# Gütemaße und Kriterien bei der Anwendung von Propensity Scores

**Oliver Kuß**

Institut für Biometrie und Epidemiologie, Deutsches Diabetes-Zentrum (DDZ), Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf

und

Centre for Health and Society (chs), Medizinische Fakultät der Heinrich-Heine-Universität Düsseldorf

# Zwei Gütemaße: Balanciertheit und Overlap

DDZ
German Diabetes Center

- **Balanciertheit**
  Ähnlichkeit der Verteilung der **Kovariablen** in den beiden Behandlungsgruppen
- **Overlap**
  Ähnlichkeit der Verteilung des **PS** in den beiden Behandlungsgruppen

# Zwei Gütemaße: Balanciertheit und Overlap

DDZ
German Diabetes Center

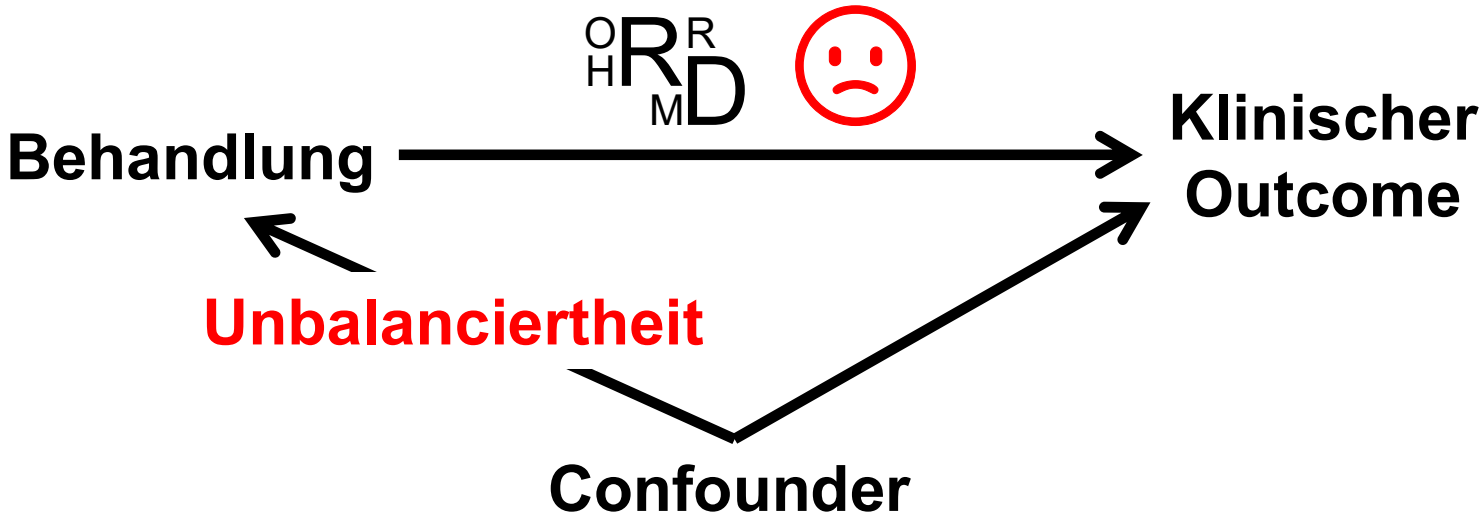
- **Warum ist Balanciertheit ein sinnvolles Gütemaß?**

# Zwei Gütemaße: Balanciertheit und Overlap

DDZ German Diabetes Center

- **Warum ist Balanciertheit ein sinnvolles Gütemaß?**

In einer nicht-randomisierten Studie:

