Aus dem Institut für Biometrie und Epidemiologie
der Heinrich-Heine-Universität Düsseldorf
Direktor: Prof. Dr. sc. hum. Oliver Kuß

Wie relevant sind unbekannte Confounder in gematchten Propensity Score Analysen?
Eine Auswertung randomisierter Studien mit Hilfe des Propensity Scores

Dissertation

zur Erlangung des Grades eines Doktors der Medizin der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Matthaeus Miller
2020

Als Inauguraldissertation gedruckt mit Genehmigung der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf

gez.:

Dekan: Prof. Dr. med. Nikolaj Klöcker

Erstgutachter: Prof. Dr. sc. hum. Oliver Kuß

Zweitgutachter: Prof. Dr. phil. Nico Dragano

Diese Dissertation widme ich Jesus Christus und meiner Familie.

Teile dieser Arbeit wurden in folgendem Paper veröffentlicht:

Kuss O, Miller M. Unknown confounders did not bias the treatment effect when improving balance of known confounders in randomized trials. Journal of Clinical Epidemiology. Oktober 2020;126:9–16.

Zusammenfassung

Aufgrund der einzigartigen Möglichkeit, sowohl bekannte als auch unbekannte Patientenmerkmale homogen zu verteilen (1), werden randomisiert kontrollierte Studien (RCTs) als Goldstandard zur Beantwortung medizinischer Fragestellungen gehandhabt. Ein Vergleich mehrerer *Reviews* ergab, dass es „im Mittel [...] wenig Anzeichen für signifikante Unterschiede in der Effektschätzung zwischen Beobachtungsstudien und RCTs [gibt]“ (2). Nur bei fehlender Assoziation mit bekannten Störgrößen sind unbekannte Störgrößen problematisch (3). Die Assoziation zwischen bekannten und unbekannten Merkmalen wird in dieser Arbeit indirekt beurteilt, indem RCT-Analysen mit PS-gematchten Analysen derselben RCTs verglichen werden (4). Durch das *Matching* wird die Verteilung bekannter Patientenmerkmale verbessert (5). Eine fehlende Assoziation zwischen bekannten und unbekannten Merkmalen liegt dann vor, wenn eine bessere Verteilung bekannter Merkmale nicht zu einer besseren Verteilung der unbekannten Merkmale führt und schließlich in einer verzerrten Effektschätzung resultiert (4).

Grundlage für die Analysen dieser Arbeit sind 26 RCTs von Kent et al. (2016) (6). Die Patientenmerkmale aus *table 1* der Publikationen wurden in das PS-Modell einbezogen und für das PS-*Matching* nach Austins Empfehlung ein *Nearest Neighbor Caliper Matching* ohne *Replacement* verwendet (7). Ordinale, binäre und Ereignis-Zeit Zielgrößen wurden auf der log-Skala und stetige Zielgrößen mithilfe von Cohen's d geschätzt. Eine schrittweise Reduzierung der Caliperweiten intensivierte das PS-Modell, die PS-Analysen wurden vorab und blind für die RCT-Analyse durchgeführt. (4)

In 96,71% der Fälle liegen die Effektschätzer der PS-gematchten Analyse in den 95%-KI der Effektschätzer der RCT-Analyse. Der McNemar Test zeigt keinen signifikanten Unterschied der Schätzung signifikanter Effekte zwischen beiden Schätzmethoden; $\chi^2=1,0;\chi^2_{1;0,95}=3,841$; $p=0,3173$. Es besteht dabei eine sehr gute Übereinstimmung; $\kappa=0,81$ [0,7217; 0,8994]. Der T-Test ($\alpha=0,05$) zeigt keinen signifikanten Unterschied mit einem Mittelwert der Differenzen von -0,00117 [-0,0231; 0,0208]; $t=-0,10$; $p=0,9166$.

Insgesamt zeigen die Analysen keine signifikanten Unterschiede zwischen den beiden Schätzmethoden. Somit verursachen unbekannte Störgrößen keine signifikante Veränderung der Effektschätzer und sind somit sehr wahrscheinlich mit bekannten Störgrößen assoziiert (4). Die unbekannten Störgrößen, welche nur durch RCTs gleichmäßig verteilt werden können (8), verlieren somit an Gewicht, sodass der Einsatz gut konzipierter Beobachtungsstudien in Zukunft an Bedeutung gewinnen und noch vergleichbarere Ergebnisse zu RCTs liefern kann.

Abstract

Due to the exceptional potential of distributing known and unknown patient characteristics similarly (1) randomized controlled trials (RCTs) are treated as the gold standard design for answering medical questions. A comparison of many reviews revealed that "on average, there is little evidence for significant effect estimate differences between observational studies and RCTs [...]“ (2). Only having no association with known patient characteristics makes unknown patient characteristics difficult to handle (3). The association of known and unknown patient characteristics will be assessed indirectly in this work comparing RCT-analyses with PS-matched analyses of the same RCT (4). PS-matching will increase balance of known patient characteristics (5). An absence of association between unknown and known patient characteristics will be present if increased balance of known patient characteristics does not lead to an increased balance of unknown patient characteristics and finally results in confounded effect estimations (4).

In the analyses 26 RCTs of Kent et al (2016) are used (6). All the patients characteristics from table 1 of the articles are included for PS-matching which is using a nearest neighbor caliper matching without replacement as a matching algorithm following a recommendation of Austin (7). Ordinal, binary and time-to-event outcomes will be analyzed on the log-scale, while continuous outcomes will be estimated by Cohen's d. Diminishing step by step the caliper width will increase the degree of matching. All PS analyses are performed in advance and without referring to the outcomes of the full RCT analyses. (4)

96.71% of the PS analyses' treatment effects lie in the 95% CI of the full RCT's treatment effects. There is no significant difference in estimating significant treatment effects between the two methods according to McNemar's test; $\chi^2=1.0$,$\chi^2_{1;0.95}=3.841$, $p=0.3173$. At the same time there is very good agreement; $\kappa=0.81$ [0.7217; 0.8994]. Furthermore T-Test ($\alpha=0.05$) shows no significant difference with a mean difference of -0.00117 [-0.0231; 0.0208]; $t=-0.10$; $p=0.9166$.

In a nutshell this work's analyses detect no significant differences between the two methods. It follows that unknown patient characteristics do not change the treatment effects significantly and therefore they are likely to be associated with known patient characteristics (4). Thus unknown patient characteristics which can only be distributed equally by RCTs (8) lose weight so that usage of well concepted observational trials will gain importance in the future and deliver even more comparable results to RCTs.

Abkürzungsverzeichnis

ACCORD	*Action to Control Cardiovascular Risk in Diabetes*
ACCORDLIP	ACCORD Studie mit den Therapiearmen Placebo oder Fenofibrat
ACE	*Angiotensin Converting Enzyme*
AFFIRM	*Atrial Fibrillation Follow-up Investigation of Rhythm Management*
ALLHAT	*Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial*
ALT	Alanin-Amino-Transferase
AMIS	*Aspirin Myocardial Infarction Study*
ARDS	*Acute Respiratory Distress Syndrome*
ATN	*Acute Renal Failure Trial Network*
AUA	*American Urological Association*
BARI	*Bypass Angioplasty*
	Revascularization Investigation
BCG	Bacille Calmette Guerin
BDI	Beck-Depressions-Inventar
BEST	*Beta-Blocker Evaluation of Survival Trial*
BHAT	*Beta-Blocker Heart Attack Trial*
BMI	*Body Mass Index*
BPERFU	Tage bis zur spontan bakteriellen Peritonitis
CAST	*Cardiac Arrhythmia Suppression Trial*
CABG	*Coronary Artery Bypass Graft*
CI	*Confidence Interval*
DCCT/EDIC	*Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications*
DIG	*Digitalis Investigation Group*
DPP	*Diabetes Prevention Program*
DPPPM	DPP mit den Therapiearmen Placebo und Metformintherapie
DPPPL	DPP mit den Therapiearmen Placebo und Lebensstiländerung
ENRICHD	*Enhancing Recovery in Coronary Heart Disease Patients Randomized Trial*
EKG	Elektrokardiogramm
ESSI	*ENRICHD Social Support Instrument*
EVENT	invasive Therapie wegen benigner Prostatahyperplasie
FAVORIT	*Folic Acid for Vascular Outcome Reduction in Transplantation*
funfmi3m	Tage bis zum nicht tödlichen Myokardinfarkt innerhalb von 3 Monaten
FUTURA	*FondaparinUx Trial with Unfractionated Heparin during Revascularization in Acute Coronary Syndromes*
GFR	glomeruläre Filtrationsrate
GOSE	*Extended Glasgow Outcome Scale*
GRACE	*Genomics to combat Resistance Against Antibiotics in Community-Acquired lower respiratory-tract infection in Europe*
GSK	GlaxoSmithKline plc.
HALTC	*Hepatitis C Antiviral Long-Term Treatment against Cirrhosis trial*
HbA1c	glykiertes Hämoglobin
HCC	Hepatozelluläres Karzinom
HDFP	*Hypertension Detection and Follow-up Program*
HDL	*High density Lipoprotein*
HEMO	*Hemodialysis Study Group*
HRSD	*Hamilton Rating Scale for Depression*

HTN	Hypertonie
ICC	Intra-Klassen-Koeffizienten
ICED	*Index of Coexistent Disease*
ICU	*Intensive Care Unit*
IDS	*Index of Disease Severity*
IPTW	*Inverse Probability of Treatment Weighting*
IST	*International Stroke Trial*
ISTASP	IST Studie mit den Therapiearmen Placebo und Aspirin
KHK	koronare Herz Krankheit
KI	Konfidenzintervall
LDL	*Low density Lipoprotein*
LLT	*Lipid Lowering Trial*
LRC-CPPT	*Lipid Research Clinics Coronary Primary Prevention Trial*
LVEF	Linksventrikuläre Ejektionsfraktion
MAGIC	*Magnesium in Coronaries Trial*
MODS	*Multi-Organ Dysfunction Score*
MRFIT	*Multiple Risk Factor Intervention Trial*
MTOPS	*Medical Therapy of Prostatic Symptoms Study*
MTOPSPD	MTOPS Studie mit den Therapiearmen Placebo und Doxazosin
MTOPSPC	MTOPS Studie mit den Therapiearmen Doxazosin/Finasterid-Kombinationstherapie und Placebo
NHLBI	*National Heart, Lung, and Blood Institute*
NIDDK	*National Institute of Diabetes and Digestive and Kidney Diseases*
NUTDEC	Urosepsis
NYHA	*New York Heart Association*
nZDiff	Anzahl der berechneten z-Differenzen
OCCODE	Zustand des Patienten nach 6 Monaten
OAT	*Occluded Artery Trial*
OVL	*Overlapping Coefficient*
PCI	*Percutaneous Coronary Intervention*
PCORI	*Patient-Centered Outcome Research Institute*
PEACE	*Prevention of Events with Angiotensin Converting Enzyme Inhibition (PEACE) Trial*
PS	*Propensity Score*
PSA	Prostataspezifisches Antigen
PSSS	*Perceived Social Support Scale*
PTCA	Perkutane Transluminale Coronar-Angioplastie
RCT	*Randomized Controlled Trial*
ROC-HS/TBI	*Resuscitation Outcomes Consortium (Hypovolemic Shock/Traumatic Brain Injury)*
ROR	*Ratio of Odds Ratios*
RR	Relatives Risiko
SAS	*Statisticsl Analysis Software*
SHEP	*Systolic Hypertension in the Eldery Program*
SOFA	*Sequential Organ Failure Assessment*
SOLVD	*Studies of Left Ventricular Dysfunction*
T_DTH	Tage bis zum Tod
TIA	*Transient Ischemic Attack*
TIMI	*Thrombolysis in Myocadial Infarction*
TIMI II	*Thrombolysis in Myocardial Infarction Phase II Trial*
WHICaD	*Women's Health Initiative Calcium and Vitamin D*

1 Inhaltsverzeichnis

1 Einleitung

1.1 Hinführung

RCTs bieten die spezielle Möglichkeit nicht nur bekannte sondern auch unbekannte Patientenmerkmale homogen zu verteilen, was sie zum Goldstandard bei der Beantwortung medizinischer Fragestellungen macht (1,8). Im Gegensatz zu RCTs, welche eine geringe externe Validität besitzen und nur kurze Beobachtungszeiten erlauben (9,10), werden Beobachtungsstudien eben u.a. für eine bessere Übertragbarkeit auf die tatsächliche Population der Erkrankten und die Möglichkeit langer Beobachtungszeiten gelobt (11). Da jedoch im Studiendesign von Beobachtungsstudien auf eine Randomisierung gänzlich verzichtet wird, kommt es zu einer inhomogenen Verteilung bekannter und unbekannter Patientenmerkmale auf die zu vergleichenden Gruppen (8).

Angesichts des Spannungsfeldes zwischen randomisierten und nicht-randomisierten Studien sollte diskutiert werden, ob Beobachtungsstudien nicht zu RCT-gleichwertigen Ergebnissen führen können, wenn es um die Beantwortung derselben Frage geht (4). Die Relevanz der Frage widerspiegelnd, liegt eine Reihe systematischer Vergleiche zwischen beiden Studientypen vor, die auch schon in einem Cochrane *Review* zusammengefasst sind. Die Ergebnisse dieses *Reviews* sind möglicherweise überraschend, aber nichtsdestotrotz eindeutig, da es „wenig Anzeichen für signifikante [...] Unterschiede zwischen Beobachtungsstudien und RCTs [gibt], unabhängig vom spezifischen [...] Studiendesign, der Heterogenität oder der Einbeziehung von Studien mit pharmakologischen Interventionen“ (2). Die verbesserte methodische Adjustierung in Beobachtungsstudien reduzierte diese Unterschiede ebenfalls in der jüngeren Vergangenheit (2). Ebenso führt eine Verbesserung der Qualität der Beobachtungsstudien und eine verbesserte Kontrolle von Störgrößen zu einer Verringerung der Unterschiede (12).

1.2 Studiendesigns

Zum Erkenntnisgewinn in der Medizin dienen sowohl Primär- als auch Sekundärforschung. Ein Beispiel für die Primärforschung ist die klinische Studie, welche sowohl interventionell als auch nicht interventionell durchgeführt werden kann und sich u.a. mit der Beurteilung von Medizinprodukten, neuen chirurgischen oder medikamentösen Therapien auseinandersetzt. (13) Im weiteren Verlauf werden sowohl RCTs (interventionell) als auch Beobachtungsstudien (nicht interventionell) näher beleuchtet.

1.2.1 Randomisiert kontrollierte Studien

RCTs gehören zu den interventionellen Studien und werden als Goldstandard zur Bewertung von Therapieeffekten gehandhabt. Beispielsweise kann mit RCTs darüber entschieden werden, ob ein Medikament zugelassen wird oder nicht. Die festgelegten Therapien werden dabei zufällig d.h. randomisiert zugeteilt; entweder erhält der Studienteilnehmer die zu testende Therapie oder er erhält eine schon bewährte Therapie oder ein Placebo. (14) Letzteres darf angewandt werden, wenn der Teilnehmer dadurch keinen Nachteil oder Schaden erfährt (15). Das Festlegen von Ein- und Ausschlusskriterien führt dazu, dass ein bestimmtes Patientenkollektiv untersucht wird, welches homogen in Bezug auf die Patientenmerkmale ist (14).

Um Verzerrungen zu vermeiden und Aussagen darüber zu machen, ob die angewandte Therapie kausal für die Unterschiede der Effektschätzer zwischen den Behandlungsgruppen ist, können mehrere Maßnahmen verfolgt werden. Zum einen wird eine Randomisierung durchgeführt. Dabei werden die festgelegten Therapien den Studienteilnehmern zufällig zugeordnet. Das Resultat davon ist einerseits eine Strukturgleichheit, was bedeutet, dass beide Gruppen homogen bezüglich der Patientenmerkmale und der Störgrößen sind, andererseits eine Verhinderung des Selektionsbias. Dadurch kann die Aussage gemacht werden, dass der Unterschied der Effektschätzer der verglichenen Gruppen durch die Therapie allein entstand. (14)

Daneben wird auch eine Verblindung durchgeführt (einfach, doppelt oder offen/unverblindet). Bei der Verblindung soll erreicht werden, dass Arzt und Patient nicht wissen, welchem Therapiearm der Patient zugeordnet wurde. Dabei ist bei einfacher Verblindung entweder der Arzt oder der Patient verblindet, bei der doppelten Verblindung beide, bei der offenen Variante keiner von beiden. Ziel des Ganzen ist, dass keine subjektive Erwartungshaltung entwickelt wird, was Einfluss auf die *Compliance* einerseits und auf die Therapiebeurteilung andererseits haben könnte. (13,14)

Um RCTs noch besser an die gegebenen Fragestellungen anzupassen und so die Aussagekraft in verschiedenen Situation zu erhöhen, können bestimmte Designs gewählt werden z.B. die Minimisierung, *Cross-over*-Studie und das faktorielle Design. (16)

Häufiger Kritikpunkt von RCTs ist die geringe externe Validität. Durch die Einschlusskriterien wird die Zielpopulation der Studie festgelegt. Folglich unterscheidet sie sich von dem Patientenkollektiv im medizinischen Alltag beispielsweise in der Anzahl der Erkrankungen und der einzunehmenden Medikamente. Das realitätsentfernte Szenario führt dazu, dass Ergebnisse schlecht auf ein breites Patientenkollektiv übertragen werden können. (14)

Um dieses Problem anzugehen, können beispielsweise pragmatische Studien verwendet werden. Dabei soll ein breites Patientenkollektiv beispielsweise durch Reduzierung des Studiendesigns bezüglich der Ein-, Ausschlusskriterien und der Endpunkte erreicht und somit das Problem der geringen externen Validität angegangen werden. Auf der anderen Seite führt die Vereinfachung des Designs aber dazu, dass durch das Ausscheiden zusätzlicher Informationen und Nebenfragestellungen ein möglicher weiterer Erkenntnisgewinn gedrosselt wird. (16)

Nach Rothwell „[kann] von RCTs und systematischen *Reviews* [...] nicht erwartet werden, Ergebnisse zu produzieren, welche unmittelbar relevant für alle Patienten und alle Situationen sind, jedoch sollten sie - um extern valide zu sein - zumindest in so einer Art geplant und berichtet werden, welche Patienten und Klinikern erlaubt, zu beurteilen, auf wen sie vernünftigerweise zu beziehen sind“ (9).

1.2.2 Beobachtungsstudien

Beobachtungsstudien, anders auch nicht-interventionelle klinische Studien, eignen sich sehr gut nach Medikamentenzulassung zur Überwachung der Sicherheit dieser Medikamente oder zur Bewertung vom Risiko und Einfluss verschiedener Faktoren. Weiter können Neben- und Wechselwirkungen evaluiert und Informationen zum Risiko und Nutzen beim *Off Lable Use* gesammelt werden. (17) Außerdem sind „Beobachtungsstudien [...] oft das einzige Mittel der Wahl, wenn es um lange Beobachtungszeiträume oder seltene Ereignisse geht oder wenn experimentelle Studien unethisch wären“ (18).

Dabei liegt keine Randomisierung vor, d.h. der Arzt entscheidet für jeden Patienten selbst, welche Therapie angemessen erscheint (13). Durch die daraus resultierende geringe Homogenität der Behandlungsgruppen erlaubt dieses Studiendesign keine Aussage darüber, ob die Therapie für den Behandlungserfolg kausal ist oder nicht.

Die Ergebnisse können deskriptiv gedeutet und dazu benutzt werden Hypothesen zu erarbeiten. Außerdem kann in Beobachtungsstudien untersucht werden, welche Auswirkung die Therapie auf die Patienten hat, welche in RCTs ausgeschlossen wurden. Beobachtungsstudien können ggf. Antworten zu in RCTs nicht beantwortbaren Fragen geben und liefern neben der „Erhebung des therapeutischen Langzeitnutzens“ aufgrund der hohen externen Validität auch die Möglichkeit der „Reproduzierbarkeit der Ergebnisse [...] unter Alltagsbedingungen“. (17)

Neben der aufgrund fehlender Randomisierung entstandenen systematischen Verzerrung gibt es auch weitere Verzerrungen, beispielsweise das Informationsbias, das *Confounding* und das Simpsons Pradoxon, welche bis zu einem bestimmten Grad mithilfe eines

guten Studiendesigns angegangen und minimiert werden können. Beispielsweise führt eine hohe Teilnehmerrate zu einer hohen Repräsentativität der Studienpopulation. Weitere Probleme sind darüber hinaus ungenau oder fehlerhaft erhobene Daten. Lösungsmöglichkeiten liegen in einer guten Studienplanung mit standardisiertem Vorgehen. (18)

Als drittes hier aufgeführtes Problem sollen Störgrößen dienen. *Confounding* liegt vor, falls eine Variable oder ein Merkmal sowohl mit der Exposition oder dem Risiko, als auch mit der Zielgröße assoziiert ist, aber nicht den kausalen Weg zwischen beiden darstellt. In Beobachtungsstudien ist aufgrund fehlender Randomisierung die gleichmäßige Verteilung von bekannten und unbekannten Störgrößen nicht zu erwarten. Bekannte Störgrößen können beispielsweise durch den Vergleich zwischen rohen Effektschätzern und den stratifizierten Effektschätzern erkannt werden. Durch sogenanntes *Matching* kann das Problem der ungleichen Verteilung bekannter Störgrößen kausal angegangen werden. Durch Pärchenbildung wird nachträglich eine Strukturgleichheit für bekannte Patientenmerkmale geschaffen (s.u. 1.3.1). (18)

1.3 Propensity Score

Eine Methode zur gleichmäßigen Verteilung bekannter Patientenmerkmale in Beobachtungsstudien ist der *Propensity Score* (PS). Der PS, vergleichbar mit einem Komorbiditäten-*Score*, bezeichnet die Wahrscheinlichkeit, dass ein Patient unter Einschluss seiner bekannten Patientenmerkmale, die zu testende Intervention oder Therapie erhält. Im Gegensatz zu RCTs mit zwei gleich großen Gruppen, wo der wahre PS 0,5 beträgt, ist dieser in Beobachtungsstudien unbekannt und muss zuerst unter Zuhilfenahme logistischer Regressionsmodelle geschätzt werden. Bei diesem Modell stellen bekannte Patientenmerkmale - unbekannte Patientenmerkmale können hierbei nicht verwendet werden - die unabhängige Variable, die Therapiezuordnung dementsprechend die abhängige Variable dar. Nach der Berechnung des PS kann dann der Therapieeffekt geschätzt werden. (8) Es müssen also nicht alle möglichen *Confounder* einzeln analysiert werden, sondern der Therapieeffekt kann unter Berücksichtigung eines einzelnen Wertes geschätzt werden.

