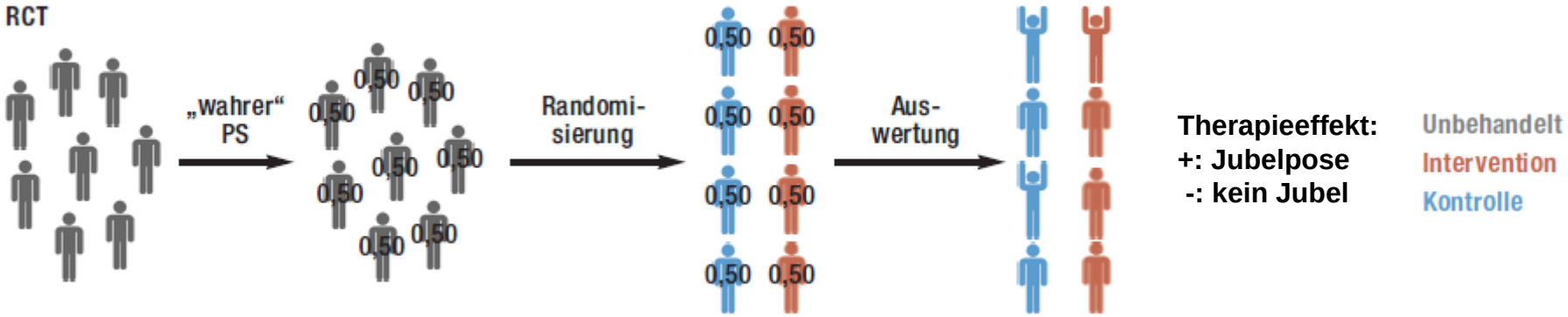

- Multivariable lineare Regression mit klinischer Zielgröße als abhängige Variable und Therapie sowie die Kovariablen als unabhängige Variablen.
  - ⇒ Robust bei extremen Propensity Scores.

1-Schritt-Verfahren mit direkter Adjustierung im Regressionsmodell.

# Propensity-Score-Methode

Deutsches Ärzteblatt | Jg. 113 | Heft 35–36 | 5. September 2016

## RCT: Goldstandard – Randomisierte kontrollierte klinische Studie




- 1. Schritt: Randomisierung:
  - ⇒ Damit ist die „Zuteilungswahrscheinlichkeit" bekannt.
  - ⇒ Hier (1:1)-Ratio, d.h. PS-Schätzer ist 0,5.
  - ⇒ Vergleichbare Gruppen.
- 2. Schritt: Gruppenvergleich

Auswertung mit Hilfe aller geeigneten klassischen ungewichteten Verfahren.

# Ausblick

Goldstandard für kontrollierte Interventionsstudien bleibt die Randomisierung.

Aber es gibt Verbesserungsmöglichkeiten (Collins et al. NEJM, 2020):

- Durch Anpassung der Studien-Guidelines, verblindete Randomisierung, Adhärenz zur Studienbehandlung, komplettes Follow-up, ITT-Analysen ⇒ wesentliche Verbesserung der Verlässlichkeit der Studienergebnisse.

# Ausblick

Weitere Verbesserungsmöglichkeiten (Collins et al. NEJM, 2020):

- Verbesserte Rekrutierung:
  Verwendung von eRegisterDaten ⇒ schneller/besser vorhersagbar.
  Weniger Ein- und Ausschlusskriterien ⇒ breiter/besser verallgemeinerbar.
- Verbesserte Qualität:
  Verwendung eines interaktiven eCRFs und Einrichtung eines zentralisierten Monitorings.
- Effektives Follow-up: Nutzung von eRegister-Daten und „Extended Outcomes“ via Smartphone oder digitalem Sensor.

# Ausblick

Für nicht-randomisierte kontrollierte Studien ist die Propensity-Score-Methode eine gute Alternative.

- Cave: Die Propensity-Score-Methode kann genau wie die herkömmliche Methode (multiple Regression) nur für die bekannten und tatsächlich erhobenen Patientenmerkmale adjustieren.

⇒ Nachfrage zur Evidenz von nicht-randomisierten Studien steigt.

⇒ Die Propensity-Score-Methode wird immer häufiger herangezogen.

# Referenzen


- Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.
- Oliver Kuss, Maria Blettner, and Jochen Börgermann. Propensity Score – eine alternative Methode zur Analyse von Therapieeffekten. Dtsch Ärztebl Int, 113:597–603, 2016.
- Peter C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3):399–424, 2011. PMID: 21818162.
- Jürgen Windeler. External validity. Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, 102(4):253–259, Jul 2008.
- N. Black. Why we need observational studies to evaluate the effectiveness of health care. BMJ (Clinical research ed.), 312:1215–8, May 1996.
- M. McKee, A. Britton, N. Black, K. McPherson, C. Sanderson, and C. Bain. Interpreting the evidence: choosing between randomised and non-randomised studies. BMJ (Clinical research ed.), 319:312–5, Jul 1999.
- Rory Collins, Louise Bowman, Martin Landray, and Richard Peto. The magic of randomization versus the myth of real-world evidence. New England Journal of Medicine, 382(7):674–678, 2020.



# Vielen Dank für Ihre Aufmerksamkeit!

# Add-on

- PS 1: PS-Matching (1:1 nearest neighbor matching)
  - Schätzung ATT (Average treatment effect of the treated ): ja
  - Schätzung ATE (Average treatment effect): nein
- PS 2: Inverse probability of treatment weighting (IPTW)
  - Schätzung ATT: ja (modifizierte Adjustierung)
  - Schätzung ATE: ja
- PS 3: Stratifizierung oder Subklassifikation
  - Schätzung ATT: ja (Gewichts-adjustiert)
  - Schätzung ATE: ja (gleiche Gewichte)
- PS 4: Regressionsadjustierung für den PS
  - Schätzung ATT: nein
  - Schätzung ATE: nein

# Propensity-Score-Methode

Deutsches Ärzteblatt | Jg. 113 | Heft 35–36 | 5. September 2016

**TABELLE 1**

**Eigenschaften der vier verschiedenen Methoden zur Berücksichtigung des Propensity Scores (PS) und der herkömmlichen Regressionsanalyse bei der Analyse von nichtrandomisierten Therapiestudien**

| | Methode | | | | |
|---|---|---|---|---|---|
| | PS-Methode | | | | herkömmliche Regressions-adjustierung |
| | PS-Matching | IPTW-Schätzung | Stratifizierung | Regressionsadjus-tierung für den PS | |
| ermöglicht eine leichte Beurteilung der Vergleichbarkeit von behandelten und unbehandelten Patienten | + | (+) | (+) | – | – |
| ermöglicht eine Beurteilung der Balanciert-heit der Merkmale im Auswertungsdatensatz | + | + | (+) | – | – |
| nutzt den vollständigen Datensatz (kleinere Varianz des Therapieeffekts bei größerer Gefahr für Bias) | – | + | + | + | + |
| ähnelt von der Vorgehensweise einem RCT (generiere vergleichbare Gruppen und ignoriere dabei die Zielgrößen) | + | (+) | (+) | – | – |
| ist robust gegenüber Patienten mit extremem PS | + | – | + | + | + |
| kommt insgesamt mit weniger statistischen Modellannahmen aus | + | + | (+) | – | – |

IPTW, „inverse probability of treatment weighting"; PS, Propensity Score; RCT, randomisierte kontrollierte Studie;
„+" bedeutet „ja" oder „ist gegeben", „–" bedeutet „nein" oder „ist nicht gegeben", „(+)" bedeutet „teilweise" oder „ist zum Teil gegeben"