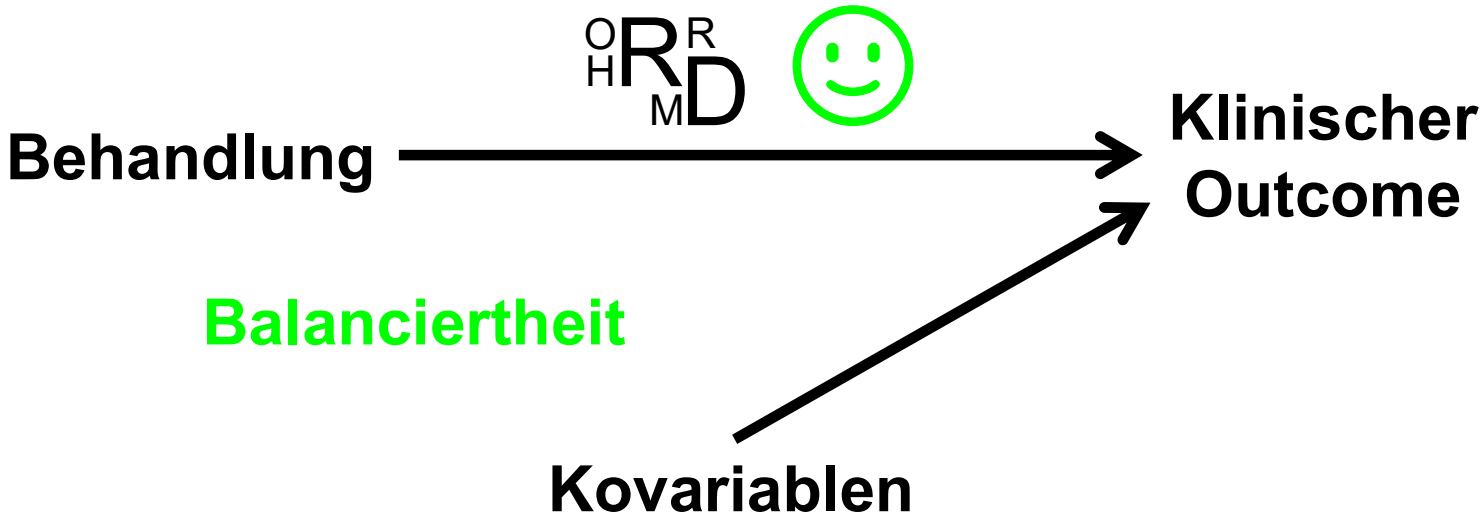


# Zwei Gütemaße: Balanciertheit und Overlap

DDZ German Diabetes Center

- **Warum ist Balanciertheit ein sinnvolles Gütemaß?**

In einer randomisierten Studie:

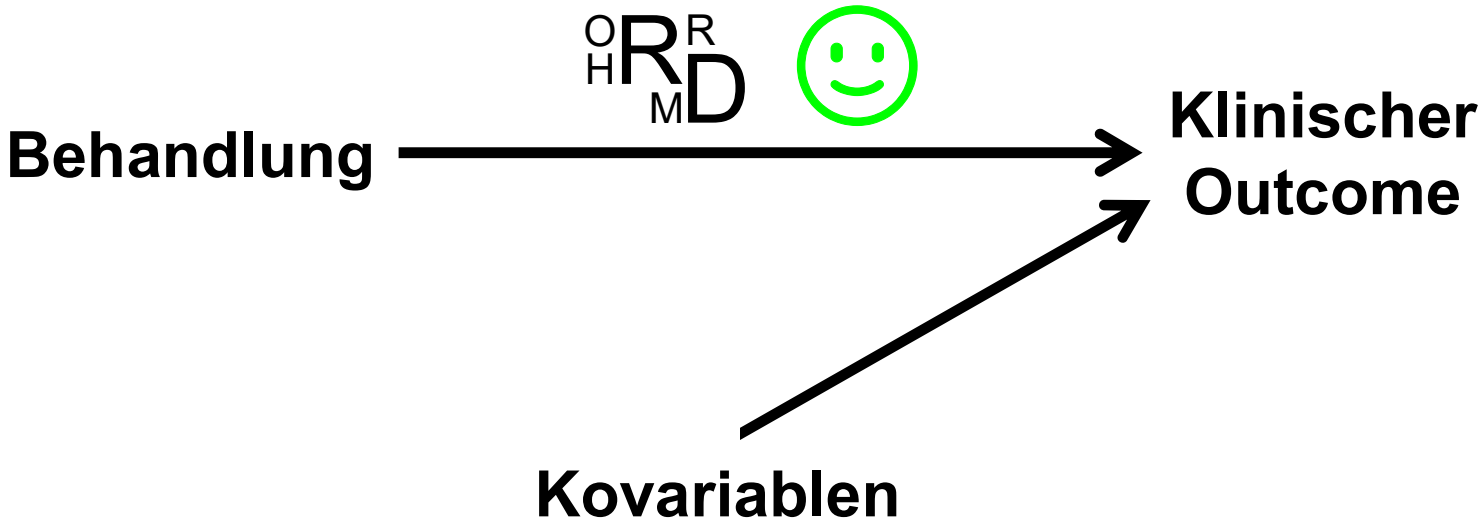


# Zwei Gütemaße: Balanciertheit und Overlap

DDZ German Diabetes Center

- **Warum ist Balanciertheit ein sinnvolles Gütemaß?**

Propensity Score-Prinzip:



# Zwei Gütemaße: Balanciertheit und Overlap

DDZ German Diabetes Center

- **Warum ist Overlap ein sinnvolles Gütemaß?**
  - In Bereichen, wo es keinen Overlap der Verteilungen des PS gibt, ist die Behandlung vordeterminiert. Dann hat man aber auch keine Variabilität der Behandlungen und kann daher auch keine validen Effekte schätzen.
  - Grundlegende Annahme für PS-Analysen: **Positivität**, d.h. für alle Beobachtungen/Kovariablenmuster muss gelten: $0 < PS < 1$
    - **Problem:** PS ist unbekannt
    - **Frage:** Ab wann ist ein geschätzter PS zu nahe an 0/1? Ab 0,05/0,95? Oder erst ab 0,01/0,99? …

# Zwei Gütemaße: Balanciertheit und Overlap

DDZ German Diabetes Center

- **Referenz: RCT**
  - Balanciertheit perfekt, da Verteilung der Kovariablen (sogar der ungemessenen) nicht nur ähnlich, sondern identisch
  - Overlap perfekt, da PS=0,5
  - „Je besser Balanciertheit/Overlap, desto ähnlicher ist eine PS-Analyse einem RCT“

# Zwei Gütemaße: Balanciertheit und Overlap

DDZ
German Diabetes Center

- **Frage:** Sind Balanciertheit und Overlap proportional / zwei Seiten der selben Medaille?

- Belitser et al. meinen: „Ja“!
  *“Assessment of balance of confounders between exposure groups is a key issue in PS methods, and possible balance measures include the overlapping coefficient, …”*

Belitser SV, Martens EP, Pestman WR, Groenwold RH, de Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf. 2011 Nov;20(11):1115-29.