Bei der PS-Methode wird zu Beginn festgelegt, welche Therapie und welche Kontrolltherapie verteilt werden. Zudem müssen *Confounder*, andere Patientenmerkmale und Zielgrößen identifiziert werden. „Idealerweise wurden *Confounder* vor der Therapie und die Zielgrößen nach der Therapie gemessen [...].“ Nach der Identifikationsphase muss das Modell zunächst geschätzt, danach beurteilt werden und zu allerletzt können dann die Zielgrößen geschätzt werden. (19)

Einer der Vorteile der PS-Methode ist der, dass ähnlich wie bei RCT-Analysen im ersten Schritt zunächst durch Schätzung des PS und z.B. anschließendes *Matching* eine Homogenität der Gruppen bezüglich bekannter Patientenmerkmale geschaffen wird und danach erst in einem weiteren Schritt die Schätzung der Zielgrößen erfolgt. Eine methodische Schwierigkeit, die uns beispielsweise bei Regressionsmodellen begegnet, ist, dass hier auch trotz Gruppenunterschiede Therapieeffekte geschätzt werden, wozu „[...] Informationen von Nichtbehandelten benutzt [werden], die unter Umständen vollkommen anders sind als die Behandelten“. (8)

Zwar ist der PS kein Ersatz für die wertvolle Randomisierung, er ist aber sehr wohl eine gute Methode, um eine homogene Verteilung bekannter Patientenmerkmale auf die Behandlungsgruppen zu erzielen. Die gleichmäßige Verteilung unbekannter Patientenmerkmale bleibt Domäne der RCTs. (8) Weitere Fragestellungen, die in Zukunft auf dem Gebiet der PS-Methode angegangen werden müssen, sind u.a. fehlende Werte bei der Erhebung von Patientenmerkmalen und das Problem nicht binärer Therapiezuteilung, sodass eine nachträgliche Binarisierung durchgeführt werden könnte oder andere PS-Modelle benutzt werden müssten (19).

1.3.1 PS-Matching

Nach der Schätzung des PS-Wertes gibt es mehrere Möglichkeiten den Effekt der Behandlung zu schätzen. Neben der Stratifizierung, der Regressionsadjustierung für den PS und der Schätzung mittels IPTW wird das PS-*Matching* bevorzugt angewandt. (8,20,21) Für das PS-*Matching* wird nicht der eigentliche PS-Wert, sondern aus statistischen Gründen der logit, also der Logarithmus des *Odds* des PS, benutzt. (19) Beim PS-*Matching* wird jedem Patienten aus der Behandlungsgruppe ein Patient aus der Kontrollgruppe zugeordnet. Dabei werden immer solche Paare gebildet, bei denen die Patienten denselben oder sehr ähnlichen PS haben. Die Verteilung der Patientenmerkmale kann für beide Gruppen vor und nach PS-*Matching* miteinander verglichen werden und somit der direkte Effekt des *Matchings* dargestellt werden. Nachteil ist, dass durch das PS-*Matching* nur die Individuen einbezogen werden, welche auch einen passenden Partner haben. Somit fallen all diejenigen Individuen heraus, die kein Paar bilden können. Folglich wird die zu analysierende Population umso kleiner, je kleiner die Toleranz beim PS-*Matching* gewählt wird. Das damit einhergehende Absinken der *Power* ist somit ein weiterer Nachteil. (8)

Beim PS-*Matching* selbst gibt es noch unterschiedliche Methoden, welche sich u.a. in der Anzahl der nach dem *Matching* übriggebliebenen Individuen und dem den jeweiligen Individuen zukommenden Gewicht unterscheiden. Neben Subklassifikation, *Full Matching* oder

der Gewichtung wird vor allem das *Nearest Neighbor Matching* benutzt. Hierbei werden Pärchen gebildet aus jeweils einem Studienteilnehmer der beiden Gruppen. Bei der Pärchenbildung werden die Studienteilnehmer so ausgewählt, dass die beiden PS-Werte so nah wie nur möglich beieinander sind. Die Gefahr für ein Individuum aus der Behandlungsgruppe keinen Partner zu finden, sogenanntes *Poor Matching*, kann durch die Wahl unterschiedlicher Caliperweiten (s.u. 1.3.2) umgangen werden. (3)

Das optimale *Matching* ist eine Variante, welche dazu führt, dass nicht primär die Gruppen, sondern die Paare an sich sehr ähnlich sind (3). Eine weitere Möglichkeit ist es, pro behandelte Person mehrere *Matches* (festgelegte oder variable Anzahl) zu vollziehen, was zwar zu einer größeren Verzerrung, aber auch zu einer verminderten Varianz führt (22,23). Es besteht außerdem die Möglichkeit das *Matching* mit *Replacement* durchzuführen. Dabei können passende Kontrollen öfter zugeordnet werden. Die Anzahl davon wird dann mitberücksichtigt, indem diese als Gewichtung miteinbezogen wird. (3)

Insgesamt 12 verschiedene Matchingalgorithmen wurden von Austin untereinander verglichen. Dabei empfiehlt er zusammenfassend das *Nearest Neighbor Matching* ohne *Replacement*, da diese Methode im Vergleich zu den kleinsten Verzerrungen führe. (7)

1.3.2 Caliperweite

Mit Zuhilfenahme der Patientenmerkmale wird zuerst die Wahrscheinlichkeit für das Erhalten der Therapie errechnet, was der PS ist. Zum weiteren *Matching* wird der PS-logit benutzt. (8,19) Da nicht immer gewährleistet ist, dass die zwei Individuen eines Paares genau denselben PS-Wert haben, kann das PS-*Matching* mit unterschiedlicher Toleranz durchgeführt werden. Der dazu benutzte Wert zur Quantifizierung dieser Toleranz oder Abweichung ist die Caliperweite. Je kleiner die Caliperweite gewählt wird, desto genauer wird das *Matching* durch homogenere Gruppen. Die Verzerrung wird zwar kleiner, aber gleichzeitig auch die Studienpopulation, da mehr Individuen ohne passende Partner herausfallen. Als optimal wird von Austin eine Caliperweite gesehen, die das „0,2-fache der Standardabweichung des PS-logits“ darstellt. (24)

1.3.3 Validierung des PS-Matchings

„Wichtig ist, dass das vorgeschlagene Studiendesign einer Beobachtungsstudie nicht daran gemessen werden sollte, wie stark das PS-Modell zu den Daten oder dem vermuteten wahren Entscheidungsprozess passt. [...] [Sondern] das beste PS-Modell ist jenes, welches zu dem Design mit der besten Variablenbalance führt“ (25). Mit Balance ist eine gleichmäßige Verteilung der Patientenmerkmale auf die Behandlungsgruppen gemeint, damit der Behandlungseffekt nicht durch Patientenmerkmale verzerrt geschätzt wird.

Belitser et al. (26) bewerteten zur Beurteilung der Balanciertheit in PS gematchten Studien unter anderem den OVL, die Kolmogorov-Smirnov Distanz und die Levy Distanz. Zusammenfassend bewerten Belitser et al. „[...] Maße, die auf Mittelwertsvergleichen basieren, [...] etwas besser als die Kolmogorov-Smirnov und Levy Distanz“ (26). Insgesamt lieferte die T-Statistik die besten Ergebnisse. Kritisch zu sehen war dabei jedoch die Fallzahlabhängigkeit, welche bei den anderen drei Methoden nicht so stark ausgeprägt war. (26,27)

Bei Teststatistiken ist bei Fallzahlreduktion ein Powerverlust zu verzeichnen. Diese Fallzahlabhängigkeit sollte aber keine Eigenschaft eines Maßes zur Beurteilung der Balanciertheit sein. (28,29) Eine weiteres Mittel ist die standardisierte Differenz. Diese besitzt an sich zwar nicht den Nachteile einer Fallzahlabhängigkeit, deren Verteilung hingegen schon. (29) Außerdem „ist es unmöglich standardisierte Differenzen von Variablen verschiedener Skalenniveaus miteinander zu vergleichen [...]“ (27).

Kuss (27) stellte 2013 die z-Differenz als ein Maß zur Evaluierung der Balanciertheit von Patientenmerkmalen mithilfe einer normalverteilten Statistik vor. Diese z-Differenz kann bei stetigen aber auch nach einem weiteren Schritt bei binären und ordinalen Merkmalen angewandt werden. Dazu muss lediglich der Quotient aus der Risikodifferenz (binär) bzw. der Wilcoxon Statistik (ordinal) und dem Standardfehler gebildet werden. Nominale Merkmale können aufgrund des Fehlens eines normalverteilten Maßes nicht direkt berücksichtigt werden, sondern müssten zerlegt und binarisiert werden. Die Vorteile der z-Differenz sind einerseits der Vergleich zwischen mehreren Skalenniveaus, andererseits die fallzahlunabhängige Verteilung der z-Differenz (jedoch mit Fallzahlabhängigkeit der z-Differenz an sich). Hinzu kommt der Vorteil der Visualisierung, mithilfe derer ein Vergleich vor und nach *Matching* möglich ist. Somit kann der unmittelbare Effekt des *Matchings* dargestellt werden. Außerdem kommt die Tatsache hinzu, dass die Berechnung der z-Differenz nicht viel komplizierter ist als die der standardisierten Differenz. Um auf die Bewertung von Belitser et al. (26,27) zurückzukommen, ist es wichtig zu unterstreichen, dass auch die z-Differenz eine T-Statistik ist und somit die Ergebnisse von Belitser et al. auch für sie gelten, wobei nicht direkt die z-Differenz aber deren Verteilung sogar fallzahlunabhängig ist. Die Übertragbarkeit der Ergebnisse müsste natürlich erst einmal speziell für die z-Differenz geprüft werden. Weiter muss noch genauer evaluiert werden, inwiefern die z-Differenz besser auf nominale Merkmale angewandt werden kann und wie gut sich die z-Differenz auf andere PS-Methoden als das *Matching* übertragen lässt. (27)

Nachdem die z-Differenz für jedes ins PS-Modell einbezogene Patientenmerkmal berechnet worden ist, kann schließlich die Summe der quadrierten z-Differenzen berechnet

werden und somit als Gesamtmaß für die Balanciertheit einer PS-gematchten Studie zu einer bestimmten Caliperweite dienen. Wird das *Matching* mit mehreren Caliperweiten durchgeführt, kann die Balanciertheit mithilfe der Summe der quadrierten z-Differenzen bewertet und die Caliperweite mit der besten Balanciertheit gewählt werden.

1.4 Ziele der Arbeit

In dieser Arbeit wird eine weitere empirische Methode zur Messung von Unterschieden zwischen randomisierten und nicht-randomisierten Studien vorgeschlagen, indem der Einfluss unbekannter Störgrößen mit Hilfe des PS evaluiert wird (4). Diese bedrohen bei fehlender Assoziation mit bekannten Störgrößen die Schätzung des wahren Therapieeffektes. Ansonsten könnte eine einfache Adjustierung für bekannte Störgrößen diesen Störeinfluss mitadjustieren und neutralisieren. (3)

Ob eine Assoziation zwischen unbekannten und bekannten Störgrößen vorliegt, kann zwar nicht beobachtet, jedoch indirekt beurteilt werden. Bei einer PS-Analyse einer randomisiert kontrollierten Studie, in welcher unbekannte und bekannte Störgrößen schon gleichmäßig verteilt sind, erhöht sich durch Reduzierung der Caliperweite - also durch Intensivierung des *Matching* - die Balance der bekannten Störgrößen (5). Bei vorliegender Assoziation würde eine Balanceverbesserung der bekannten Störgrößen ebenfalls die Balance der unbekannten Störgrößen verbessern, sodass sich der wahre Behandlungseffekt aus der RCT-Analyse nicht signifikant von dem aus der PS-Analyse unterscheiden würde. Bei fehlender Assoziation jedoch käme es zu keiner Balancesteigerung der unbekannten Störgrößen, was in der Schätzung der Behandlungseffekte zu einem Unterschied zwischen RCT- und PS-Analyse führen würde. (4)

Die Aussage darüber, ob bekannte und unbekannte Patientenmerkmale miteinander assoziiert sind, würde wichtige Informationen hinsichtlich zukünftiger Studienplanung und -durchführung liefern. Falls bekannte und unbekannte Patientenmerkmale miteinander assoziiert wären, könnte sich die Studienplanung gegebenenfalls in Richtung von Beobachtungsstudien verschieben, die kostengünstiger, leichter durchzuführen sind, eine längere Beobachtungsdauer und größere Stichproben erlauben und zudem eine bessere externe Validität vorweisen als RCTs (11). Die gleichmäßige Verteilung der unbekannten Patientenmerkmale, welche den RCTs vorbehalten ist (8), wäre dann nicht mehr so stark relevant, da eine Assoziation zu den bekannten Patientenmerkmalen bestehen würde. Methoden, die bekannte Patientenmerkmale nachträglich gleichmäßig verteilen können, wie der *PS*, würden an Wichtigkeit gewinnen. Durch Assoziation mit den unbekannten Patientenmerkmalen würde eine bessere Verteilung der bekannten Patientenmerkmale

mithilfe des PS auch zu einer besseren Verteilung der unbekannten Patientenmerkmale führen. So könnten durch Beobachtungsstudien und „nachträgliche Randomisierung“ noch ähnlichere Ergebnisse zu RCTs erzielt werden und die Unterschiede zwischen den beiden Studientypen noch weiter minimiert werden. Bestehe zwischen bekannten und unbekannten Patientenmerkmalen keine Assoziation, würde dies auch einen erkenntnistheoretischen Informationsgewinn bringen und die RCTs in der Eigenschaft, als einziger Studientyp bekannte und unbekannte (1,11) Patientenmerkmale gleichermaßen verteilen zu können, bekräftigen.

2 Material und Methoden

2.1 Benutzte randomisierte Studien und Artikel

Die für die Analyse zugrunde liegenden randomisierten Studien entstammen aus dem von Kent et al. (2016) verfassten Artikel „*Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials*“. In diesem Artikel wurden die „Risikoheterogenität mittels *extreme quartile risk ratio* (EQRR [...]), die statistische Schiefe mittels *median mean risk ratio* (MMRR [...]) [und] [...] die Heterogenität des Behandlungseffektes [...] über Risikoschichten [anhand 32 RCTs untersucht]“. (6)

Diese Studien sind in verschiedenen Datenbanken zu finden. In der Datenbank des NIDDK sind 7 Datensätze erhältlich: ATN, DCCT/EDIC, DPP, FAVORIT, HALT-C, HEMO und MTOPS. Weitere 20 Datensätze sind im Datenregister des NHLBI erhältlich: ACCORD, AFFIRM, ALLHAT, AMIS, BARI, BEST, BHAT, CAST, DIG, ENRICHD, HDFP, LRC-CPPT, MAGIC, MRFIT, OAT, PEACE, ROC-HS/TBI, SHEP, SOLVD und TIMI II. Die IST-Studie und die von GSK gesponserte FUTURA-Studie sind in keinem der beiden Register aufzufinden, jedoch frei im Internet verfügbar. (30,31)

Für ein positives Ethikvotum waren Informationen nötig, die über die Art der Entidentifizierung der Studienpopulationen Auskunft geben. Sowohl die Studiendatensätze des NHLBI, des NIDDK als auch der IST-Studie sind entweder anonymisiert oder pseudonymisiert. Eine vergleichbare Information liegt zu der FUTURA-Studie nicht vor, sodass diese deswegen aus der weiteren Analyse ausgeschlossen wird (4).

Zu den in Kent et al. (2016) aufgelisteten Studien sind auch die jeweiligen Artikel aus medizinischen Fachzeitschriften wie beispielsweise dem *New England Journal of Medicine* oder dem *Journal of the American Medical Association* angegeben. Diese Artikel sind entweder auf den Internetseiten der jeweiligen Fachzeitschrift zugängig (32,33) oder bei Elsevier mithilfe der Campuslizenz erhältlich (34). Nicht frei zugängig sind Artikel zu den Studien BHAT, CPPT, HDFP, MRFIT, SHEP und TIMI II. Diese können aus Sammelwerken der Universitäts- und Landesbibliothek Düsseldorf entnommen werden.

Für die Anwendung des PS sind sowohl Patientenmerkmale als auch Zielgrößen notwendig. Patientenmerkmale können aus *table1*, Zielgrößen aus weiteren Tabellen des jeweiligen Artikels entnommen werden. In vereinzelten Fällen liegen in den von Kent et al. (2016) angegebenen Artikeln keine Patientenmerkmale vor, wie z.B. in CAST (35) oder CPPT (36). In diesen Fällen werden Patientenmerkmale genommen, welche in anderen Artikeln dieser Studie enthalten sind (37,38). Da die ACCORD Studie aus mehreren Unterstudien besteht, wird zu jeder Unterstudie ein separater Artikel mit eigenen

Patientenmerkmalen und Zielgrößen benutzt (39–41). Die Titel dieser Artikel sind alle in einer PDF-Datei (*ACCORD Public Use Data - Overview - 1. Synopsis*) im Datensatz der ACCORD-Studie aufgeführt. In der DCCT-Studie sind die Zielgrößen aus dem in Kent et al. (2016) angegebenen Artikel aufgrund der fehlenden Angabe zur Anzahl der eingetroffenen Ereignisse nicht brauchbar (42). Die zugehörigen Werte werden deswegen aus den Dateien *„MacrovascularCODEBOOK_Rev5“* und *„RetinopathyCODEBOOK“* entnommen. Weitere Zielgrößen befinden sich in einem im Datensatz vorliegenden Artikel zu makrovaskulären Ereignissen (43).

2.2 Erklärung

Die ATN, DCCT/EDIC, DPP, FAVORIT, HEMO, MTOPS, HALT-C wurden durchgeführt von den ATN, DCCT/EDIC, DPP, FAVORIT, HEMO, MTOPS, HALT-C Forschungsgruppen und unterstützt von dem *National Institute of Diabetes and Digestive and KidneyDiseases* (NIDDK). Die o.g. Datensätze wurden durch die *NIDDK Central Repositories* zur Verfügung gestellt. Diese Dissertation wurde nicht in Zusammenarbeit mit den o.g. Forschungsgruppen angefertigt und spiegelt nicht notwendigerweise die Ansichten und Meinungen derer oder die des NIDDK wider.

Weiter wurde diese Dissertation angefertigt mithilfe der Datensätze von ACCORD, AFFIRM, ALLHAT, AMIS, BARI, BEST, BHAT, CAST, DIG, ENRICHD, HDFP, LRCPPT, MAGIC, MRFIT, OAT, PEACE, ROCHS_TBI, SHEP, SOLVD, TIMI2, welche durch das NHLBI *Biologic Specimen and Data Repository Information Coordination Center* zur Verfügung gestellt wurden. Sie spiegelt nicht notwendigerweise die Meinungen und Ansichten von ACCORD, AFFIRM, ALLHAT, AMIS, BARI, BEST, BHAT, CAST, DIG, ENRICHD, HDFP, LRCPPT, MAGIC, MRFIT, OAT, PEACE, ROCHS_TBI, SHEP, SOLVD, TIMI2 oder NHLBI wider.

2.3 Datenvorbereitung

Die zugrunde liegenden Artikel werden zunächst auf Patientenmerkmale und Zielgrößen untersucht. In den Studiendatensätzen befindet sich meist eine Datei namens „*Data Dictionary*“, mit der abgeglichen werden kann, wie die benötigten Variablen benannt und in welcher Datei sie zu finden sind. Die Studiendatensätze unterscheiden sich in der Anzahl der im Datensatz enthaltenen Dateien. Einige wenige Studien besitzen eine einzige Datei mit allen Variablen, der Großteil der Studien jedoch besitzt viele Einzeldateien. Für die Aufschlüsselung der verschiedenen Variablen gibt es in den meisten Fällen eine Formattextdatei, aus welcher hervorgeht, welche numerische Variablenausprägung welchem Inhalt zugeordnet ist. In Ausnahmefällen müssen die vom Patienten auszufüllenden Studienformulare gesichtet und

anhand derer die Aufschlüsselung manuell ermittelt werden. Wurden die im Artikel aufgelisteten Variablen in den Dateien nicht gefunden, können sie für die weiteren Analysen nicht verwendet werden.

Als nächstes werden die Studien in SAS (*University Edition*2.7.9.4 M5) eingelesen und bearbeitet. Zu dieser Bearbeitung gehört die Umwandlung binärer Variablen in ein einheitliches Format, das als einzige Variablenausprägungen die Werte 0 und 1 enthält. Weiter gehört die Umwandlung alphanumerischer Variablen in numerische Variablen dazu, die Zuweisung von Etiketten in deutscher Sprache zu jeder Variable und zum Schluss die Entfernung der Formate, welche den Rohdateien zugewiesen waren. Ist die Datei so weit vorbereitet, können die Reproduktion der Effektschätzer aus den Artikeln und die Berechnung fehlender Werte bei den Patientenmerkmalen stattfinden. Zum Schluss werden anhand festgelegter Regeln Variablen aussortiert, sodass finale Dateien entstehen, welche schließlich für die Auswertung benutzt werden können.

2.3.1 Validierung der Zielgrößen

Nach erfolgreicher Datenextraktion und -bearbeitung können die Zielgrößen reproduziert werden und mit den Ergebnissen der Artikel verglichen werden. Zur Reproduktion der Zielgrößen werden für binäre Zielgrößen „PROC LOGISTIC“, für Überlebensvariablen „PROC PHREG“, für ordinale Zielgrößen „PROC GLIMMIX“, für stetige Zielgrößen „PROC MIXED“ benutzt und für jede Studie als Schleife programmiert, sodass alle Zielgrößen einer Studie in einem Durchlauf reproduziert werden können.

2.3.2 Regeln zur Aussortierung von Variablen

Vor der Auswertung der Datensätze mithilfe des PS-Makros wurden *a priori* Regeln zum Aussortieren von Patientenmerkmalen und Zielgrößen festgelegt. Ziel davon ist, nur die Zielgrößen mit der besten Reproduktion für die weiteren Analysen zu benutzen. Ein weiteres Interesse liegt darin, möglichst keine Korrelation unter den Zielgrößen zuzulassen und eine genügend große *Power* zur Messung des Behandlungseffekts zu erzielen. In Einzelfällen muss auch abgewogen werden, welche Zielgröße die patientenrelevanteste ist (4).

Bei Zielgrößen, die zu mehreren Zeitpunkten aufgenommen wurden, werden diejenigen in die weiteren Analysen eingeschlossen, welche zum frühesten Zeitpunkt aufgezeichnet wurden. Somit wird die Anzahl fehlender Werte, die mit der Dauer zunehmen würde, minimal gehalten. (4)

Binäre Zielgrößen, die Teilinformationen von bestehenden stetigen Variablen enthalten, werden ausgeschlossen. Wurden mehrere Zielgrößen unter einer Gesamtvariable zusammengefasst, werden diese Gesamtvariablen ausgeschlossen. Auch bei kombinierten

Endpunkten, wird auf diese verzichtet, falls die Zielgrößen auch einzeln vorhanden sind. Falls nicht alle Zielgrößen des kombinierten Endpunktes auch einzeln vorhanden sind, werden diejenigen heraus gerechnet, welche schon als separate Zielgrößen bestehen. (4)

Untereinander konkurrierende Zielgrößen, wie beispielsweise verschiedene Todesursachen, werden ausgeschlossen. Dadurch, dass ein Patient nicht an zwei verschiedenen Ursachen sterben kann, würden die beiden Zielgrößen miteinander korrelieren. Weiter werden kombinierte Endpunkte, die einen letalen Ausgang mitberücksichtigen, so verändert, dass der letale Anteil herausgenommen wird, falls dieser schon als eigenständige Zielgröße verfügbar ist. (4)

Weichen die Ergebnisse aus Reproduktion und Paper in beiden Behandlungsgruppen jeweils um mehr als 20% voneinander ab, werden diese Zielgrößen ausgeschlossen. Patientenmerkmale mit über 10% fehlenden Werten werden ebenfalls ausgeschlossen. Ein weiteres Kriterium für den Einschluss der Studien ist die 1:1-Randomisierung. Hat eine Studie mehrere Therapiearme, werden nur die verwendet, welche eben das Kriterium einer 1:1-Randomisierung erfüllen. Bei Abweichungen von einer 1:1-Randomisierung würden nämlich beim *Matching* direkt zu Beginn zu viele Beobachtungen wegfallen, da schon vom Studiendesign her nicht jedem Patienten der Therapiegruppe ein Partner aus der Kontrollgruppe zugeordnet werden kann. (4)

2.3.3 Aussortierte Studien

Vollkommen aus der weiteren Analyse ausgeschlossen werden sowohl BARI als auch HDFP. Bei BARI steht nur eine Datei zur Verfügung, in welcher 6 Zielgrößen zur weiteren Analyse benutzt werden könnten. Da jedoch keinerlei Vergleichswerte zu diesen Zielgrößen in den Artikeln auffindbar sind, muss diese Studie ausgeschlossen werden. Bei HDFP gibt es auch nur eine Datei. Eine Auflösung der „dat“-Datei ist aufgrund fehlender Informationen zu den Ausprägungen der Variablen nicht möglich. (4)

2.4 Übersicht der zur weiteren Analyse benutzten Studien

Nachfolgend sind alle Studien aufgelistet inklusive der Patientenmerkmale und Zielgrößen, die für das PS-*Matching* und die weiteren Analysen benutzt werden. Aus 26 Studien können insgesamt aufgrund von Unterstudien oder dem Vorhandensein mehrerer Studienarme eine Gesamtzahl von 37 Studiendatensätzen, jeweils mit zwei Therapiearmen, erstellt werden. Diese beinhalten Informationen zu 193.620 Patienten und stellen, nach erfolgter Aussortierung, 713 Patientenmerkmale und 213 Zielgrößen zur weiteren Analyse zur Verfügung. (4)

2.4.1 ACCORD

In der 2008 veröffentlichten randomisierten ACCORD-Studie wurde der Einfluss verschiedener Therapieregime bei Patienten mit Diabetes mellitus Typ II und hohem Risikoprofil auf unterschiedliche Zielgrößen evaluiert. Im Blutzuckertherapieregime, bei dem für die Analyse dieser Arbeit Daten von 10.251 Patienten, eine Zielgröße und 31 Patientenmerkmale zur Verfügung stehen, wurde eine intensive Blutzuckertherapie mit einer Standardblutzuckertherapie verglichen. Zu den Patientenmerkmalen gehören Alter, Geschlecht, Körpergröße, Körpergewicht, BMI, Hüftumfang, Bildungsgrad des Patienten, die Erkrankungsdauer des Diabetes mellitus, kardiovaskuläre Vorerkrankungen, Herzinsuffizienz als Vorerkrankung, Informationen zur aktuellen Medikation des Patienten, ethnische Zugehörigkeit, Raucherstatus, Blutdruck und Laborparameter wie HbA1c, Gesamtcholesterin und -triglyzeride, LDL, HDL, Nüchternblutzucker, Kalium und Serumkreatinin. Die Ereignis-Zeit-Zielgröße beschreibt die Zeit bis zu einer Hypoglykämie, die medizinische Hilfe benötigt. (40)

Bei dem Blutdrucktherapieregime stehen Daten von 4.733 Patienten, 6 Zielgrößen und 24 Patientenmerkmale zur Verfügung. Hier wurde eine intensive Blutdrucktherapie oder eine Standardblutdrucktherapie angewandt. Zu den Patientenmerkmalen gehören Alter, Geschlecht, Körpergröße, Körpergewicht, BMI, Bildungsgrad des Patienten, die Erkrankungsdauer des Diabetes mellitus, kardiovaskuläre Vorerkrankungen, Herzinsuffizienz als Vorerkrankung, Anzahl der aktuell vom Patienten eingenommenen Antihypertonika, ethnische Zugehörigkeit, Raucherstatus, Blutdruck und Laborparameter wie HbA1c, Gesamtcholesterin und -triglyzeride, LDL, HDL, Nüchternblutzucker, Kalium, Serumkreatinin, GFR und Harnkreatinin. Bei den 6 Zielgrößen handelt es sich um Ereignis-Zeit-Zielgrößen. Dabei wurde die Zeit bis zum Auftreten von Tod, nicht-tödlichem Myokard- und Hirninfarkt, nicht-tödlicher Herzinsuffizienz, unstabiler Angina und Nephropathie gemessen. (39)

Im Blutfettherapieregime wurde entweder ein Placebo oder Fenofibrat verabreicht. Zur weiteren Analyse ergeben sich hierbei Informationen zu 5.518 Patienten, darunter 34 Patientenmerkmale und 5 Zielgrößen. Zu den Patientenmerkmalen gehören Alter, Geschlecht, Körpergröße, Körpergewicht, BMI, Bildungsgrad des Patienten, die Erkrankungsdauer des Diabetes mellitus, kardiovaskuläre Vorerkrankungen, Herzinsuffizienz als Vorerkrankung, Amputation einer unteren Extremität in der Vorgeschichte, Informationen zur aktuellen Medikation des Patienten, ethnische Zugehörigkeit, Raucherstatus, Blutdruck und Laborparameter wie HbA1c, Gesamtcholesterin und -triglyzeride, LDL, HDL, Nüchternblutzucker, Kalium, GFR und Serumkreatinin. Bei den 5 Zielgrößen handelt es sich um Ereignis-Zeit-Zielgrößen. Dabei wurde die Zeit bis zum Auftreten von Tod, nicht-tödlichem

Myokard- und Hirninfarkt, nicht-tödlicher Herzinsuffizienz und unstabiler Angina verzeichnet. (41)

2.4.2 AFFIRM

In der randomisierten AFFIRM-Studie (2002) wurde der Einfluss zweier Therapieansätze bei Patienten mit Vorhofflimmern und hohem Schlaganfalls- oder Mortalitätsrisiko auf verschiedene Zielgrößen evaluiert. 6 Zielgrößen und 8 Patientenmerkmale von 4.060 Patienten können für das PS-*Matching* und die anschließenden Analysen verwendet werden. Therapiearme dieser Studie sind Rhythmus- und Frequenzkontrolle. Die Patientenmerkmale, welche für das *Matching* benutzt werden, sind Alter, Geschlecht, Herzinsuffizienz als Vorerkrankung, Dauer des qualifizierenden Vorhofflimmerns, Vorliegen einer fehlgeschlagenen Antiarrhythmikatherapie vor Randomisierung und die Frage danach, ob es sich um die erste Episode des Vorhofflimmerns handelt. Außerdem werden auch noch berücksichtigt die Art der dominierenden Herzerkrankung und die Zugehörigkeit zu einer ethnischen Minderheit. Zu den Zielgrößen gehören 5 Ereignis-Zeit-Zielgrößen - Tod, Arrhythmie, starke Blutung, ischämischer Hirninfarkt, ein anderes Ereignis - und eine binäre Zielgröße, nämlich die Frage nach Hospitalisation. (44)

2.4.3 ALLHAT HTN

Der Einfluss von Amlodipin und Lisinopril bei Patienten mit Hypertonie ersten oder zweiten Grades und mindestens einem kardiovaskulären Risikofaktor auf unten genannte Zielgrößen wurde in der 2002 veröffentlichten randomisierten ALLHAT-HTN-Studie evaluiert. Durch die Voraussetzung einer 1:1-Randomisierung entfällt die Chlorthalidongruppe hierbei, da ansonsten die Gruppengrößen zu unterschiedlich wären. Es stehen Daten von 18.102 Patientenzur Verfügung, dazu 16 Zielgrößen und 21 Patientenmerkmale. Patientenmerkmale, welche mittels PS für das *Matching* benutzt werden, sind neben Alter, Geschlecht, BMI und Blutdruck auch das Vorliegen antihypertensiver Medikation, eines Myokard- oder Hirninfarktes, einer koronaren Revaskularisierung, ST-Senkung, T-Negativierung, koronaren Herzerkrankung, anderer atherosklerotischer Gefäßerkrankungen oder eines Diabetes mellitus in der Vorgeschichte. Zudem auch noch der Raucherstatus, Bildungsgrad, die ethnische Zugehörigkeit, ein erniedrigter HDL in den letzten 5 Jahren, eine per EKG oder Echokardiographie nachgewiesene linksventrikuläre Hypertrophie in den letzten 2 Jahren, die Einnahme von Aspirin oder Östrogenpräparaten und die Zugehörigkeit zur Fettsenkerstudie. 10 Zielgrößen sind Ereignis-Zeit-Zielgrößen und beschäftigen sich mit dem Eintritt von Tod, Hirninfarkt, Herzinsuffizienz, Krebs, Hospitalisation wegen Herzinsuffizienz, koronarer Revaskularisierung, paVK, terminaler Nierenerkrankung, Hospitalisation aufgrund Angina

pectoris oder gastrointestinaler Blutungen. Weitere 6 sind stetige Zielgrößen: Blutdruck nach einem Jahr und GFR, Kalium, Gesamtcholesterin und Nüchternblutzucker nach 2 Jahren. (45)

In der im Jahre 2002 veröffentlichten ALLHAT-LLT-Studie wurde ebenfalls der Einfluss zweier Therapiearme bei Patienten mit Hypertonie und Hypercholesterinämie evaluiert. Daten von 10.355 randomisierten Patienten bilden die zur Verfügung stehende Grundlage, darunter 6 Zielgrößen und 18 Patientenmerkmale. Therapeutisch wurde Pravastatin oder eine *usual care* Therapie angewandt. Patientenmerkmale, die genutzt werden können sind Alter, Geschlecht, BMI, Blutdruck, Raucherstatus und Bildungsgrad der Patienten. Außerdem noch die ethnische Zugehörigkeit, spanische Nationalität, das Vorliegen von antihypertensiver Therapie, die Einnahme von Aspirin oder Östrogenpräparaten, Diabetes oder koronaren Herzerkrankung und Laborwerte wie Gesamtcholesterin, LDL und HDL. Neben der stetigen Zielgröße Cholesterin nach 2 Jahren, werden die Ereignis-Zeit-Zielgrößen mit der Frage nach Tod, nicht-tödlichem Myokard-, Hirninfarkt, hospitalisierter Herzinsuffizienz und Krebs evaluiert. (46)

2.4.4 AMIS

In 1980 wurde die randomisierte AMIS-Studie veröffentlicht, in welcher der Einfluss von Aspirin unter anderem auf Überlebenszielgrößen bei Patienten nach Myokardinfarkt evaluiert wurde. Für die Analyse dieser Arbeit stehen Daten von 4.524 Patienten, 7 Zielgrößen und 45 Patientenmerkmale zur Verfügung. Therapeutisch wurde entweder Aspirin oder ein Placebo verabreicht. Neben Alter, Geschlecht, Körpergewicht, Blutdruck, Herzfrequenz, Cholesterin und Triglyzeriden, ethnischer Zugehörigkeit, Anzahl der erlebten Myokardinfarkte und dem Alter beim jüngsten Myokardinfarkt werden auch Patientenmerkmale wie Hirninfarkt, Lungenembolie, Kardiomegalie, Herzinsuffizienz mit NYHA Klassifikation und Angina pectoris in der Vorgeschichte für die PS-Analyse benutzt. Außerdem gehören auch noch verschiedene EKG-Veränderungen, der Raucherstatus inklusive Anzahl gerauchter Zigaretten bei aktuellen Rauchern und die aktuellen Medikamente zu den Patientenmerkmalen. Alle 7 Zielgrößen sind Ereignis-Zeit-Zielgrößen: Tod, Myokard-, Hirninfarkt, *intermittent cerebral ischemic attack,* periphere arterielle Verengung, Lungenembolie und kardiovaskuläre Operation. (47)

2.4.5 ATN

In der randomisierten ATN-Studie (2008) wurde der Einfluss von Nierenersatztherapieverfahren auf Überlebens- und Erholungszielgrößen bei Patienten mit akuter Niereninsuffizienz oder Sepsis evaluiert. Die Studienpopulation für das PS-*Matching* und die Analyse besteht aus 1.124 Patienten. 6 Zielgrößen und 21 Patientenmerkmale können dabei verwendet werden. Die Nierenersatzverfahren bestanden aus einer intensiven oder weniger intensiven Strategie. Patientenmerkmale, die für das *Matching* benutzt werden, sind

Alter, Geschlecht, Gewicht, ethnische Zugehörigkeit, Oligurie, künstliche Beatmung, Primärbehandlung, Sepsis, Grund für die akute Nierenerkrankung und Dialyse vor Randomisierung, sowie Serum Kreatinin, verschiedene *SOFA Scores* bzw. *Total Apache Score* zum Anfangszeitpunkt und der Blut-Harnstoff-Stickstoff. 2 Ereignis-Zeit-Zielgrößen betrachten die Ereignisse Tod und die Entlassung nach Hause innerhalb von 60 Tagen, 3 stetige Zielgrößen messen die Tage, die ein Patient frei von Nierenersatztherapie, ICU oder Krankenhaus bewältigte und die ordinale Zielgröße beschreibt den Grad der Erholung der Niere. (48)

2.4.6 BEST

In der in 2001 veröffentlichten BEST-Studie wurde der Einfluss des Betablockers Bucindolol auf Überlebensvariablenzielgrößen, Hospitalisation und kardiovaskuläre Ereignisse bei Patienten mit Herzinsuffizienz im Stadium III oder IV evaluiert. Informationen zu 2.707 randomisierten Patienten bilden den Datensatz für das *Matching*, wobei eine Zielgröße und 23 Patientenmerkmale zur Verfügung stehen. Die Kontrollgruppe erhielt ein Placebo. Die Patientenmerkmale sind Alter, Geschlecht, ethnische Zugehörigkeit, Körpergewicht, -größe, BMI, Herzfrequenz, Blutdruck, Raucherstatus inklusive der Dauer des Rauchens, Dauer, NYHA-Klassifikation und Grund der Herzinsuffizienz, LVEF und Angaben zu diversen Vorerkrankungen. Die Ereignis-Zeit-Zielgröße betrifft den Tod. (49)

2.4.7 BHAT

Der Einfluss von Propranolol auf Überlebenszielgrößen bei Patienten mit akutem Myokardinfarkt wurde in der in 1982 veröffentlichten BHAT-Studie Placebo kontrolliert evaluiert. Hierbei stehen Daten von 3.837 randomisierten Patienten, eine Zielgröße und 28 Patientenmerkmale zur Verfügung. Die Patientenmerkmale beinhalten Alter, Geschlecht, ethnische Zugehörigkeit, Gewicht, Raucherstatus, Blutdruck, Herzfrequenz, Cholesterin, Herz-Thorax-Quotienten, sowie Angaben zu ventrikulären Tachykardien, Vorhofflimmern oder Herzinsuffizienz im Krankenhaus vor Randomisierung, diversen Vorerkrankungen, der Einnahme bestimmter Medikamente und der Lokalisation des Myokardinfarktes. Der Tod beschreibt das Ereignis in der Ereignis-Zeit-Zielgröße. (50)

2.4.8 CAST

In der CAST-Studie (1991) wurde der Einfluss von Encainid und Flecainid auf die Mortalität und Morbidität von Patienten mit Myokardinfarkt evaluiert. 1.498 randomisierte Patienten bilden die Population dieser Studie. Dabei können 2 Zielgrößen und 20 Patientenmerkmale zur weiteren Analyse benutzt werden. Die Kontrollgruppe erhielt hierbei ein Placebo. Die Patientenmerkmale setzen sich zusammen aus Alter, Geschlecht, Raucherstatus, Herzfrequenz,

LVEF, Zeitpunkt des qualifizierenden Myokardinfarktes und diversen Angaben zu Vorerkrankungen und der Einnahme bestimmter Medikamente. Die Effekte von 2 Ereignis-Zeit-Zielgrößen werden berechnet: allgemeine Mortalität und Therapieabbruch aufgrund einer Herzinsuffizienz. (35,37)

2.4.9 CPPT

In 1984 wurde die randomisierte CPPT-Studie veröffentlicht, in welcher der Einfluss von Cholestyramin auf die Mortalität und das Risiko einer koronaren Herzerkrankung bei Patienten mit Hypercholesterinämie Placebo-kontrolliert evaluiert wird. Hierbei stehen 10 Zielgrößen und 7 Patientenmerkmale aus einem Kollektiv von 3.806 Patienten zur Verfügung. Patientenmerkmale, die für das PS-*Matching* benutzt werden, sind Alter, Blutdruck, Gesamtcholesterin, HDL, LDL und Raucherstatus. Der Effekt von 10 Ereignis-Zeit-Variablen, welche Ereignisse wie den Tod, nicht-tödlichen Myokardinfarkt, Angina, koronare Bypasschirurgie, Herzinsuffizienz, intraoperativen Myokardinfarkt, wiederbelebten koronaren Kollaps, Verdacht auf TIA oder atherothrombotischen Hirninfarkt und Claudicatio intermittens beschreiben, können dann berechnet werden. (36,38)

2.4.10 DCCT

In der in 1993 veröffentlichten DCCT-Studie wurde der Einfluss der Blutzuckereinstellung auf die Entwicklung von beispielsweise Retino-, Neuro-, Nephropathien und andere Zielgrößen bei Patienten mit insulinpflichtigem Diabetes mellitus randomisiert evaluiert. Hierbei stehen Daten von 1.441 Patienten zur Verfügung, darunter 6 Zielgrößen und 19 Patientenmerkmale. Es wurde entweder eine konventionelle oder intensive Diabetestherapie durchgeführt. Zu den Patientenmerkmalen gehören Alter, Geschlecht, Körpergewicht, Blutdruck, Raucherstatus, ethnische Zugehörigkeit, Laborparameter wie Albuminurie, Kreatininclearance, Serumcholesterin, -triglizeride, HDL, LDL, Blutzucker, HbA1c und außerdem noch Aussagen zur Dauer des insulinpflichtigen Diabetes, Insulindosis, Einstufung des Retinopathielevels, Vorliegen einer Retinopathie oder einer klinischen Neuropathie zum Ausgangszeitpunkt. Zu den 6 Zielgrößen gehören 4 Ereignis-Zeit-Zielgrößen, nämlich persistierende Retinopathie mit einer Veränderung um mehr als 3 Stufen, klinisch signifikantes Makulaödem, Übergewicht, Hypertonie, und 2 binäre Zielgrößen mit Hypercholesterinämie und einem großen Unfall als Ereignisse. (42,43)

2.4.11 DIG

In der randomisierten DIG-Studie von 1997 wurde der Einfluss von Herzglykosiden u.a. auf die Mortalität chronisch herzinsuffizienter Patienten evaluiert. Dabei besitzt der Datensatz

Informationen zu 7.788 Patienten, 13 Zielgrößen und 26 Patientenmerkmalen. Therapeutisch wurde ein Placebo oder Digoxin verabreicht. Alter, Geschlecht, ethnische Zugehörigkeit, Ejektionsfraktion und deren Ermittlungsmethode, Herz-Thorax-Quotient, Dauer, Klassifikation und Hauptgrund der Herzinsuffizienz, sowie Informationen zu aktuellen Symptomen, Medikamenten und Vorerkrankungen sowie die tägliche Dosis des Studienmedikamentes gehören zu den Patientenmerkmalen. 12 Ereignis-Zeit-Zielgrößen, darunter Tod, Gründe für eine Hospitalisation wie Verschlechterung der Herzinsuffizienz, ventrikuläre Arrhythmien, Digoxinintoxikation, Myokardinfarkt, unstabile Angina pectoris, Hirninfarkt, Herztransplantation, andere kardiovaskuläre Ereignisse, Atemwegsinfekte, nicht kardiovaskuläre oder nicht weiter spezifizierte Gründe und eine stetige Zielgröße, nämlich die Anzahl der Krankenhausaufenthalte, sind die Zielgrößen, die zum Therapievergleich benutzt werden. (51)

2.4.12 DPP

Der Einfluss von zwei Therapieansätzen auf die Entwicklung von Diabetes bei Patienten mit erhöhtem Diabetesrisiko wurde in der im Jahre 2002 veröffentlichten DPP-Studie randomisiert evaluiert. Hierbei liegen Daten von 2.058 Patienten vor, 7 Zielgrößen und 14 Patientenmerkmale. Dabei wurde die Placebotherapie mit einer Metformintherapie verglichen. Sowohl Alter, Geschlecht, ethnische Zugehörigkeit, BMI, Taillen-, Hüftumfang, Gewicht, als auch Informationen zu einem bekannten Diabetes in der Familie und Laborparametern wie Nüchternblutzucker, 2h-Blutzucker und HbA1c sind als Patientenmerkmale in die PS-gestützte Berechnung eingegangen. Dabei werden 7 stetige Zielgrößen wie Nüchterninsulin, -proinsulin, -blutzucker, systolischer und diastolischer Blutdruck, 2h-Blutzucker und das Körpergewicht jeweils alle nach einem Jahr, zum Therapievergleich benutzt. Der Therapiearm *lifestyle* und somit der Datensatz DPPPL wird aus der weiteren Analyse gestrichen (vgl. 2.7). (52)

2.4.13 ENRICHD

In der im Jahre 2003 veröffentlichten randomisierten ENRICHD-Studie wurde der Einfluss von kognitiver Verhaltenstherapie und Serotoninwiederaufnahmehemmern auf die Mortalität, Reinfarkte und die Entwicklung einer Depression bei Patienten mit Myokardinfarkt evaluiert. 2.481 Patienten stehen hierbei mit 8 Zielgrößen und 30 Patientenmerkmalen zur Verfügung. Unterschieden wurde eine Verhaltenstherapie in Kombination mit einer medikamentösen Therapie von einer *usual care* Therapie. Die dabei berücksichtigten Patientenmerkmale sind Alter, Geschlecht, ethnische Zugehörigkeit, Bildungsgrad, Familienstand, BMI, Raucherstatus, systolischer Blutdruck, DHoore Komorbiditäten Index, Angaben zu Vorerkrankungen,

Medikamenten, koronaren Interventionen in der Vorgeschichte, psychologischen Risikofaktoren und der Art der Behandlung des Indexmyokardinfarktes. Zu den Zielgrößen gehören 4 stetige Zielgrößen in Form von verschiedenen Scores, wie HRSD, BDI, ESSI, PSSS und 4 Ereignis-Zeit-Zielgrößen, nämlich Tod, Revaskularisierung, Hospitalisation aufgrund kardiovaskulärer Ereignisse und nicht-tödlicher Myokardinfarkt. (53)

2.4.14 FAVORIT

Die randomisierte FAVORIT-Studie (2011) untersuchte den Einfluss von Homocystein-senkenden Maßnahmen auf die Entwicklung kardiovaskulärer Ereignisse bei Patienten nach Nierentransplantation. Hierbei stehen 5 Zielgrößen und 18 Patientenmerkmale von 4.110 Patienten zur Verfügung, denen entweder eine hohe oder niedrige Dosis von Folsäure und Vitaminen B6 und B12 verabreicht wurde. Zu den für den PS benutzten Patientenmerkmalen zählen Alter, Geschlecht, Raucherstatus, ethnische Zugehörigkeit, BMI, Herkunftsland, Transplantatalter, Laborwerte wie Kreatinin, GFR, LDL, HDL, Gesamtcholesterin, Triglizeride und Angaben zu Vorerkrankungen. Zu den 5 Ereignis-Zeit-Zielgrößen gehören die Ereignisse Tod, Rückkehr zur Dialyse, koronare Revaskularisierung, Amputation oder Revaskularisierung der unteren Extremität, nicht-tödlicher Myokard- und Hirninfarkt. (54)

2.4.15 HALT-C

Im Jahre 2008 wurde in der veröffentlichten HALT-C-Studie der Einfluss einer langfristigen antiviralen Therapie auf die Mortalität, die Verschlechterung einer Leberzirrhose und die Entwicklung von HCC bei Patienten mit chronischer Hepatitis C Infektion ohne Ansprechen auf eine antivirale Therapie randomisiert evaluiert. Eine Therapie mit Peginterferonα-2a wurde nach 3,5 Jahren entweder fortgeführt oder beendet. Für das PS-*Matching* und die statistische Analyse stehen Daten von 1.050 Patienten, mit 6 Zielgrößen und 20 Patientenmerkmalen zur Verfügung. Die für das PS-Modell benutzten Patientenmerkmale beinhalten Alter, Geschlecht, ethnische Zugehörigkeit, Gesamtzahl der im Leben konsumierten alkoholischen Getränke, Laborparameter wie ALT, Gesamtbilirubin, Albumin, Prothrombinzeit, ALT-Quotient, Nachweisbarkeit der Zirrhose in der Biopsie, Probenlänge der Biopsie, Dauer der Infektion und deren Genotyp, Ishak Fibrose- und Entzündungsscore. Zu den 7 Ereignis-Zeit-Zielgrößen gehören das Eintreten von Tod, HCC, Varizenblutung, Aszites und hepatischer Enzephalopathie und ein Anstieg von über 7 Punkten im *Child Pugh Score*. (55)

2.4.16 HEMO

In der randomisierten HEMO-Studie (2002) wurde der Einfluss von verschiedenen Dialyseregimen auf die Mortalität und Morbidität bei dialysepflichtigen Patienten evaluiert.