# Wie misst man Balanciertheit **für eine einzelne Kovariable**?

DDZ German Diabetes Center

- **Quasi-Standard:** Standardisierte Differenz
  - Differenz der Mittelwerte oder Anteile in beiden Gruppen, dividiert durch gemeinsame Standardabweichung
  - **Faustregel:** Ein Wert von 10 % zeigt gute Balanciertheit an
  - **Zwei Probleme:**
    1. Verteilung hängt von der Stichprobengröße ab [Austin2009]
    2. Kein Vergleich von Kovariablen auf verschiedenen Skalen (stetig, binär, ordinal, nominal)

Austin PC. Stat Med 2009;28:3083-107.

# Wie misst man Balanciertheit **für eine einzelne Kovariable**?

DDZ German Diabetes Center

- **z-Differenz** [Kuss2013]
  - Maß für Balanciertheit, das für metrische, binäre und ordinale Kovariablen definiert und vergleichbar ist
  - **Prinzip:** Standardisiere das jeweilige Unterschiedsmaß (Mittelwertsdifferenz, Risikodifferenz, Wilcoxon-Statistik) mit seinem Standardfehler („z-Standardisierung“)
  - **Update I:** Inzwischen gibt es auch eine Version für nominale Merkmale (PIT der herkömmlichen $\chi^2$-Statistik)
  - **Update II:** Inzwischen gibt es auch eine gewichtete z-Differenz für gewichtete PS-Analysen

Kuss O. J Clin Epidemiol. 2013 Nov;66(11):1302-7.

**The weighted z-difference can be used to measure covariate balance in weighted propensity score analyses**

Filla T[1], Kuss O[1,2]

[1] Institute of Medical Statistics, Medical Faculty, Heinrich Heine University Düsseldorf

[2] Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf

# Wie misst man Balanciertheit **für eine einzelne Kovariable**?

DDZ German Diabetes Center

- **z-Differenz** [Kuss2013]
  - In einem RCT sind die z-Differenzen standard-normalverteilt (N(0,1))
  - In einer perfekt gematchten Studie sind die z-Differenzen N(0,½)-verteilt [RubinThomas1996] (und damit besser als in einem RCT!)

Kuss O. J Clin Epidemiol. 2013 Nov;66(11):1302-7;
Rubin DB, Thomas N. Biometrics 1996 Mar;52(1):249-64.

# Wie misst man die **globale** Balanciertheit?

DDZ
German Diabetes Center

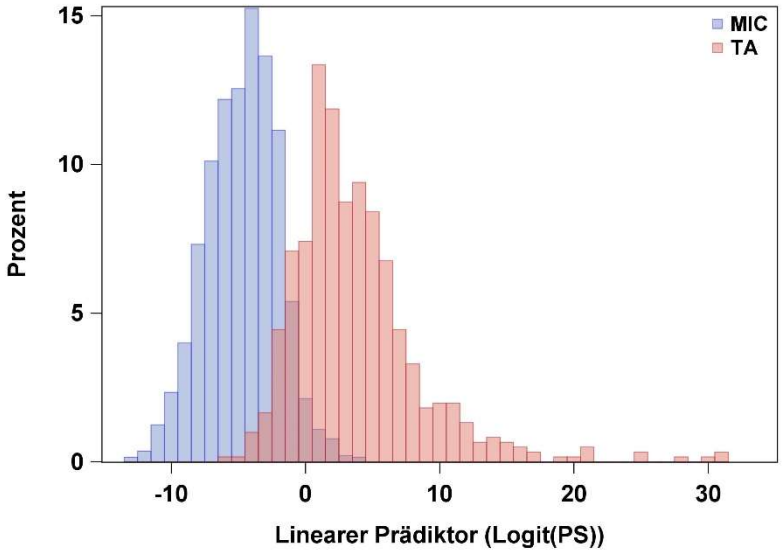
- **Vorschlag:** Summe der quadrierten z-Differenzen
  - Wenn die z-Differenzen von k Merkmalen alle standard-normalverteilt sind, dann ist die Summe der quadrierten z-Differenzen $SSQ_{zDiff}$ (zumindest approximativ) $\chi^2_k$-verteilt.
  - In einem RCT gilt: $E(SSQ_{zDiff}) = k$
  - In einer perfekt gematchten PS-Studie gilt: $E(SSQ_{zDiff}) = k/2$

Kuss O. J Clin Epidemiol. 2013 Nov;66(11):1302-7.

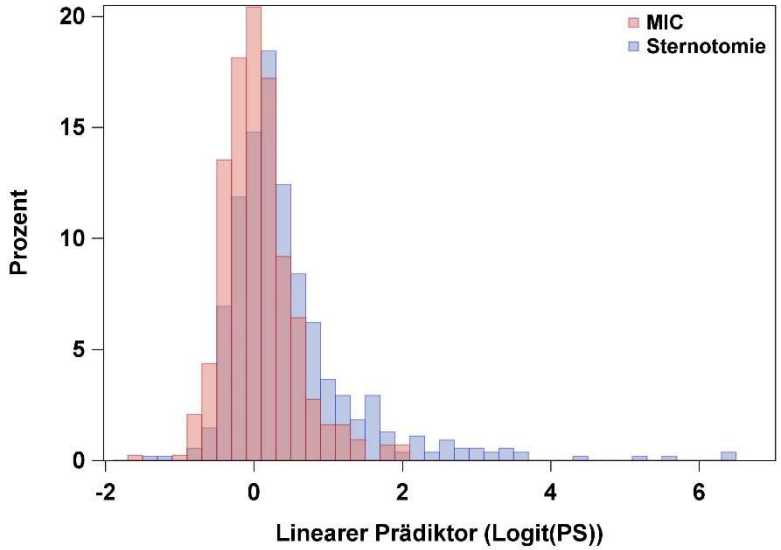
# Wie misst man Overlap?