Hierbei stehen in zwei eigens benannten Kohorten, FLUX und KTV, jeweils Daten von 1.846 Patienten, 4 Zielgrößen und 22 Patientenmerkmale zur Verfügung, bei denen die Dialyse einerseits mit hoher oder normaler Standarddosis (KTV), andererseits mit hohem oder niedrigem Fluß (FLUX) durchgeführt wurde. Patientenmerkmale, welche für die PS-Analyse berücksichtigt werden, sind Alter, Geschlecht, ethnische Zugehörigkeit, spanische Nationalität, Gewicht, Blutdruck, Laborparameter wie Serumkreatinin, Gesamtcholesterin, Albumin, Rest-Harnstoff Clearance, Angaben zur Dialyse, jeweils ein IDS Score zu ischämischer Herzerkrankung, Herzinsuffizienz, Herzrhythmusstörungen, Diabetes und anderen Herzerkrankungen und der ICED Score. Die 4 Ereignis-Zeit-Zielgrößen beschreiben das Auftreten von Tod, Albuminabfall und Hospitalisation aus kardiovaskulärer Ursache oder aufgrund einer Infektion. (56)

2.4.17 IST

In der IST-Studie von 1997 wurde der Einfluss antithrombotischer Therapie auf die Mortalität und den gesundheitlichen Zustand nach 6 Monaten bei Patienten mit ischämischem Schlaganfall randomisiert evaluiert. Hierbei stehen drei eigens benannte Kohorten, ASP, HEP und HEPDOS, zur Verfügung mit jeweils 3 Zielgrößen und 12 Patientenmerkmalen. Die ersten beiden Kohorten (ASP und HEP) bestehen jeweils aus 19.435 Patienten, bei denen jeweils die Hälfte der Patienten ein Placebo verabreicht bekam, die andere Hälfte je nach Kohorte Aspirin oder Heparin. Die dritte Kohorte (HEPDOS) enthält alle 9.717 Patienten, welche Heparin verabreicht bekamen, und unterscheidet zwischen niedriger und mittlerer Heparindosis. Die Patientenmerkmale sind Alter, Geschlecht, Blutdruck, Vigilanzzustand und Angaben zur Lokalisation des Hirninfarkt Syndroms, einer Schwäche im Bein, der Sichtbarkeit im CT und zur Einnahme von Heparin und Aspirin in den letzten 24h bzw. 72h. Zu den 2 Ereignis-Zeit-Zielgrößen gehören Tod und Lungenembolie, die ordinale Zielgröße zeigt den Zustand des Patienten nach 6 Monaten an. (57)

Die Studie wurde durchgeführt von Peter AG Sandercock, Maciej Niewada und Anna Członkowska, der „*International Stroke Trial Collaborative Group*“. Erhältlich ist der Datensatz unter folgendem Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104487/. Finanziert wurde die Studie von *UK Medical Research Council, the UK Stroke Association, European Union BIOMED-1 program,* Julius Brendel Trust und dem *Lottery Grants Board*. Unterstützt wurde die Studie durch Eli Lilly, Sterling Winthrop, *Sanofi, Bayer UK, National Heart Foundation grant, Nova Scotia Heart and Stroke Foundation grant, IGA Ministry of Health grant, McMaster INCLEN program, All India Institute of Medical Sciences, Norwegian Council on Cardiovascular Disease and Nycomed.* (57)

2.4.18 MAGIC

In der im Jahre 2002 veröffentlichten randomisierten MAGIC-Studie wurde der Einfluss von Magnesium unter anderem auf die Mortalität von Patienten mit akutem Myokardinfarkt evaluiert. Zur weiteren Analyse stehen Daten von 6.213 Patienten zur Verfügung mit 4 Zielgrößen und 13 Patientenmerkmalen. Die Therapien, die angewandt wurden, bestanden entweder aus der Gabe intravenösen Magnesiumsulfats oder eines Placebos. Die Patientenmerkmale, welche ins PS-Modell miteinbezogen werden, sind Alter, Geschlecht, systolischer Blutdruck, Herzfrequenz, Dauer zwischen Myokardinfarkt und Bolusgabe, Angaben zu Vorerkrankungen und Koronarinterventionen in der Vorgeschichte, Infarktlokalisation und TIMI Risikoscore. Die Ereignis-Zeit-Zielgröße beinhaltet den Tod als Ereignis, die 3 binären Zielgrößen beziehen sich auf den Abbruch der Medikamenteneinnahme wegen Hypotonie bzw. Bradykardie und auf eine Defibrillation aufgrund Kammerflimmerns oder anhaltender ventrikulärer Tachykardie. (58)

2.4.19 MRFIT

In der randomisierten MRFIT-Studie von 1982 wurde der Einfluss von zwei Behandlungsansätzen auf die Mortalität durch eine KHK bei Patienten mit erhöhtem KHK Risikoprofil evaluiert. Den 12.866 Patienten wurde entweder eine spezielle Intervention mit Beratung und Diät angeboten oder eine *usual care* Therapie. 4 Zielgrößen und 19 Patientenmerkmale stehen zur Auswertung zur Verfügung. Zu den Patientenmerkmalen gehören Alter, Geschlecht, Körpergewicht, ethnische Zugehörigkeit, Raucherstatus, Anzahl der gerauchten Zigaretten am Tag, Blutdruck, Laborparameter wie Triglizeride und Cholesterin im Plasma, HDL, LDL, Serum Thiozyanate, pathologische EKG-Veränderungen, Vorliegen einer Belastungsischämie und Angaben zu den zugeführten Kalorien. Zu den 2 stetigen Zielgrößen gehören Serumcholesterin und diastolischer Blutdruck nach einem Jahr, die eine binäre Zielgröße beschäftigt sich mit dem Raucherstatus und die Ereignis-Zeit-Zielgröße bezieht sich auf den Tod. (59)

2.4.20 MTOPS

Die im Jahre 2003 veröffentlichte MTOPS-Studie evaluierte der Einfluss von α-Blockern und 5α-Reduktasehemmern auf die klinische Progression von benigner Prostatahyperplasie. Hierbei stehen Daten von drei eigens benannten Kohorten, PD, PF und PC, zur Verfügung mit jeweils 4 Zielgrößen und 8 Patientenmerkmalen (die PD Kohorte hat 5 Zielgrößen). Die PD-Kohorte beinhaltet 1.599 Patienten mit den Thearpiearmen Placebo und Doxazosin, die PF-Kohorte 1.646 Patienten mit den Therapiearmen Placebo und Finasterid, die PC-Kohorte 1.634 Patienten mit den Therapiearmen Doxazosin/Finasterid-Kombinationstherapie und Placebo. Zu

den Patientenmerkmalen werden Alter, ethnische Zugehörigkeit, Serumkreatinin, maximale Urinflussrate, Restharnvolumen, Serum PSA, Prostatavolumen und der AUA *Symptom Score* gezählt. Zu den 5 binären Zielgrößen gehören das Vorliegen eines Anstieges um mehr als 3 Punkte im AUA *Symptom Score*, einer Retention, Urosepsis, Inkontinenz und einer invasiven Therapie wegen benigner Prostatahyperplasie. (60)

2.4.21 OAT

In der OAT-Studie (2006) wurde der Einfluss einer PCI auf die Mortalität, Reinfarktrate und das Entstehen von hochgradigen Herzinsuffizienzen bei Patienten mit Myokardinfarkt randomisiert evaluiert. 7 Zielgrößen und 28 Patientenmerkmale aus einer Studienpopulation von 2.166 Patienten können für das PS-*Matching* und die weiteren Analysen verarbeitet werden. Dabei wurde entweder eine optimale medikamentöse Therapie oder zusätzlich dazu eine PCI durchgeführt. Zu den Patientenmerkmalen zählen Alter, Geschlecht, Raucherstatus, ethnische Zugehörigkeit, GFR, Ejektionsfraktion, Ausbildung von Kollateralen, Vorliegen von Mehrgefäßerkrankung, Identifikation der infarktversorgenden Arterie, Angaben zum Insulingebrauch, durchgeführtem Stresstest, thrombolytischer Therapie innerhalb der ersten 24h nach Myokardinfarkt, Dauer zwischen Myokardinfarkt und Randomisierung, Killip Stufe des Indexmyokardinfarktes, NYHA Klassifikation vor Randomisierung, und Angaben zu Vorerkrankungen und Interventionen am Herzen in der Vorgeschichte. Zu den 6 Ereignis-Zeit-Zielgrößen gehören Tod, nicht-tödlicher Reinfarkt, Hirninfarkt, Hospitalisation oder Behandlung jeglicher Herzinsuffizienz, Revaskularisierung, CABG. Die binäre Zielgröße gibt Auskunft über das Auftreten wiederholter Erhöhung von kardialen Markern innerhalb 48h nach Randomisierung. (61)

2.4.22 PEACE

In der in 2004 veröffentlichten randomisierten PEACE-Studie wurde der Einfluss des ACE-Hemmers Trandolapril auf Mortalität, kardiovaskuläre Ereignisse und Revaskularisierung bei Patienten mit koronarer Herzerkrankung Placebo-kontrolliert evaluiert. Hierbei stehen Daten von 8.290 Patienten mit 7 Zielgrößen und 26 Patientenmerkmalen zur Verfügung. Die für das PS-*Matching* benutzten Patientenmerkmale beinhalten Informationen zu Alter, Geschlecht, Raucherstatus, Blutdruck, LVEF, Serum Cholesterin, Interventionen am Herzen in der Vorgeschichte und Angaben zu Vorerkrankungen und zur Einnahme bestimmter Medikamente. Zu den 7 Ereignis-Zeit-Zielgrößen zählen Tod, nicht-tödlicher Myokardinfarkt, PTCA, CABG, Hirninfarkt, Hospitalisation aufgrund einer Herzinsuffizienz und das Neuauftreten von Diabetes. (62)

2.4.23 ROC

In der ROC-TBI-Studie aus dem Jahre 2010 wurde der Einfluss hypertoner Kochsalzlösung u.a. auf die Mortalität, den Flüssigkeits- und Blutkonservenbedarf und das Auftreten nosokomialer Infektionen bei Patienten mit traumatischer Gehirnverletzung randomisiert evaluiert. Aufgrund der Voraussetzung der 1:1-Randomisierung wurde der Therapiearm mit isotoner Kochsalzlösung ausgeschlossen. Daten von 700 Patienten mit 16 Zielgrößen und 13 Patientenmerkmalen stehen zur Verfügung. Zu den Patientenmerkmalen gehören Alter, Geschlecht, Informationen über das Vorliegen eines stumpfen Traumas, mehrere Trauma- bzw. Verletzungsscores, Angaben zur erhaltenen Gesamtflüssigkeit außerhalb des Krankenhauses, Dauer zwischen Polizeianruf und Infusion, Zeit, die außerhalb des Krankenhauses verbracht wurde und die Auskunft, ob ein Lufttransport stattfand und ob die Atemwege erfolgreich sichergestellt wurden. Zu den 6 stetigen Zielgrößen gehören die Anzahl der Tage, die lebend außerhalb der ICU, des Krankenhauses oder ohne künstliche Beatmung verbracht wurden. Weitere Patientenmerkmale sind der schlechteste MODS *Score*, Anzahl der Erythrozytenkonzentrate innerhalb der ersten 24h und die erhaltene Gesamtflüssigkeit innerhalb der ersten 24h. Zu den 12 binären Zielgrößen gehören das Überleben von 28 Tagen, die Entlassung (lebend), ARDS freies Überleben der ersten 28 Tage, eine Hypernatriämie mit Interventionsbedarf, nosokomiale Infektionen, Pneumonie, Sepsis, Harnwegs-, Wundinfekt, intrakranielle Blutung, Anfall innerhalb von 24h und ein GOSE über 3 nach 6 Monaten. (63)

Die im Jahre 2011 veröffentlichte randomisierte ROC-HS-Studie evaluiert den Einfluss hypertoner Kochsalzlösung auf den GOSE und auf u.a. die Mortalität und das Auftreten nosokomialer Infektionen bei Patienten mit hypovolämischem Schock. Aufgrund der Voraussetzung der 1:1-Randomisierung wurde der Therapiearm mit isotoner Kochsalzlösung ausgeschlossen. In dieser Studie liefern 475 Patienten Informationen, darunter 14 Zielgrößen und 16 Patientenmerkmale. Zu den im PS-Modell berücksichtigten Patientenmerkmalen gehören Alter, Geschlecht, Informationen über das Vorliegen eines stumpfen und penetrierenden Traumas, mehrere Trauma- bzw. Verletzungsscores, qualifizierender systolischer Blutdruck, Dauer zwischen Polizeianruf und Infusion, Auskunft, ob ein Lufttransport stattfand und ob die Atemwege erfolgreich sichergestellt wurden. Zu den 6 stetigen Zielgrößen gehören die Anzahl der Tage, die lebend außerhalb der ICU, des Krankenhauses, ohne künstliche Beatmung verbracht wurden. Außerdem der schlechteste MODS *Score*, Anzahl der Erythrozytenkonzentrate innerhalb der ersten 24h und die erhaltene Gesamtflüssigkeit innerhalb der ersten 24h. Zu den 12 binären Zielgrößen gehören das Überleben von 28 Tagen, der Tod vor Ort, zur Zeit des Transportes oder im Krankenhaus oder innerhalb von 6h, die Entlassung (lebend), ARDS freies Überleben der ersten 28 Tage, eine

Hypernatriämie mit Interventionsbedarf, nosokomiale Infektionen, Pneumonie, Sepsis, Harnwegs-, Wundinfekt, intrakranielle Blutung und ein angestiegenes Natrium (>145mEq/l) in den ersten 4h. Bei beiden Studien wurde entweder eine normotone oder hypertone Kochsalzlösung verabreicht. (64)

2.4.24 SHEP

In der im Jahre 1991 veröffentlichten SHEP-Studie wurde der Einfluss einer antihypertensiven Medikation u.a. auf das Auftreten tödlicher oder nicht-tödlicher Schlaganfälle und die kardiovaskuläre Mortalität und Morbidität bei Patienten mit isoliertem systolischen Hypertonus randomisiert evaluiert. Das Kollektiv von 4.743 Patienten liefert Informationen zu 4 Zielgrößen und 15 Patientenmerkmalen. Es wurde entweder ein Placebo oder Chlorthalidon bzw. Atenolol verabreicht. Zu den Patientenmerkmalen gehören Alter, Geschlecht, ethnische Zugehörigkeit, höchster Schulabschluss, Raucher- und Trinkstatus, Herzfrequenz, Blutdruck, BMI, Vorliegen von Strömungsgeräuschen über den Karotiden, vorhandene antihypertensive Behandlung und Vorerkrankungen. Zu den 3 binären Zielgrößen gehören TIA, Linksherzversagen, Hirninfarkt und zu der Ereignis-Zeit-Zielgröße der Tod. (65)

2.4.25 SOLVD

In der SOLVD-Intervention-Studie (1991) wurde der Einfluss von Enalapril u.a. auf die Mortalität und Hospitalisation bei Patienten mit reduzierter LVEF und Herzinsuffizienz randomisiert evaluiert. Hierbei stehen Daten von 2.568 Patienten mit 2 Zielgrößen und 25 Patientenmerkmalen zur Verfügung. Zu den hierbei verwendeten Patientenmerkmalen gehören Alter, Geschlecht, ethnische Zugehörigkeit, Raucherstatus, Körpergewicht, Herzfrequenz und Blutdruck, Ejektionsfraktion, NYHA Klassifikation, Laborparameter wie Natrium, Kalium, Kreatinin und Angaben zu Vorerkrankungen und der Einnahme bestimmter Medikamente. 2 Ereignis-Zeit-Zielgrößen werden evaluiert: Tod und erste Hospitalisation aufgrund einer Herzinsuffizienz. (66)

In der in 1992 veröffentlichten randomisierten SOLVD-Prävention-Studie wurden Patienten mit reduzierter LVEF im Hinblick auf den Einfluss von Enalapril auf die Mortalität, Entwicklung von Herzinsuffizienz und Hospitalisation untersucht. Hierbei liegen 3 Zielgrößen und 25 Patientenmerkmale von 4.225 Patienten zur Analyse bereit. Hierbei stehen dieselben Patientenmerkmale wie in der Interventionsstudie zur Verfügung. Zu den beiden Zielgrößen aus der Interventionsstudie kommt auch noch die Ereignis-Zeit-Zielgröße hinzu, welche die Verschlechterung der Herzinsuffizienz beinhaltet. Es wurde bei beiden Studien entweder ein Placebo oder Enalapril verabreicht. (67)

2.4.26 TIMI-II

Im Jahre 1989 wurde die randomisierte TIMI-II-Studie veröffentlicht, in welcher der Einfluss von einer invasiven Therapie mit PCI u.a. auf die Mortalität, das Auftreten von Reinfarkten und die Ejektionsfraktion bei Patienten mit akutem Myokardinfarkt evaluiert wurde. 3.339 Patienten bilden die Studienpopulation und liefern Informationen zu 8 Zielgrößen und 11 Patientenmerkmalen. Hierbei wurde entweder eine invasive Therapie mit PCI oder eine konservative Therapie, in Ausnahmefällen auch mit PCI, durchgeführt. Zu den im PS-Modell verwendeten Patientenmerkmalen zählen Alter, Geschlecht, ethnische Zugehörigkeit, die Dauer von Symptombeginn bis Studieneintritt und Vorerkrankungen. Die 2 stetigen Zielgrößen beschreiben die Differenz zwischen Belastungs- und Ruheejektionsfraktion und die Ruheejektion jeweils bei Entlassung. Die binäre Zielgröße zeigt an, ob der Belastungstest bei Entlassung positiv war. Zu den 5 Ereignis-Zeit-Zielgrößen gehören die Ereignisse Transfusion, Schlaganfall, Tod, nicht-tödlicher Reinfarkt und die Durchführung einer CABG. (68)

In einer Unterstudie, eigens TIMI-B benannt, wurde bei einer Gruppe von 1.434 Patienten der Einfluss einer sofortigen Betablockergabe auf die Mortalität, das Auftreten von Reinfarkten und die Ejektionsfraktion untersucht. Dabei wurde die Betablockertherapie sofort oder verzögert eingeleitet. Für das PS-*Matching* und die weiteren Analysen liegen 2 Zielgrößen und 11 Patientenmerkmale vor. Dabei stehen dieselben Patientenmerkmale wie beim ersten TIMI-II-Datensatz zur Verfügung. Die beiden Ereignis-Zeit-Zielgrößen sind Tod und nicht-tödlicher Reinfarkt. (68)

2.5 Ethikvotum

Nach Erfüllen der Auflagen und Prüfung durch die Ethikkommission der Medizinischen Fakultät der Heinrich-Heine-Universität Düsseldorf bestehen keine Bedenken gegen die Durchführung der retrospektiven pseudonymisierten Datenanalyse. Ein positives Ethikvotum (5986R, 11.08.2017) liegt vor.

2.6 Statistische Methodik

Die Datensätze werden mithilfe eines „PSinRCT“-Makros bearbeitet und analysiert. Dieses von Prof. Kuß zur Verfügung gestellte Makro basiert auf einer abgeänderten Version des *PSMatching*-Makros von Coca-Perraillon (69). Zusätzlich beinhaltet das „PSinRCT“-Makro noch ein *Caliper*-Makro, das innerhalb des PS-Makros wiederholt wird. In Abhängigkeit von jeweils einer Caliperweite wird dann das PS-Makro aufgerufen und für eine Vielzahl von Caliperweiten wiederholt. Der einzige Parameter des *Caliper*-Makros ist die Caliperweite "*_tenthousandthSTDcaliper*“, die in 1/10.000 vom Vielfachen der Standardabweichung des

linearen Prädiktors gemessen wird, sodass *„_tenthousandthSTDcaliper*“ mit dem Wert 2000 der Standardempfehlung von Austin entspricht (7). Für das PS-*Matching* wird, einer Empfehlung von Austin folgend, ein *Nearest Neighbor Caliper Matching* ohne *Replacement* verwendet (7). Die Intensivierung des PS-*Matchings* wird durch eine schrittweise Reduzierung der Caliperweite in der Menge 1; 0,9; … 0,1; 0,09; … 0,01; 0,009; ... 0,001; 0,0009; ... 0,0001 erreicht**.** (4)

In Abwandlung des PS-Makros werden nicht alle Zielgrößen automatisch ausgewertet, sondern es wird nur eine Zielgröße ausgearbeitet. Deren Skalenniveau muss mit den Makro-Variablen *„_Binary“, „_Ordinal“, „_Continuous“* oder *„_Survival/_SurvivalCensVar/_CensValue“* vorgegeben werden. Binäre, ordinale und Ereignis-Zeit-Zielgrößen werden auf der log-Skala ausgegeben. Die Schätzer stetiger Zielgrößen werden durch die Standardabweichung der beobachteten Werte dividiert (Cohen's d für gleiche Gruppengrößen und unterschiedliche Varianzen), sodass ein Vergleich aller Zielgrößen möglich ist. Die Konfidenzintervalle werden ebenfalls auf der log-Skala ausgegeben. Dies gilt jedoch nur für binäre und ordinale Zielgrößen. Konfidenzintervalle von Ereignis-Zeit-Zielgrößen werden auf der Hazard Skala ausgegeben und müssen dann noch nachberechnet werden (*LowerCL=Estimate-probit(0.975)*StdErr; UpperCL=Estimate+probit(0.975)*StdErr).* Die Konfidenzintervalle stetiger Zielgrößen werden ebenfalls wie oben in Anlehnung an Cohen's d durch die Standardabweichung dividiert. (4)

Konkret wird für binäre Zielgrößen eine stratifizierte logistische Regression mit dem *Matching Stratum* als Schichtvariable mittels PROC LOGISTIC benutzt. Für stetige Zielgrößen wird ein gemischtes Modell mit zufälligem *Intercept* für das *Matching Stratum* mittels PROC MIXED verwendet. PROC GLIMMIX wird bei ordinalen Zielgrößen angewandt, ein gemischtes *Proportional-Odds*-Modell mit zufälligem *Intercept* für das *Matching Stratum*. Zuletzt wird für Ereignis-Zeit-Zielgrößen ein stratifiziertes *Proportional-Hazard*-Modell mittels PROC PHREG benutzt, auch hier mit dem *Matching Stratum* als Schichtvariable. (70)

Nach Eingabe des Datensatznamens, der Variablen (Therapieregime, Patientenmerkmale gruppiert nach Skalenniveau, Patientenidentifikationsnummer) und zuletzt der Zielgröße kann das „PSinRCT“-Makro für jede Zielgröße aufgerufen und alle zuvor festgelegten Caliperweiten für das *Matching* durchlaufen werden. Pro Zielgröße wird eine Datei erstellt, welche als Variablen die Bezeichnung der Zielgröße (*Outcome*), des Outcome-Typs und des Datensatzes, die Caliperweite, die Anzahl der gematchten Beobachtungen und der Kovariablen im PS-Modell, die Summe der quadrierten z-Differenzen VOR dem *Matching*, die Summe der quadrierten z-Differenzen NACH dem *Matching* und den Effektschätzer samt 95%-Konfidenzintervalle enthält. Alle PS-Analysen werden vorab und blind für die Ergebnisse der RCT-Analyse durchgeführt (4).

Pro Datensatz wird die Caliperweite ausgesucht, welche die kleinste Summe der quadrierten z-Differenzen nach *Matching* besitzt, sodass die Effektschätzer zur besten Balanciertheit bewertet werden können. Diese Beobachtungen werden dann zusammengelegt, sodass ein Datensatz mit allen Zielgrößen aller Studiendatensätze zur jeweils optimalen Caliperweite entsteht. (4)

Danach wird die RCT Analyse durchgeführt. Hierzu wird ein RCT-Makro benutzt, welches auf dem „PSinRCT"-Makro basiert, die Zielgrößen auch nach derselben Methodik analysiert. Dieses verzichtet jedoch auf das *Matching* und ist von allen Variablen befreit, die sich auf das *Matching* oder *Matching Stratum* beziehen.

Zum Schluss werden sowohl die Beobachtungen aus der RCT-Analyse, als auch die aus der PS-Analyse zusammengeführt und miteinander verglichen. Dabei wird analysiert, wie viele Schätzer aus den PS-gematchten Analysen sich im 95%-Konfidenzintervall der Schätzer der zugehörigen RCT-Analyse befinden. Weiter wird mit dem McNemar Test einschließlich Edward- und Yates-Korrektur untersucht, inwiefern es eine Veränderung in der Ausgabe von signifikanten Ergebnissen zwischen den beiden Schätzmethoden gibt, bzw. mithilfe des Kappa-Korrelationskoeffizienten wie groß die Übereinstimmung des Auftretens von signifikanten Effekten beider Schätzmethoden ist. Die Effektschätzer von PS-gematchter und RCT-Analyse werden gegeneinander in einer Grafik dargestellt und bei Vermutung eines linearen Zusammenhanges eine lineare Regressionsanalyse durchgeführt. Der Mittelwertvergleich wird mit dem paramatrischen T-Test für verbundene Stichproben und als nicht-parametrischer *Wilcoxon Signed Rank Test* durchgeführt.

Zu jeder RCT wird außerdem eine zufällige Vergleichsprobe genommen, welche genauso viele Beobachtungen hat wie die zugehörige PS-Analyse zur optimalen Caliperweite. Der Behandlungseffekt kann durch die gleichmäßige Verteilung von bekannten und unbekannten Störgrößen wie in dem vollen Datensatz auch in der Vergleichsprobe unverzerrt geschätzt werden. Der Vergleich der drei Methoden wird im Scatterplot und Bald-Altman-Diagramm dargestellt. Weiter werden Intra-Klassen-Koeffizienten (ICC) berechnet (ICC3,1 nach Shrout/Fleiss) (71). Dabei wird eine zusätzlich genormte inverse Varianz-Gewichtung benutzt um die unterschiedlich vorliegende Präzision der Effektschätzung der einzelnen Studien zu beachten. „Lineare gemischte Modelle mit zufälligem *Intercept* für die Zielgrößen und festgelegtem binären Vergleichseffekt“ werden für die Konfidenzintervalle der ICCs benutzt. (4)

2.7 Fehlersuche, -behebung und nachträglicher Ausschluss von DPPPL

Bei mehreren Ereignis-Zeit Zielgrößen fehlten häufig die zu den zensierten Beobachtungen zugehörigen Zeitangaben. Diese fehlenden Werte konnten dann wieder mithilfe anderer passender Variablen aufgefüllt werden. Beispielsweise konnte dafür eine Variable benutzt werden, die den letzten *follow-up* Termin oder den letzten Tag, an dem der Patient lebend gesehen wurde, anzeigte. Andernfalls konnte bei einer komplett vollständigen Ereignis-Zeit Zielgröße mit dem Tod als Ereignis die Zeitangabe zu zensierten Beobachtungen für andere Ereignis-Zeit Zielgrößen übernommen werden. Dies gelang in fast allen Fällen. Einzig bei den Studiendatensätzen von MTOPS konnte keine adäquate Zeitvariable zum Auffüllen gefunden werden, sodass einige Zielgrößen nur als binäre Zielgrößen in die weiteren Berechnungen mit einbezogen werden können.

Bei der Berechnung der Parameterschätzer fielen mehrere Warnungen und Fehler auf. Im Folgenden werden diese selbst und der Umgang mit ihnen beschrieben.

Die Warnung *„There is possibly a quasi-complete separation of data points. The maximum likelihood estimate may not exist“* kam im Schritt *„Compute PS-Model*“ z.B. bei Ereignis-Zeit Zielgrößen der Studie FAVORIT vor. Da in diesem Schritt jedoch keine Paramterschätzer gebraucht werden, sondern nur die geschätzten Ereigniswahrscheinlichkeiten, die dann in den PS umgerechnet werden, konnte diese Warnung im weiteren Verlauf vernachlässigt werden.

Bei der Berechnung der z-Differenzen für binäre Variablen kam beispielsweise bei stetigen Zielgrößen des Datensatzes DPPPM die Warnung *„Output 'RiskDiffCol1' was not created*“ vor. Die Meldung vor der Warnung zeigte entweder *„RISKDIFF statistics are computed only for 2x2 tables*“ oder andererseits *„No statistics are computed for X * Y because X has fewer than 2 nonmissing levels*“ an. In letzterem Fall bestand das Problem entweder darin, dass eine binäre Variable zwar zwei Variablenausprägungen haben kann, aber nur ein Wert angenommen wurde. Oder aber beide Ausprägungen kamen vor, ab einer gewissen Caliperweite war jedoch die Fallzahl so gering, dass eine der beiden Ausprägungen dann nicht mehr angenommen wurde. Beide Fälle kamen einmal vor. Im ersten Fall wurde die binäre Variable ausgeschlossen, im letzteren Fall taucht die Warnung erst bei späteren Caliperweiten auf, sodass schon weit vorher die optimale Caliperweite erreicht wurde und somit das Problem vernachlässigt werden konnte. Im Fall der *„2x2 tables“* konnte mithilfe von PROC FREQ gezeigt werden, dass Variablen als binär eingetragen wurden, welche jedoch eine dritte Ausprägung haben. Diese dritte Ausprägung gab aber in jedem Fall das Ereignis „unbekannt“ an, sodass es durch einen Leerwert ersetzt werden konnte. Da statt zwei Ausprägungen drei angeboten

wurden, konnte keine z-Differenz berechnet werden, da es für nominale Variablen keine z-Differenzen (27) gibt. Nach Umtausch durch einen Leerwert, konnte auch diese Warnung behoben werden.

Ebenfalls gab es die Warnung *„Output 'ParameterEstimates' was not created“* mit der vorausgehenden Meldung *„Did not converge“* bei der ordinalen Zielgröße OCCODE der Studie ISTASP. Hier reichte es die Option „MAXOPT=500“ bei PROC GLIMMIX zu ergänzen, wodurch der Algorithmus konvergierte und die Warnung verschwand.

Bei mehreren Ereignis-Zeit (z.B. funfmi3m bei FAVORIT) und binären Zielgrößen kam es zur Warnung *„Convergence was not attained in 25 iterations"*. Diese beiden Zielgrößen wurden entweder mit PROC PHREG (Ereignis-Zeit) oder mit PROC LOGISTIC (binär) geschätzt. Im *Model*-Befehl gibt es eine Option, welche „MAXITER=“ heißt. Sie legt die Anzahl der Iterationen im finalen Schätzalgorithmus fest und ist in der Grundeinstellung auf den Wert 25 festgelegt. Diese Einstellung ist in manchen Fällen zu klein, sodass der Algorithmus nicht konvergiert und die beschriebene Warnung anzeigt. Mit der Einstellung „MAXITER=1000“ konnte diese Warnung in allen Fällen beseitigt werden.

Mit der Reduktion der Fallzahl durch sinkende Caliperweiten können die Modelle ab einer bestimmten Caliperweite nicht mehr konvergieren oder eine Separation tritt auf. In einigen Fällen (z.B. EVENT bei der Studie MTOPSPD) kam dann der Fehler *„All explanatory variables are dependent on the strata“* vor. Dass die Modelle irgendwann nicht mehr konvergieren, war zu erwarten. Dies trat aber erst dann auf, als die optimale Caliperweite schon erreicht wurde, sodass sich diese Problematik auf die Analyse der Ergebnisse nicht auswirkte. Bei manchen Zielgrößen konnte es auch so sein, dass so wenige Informationen vorhanden sind, dass zwar das Modell konvergiert, die Konfidenzintervalle am Ende dann aber so groß sind, dass aus diesen Zielgrößen dann keine Schlüsse gezogen werden können (s.u.).

Die Anzahl der berechneten z-Differenzen gleicht ungefähr der Summe der Patientenmerkmale, wobei es für nominale Patientenmerkmale keine z-Differenzen gibt und diese somit in der Summe nicht berücksichtigt werden. Nach ersten Berechnungen gab es keine Übereinstimmung der beiden Werte. Falls ein Skalenniveau leer war, wurde nicht der Wert 0 angezeigt, sondern der von einem vorhandenen Skalenniveau. Der Code musste dahingehend verändert werden, dass jedes Skalenniveau getrennt abgefragt wird, ob überhaupt ein Patientenmerkmal vorhanden ist, und dann je nachdem, entweder eine leere Datei erstellt wird oder die korrekte Berechnung der z-Differenzen erfolgt. Trotzdem wurden nicht die korrekten Werte angezeigt. Dies lag an dem Schritt, in welchem die z-Differenzen zusammengelegt wurden. In einer Bedingung wurde nämlich Folgendes verlangt: *„odsoutput RiskDiffCol1=_RiskDiffOut (where=(Row="**Difference**") keep=Risk ASE Row);“* Da mit einer

deutschen Version von SAS gearbeitet wurde, konnten somit keine Beobachtungen gefunden werden, die „*Difference“* enthielten, es gab nur welche mit „Differenz". So wurde eine leere Datei generiert, was folglich zu falschen Werten führte. Durch den Wechsel in die englische Version von SAS, konnte auch dieser Fehler behoben werden.

Nach Zusammenstellen der finalen Ergebnisdatei fiel Zweierlei auf: Es gab Zielgrößen, die bezüglich der Konfidenzintervalle starke Ausreißer waren und es gab Studien, bei denen sich die Anzahl der berechneten z-Differenzen stark von der Summe der quadrierten z-Differenzen vor *Matching* unterschieden. Bei RCTs sollten diese beiden letzteren Werte nämlich annähernd gleich sein.

Das Problem der weiten Konfidenzintervalle kam bei drei Zielgrößen aus drei verschiedenen Studien vor: DCCT mit der Zielgröße T_DTH, HALT-C mit der Zielgröße BPERFU, und MTOPSPC mit der Zielgröße NUTDEC. Bei den ersten beiden Zielgrößen kam dann auch folgende Warnung vor: *The likelihood ratio test for strata homogeneity is questionable since some strata have noevents*. Bei der Zielgröße NUTDEC in MTOPSPC kam es zu einer Fehlermeldung: *All explanatory variables are dependent on the strata.* Das Problem an diesen Zielgrößen war, dass sie nur in einem Therapiearm ein Ereignis hatten, der andere Therapiearm hingegen keins. Diese Zielgrößen wurden dann aus der weiteren Analyse ausgeschlossen, da sie aufgrund der wenigen Ereignisse keinen Informationsgewinn darstellten und durch die ausreißenden Konfidenzintervalle die Statistiken stark verzerrten.

Das zweite Problem wurde so angegangen, dass die 95%-Konfidenzintervalle für die Summe der quadrierten z-Differenzen vor *Matching* berechnet wurden und geprüft wurde, welche Anzahl der berechneten z-Differenzen (nZDiff) von diesen Konfidenzintervallen abweichen. Befinden sich die Anzahlen der berechneten z-Differenzen in den jeweiligen 95%-Konfidenzintervallen, heißt dies, dass diese Abweichungen mit 95% Wahrscheinlichkeit zufällig entstanden sind. Bei 4 Studien gab es eine Abweichung: DPPPL mit „nZDiff“ 13 und dem 95%-Konfidenzintervall [31,6817; 156,459], ROCTBI mit „nZDiff“ 14 und dem 95%-Konfidenzintervall [14,4581; 67,090], AMIS mit „nZDiff“ 34 und dem 95%-Konfidenzintervall [39,0000; 102,325], ACCORDLIP mit „nZDiff“ 32 und dem 95%-Konfidenzintervall [10,63; 28,76]. Die Abweichungen vom Konfidenzintervall sind hierbei nur im Falle von DPPPL sehr deutlich. Bei den anderen, ist die Anzahl der berechneten z-Differenzen nicht so weit von den Grenzen der Konfidenzintervalle entfernt und somit auch noch gut mit dem Zufall zu erklären. Somit wurde der gesamte Datensatz DPPPL von den weiteren Analysen ausgeschlossen. (4)

3 Ergebnisse

3.1 Deskriptive Statistik

Insgesamt wird ein Datensatz mit 213 Zielgrößen zu jeweils 106 Caliperweiten aus insgesamt 37 Studiendatensätzen analysiert. Dabei stehen 128 Ereignis-Zeit, 38 stetige, 43 binäre und 4 ordinale Zielgrößen zur Verfügung. Die Anzahl der Beobachtungen pro Studie reicht von 475 bis 19.435, im Mittel sind es 5.233 Beobachtungen (Median: 3.806). Nach *Matching* liegen im Mittel noch 89,6% (Median: 93,8%) der Beobachtungen vor, im Durchschnitt also 4.925 Beobachtungen (Median: 3.566). Für das PS-Modell werden im Median 17 z-Differenzen pro Studie berechnet. Dabei reicht die Anzahl der berechneten z-Differenzen von 6 bis 34. (4) Die Summe der quadrierten z-Differenzen fällt nach *Matching* auf durchschnittlich 13,2% des Wertes vor *Matching* (Median: 8,9%). Die Summe der quadrierten z-Differenzen vor *Matching* weicht von der Anzahl der berechneten z-Differenzen ab. Hierbei kommen relative Werte von 42% bis 175% im Verhältnis zu der Anzahl der berechneten z-Differenzen vor, im Durchschnitt 87%. Im Mittel beträgt der Schätzer der PS-gematchten Analyse -0,093, der der RCT-Analyse -0,094 und sie unterscheiden sich im Mittel um 0,001. Weitere Werte der MEANS Prozedur sind in Tabelle 1 (S.32) und Tabelle 2 (S.33) veranschaulicht.

Tabelle 1: Deskriptive Statistik der Hauptcharakteristika aller Studien.

Variable	Mean	Median	Lower Quartile	Upper Quartile	Maximum
Allobs	5232,973	3806	1646	6213	19435
Nmatchedobs	4925,351	3566	1416	5984	19174
ProportionReducedObs	0,896	0,938	0,892	0,953	0,987
nZDiffSquared	16,703	17	10	22	34
ProportionSumZDiffSquaredNumber	0,866	0,805	0,601	1,054	1,753
ProportionReducedSumZDiffSquared	0,132	0,089	0,063	0,181	0,405

Allobs: Anzahl der Beobachtungen; Nmatchedobs: Anzahl der Beobachtungen nach *Matching*; ProportionReducedObs: Anteil der gematchten Beobachtungen an der Gesamtzahl der Beobachtungen ohne *Matching*; nZDiffSquared: Anzahl der quadrierten z-Differenzen; ProportionSumZDiffSquaredNumber: Die Summe der quadrierten z-Differenzen vor *Matching* in Relation zu der Anzahl der berechneten z-Differenzen; ProportionReducedSumZDiffSquared: Summe der quadrierten z-Differenzen nach *Matching* in Relation zu der Summe der quadrierten z-Differenzen vor *Matching*

Tabelle 2: Deskriptive Statistik zu allen Schätzern nach RCT- und PS-gematchter Analyse

Variable	Mean	Median	Minimum	Lower Quartile	Upper Quartile	Maximum
Estimate	-0,093	-0,027	-1,792	-0,248	0,079	1,338
LowerCL	-0,494	-0,311	-3,909	-0,641	-0,161	0,857
UpperCL	0,308	0,2	-0,745	0,064	0,462	2,772
EstimateRCT	-0,094	-0,042	-1,815	-0,213	0,051	1,188
LowerCLRCT	-0,435	-0,274	-3,934	-0,545	-0,15	0,853
UpperCLRCT	0,246	0,164	-0,823	0,024	0,353	3,123
absDif	0,001	0,01	-0,759	-0,034	0,055	0,882

Estimate: Schätzer nach PS-Analyse; LowerCL: unteres Konfidenzintervall des PS-Schätzers; UpperCL: oberes Konfidenzintervall des PS-Schätzers; EstimateRCT: Schätzer nach RCT-Analyse; LowerCLRCT: unteres Konfidenzintervall des RCT-Schätzers; UpperCLRCT: oberes Konfidenzintervall des RCT-Schätzers; absDif: absolute Differenz zwischen PS-Schätzer und RCT-Schätzer

Der relative Anteil der PS-gematchten Beobachtungen an der Gesamtzahl der Beobachtungen jeder Studie ist in Abbildung 1 (S.33) abhängig von der Caliperweite dargestellt. Mit sinkender Caliperweite von 10.000 bis 1 (in 1/10.000) nimmt auch die Anzahl der gematchten Beobachtungen ab. Dabei bleibt die Relation bis zur Caliperweite von 300 fast gleich und fällt dann bis zur Caliperweite von 10 sehr stark ab.

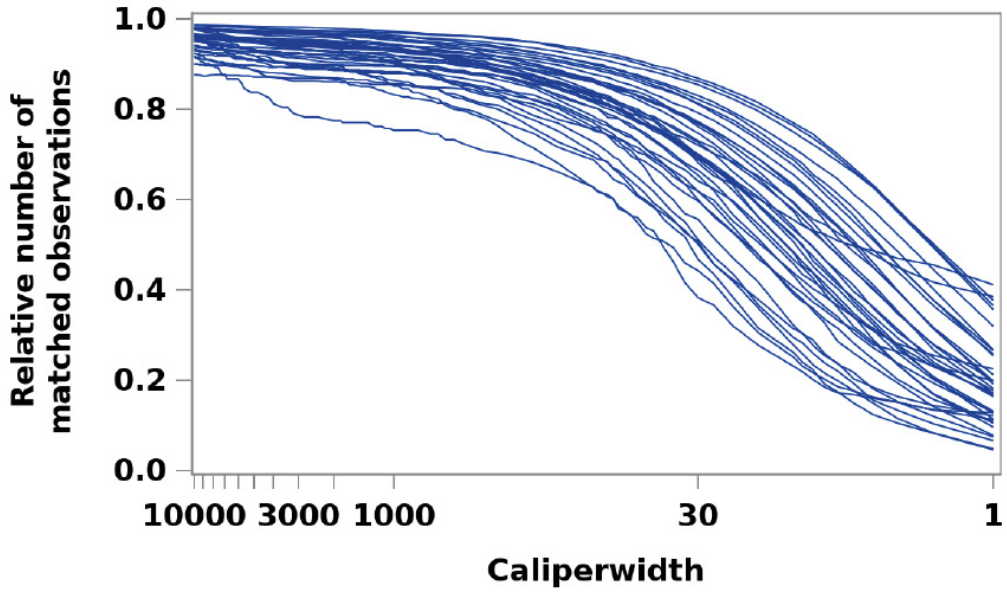

Abb.1: Reduktion der gematchten Beobachtungen. Verlauf des Anteils der Anzahl der gematchten Beobachtungen an der Gesamtbeobachtungszahl im Verlauf abhängig von der jeweiligen Caliperweite (in 1/10.000). Jede Linie veranschaulicht dabei jeweils eine Studie.