German Diabetes Center

- Es gibt kein akzeptiertes Maß (außer der grafischen Darstellung)




Schlechter Overlap




Guter Overlap

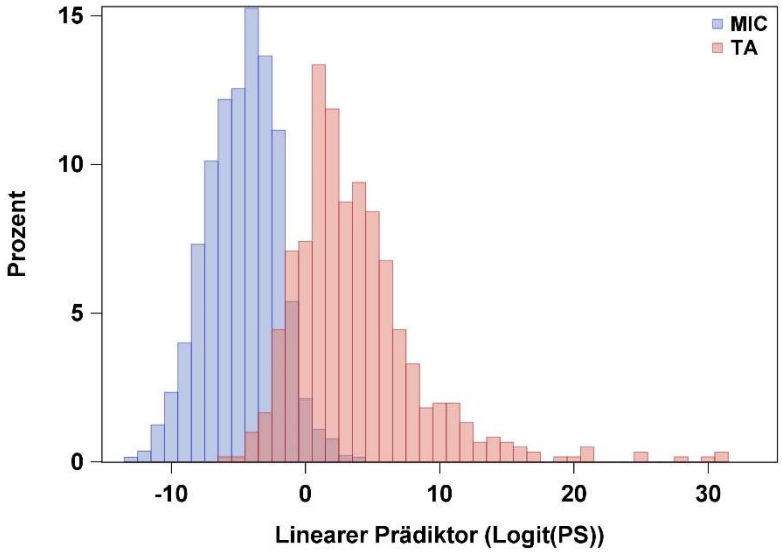
**Aus:** Furukawa N, Kuss O, … J Thorac Cardiovasc Surg, 2018 Nov;156(5):1825-1834.

**Aus:** Furukawa N, Kuss O, … Eur J Cardiothorac Surg. 2014 Aug;46(2):221-6.

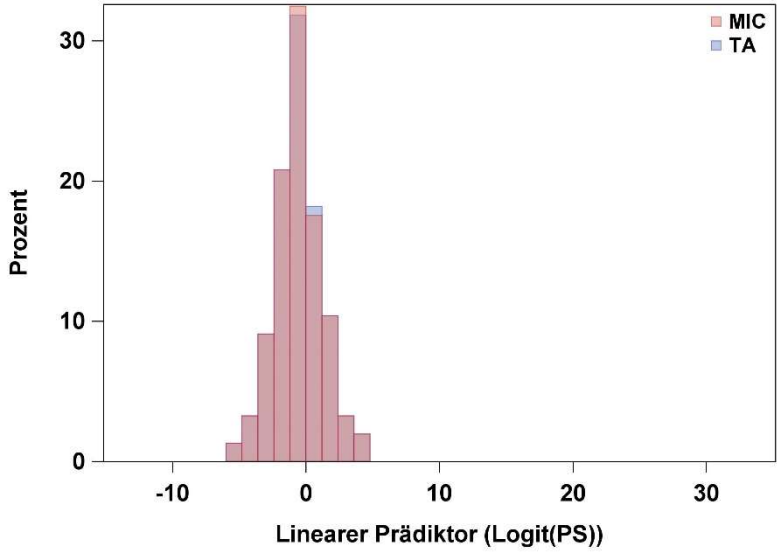
# Wie misst man Overlap?

DDZ German Diabetes Center

- **Meine Lösung:** Verwende PS-Matching, generiert per Definition guten Overlap




Aus schlechtem Overlap vor dem PS-Matching …




… wird perfekter Overlap nach dem PS-Matching

**Aus:** Furukawa N, Kuss O, … J Thorac Cardiovasc Surg, 2018 Nov;156(5):1825-1834.

# Ein Beispiel aus der Herzchirurgie

DDZ German Diabetes Center

- Publizierte PS-Analyse aus der Aortenklappenchirurgie [Furukawa2018]
- 3.809 PatientInnen, denen zwischen Juli 2009 und Juli 2017 am Herz- und Diabeteszentrum NRW in Bad Oeynhausen minimal invasiv eine neue Aortenklappe eingesetzt worden war.
- **Im Paper:** Vergleich konventionelle OP (Ministernotomie, MIC, N = 1.929), und zwei katheder-basierte Behandlungen (transapikal, TA, N=607 und transfemoral, TF, N = 1.273)

Furukawa N, Kuss O, Emmel E, Scholtz S, Scholtz W, Fujita B, Ensminger S, Gummert JF, Börgermann J, Minimally invasive versus transapical versus transfemoral aortic valve implantation: A one-to-one-to-one propensity score-matched analysis, J Thorac Cardiovasc Surg, 2018 Nov;156(5):1825-1834,

# Ein Beispiel aus der Herzchirurgie

DDZ German Diabetes Center

- **Hier:** Vergleich MIC vs. TA
- Entscheidung bzgl. Behandlung durch Konsensus des TAVI-Teams (unter Beteiligung von Kardiochirurgie, Kardiologie und Anästhesiologie)

Furukawa N, Kuss O, Emmel E, Scholtz S, Scholtz W, Fujita B, Ensminger S, Gummert JF, Börgermann J, Minimally invasive versus transapical versus transfemoral aortic valve implantation: A one-to-one-to-one propensity score-matched analysis, J Thorac Cardiovasc Surg, 2018 Nov;156(5):1825-1834,

# Ein Beispiel aus der Herzchirurgie

DDZ German Diabetes Center

- Schätzung des PS-Modell durch logistisches Regressionsmodell mit 23 Kovariablen
  Alter, Geschlecht, OP-Jahr, Körpergröße, Körpergewicht, LVEF, eGFR, euroScore II, STS score, German Aortic Valve Score, Bluthochdruck, vorherige Aortenklappen-OP, Diabetes, COPD, pulmonarer Hochdruck, Z,n, Schlaganfall, PAOD, Zerebrovaskuläre Krankheit, Z,n, AF, koronare Herzkrankheit, Z,n, Herzinfarkt, NYHA-Klasse, Priorität

Furukawa N, Kuss O, Emmel E, Scholtz S, Scholtz W, Fujita B, Ensminger S, Gummert JF, Börgermann J, Minimally invasive versus transapical versus transfemoral aortic valve implantation: A one-to-one-to-one propensity score-matched analysis, J Thorac Cardiovasc Surg, 2018 Nov;156(5):1825-1834,

# Ein Beispiel aus der Herzchirurgie

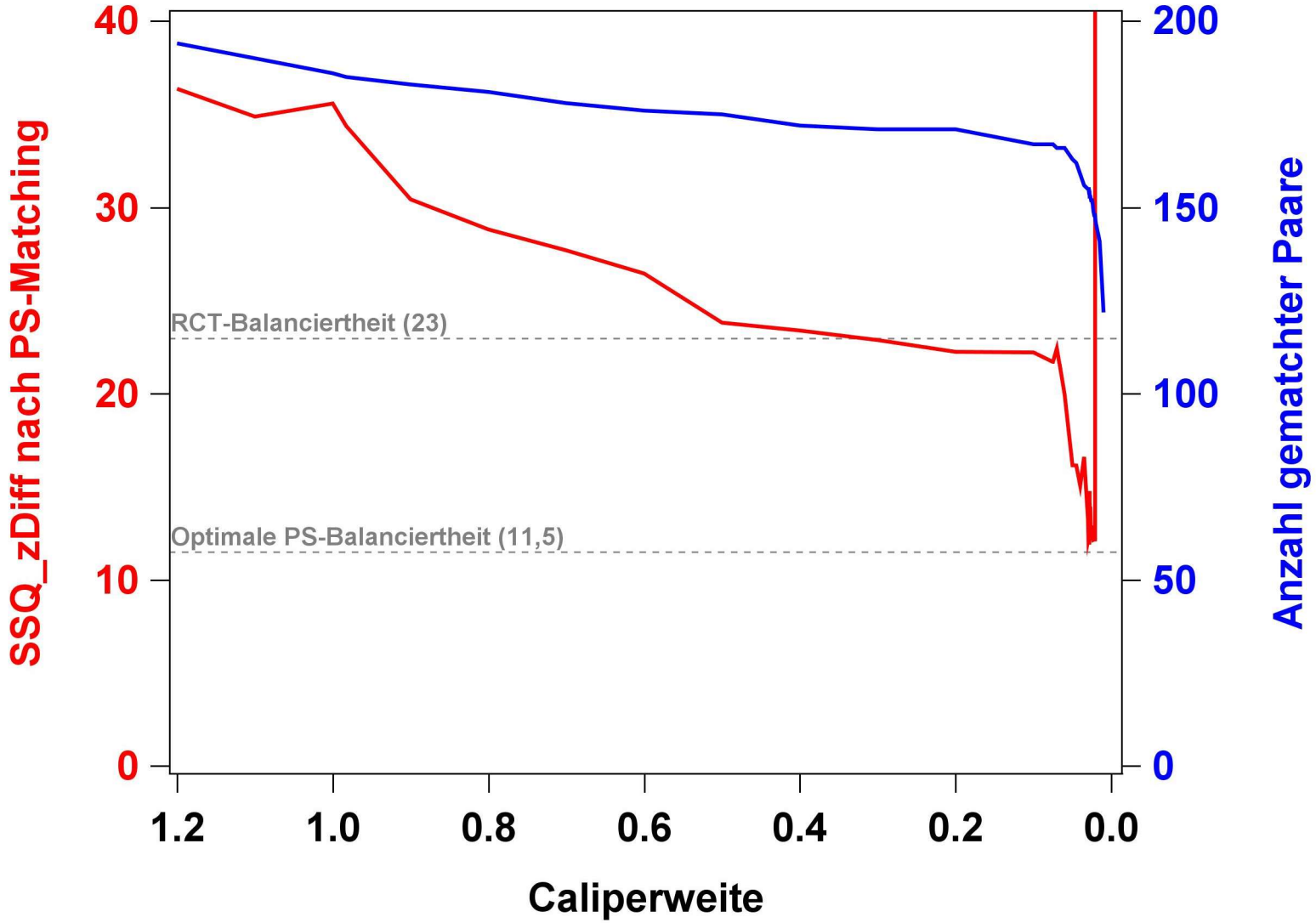
DDZ German Diabetes Center

- 1:1-PS-Matching mit einem “optimal matching algorithm” mit optimierter Caliper-Breite und dem logit-transformierten PS
- Optimierung der Caliper-Breite durch Trade-Off zwischen
  - Anzahl der gematchten Paare
  - Maximale Balanciertheit der Kovariablen (Minimierung der $SSQ_{zDiff}$)