Tabelle 3: Anzahl der Beobachtungen. Pro Studie ist angegeben, wie viele Beobachtungen vor und nach *Matching* vorhanden sind und auf welchen Anteil diese Anzahl gesunken ist.

Daten	Allobs	Nmatchedobs	ProportionReducedObs
ACCORDBP	4733	4238	0,89542
ACCORDGLY	10251	9730	0,94918
ACCORDLIP	5518	5272	0,95542
AFFIRM	4060	3834	0,94433
ALLHATAL	18102	17522	0,96796
ALLHATLLT	10355	10046	0,97016
AMIS	4524	4060	0,89744
ATN	1124	1054	0,93772
BEST	2707	2456	0,90728
BHAT	3837	3566	0,92937
CAST	1498	1416	0,94526
CPPT	3806	3660	0,96164
DCCT	1441	340	0,23595
DIG	7788	7414	0,95198
DPPPM	2058	1835	0,89164
ENRICHD	2481	2302	0,92785
FAVORIT	4110	3856	0,9382
HALTC	1050	912	0,86857
HEMOFLUX	1846	1404	0,76056
HEMOKTV	1846	1752	0,94908
ISTASP	19435	19174	0,98657
ISTHEP	19435	19150	0,98534
ISTHEPDOS	9717	9394	0,96676
MAGIC	6213	5984	0,96314
MRFIT	12866	12410	0,96456
MTOPSPC	1634	1403	0,85863
MTOPSPD	1599	976	0,61038
MTOPSPF	1646	1432	0,86999
OAT	2166	1930	0,89104
PEACE	8290	7902	0,9532
ROCHS	475	398	0,83789
ROCTBI	700	636	0,90857
SHEP	4743	3912	0,82479
SOLVDINT	2568	2400	0,93458
SOLVDPRE	4225	4020	0,95148
TIMI	3339	3132	0,93801
TIMIB	1434	1316	0,91771

Allobs: Anzahl der Beobachtungen; Nmatchedobs: Anzahl der Beobachtungen nach *Matching*; ProportionReducedObs: Anteil der gematchten Beobachtungen an der Gesamtzahl der Beobachtungen ohne *Matching*

In Tabelle 3 (S.32) ist eine Übersicht gegeben, wie viele Beobachtungen pro Studie für die RCT- und PS-*Matching*-Analyse zur Verfügung stehen. Außerdem ist jeweils bei optimaler Caliperweite der relative Anteil der gematchten Beobachtungen an der Gesamtzahl der Beobachtungen aufgezeigt. Dabei fällt auf, dass in ca. 68% der benutzten Studien noch über 90% der Gesamtzahl der Beobachtungen nach *Matching* vorhanden ist, in ca. 92% sind über 80% der Gesamtbeobachtungen vorhanden. Im Median können bei optimaler Caliperweite 93,8% der Ausgangsbeobachtungen für das *Matching* benutzt werden. In Abbildung 4 (S.37) sind die absoluten Anzahlen der für die Analyse benutzten Beobachtungen vor und nach *Matching* grafisch veranschaulicht.

In Abbildung 2 (S.35) ist der relative Anteil der Summe der quadrierten z-Differenzen nach *Matching* an der Summe der quadrierten z-Differenzen vor *Matching* dargestellt. Hierbei fällt auf, dass die relativen Anteile wieder etwa ab der Caliperweite 100 stark ansteigen. Davor zeigt sich nach dem initialen Absinken ein monotoner, undulierend leicht steigender und fallender Verlauf.

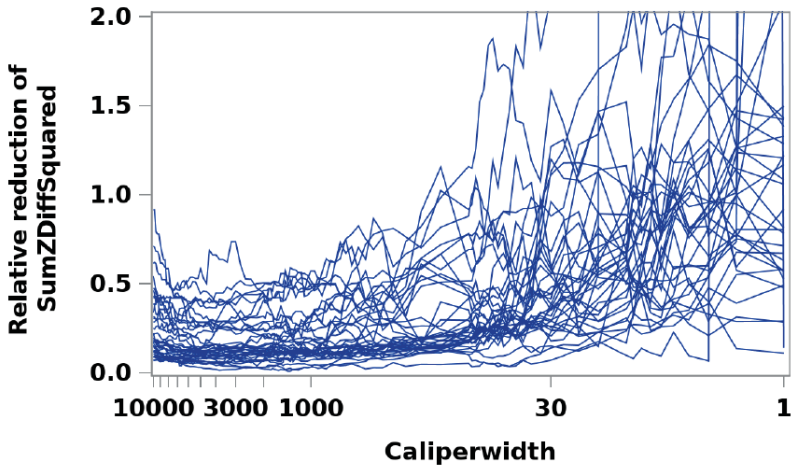

Abb.2: Quadrierte z-Differenzenreduktion. Quotient der Summen der quadrierten z-Differenzen nach *Matching* durch die Summen der quadrierten z-Differenzen vor *Matching* (SumZDiffSquared) für jede Studie mit sinkender Caliperweite

In Tabelle 4 (S.36) ist für jede Studie separat die optimale Caliperweite aufgelistet, die jeweilige Anzahl der berechneten z-Differenzen und die dazugehörigen Summen der quadrierten z-Differenzen vor und nach *Matching*. Dabei wird bei über 50% der Studien eine Reduktion auf unter 8,9% der Ausgangssumme erzielt, bei 90% der Beobachtungen ist die Reduktion auf unter 30% der Ausgangssumme und bei 5% der Beobachtungen wird die Summe der quadrierten z-Differenzen sogar auf unter 1,5% reduziert. Im Mittel erfolgt eine Reduktion von 15,7 auf 2,2 *(4)*. Grafisch festgehalten ist dies in Abbildung 3 (S.37), wo die Summen der berechneten z-Differenzen vor dem *Matching* und zur optimalen Caliperweite aufgetragen sind.

Tabelle 4: Optimale Caliperweiten. Zu jeder Studie ist die optimale Caliperweite aufgeführt mit den zugehörigen Summen der quadrierten z-Differenzen vor und nach *Matching*. Die optimale Caliperweite reicht dabei von 6 zu 9000, im Mittel (Median) 3509 (3600). (4)

Daten	nZDiffSquared	sumZDiffSquaredBefore	sumZDiffSquaredAfter	Caliperwidth
ACCORDBP	22	23,1812	3,03445	400
ACCORDGLY	29	26,3174	2,56542	3600
ACCORDLIP	32	16,4378	3,56515	3000
AFFIRM	7	5,7022	0,26339	3600
ALLHATAL	17	20,3259	1,04951	2800
ALLHATLLT	12	10,0601	0,8172	4800
AMIS	34	59,608	7,12357	1400
ATN	19	11,4205	2,57162	7800
BEST	21	24,4375	1,86015	1150
BHAT	25	19,6459	1,73682	850
CAST	18	15,5879	1,39244	8400
CPPT	6	2,5209	0,09531	2200
DCCT	17	18,1814	6,40793	6
DIG	18	11,4836	1,40804	6600
DPPPM	13	15,3463	1,23469	5800
ENRICHD	25	25,6264	9,32671	4600
FAVORIT	16	11,5	2,08167	1500
HALTC	17	27,956	2,18235	3600
HEMOFLUX	20	10,5975	2,62135	50
HEMOKTV	20	12,3523	3,64164	6600
ISTASP	10	5,0579	0,33383	8000
ISTHEP	10	7,561	0,61032	9000
ISTHEPDOS	10	12,8283	0,18416	3800
MAGIC	12	6,7609	0,07126	1400
MRFIT	18	10,9784	2,51931	1050
MTOPSPC	7	3,7522	1,51919	450
MTOPSPD	7	3,873	0,56922	60
MTOPSPF	7	4,9574	0,26141	3600
OAT	25	20,4431	2,71513	350
PEACE	25	34,4142	2,54634	7600
ROCHS	16	12,8876	1,62198	5200
ROCTBI	13	6,1321	1,75257	5200
SHEP	10	7,2716	0,25762	55
SOLVDINT	24	23,4821	2,4574	8600
SOLVDPRE	24	20,5342	1,0255	4200
TIMI	6	10,197	0,15688	1200
TIMIB	6	2,8293	0,17867	1300

nZDiffSquared: Anzahl der berechneten z-Differenzen; sumZDiffSquaredBefore: Summe der quadrierten z-Differenzen vor *Matching*; sumZDiffSquaredAfter: Summe der quadrierten z-Differenzen nach *Matching*; Caliperwidth: Caliperweite

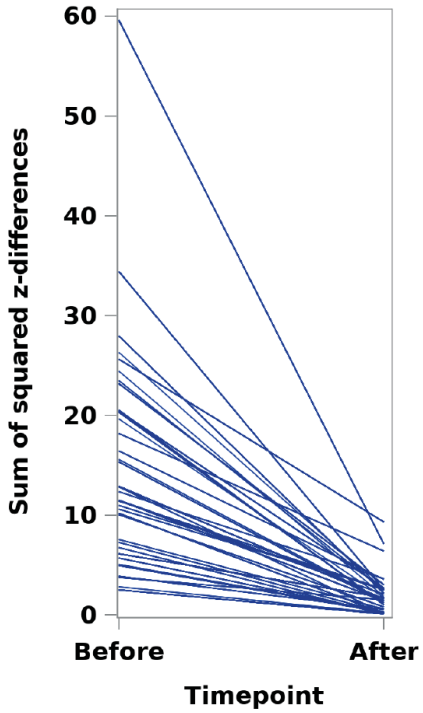

Abb.3: Reduktion der Summe der quadrierten z-Differenzen nach *Matching* bei optimaler Caliperweite

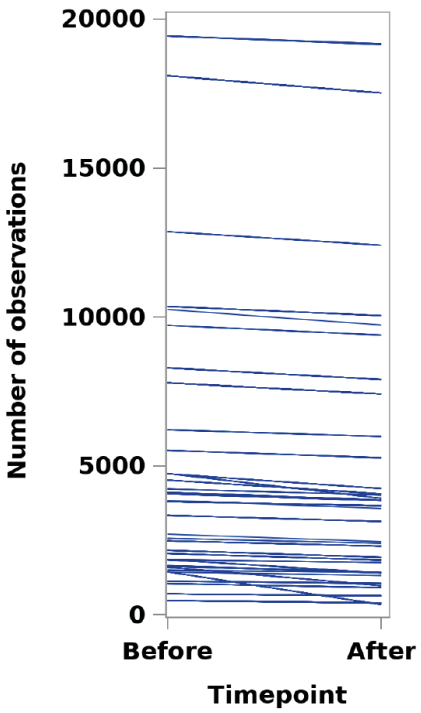

Abb.4: Darstellung der Anzahl an benutzten Beobachtungen vor und nach *Matching*

3.2 Vergleich zwischen PS-gematchter Analyse und RCT-Analyse

Abbildungen 6-9 (S.39-40) zeigen, wie sich die Effektschätzer verschiedener Skalenniveaus vor und nach *Matching* verändern. Dabei fällt auf, dass die Veränderung der Effektschätzer bis auf einige Ausnahmen nur eine geringe Steigung besitzt, bzw. in einigen Fällen sogar parallel zur x-Achse verläuft, was für ein Gleichbleiben des Effektschätzers spricht.

In Abbildung 5 (S.38) wird der Schätzer der PS-gematchten Analyse abhängig vom Schätzer der RCT-Analyse dargestellt. Auffällig ist hierbei, dass sich die jeweiligen Punkte augenscheinlich in der Mitte der durch die 95%-Konfidenzintervalle entstandenen Kreuze befinden. Dies spricht dafür, dass sich beide Schätzer etwa in der Mitte beider Konfidenzintervalle befinden. Die rot eingetragene Linie ist die Winkelhalbierende (y=x) und stellt eine perfekte Übereinstimmung beider Schätzmethoden dar. Es fällt auf, dass die Werte sich stark an dieser Funktion orientieren. Mittel-, Medianwerte etc. zu den Konfidenzintervallen können aus Tabelle 2 (S.33) entnommen werden.

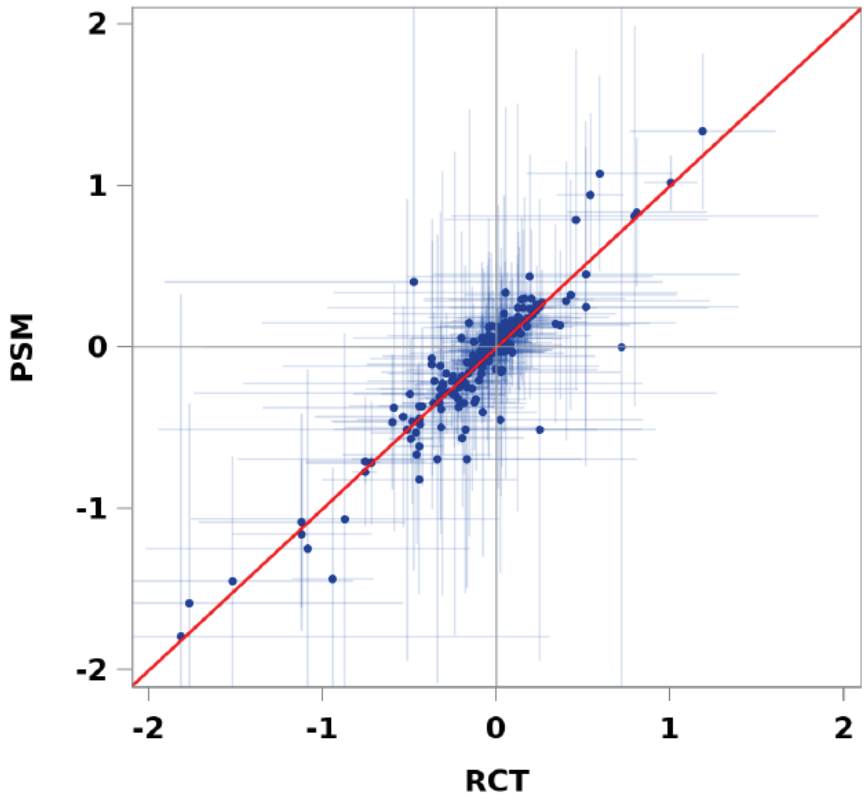

Abb.5: Schätzer von RCT und PS-Analyse. Die Schätzer (inkl. 95%-KI) beider Schätzmethoden sind gegeneinander aufgetragen. Die Werte sind insgesamt sehr nah an der Winkelhalbierenden (rot) angeordnet, auf welcher eine perfekte Übereinstimmung liegen würde (PSM: PS-*Matching*-Schätzer; RCT: RCT-Schätzer)

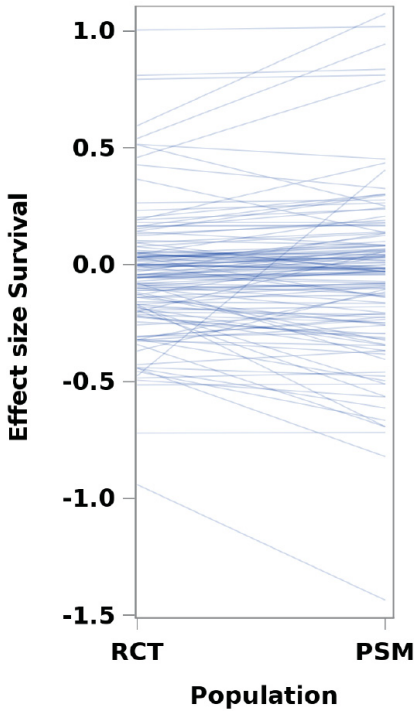

Abb.6: Ereignis-Zeit Zielgrößen. Darstellung der Größe des Effektschätzers von Ereignis-Zeit Zielgrößen in der RCT-Analyse verglichen mit der PS-gematchten Analyse (PSM: PS-*Matching*-Schätzer; RCT: RCT-Schätzer)

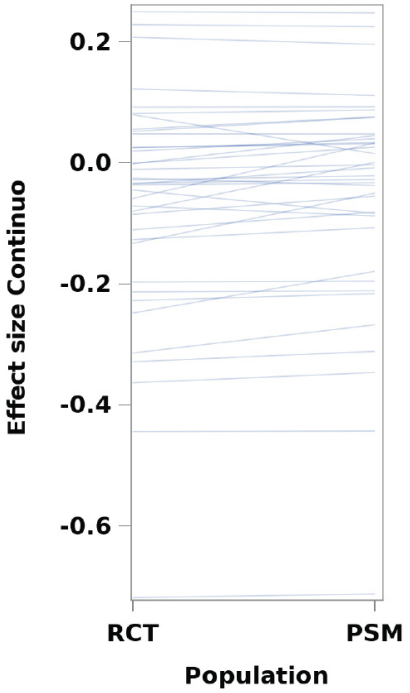

Abb.7: Stetige Zielgrößen. Darstellung der Größe des Effektschätzers von stetigen Zielgrößen in der RCT-Analyse verglichen mit der PS-gematchten Analyse (PSM: PS-*Matching*-Schätzer; RCT: RCT-Schätzer)

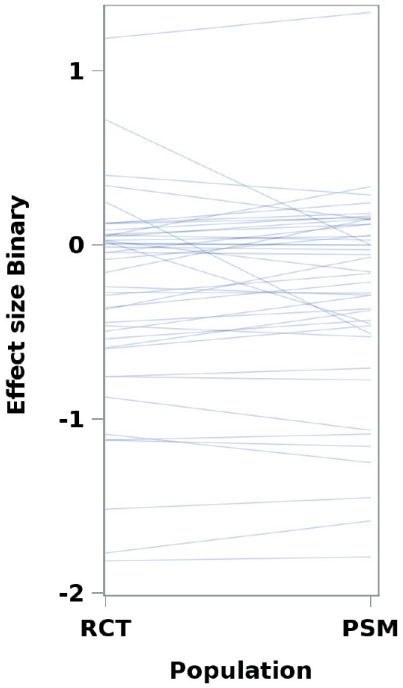

Abb.8: Binäre Zielgrößen. Darstellung der Größe des Effektschätzers von binären Zielgrößen in der RCT-Analyse verglichen mit der PS-gematchten Analyse (PSM: PS-*Matching*-Schätzer; RCT: RCT-Schätzer)

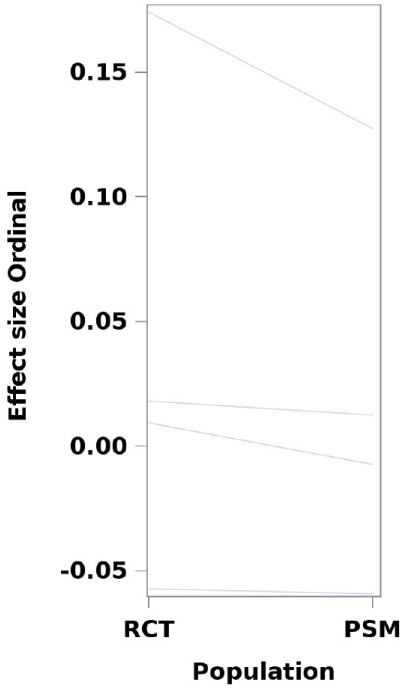

Abb.9: Ordinale Zielgrößen. Darstellung der Größe des Effektschätzers von ordinalen Zielgrößen in der RCT-Analyse verglichen mit der PS-gematchten Analyse (PSM: PS-*Matching*-Schätzer; RCT: RCT-Schätzer)

Werden die PS-gematchten Schätzer mit den Konfidenzintervallen der RCT-Schätzer verglichen, fällt auf, dass 96,71% (206/213) der PS-gematchten Effektschätzer in den zugehörigen 95%-Konfidenzintervallen der RCT-Schätzer liegen. Das spricht perfekt für eine rein zufällige Abweichung, weil unter der Hypothese, dass PS- und RCT-Schätzer gleich sind, genau in 5% der Fälle eine Abweichung erwartet werden kann.

Auch die Effekte, die für eine Schätzmethode (PS-gematcht vs. RCT) signifikant sind, für die andere nicht, sind selten. Tabelle 5 (S.41) zeigt dazu die genaue Verteilung bei RCT- und PS-gematchter Analyse. Bei einem Signifikanzniveau von α=0,05 ergibt sich für den McNemar-Test ein kritischer Wert von $\chi^2_{1;0,95}$=3,841. Der McNemar-Test ergibt eine Teststatistik von 1,0, nach Yates Korrektur ergibt sich eine Prüfgröße von 0,766, nach Edward Korrektur eine Prüfgröße von 0,563. Da alle Prüfgrößen kleiner als der kritische Wert sind, kann die Nullhypothese, welche die Ergebnisse beider Schätzmethoden als gleich betrachtet, nicht zugunsten der Alternativhypothese abgelehnt werden. Somit gibt es mit p=0,3173 keine signifikante Veränderung der Schätzung eines signifikanten Effektes nach PS-gematchter Analyse im Vergleich zur RCT-Analyse.

Tabelle 5: Verteilung der signifikanten Effekte nach RCT- und PS-gematchter Analyse

	PSMEstimate_Significant		
RCTEstimate_Significant	0	1	**Total**
0	147	6	153
	69,01	2,82	71,83
	96,08	3,92	
	93,63	10,71	
1	10	50	60
	4,69	23,47	28,17
	16,67	83,33	
	6,3	89,29	
Total	157	56	213
	73,71	26,29	100

RCTEstimate_Significant: Schätzung eines signifikanten Effektes in der RCT-Analyse; PSMEstimate_Significant: Schätzung eines signifikanten Effektes in der PS-Analyse

Es besteht eine sehr gute Übereinstimmung beider Schätzmethoden; κ=0,81 [0,7217; 0,8994]. Die Hypothese von keinerlei Übereinstimmung kann abgelehnt werden: z=11,8427; p<0,0001. Somit gibt es eine signifikante Übereinstimmung beider Schätzmethoden. Durch das Konfidenzintervall wird dies unterstützt, da hervorgeht, dass der wahre Kappa Koeffizient größer als 0 ist. In Abbildung 10 (S.42) ist die Übereinstimmung grafisch festgehalten.