Furukawa N, Kuss O, Emmel E, Scholtz S, Scholtz W, Fujita B, Ensminger S, Gummert JF, Börgermann J, Minimally invasive versus transapical versus transfemoral aortic valve implantation: A one-to-one-to-one propensity score-matched analysis, J Thorac Cardiovasc Surg, 2018 Nov;156(5):1825-1834,

# Ein Beispiel aus der Herzchirurgie

DDZ German Diabetes Center



# Ein Beispiel aus der Herzchirurgie: Balanciertheit

DDZ German Diabetes Center

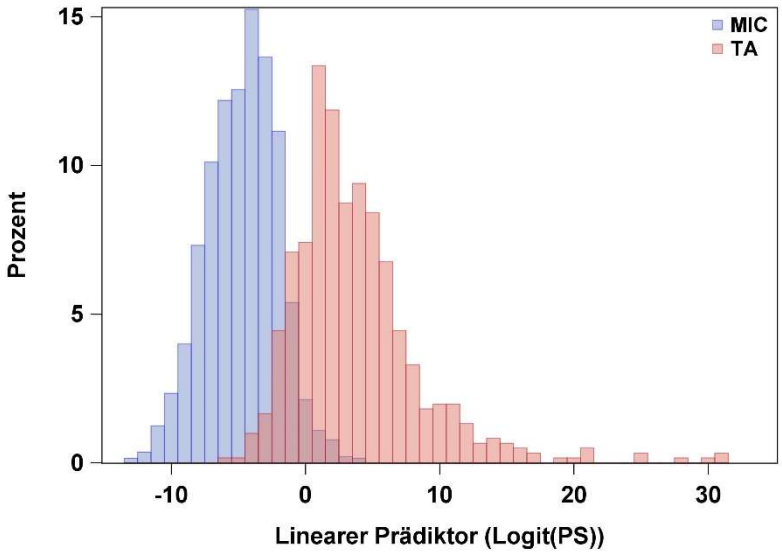
| | Alle PatientInnen (n = 2.536) | | |
|---|---|---|---|
| | MIC (n = 1,929) | TA (n = 607) | z-Differenz |
| Alter [Jahre, MW ± SD] | 67 ± 11 | 81 ± 6 | -38,2 |
| eGFR [ml/min, MW ± SD] | 79 ± 20 | 56 ± 23 | 22,2 |
| Diabetes [%] | 19 | 35 | -7,75 |