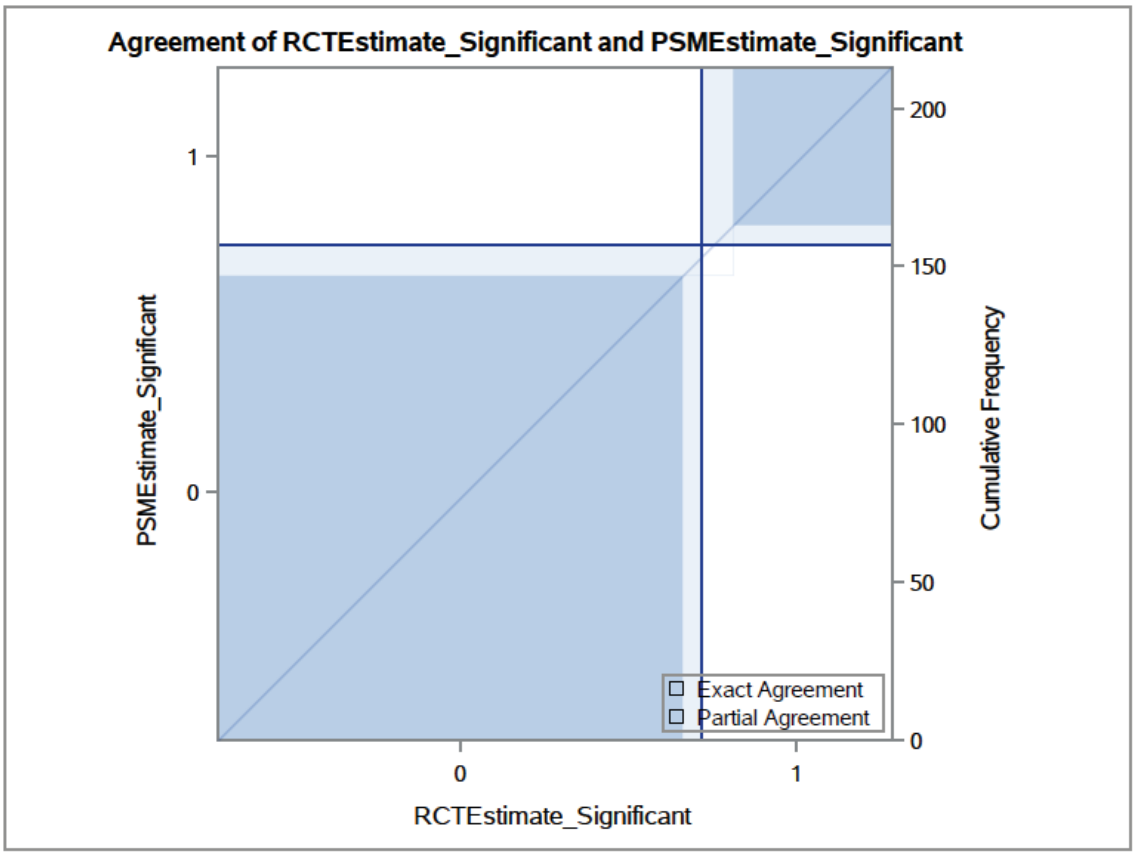

Abb.10: Übereinstimmung der Schätzmethoden. Übereinstimmung der signifikanten Effekte nach PS-gamatchter und RCT-Analyse. Dabei reichen die Grenzen der Übereinstimmung fast bis zur perfekten Übereinstimmung (blaues Quadrat reicht fast bis an die dicken blauen Linien) (RCTEstimate_Significant: Schätzung eines signifikanten Effektes in der RCT-Analyse; PSMEstimate_Significant: Schätzung eines signifikanten Effektes in der PS-Analyse)

Die Grafiken in Abbildung 12 (S.43) zeigen, dass sich die Residualwerte der Beobachtungen vor allem in dem Bereich um den Koordinatenursprung befinden (Heteroskedastizität) und dass sie nicht normalverteilt sind. Bei der Regressionsanalyse werden 213 Beobachtungen untersucht, dabei stellt der Effektschätzer aus der RCT-Analyse die unabhängige Variable, der aus der PS-gematchten Analyse hingegen die abhängige Variable dar. Die Regressionsanalyse zeigt einen starken Zusammenhang zwischen den Effektschätzern beider Methoden; $p<0{,}0001$, $R^2=0{,}8333$. Dabei unterscheidet sich der y-Achsenabschnitt der berechneten Regressionsgleichung (Tabelle 6, S.43) nicht signifikant von 0; $t=0{,}04$, $p=0{,}9691$. Die Steigung hingegen unterscheidet sich signifikant von 0; $t=32{,}48$, $p<0{,}0001$. Die zugehörige Gleichung der Regressionsgeraden in Abbildung 11 (S.43) lautet: $y=0{,}99238x+0{,}00044610$. Dabei fällt auf, dass diese fast perfekt der in Abbildung 5 (S.38) eingezeichneten Winkelhalbierenden ($y=\mathbf{1}x+\mathbf{0}$) gleicht.

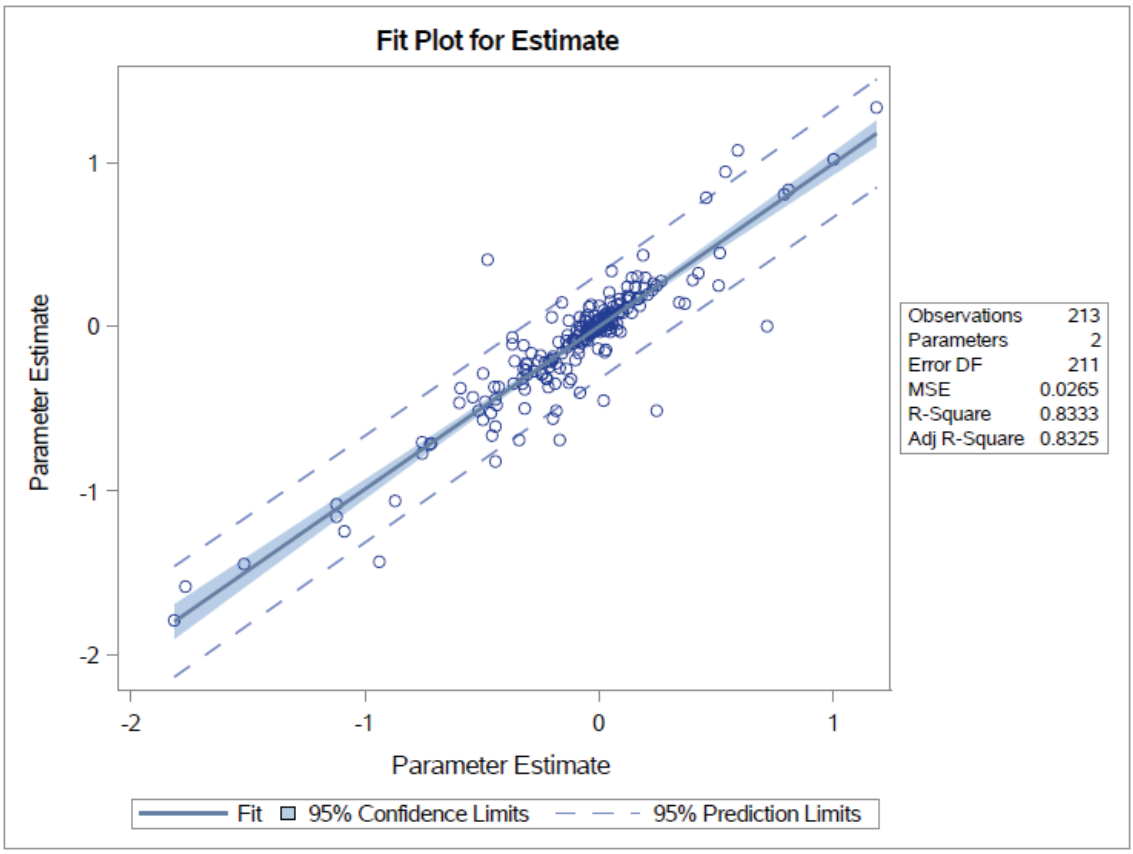

Abb.11: Regressionsgerade durch die Punktwolke der Effektschätzer beider Schätzmethoden (DF: *degree of freedom*; MSE: *mean square error*; Adj: *adjusted*)

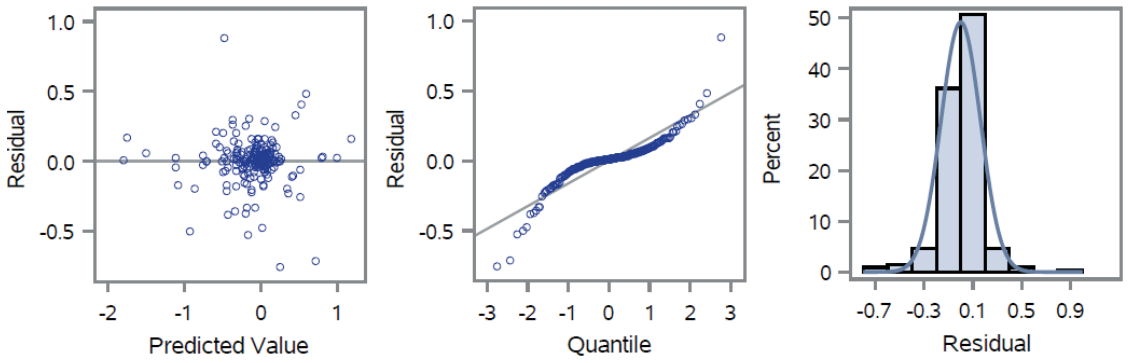

Abb.12: Grafiken zur Beurteilung von Homoskedastizität und der Normalverteilungsannahme der Residualwerte.

Tabelle 6: Teststatistik, p-Werte und Standardfehler des Regressionsmodells

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0,0004461	0,01151	0,04	0,9691
EstimateRCT	1	0,99238	0,03055	32,48	<0,0001

DF: *degree of freedom*; *t Value*: Testwert

Werden die Verteilungen der Effektschätzer beider Methoden betrachtet, fällt auf, dass die Normalverteilungsannahme nach dem Shapiro-Wilks, Kolmogorov-Smirnov, Cramer-von Mises und Anderson-Darling Test bei beiden Schätzmethoden verletzt wird: W_{RCT}=0,869 (W_{PS}=0,886); p<0,0001; D_{RCT}=0,131 (D_{PS}=0,134); p<0,01, W-Sq_{RCT}=1,533 (W-Sq_{PS}=1,393) und A-Sq_{RCT}=8,298 (A-Sq_{PS}=7,772) jeweils p<0,005. Eine Korrelationsanalyse zwischen den beiden Effektschätzern zeigt einen signifkanten Zusammenhang; r_s=0,86675, r_P=0,91287, p<0,0001. Ein Mittelwertvergleich über den *Wilcoxon Signed Rank Test* ergibt zwischen den Schätzern nach RCT-Analyse (M=-0,094) und den Schätzern nach PS-gematchter Analyse (M=-0,093) bei einem α-Level von 0,05 keinen signifikanten Unterschied: S=1618,5; p=0,0722. Ein zum Signifikanzniveau α=0,05 durchgeführter T-Test zeigt keinen signifikanten Unterschied zwischen den beiden Schätzmethoden; t=-0,10, p=0,9166. Dabei ist der Mittelwert der Differenzen -0,00117 [-0,0231; 0,0208], die Standardabweichung 0,1623. An den 95%-Konfidenzintervallen wird deutlich, dass der wahre Mittelwert die 0 mit einschließt und die Schätzmethoden sich somit nicht signifikant voneinander unterscheiden. Die Effektratio (Effektschätzer der PS-Analyse im Verhältnis zu dem Effektschätzer der RCT-Analyse) beträgt im Mittel 0,95 [-0,143; 2,043] (Median: 0,989). Somit unterscheiden sich die beiden Schätzgrößen nicht signifikant voneinander, die 1 befindet sich mitten im 95%-Konfidenzintervall.

3.3 Zusatzvergleich mit reduzierter Zufallsprobe

Der Vergleich zwischen RCT- und PS-Analyse und der Zusatzvergleich zwischen RCT- und reduzierter RCT-Analyse ergeben beide sehr ähnliche Diagramme (Abbildungen 13, 14, S.45, 46). Vor allem bei größeren Studien gibt es jedoch Unterschiede in Abbildung 14. Das zeigt sich auch bei der Betrachtung der ICCs, welche ohne Gewichtung 91,0% (95%-KI: 88,6%; 93,3%) für den RCT-PS Vergleich und 92,7% (95%-KI: 90,8%; 94,6%) für den Vergleich zwischen RCT und reduzierten RCTs ergeben und mit Gewichtung 98,0% (95%-KI: 97,6%; 98,4%) für den RCT-PS Vergleich und 87,5% (95%-KI: 83,3%; 91,7%) für den Vergleich zwischen RCT und reduzierten RCTs. (4)

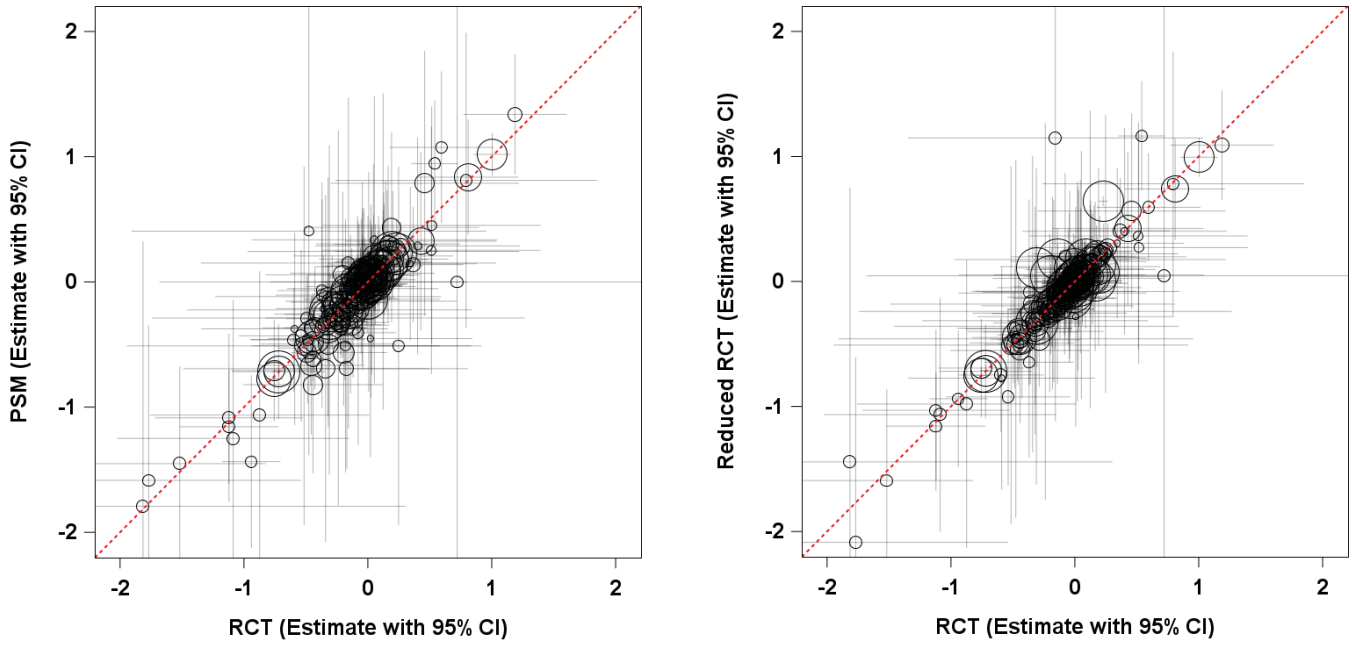

Abb. 13: Zusatzdarstellung mit zufälliger Reduktion. Die Effektschätzer sind mit 95%-KI gegeneinander aufgetragen. Die Kreisgröße stellt die Größe der Studie dar. Links: RCT-Analyse gegen PS-Analyse. Rechts: RCT-Analyse gegen reduzierte RCT-Analyse (4)

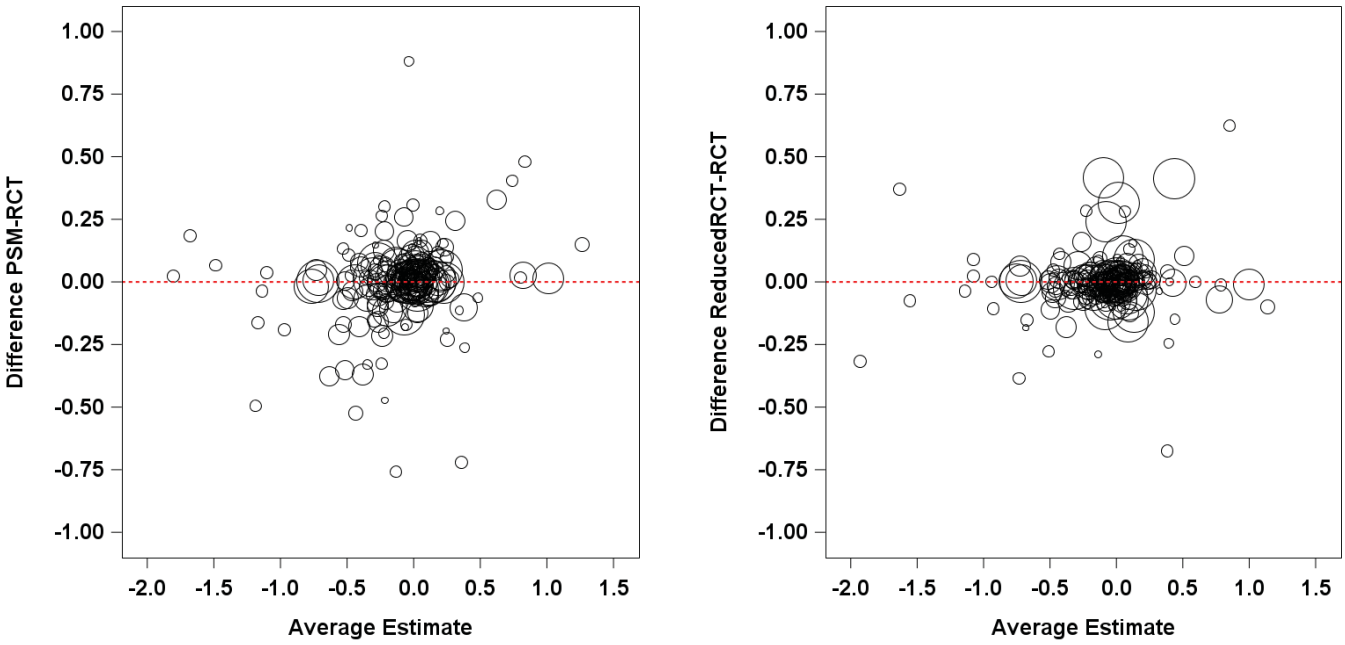

Abb. 14: Bland-Altman-Diagramme. Bland-Altman-Diagramme zur Darstellung des Mittelwertes zweier Studien gegen deren Differenz. Die Kreisgröße stellt die Größe der Studie dar. Links: RCT-Analyse und PS-Analyse. Rechts: RCT-Analyse und reduzierte RCT-Analyse (4)

4 Diskussion und Schlussfolgerungen

4.1 Interpretation der Ergebnisse

In dieser Arbeit wurde eine weitere Methode zum Vergleich zwischen randomisierten und nicht-randomisierten Studien vorgestellt. Dazu sollte der Einfluss unbekannter Störgrößen auf die Effektschätzer randomisiert kontrollierter Studien analysiert werden. Dabei wurden pro randomisiert kontrollierte Studie zwei separate Auswertungen vorgenommen: eine RCT-Analyse und eine PS-gematchte Analyse - eine Analyse, bei der nur die bekannten Patientenmerkmale noch gleichmäßiger verteilt werden. Im Anschluss wurden diese beiden Analysen auf Abweichungen oder Übereinstimmung miteinander verglichen. Bei fehlender Assoziation von unbekannten und bekannten Störgrößen würde der Behandlungseffekt der PS-gematchten Analyse verzerrt geschätzt und die Ergebnisse würden sich folglich voneinander unterscheiden. Bei Assoziation hingegen würde die bessere Verteilung der bekannten Störgrößen eine bessere Verteilung der unbekannten Störgrößen mit sich ziehen und keine Veränderung in der Schätzung des Effektes herbeiführen, sodass sich die Ergebnisse nicht signifikant voneinander unterscheiden würden. Zusätzlich wurde noch eine Zufallsstichprobe gezogen, welche genauso viele Beobachtungen hat wie beim PS-*Matching* zur optimalen Caliperweite. (4)

Zusammengefasst zeigen sich keinerlei signifikante Unterschiede zwischen den beiden Schätzmethoden. Die Zusatzauswertung mit den Zufallsstichproben unterstreicht sogar die bessere Übereinstimmung zwischen RCT- und PS-*Matching*-Analyse (4). Die Schätzer der PS-gematchten Analyse befinden sich in den 95%-Konfidenzintervallen der Schätzer der RCT-Analyse und beide Methoden stimmen signifikant in der Schätzung signifikanter Effekte überein. Der dazu benutzte McNemar-Test analysiert die Verteilung binärer Variablen bei verbundenen Stichproben und ist somit zum Vergleich der Schätzung signifikanter Effekte beider Schätzmethoden geeignet. Genauer genommen testet er auf diejenigen Felder einer 2x2 Kreuztabelle, die eine Abweichung zwischen den beiden Schätzmethoden anzeigen. Der für diesen Vergleich ebenfalls benutzte Kappa-Koeffizient zielt hingegen auf die Übereinstimmung ab und zeigt mit κ=0,81 nach Greve und Wentura (72) eine sehr gute bis ausgezeichnete, nach Landis und Koch (73) eine nahezu perfekte Übereinstimmung. Folglich kann die Kombination der Aussagen des McNemar-Tests und des Kappa-Koeffizienten eine sehr gute Aussage über Übereinstimmung und Abweichung der beiden Schätzmethoden machen. Im T-Test und *Wilcoxon Signed Rank Test* zeigen sich keine Unterschiede der beiden Schätzmethoden. Da die Normalverteilungsannahme für die zu testenden Variablen aber

verletzt wird, könnte man meinen, dass der parametrische T-Test nicht mehr anzuwenden ist und auf einen nicht-parametrischen Test ausgewichen werden muss. Der T-Test ist aber äußerst robust gegen die Verletzung der Normalverteilungsannahme, falls die zu vergleichenden Gruppen nicht zu klein oder zu unterschiedlich groß sind (74), was in unseren Analysen nicht der Fall ist. Deshalb kann der T-Test hier trotzdem angewandt werden und liefert valide p-Werte. Außerdem kann bei ihm im Gegensatz zum *Wilcoxon Signed Rank Test* ein Effektschätzer auf der Skala der Zielgröße mit Konfidenzintervallen ausgegeben werden. Diese Konfidenzintervalle liefern zusätzliche Informationen dazu, in welchem Bereich sich der wahre Mittelwert der Differenzen befindet. Da in unserem Fall die 0 mit eingeschlossen ist, ist dies eine weitere Bestätigung dafür, dass es keine signifikanten Unterschiede zwischen den beiden Schätzmethoden gibt. Weiter zeigen die Schätzer beider Methoden einen starken linearen Zusammenhang. Auch bei der linearen Regression ist das Modell robust gegen das Vorliegen von Heteroskedastizität statt Homoskedastizität, sodass der lineare Zusammenhang als valide betrachtet werden kann. Mit einer Regressionsfunktion von y=0,99238x+0,00044610 wird Folgendes deutlich: Der y-Achsenabschnitt ist vernachlässigbar gering und mit p=0,9691 nicht signifikant von 0 verschieden. Der Achsenabschnitt ist also mit sehr hoher Wahrscheinlichkeit mit 0 identisch. Dagegen ist der Regressionkoeffezient sehr wohl signifikant von 0 verschieden; p<0,0001. Wird der RCT-Schätzer um 1 Einheit erhöht, erhöht sich der PS-Schätzer um 0,99238 Einheiten. Der Zusammenhang zwischen RCT- und PS-Schätzer gleicht nahezu perfekt unter Berücksichtigung des sehr geringen Achsenabschnittes und der Steigung, welche fast 1 ist, einer Funktion von y=x. Diese Funktion würde aussagen, dass sich die Schätzer beider Schätzmethoden keineswegs unterscheiden und identisch wären.