# Ein Beispiel aus der Herzchirurgie: Overlap

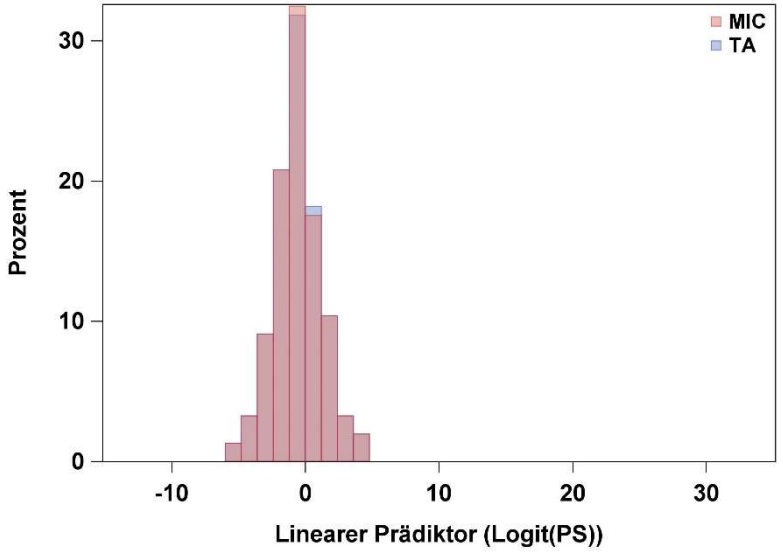
German Diabetes Center

## Linearer Prädiktor

**Vor** PS-Matching




**Nach** PS-Matching

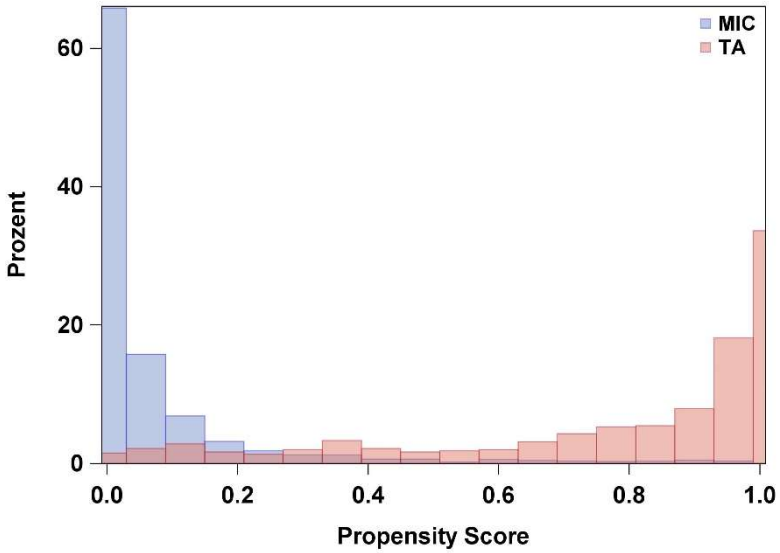


# Ein Beispiel aus der Herzchirurgie: Overlap

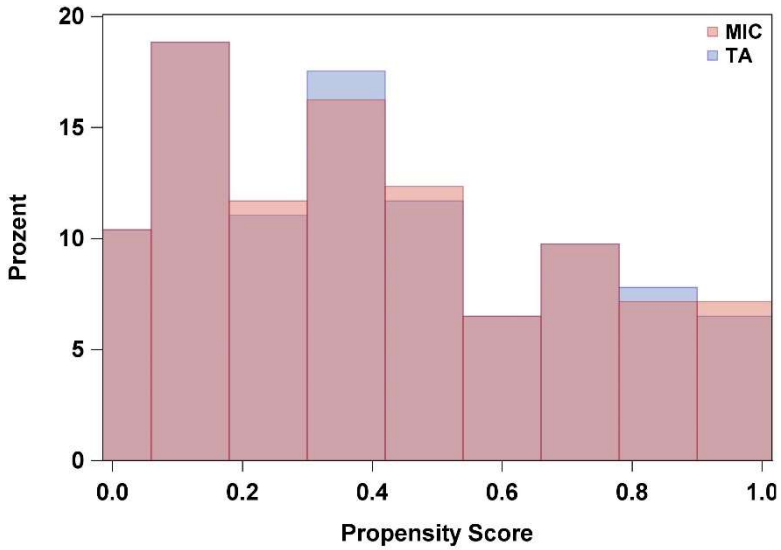
German Diabetes Center

## Propensity Score

**Vor** PS-Matching

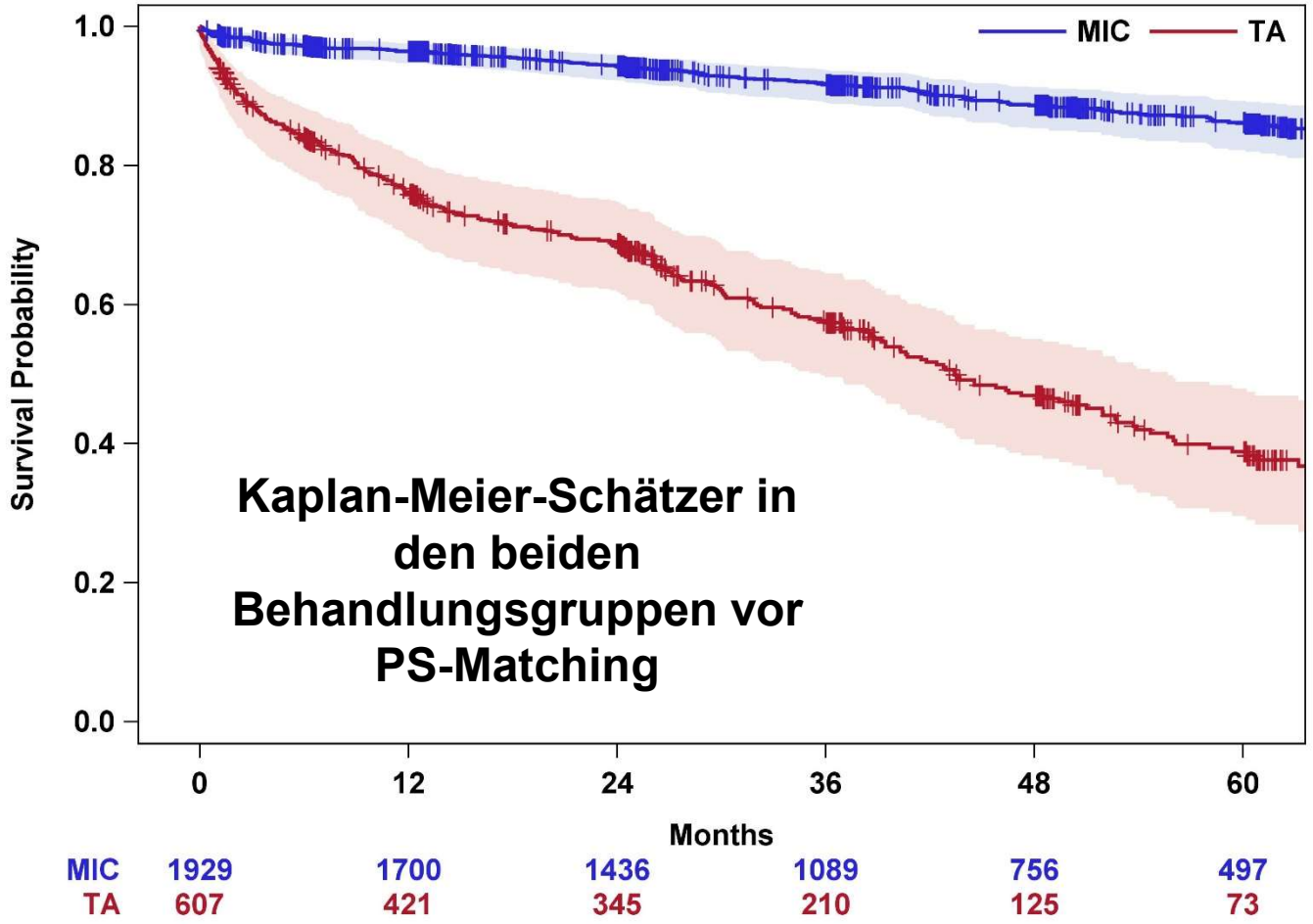



**Nach** PS-Matching



# Ein Beispiel aus der Herzchirurgie: Klinischer Outcome

DDZ German Diabetes Center



# Ein Beispiel aus der Herzchirurgie: Klinischer Outcome

DDZ German Diabetes Center

| | N / Anzahl Ereignisse | Hazard Ratio für Outcome Tod im Follow-Up [95%-KI, Ref.: MIC] |
|---|---|---|
| Unadjustiert | 2.536 / 479 | 6,40 [5,33; 7,69] |

# Zusammenfassung

DDZ German Diabetes Center

- Eine hinreichende/optimale Balanciertheit der relevanten Kovariablen in den Behandlungsgruppen ist das zentrale Qualitätskriterium einer PS-Analyse
- Aussagen über den Behandlungseffekt sollten nur für eine Population mit gutem Overlap gemacht werden
- In der Regel folgt aus einer guten Balanciertheit der Kovariablen auch ein guter Overlap des PS

# Zusammenfassung

DDZ German Diabetes Center

- Bei a priori stark unterschiedlichen Behandlungsgruppen verändert eine gute Balanciertheit die Population, über die Aussagen gemacht werden dürfen …
- … und insbesondere auch die Größe der Stichprobe

# Zusammenfassung

DDZ German Diabetes Center

- Das ist **eine Stärke des PS-Ansatzes und nicht etwa eine Schwäche** (auch wenn Power verloren geht …)
- Das schwache Verfahren ist in dieser Situation das herkömmliche Regressionsmodell, das über alles “drüberbügelt”.

# Ausblick

DDZ German Diabetes Center

- **Gesucht:** PS-Verfahren, das für Balanciertheit sorgt, aber keine Beobachtungen wegen schlechtem Overlap löscht.

- **Möglichkeit 1: Matching Weights** [Li/Greene 2013]
  - IPTW mit Gewicht

    min(PS,1-PS) / PS  in der Behandlungsgruppe

    min(PS,1-PS) / (1-PS)  in der Kontrollgruppe

    → Gewichte liegen immer zwischen 0 und 1

    → Keine Probleme mit extremen Gewichten wie beim Standard-IPTW-Verfahren

Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. Int J Biostat. 2013 Jul 31;9(2):215-34.

# Ausblick

DDZ German Diabetes Center

- **Gesucht:** PS-Verfahren, das für Balanciertheit sorgt, aber keine Beobachtungen wegen schlechtem Overlap löscht.

- **Möglichkeit 1: Matching Weights** [Li/Greene 2013]
  - Asymptotisch äquivalent zu PS-Matching
  - “The matching weight lets each subject to contribute only a fraction of itself, and that fraction is the matching weight”.

Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. Int J Biostat. 2013 Jul 31;9(2):215-34.

# Ausblick

DDZ German Diabetes Center

- **Gesucht:** PS-Verfahren, das für Balanciertheit sorgt, aber keine Beobachtungen wegen schlechtem Overlap löscht.

- **Möglichkeit 2: Overlap Weights** [Li et al. 2018]
  - IPTW mit Gewicht

    1-PS in der Behandlungsgruppe

    PS in der Kontrollgruppe

    → Gewichte liegen immer zwischen 0 und 1

    → Keine Probleme mit extremen Gewichten wie beim Standard-IPTW-Verfahren

Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J. Am. Stat. Assoc. 2018;113(521): 390-400.

# Ausblick

DDZ German Diabetes Center

- Matching und Overlap Weights werden auch als “equipoise weights” bezeichnet, …
  *“… [because they] emphasize a subpopulation that exhibits better overlap in the distribution of the (measured) covariates, similar to the target population of patients for whom there is clinical equipoise used in randomized clinical trials.”* [Zhou et al. 2020]

Zhou Y, Matsouaka RA, Thomas L. Stat Methods Med Res. 2020 Dec;29(12):3721-3756.

# Ausblick

DDZ German Diabetes Center

| | N / Anzahl Ereignisse | Hazard Ratio für Outcome Tod im Follow-Up [95%-KI, Ref.: MIC] |
|---|---|---|
| Unadjustiert | 2.536 / 479 | 6,40 [5,33; 7,69] |
| Regressionsadjustierung (Cox-Modell mit 23 Kovariablen) | 2.536 / 479 | 1,64 [1,23; 2,19] |
| Stratifiziertes Cox-Modell in der PS-gematchten Population | 308 / 108 | 1,25 [0,79; 1,99] |