4.2 Limitationen

Trotz eindeutiger Ergebnisse, gibt es in der Methodik und Herangehensweise dieser Arbeit Defizite, welche nicht verschwiegen werden sollten. Eine verbesserte Herangehensweise würde noch präzisere Ergebnisse liefern und die Frage nach dem Einfluss unbekannter Störgrößen in RCTs noch besser beantworten können.

Zuerst kann die Gesamtanzahl eingeschlossener Studien kritisiert werden. In die Analyse dieser Arbeit konnten nach strengen Einschlusskriterien 26 Studien eingeschlossen werden, woraus 37 Studiendatensätzen entstanden, da einige Studien auch mehr als zwei Therapiearme besitzen. Aus diesen Datensätzen können 213 Zielgrößen extrahiert werden, welche die Grundlage der Analyse sind. Mit 213 Beobachtungen aus 26 Studien können zwar Ergebnisse erzielt werden. Vergleicht man diese Anzahl jedoch mit der Anzahl eingeschlossener Patienten innerhalb der einzelnen Studien (Minimum 475,

Maximum 19.435), wird klar, dass die Anzahl unserer Beobachtungen niedriger ist als die der kleinsten eingeschlossenen Studie. Je größer die Anzahl der eingeschlossenen Studien und der benutzten Zielgrößen ist, desto besser können mögliche Mittelwertsunterschiede aufgedeckt werden.

Mit dem Einschluss einer größeren Anzahl an Studien könnte auch eine weitere Limitation angegangen werden: Durch den Einschluss der 26 Studien werden nicht alle Bereiche der Human- bzw. erweiternd auch der Zahnmedizin abgedeckt. Die 26 Studien beschäftigen sich vorwiegend mit Erkrankungen des Herz-Kreislaufsystems, mit Nierenerkrankungen und Diabetes. (4) Einzelne Studien beschäftigen sich außerdem noch mit Hypercholesterinämie, benigner Prostatahyperplasie, traumatischer Hirnverletzung und Schlaganfällen. Es wird deutlich, dass die Art der Erkrankungen, die in den Studien untersucht werden, sehr einseitig ist. Nur 13 verschiedene Entitäten werden dabei unterschieden. Um eine bessere Verallgemeinerung anzustreben, ist es ratsam Studien zu allerlei Themen der Medizin einzuschließen, sei es die Krebsforschung, die Endoprothetik oder Studien zur Zulassung neuer Medikamente in der Neurologie. Je vielseitiger die Themen der Studien, desto repräsentativer werden die Vergleiche der beiden Schätzmethoden für die Gesamtzahl der RCTs. Zum Vergleich beschäftigte sich der Reviewvergleich von Anglemyer et al. (2014) (2) mit insgesamt 228 Entitäten. Eine ähnlich große Anzahl ist anzustreben.

Ein weiterer Punkt betrifft die Auswahl der Patientenmerkmale, welche für die Berechnung des PS-Modells gebraucht wurden. Mit dem Hintergedanken, dass von unserer Seite keinerlei Manipulation stattfinden kann, wurden bewusst nur die Patientenmerkmale genommen, welche sich in *table1* der zugehörigen Artikel befinden. (4) Leider stellt sich heraus, dass diese Anzahl der Patientenmerkmale unter den Studien stark schwankt und zum Teil sehr niedrig ist. Beispielsweise können in der CPPT Studie nur 7 Patientenmerkmale in das PS-Modell eingeschlossen werden. Hinzu kommt die Problematik, dass die Summe der quadrierten z-Differenzen vor *Matching*, welche in etwa mit der Anzahl der in das PS-Modell einbezogenen Patientenmerkmale (Summe der binären, ordinalen und stetigen Patientenmerkmale) übereinstimmen sollte, stark von dieser Anzahl abweicht. Diese Abweichung reicht von einer nahezu perfekten Übereinstimmung mit Abweichungen von 2,2% bis hin zu Abweichungen um 75,3%. Obwohl sich die Anzahl der berechneten z-Differenzen in allen Fällen im 95%-Konfidenzintervall der Summe der quadrierten z-Differenzen befindet (Ausnahmen s. 2.7), sind die Schwankungen subjektiv gesehen trotzdem relativ groß. Für die Schätzung des PS-Modells sollten am besten so viele Patientenmerkmale wie möglich einbezogen werden und nicht nur die, welche im Paper hervorgehoben werden. Dadurch könnte das PS-Modell insgesamt verbessert, verfeinert und die Übereinstimmung der Anzahl

der berechneten z-Differenzen mit der Summe der quadrierten z-Differenzen vor *Matching* optimiert werden.

Zuletzt kann auch die Anzahl der Beobachtungen nach *Matching* kritisiert werden. Auch nach *Matching* stehen noch im Mittel 89,6% der Beobachtungen der RCT-Analyse zur Verfügung. Diskutiert werden kann hierbei, ob bei einem Wegfall von nur rund 10% der Beobachtungen die unbekannten Störgrößen vielleicht gar keine Möglichkeit hatten, Einfluss auf die Effektschätzer auszuüben. Dem gegenüber steht jedoch das PS-Paradox. (4) Dieses besagt, dass nach Erreichen der besten Balanciertheit die weitere Reduktion der Beobachtungen den gegenteiligen Effekt hervorrufen würde, nämlich eine Zunahme von Imbalance und Verzerrungen (75).

Zusammenfassend kann die Anzahl der eingeschlossenen Studien und der Patientenmerkmale kritisiert werden, die Anzahl der untersuchten Entitäten und die Abweichung der Summe quadrierter z-Differenzen vor *Matching* von der Anzahl berechneter z-Differenzen. Durch Erhöhung der Studienanzahl und der Anzahl der Patientenmerkmale kann ein repräsentativeres Kollektiv zur Verfügung gestellt werden und genauere PS-Modelle geschätzt werden. Zu einer Steigerung der Repräsentativität würde auch der Einschluss vieler verschiedener Erkrankungen und medizinischer Zusammenhänge führen.

4.3 Literaturvergleich

In der Literatur wurden schon einige Studien oder *Reviews* beschrieben, in denen RCTs mit Beobachtungsstudien verglichen wurden. Beispielsweise führten Dahabreh et al. 2012 (76) einen Vergleich zum Krankheitsbild des akuten Koronarsyndroms zwischen RCTs und Beobachtungsstudien durch, bei denen PS-Methoden angewandt wurden. Dazu wurden 63 Beobachtungsstudien zu 21 RCTs gematcht, wobei darauf geachtet wurde, dass bei jedem 3:1-*Matching* die Grundbedingungen übereinstimmten (z.B. Therapie, Patientenmerkmale und Mortalitätszielgrößen). Am Ende wurden dann die Effektschätzer beider Designs in Relation gesetzt und auf Übereinstimmung getestet. Dabei ergaben „nur [...] 2 von 17 Vergleichen [...] statistisch signifikant unterschiedliche Schätzer", „jedoch zeigten sich substanzielle Unterschiede in der Effektgröße (Ratio zwischen PS und RCT-basierten RRs kleiner als 0,7) in 6 von 17 Vergleichen". Weiter führten PS-Analysen zu Überschätzungen der Effekte. Trotz der homogenen Grundbedingungen, gab es auch Schwächen. So wurde das 3:1-*Matching* mittels dreier *Reviewer* als zu subjektiv empfunden und die „gematchten Beobachtungsstudien [...] nicht als repräsentativ für alle PS-*Matching* benutzenden Beobachtungsstudien“ gesehen. Außerdem wurden die Vorgaben bei der Anwendung des PS-Modells nicht eingehalten (Einsatz

von Regression statt empfohlene Analyse mit *Matching* (8,20,21)), was gegebenenfalls zu einer größeren Übereinstimmung hätte führen können. (76)

Auch Bolland et al. (2015) (77) verglichen RCTs mit Beobachtungsstudien. Hierbei wurden aber zwei Studiendesigns in ein und derselben Studie angewandt. In der WHICaD Studie wurden 36.282 postmenopausale Frauen 7 Jahre lang beobachtet. Dabei konnte ihnen entweder ein Placebo, oder aber eine Kombinationsgabe aus Calcium und Vitamin D gegeben werden. Zusätzlich konnten die Frauen aber auch selbst entscheiden, ob sie zusätzliches Calcium und Vitamin D zu sich nehmen wollten. Daraus resultierten zwei Studiendesigns: einerseits eine RCT mit den beiden Therapiearmen Placebo oder Substitutionstherapie, in welchen aber keine der Frauen zusätzliches Calcium und Vitamin D zu sich nahmen. Andererseits ergab sich auch eine Beobachtungsstudie über alle Placebopatientinnen, hierbei jedoch sowohl Frauen mit als auch ohne zusätzliche Präparate. Die Ergebnisse dieser beiden Studiendesigns wurden miteinander verglichen, ebenfalls wurde die RCT mit einer separat durchgeführten WHI Beobachtungsstudie verglichen. Zum einen ergaben die Analysen, dass es eher keine Unterschiede zwischen den beiden Designs der WHICaD gab (jedoch zu beachten: gleiche Kontrollgruppen, keine unabhängigen Studien), zum anderen jedoch unterschieden sich die Resultate zwischen WHICaD und der WHI Beobachtungsstudie in 50% der Fälle. (77)

Einen ähnlichen Ansatz verfolgten Stuart et al. (2017) (78), als sie mit der GRACE Studie ebenfalls eine RCT mit einer Beobachtungsstudie verglichen. Bei der RCT wurde entweder ein Placebo oder Amoxicillin verabreicht, bei der Beobachtungsstudie wurden therapeutisch unterschiedliche oder keine Antibiotika verschrieben. Dabei ist das Besondere, dass für beide Studiendesigns u.a. dieselben Ein- und Ausschlusskriterien und Zielgrößen festgelegt wurden und sie sich somit exakt in den Grundbedingungen glichen. Letztendlich lieferten sowohl RCT als auch die Beobachtungsstudie vergleichbare Ergebnisse, sodass Beobachtungsstudien eine gute Alternative darstellen können. Jedoch wurde bei der Therapie kein signifikanter Effekt festgestellt, „diese Analyse sollte bei zwei Studien mit statistisch signifikantem Therapieeffekt nachgebildet werden". (78)

Die oben genannten Studien deuten leicht eine gute Übereinstimmung zwischen den Ergebnissen von RCTs und Beobachtungsstudien zu derselben Entität an. Concato et al. (2000) (79) führte einen Vergleich zwischen Metaanalysen zu insgesamt 5 Entitäten durch. Dabei enthielten die Metaanalysen entweder nur Informationen aus RCTs, nur aus Beobachtungsstudien oder aber Informationen aus beiden Studiendesigns. Zum Vergleich wurde u.a. pro Entität jeweils für die RCTs und die Beobachtungsstudien ein gepooltes relatives Risiko (RR) mit 95%-KI berechnet. Bei der Beurteilung der BCG-Impfung ergab sich ein gepooltes RR von 0,49 [0,34; 0,70] bei RCTs und 0,50 [0,39; 0,65] bei den

Beobachtungsstudien. Die Analysen zum Thema Mammographiescreening und Brustkrebs ergaben ein gepooltes RR von 0,79 [0,71; 0,88] aus den RCTs im Vergleich zu 0,61 [0,49; 0,77] aus den Beobachtungsstudien. Bei dem Zusammenhang zwischen Traumata und Cholesterinspiegel ergab sich ein gepooltes RR der RCT von 1,42 [0,94; 2,15] und bei den Beobachtungsstudien 1,40 [1,14; 1,66]. Die Entität Schlaganfall bei medikamentös behandeltem Bluthochdruck lieferte bei den RCTs ein gepooltes RR von 0,58 [0,50; 0,67] und bei den Beobachtungsstudien 0,62 [0,60; 0,65]. Schließlich ergab die Untersuchung von KHK bei medikamentös behandeltem Bluthochdruck ein gepooltes RR der RCTs von 0,86 [0,78; 0,96] im Vergleich zu 0,77 [0,75; 0,80] aus den Beobachtungsstudien. Die Ergebnisse zeigten sehr geringe Unterschiede zwischen den beiden Designs und Anzeichen dafür, dass „sehr gut designte Beobachtungsstudien nicht regelmäßig Therapieeffekte überschätzen". (79)

Anglemyer et al. (2014) (2) führten eine Analyse von 14 *Reviews* durch, welche Daten aus 1.583 Metaanalysen besaßen und sich mit 228 Krankheiten beschäftigten. Die ausgewählten *Reviews* verglichen RCTs und Beobachtungsstudien zu derselben Fragestellung. Zum Vergleich wurden die Effektschätzer in ein Verhältnis gesetzt (ROR: *ratio of odds ratios*) und schließlich aus allen RORs eine gepoolte ROR gebildet. Die gepoolte ROR aller *Reviews* betrug 1,08 [0,96; 1,22], darunter zeigten 11 von 14 *Reviews* keine signifikanten Unterschiede. Auch im Vergleich zwischen RCTs und Kohortenstudien oder RCTs und Fall-Kontrollstudien gab es keine signifikanten Unterschiede mit gepoolten RORs von 1,04 [0,89; 1,21] für den Vergleich zwischen RCTs und Kohortenstudien und 1,11 [0,91; 1,35] für den Vergleich RCTs mit Fall-Kontrollstudien. Letztendlich gab es „wenig Anzeichen für signifikante Unterschiede in der Effektschätzung zwischen Beobachtungsstudien und RCTs unabhängig von spezifischem Design der Beobachtungsstudien, der Heterogenität oder dem Einschluss von Studien mit pharmakologischen Interventionen". (2)

Gegensätzlich zu Anglemyers Ergebnissen fanden Hemkens et al. (80) heraus, dass sich Ergebnisse aus Beobachtungsstudien und aus RCTs zu der gleichen klinischen Fragestellung sehr wohl unterscheiden. Die Ergebnisse dieser Veröffentlichung werden jedoch von Franklin et al. (81) bezüglich der methodischen Herangehensweise (selektive Invertierung) kritisiert. (4)

Im Vergleich zwischen RCTs und Beobachtungsstudien wird immer wieder die Problematik der Störgrößen beschrieben. Zwar gibt es Methoden, um in Beobachtungsstudien bekannte Störgrößen zu kontrollieren (bsp. *Matching* oder Exklusion), unbekannte Störgrößen sind aber robust gegen solche Methoden und können nur durch RCTs kontrolliert und gleichmäßig auf die Gruppen verteilt werden. (1) Unbekannte Störgrößen sind zwar erst relevant, wenn eine Assoziation mit den bekannten Störgrößen fehlt (3), sie sind aber auch erst dann relevant, wenn sie „stärkstens und negativ mit sowohl Zielgröße als auch Exposition

assoziiert sind“ (82). Fraglich ist, ob solche starken Störgrößen bis jetzt immer noch unbekannt sind oder ob diese vor allem aufgrund des großen Einflusses schon gefunden wurden (4).

Auch in dieser Arbeit wurde eine Methode zum Vergleich zwischen RCTs und Beobachtungsstudien angewandt. Hierbei wurde jedoch auf das eben genannte Problem der unbekannten Störgrößen eingegangen. RCTs wurden dabei in herkömmlicher Weise analysiert und zusätzlich nach PS-*Matching*. Bei Dahabreh et al. (2012) (76) wurde schon kritisiert, dass sie sich bei der PS-Methode nicht an die Vorgaben hielten und dass sich mit Einsatz des PS-*Matching* als empfohlene Methode gegebenenfalls noch weniger Unterschiede zwischen den Studiendesigns ergeben würden. In der Durchführung unserer Studie wird nach Austins (7) Empfehlung das *Nearest Neighbor Caliper Matching* ohne *Replacement* benutzt und Patientenmerkmale in das Modell einbezogen, welche keine Verbindung mit der Therapiezuordnung, wohl aber mit den Zielgrößen haben. Bei der angewandten Matchingmethode ist die 1:1-Zuteilung die einfachste, wenngleich auch kritisiert wird, dass durch das Wegfallen vieler Individuen ein Powerverlust zu verzeichnen ist (8). Dem entgegenzusetzen sei, dass in unseren Analysen die Anzahl der Beobachtungen nach *Matching* nicht stark sank und außerdem „die *Power* [auch] steigt, wenn die Gruppen ähnlicher sind, wegen reduzierter Extrapolation und höherer Präzision“ (3,83). Weiter wurde festgestellt, dass „Effektschätzer bei 1:1 *Matching* kleinere Standardabweichungen besitzen als die bei einer linearen Regression trotz tausender im PS-*Matching* wegfallender Individuen“ (84). Die angewandte Matchingmethodik sollte die bestmöglichen Ergebnisse garantieren. Leider hätte die Anzahl der Patientenmerkmale, anstatt sich auf die Merkmale aus *table*1 zu beschränken, deutlich erhöht und damit bessere Ergebnisse erzielt werden können.

Im Vergleich zu Stuart et al. (2017) (78) wurden in dieser Arbeit sowohl Therapien mit als auch ohne signifikanten Effekten untersucht. Wo Stuart et al. (2017) (78) zwar wenige Unterschiede zwischen der Effektschätzung aus RCT und Beobachtungsstudie zeigten, waren die Ergebnisse bei Dahabreh et al. (2012) (76) und Bolland et al. (2015) (77) teils ambivalent. Repräsentative Schlüsse auf die Gesamtheit aller Studien aus Analysen zu jeweils nur einer Entität (76–78) oder weniger Entitäten (79) zu ziehen, stellt eine klare Schwierigkeit dar. Diese Schwierigkeit zeigt sich auch in dieser Arbeit. Zwar werden insgesamt 13 Entitäten aus 26 Studien behandelt, ein Vergleich mit Anglemyer et al. (2014) (2), in dem Studien zu 228 verschiedenen Entitäten verglichen wurden, zeigt deutlich, dass nur ein kleiner Teil medizinischer Fragestellungen mit dieser Arbeit abgedeckt werden konnte.

Mit dem Einschluss der Studien aus Kent et al. (6) wurden Studien analysiert, die eine Fallzahl von 475 bis 19.435, im Mittel 5.233 Beobachtungen, besitzen. Damit wird sicher gegangen, dass die Studien groß genug sind und die Randomisierung überhaupt erst wirkt und

somit unverzerrte Effekte geschätzt werden können. Nguyen et al. fanden heraus, dass dies erst bei Studien ab 1.000 Teilnehmern gegeben ist; eine Zahl, die „nicht einmal die Hälfte der Phase 3 RCTs [...] in den Top fünf der Journals mit dem höchsten Einfluss“ erreiche (85).

4.4 Schlussfolgerungen und Ausblick

Der Vergleich zwischen RCTs und PS-gematchten RCTs, durch den eine indirekte Aussage über die unbekannten Störgrößen gemacht werden kann, ergab ein verblüffendes Ergebnis. Es gab keine signifikanten Unterschiede zwischen den Schätzmethoden, sondern eine sehr gute Übereinstimmung und eine hohe Korrelation. Der Vergleich zwischen RCT- und PS-Analyse zeigt sogar kleinere Unterschiede in der Schätzung des wahren Effektes als ein Vergleich zwischen RCT- und reduzierter RCT-Analyse, in welcher *per definitionem* der Effekt unverzerrt geschätzt wird (4).

Aufgrund des Fehlens signifikanter Unterschiede zwischen den beiden Schätzmethoden ist aus den Ergebnissen dieser Arbeit zu schlussfolgern, dass sich nach *Matching* die unbekannten Patientenmerkmale genauso wie die bekannten Patientenmerkmale gleichmäßig auf Kontroll- und Interventionsgruppe verteilten und somit keine Veränderung der Effektschätzer hervorrufen konnten. Wenn die unbekannten Patientenmerkmale sich mit den bekannten gleichmäßig verteilen, sind sie mit diesen in irgendeiner Art positiv assoziiert. Sind sie wiederum positiv assoziiert, stellen sie aber für die Schätzung des wahren Behandlungseffektes im Sinne von Stuart (3) keine Gefahr mehr dar. Stellen sie keine Gefahr mehr dar, verliert einer der Hauptgründe für den Einsatz von RCTs wiederum an Gewicht. Nämlich jener, dass nur in RCTs eine gleichmäßige Verteilung von bekannten und unbekannten Patientenmerkmalen garantiert ist (1,8).

Da ohnehin schon Evidenz besteht, dass zwischen RCTs und Beobachtungsstudien wenig Unterschiede bestehen (2), kann in gut konzipierten Beobachtungsstudien ein PS-*Matching* nachträglich durchgeführt werden, sodass bekannte und unbekannte Störgrößen gleichmäßig verteilt werden und so eine Randomisierung imitiert wird. Somit können Ergebnisse erzielt werden, die sehr gut mit Ergebnissen aus RCTs vergleichbar und darüber hinaus auch extern valide sind. Somit kann die Problematik der Beobachtungsstudien, nämlich die fehlende Möglichkeit der gleichmäßigen Verteilung unbekannter Störgrößen, als weniger gravierend eingestuft werden als bisher eingeschätzt.

Das heißt für die Zukunft der Studienplanung, dass vermehrt auf Beobachtungsstudien gesetzt werden kann. Die RCTs können den höchsten Evidenzgrad medizinischer Forschung nicht mehr für sich beanspruchen und es kann vermehrt auf Alternativen zurückgegriffen werden.

Die häufig kritisierte kurze Beobachtungszeit und mangelnde externe Validität sind Hauptkritikpunkte von RCTs. Außerdem sind diese immer mit enormen Kosten verbunden und in manchen Fragestellungen gegebenenfalls ethisch gar nicht vertretbar. Beobachtungsstudien hingegen erlauben längere Beobachtungszeiten, ja sogar Langzeitbeobachtungen und vor allem werden sie für die höhere externe Validität und Patientenrelevanz geschätzt. (9–11,18)

Für praxisrelevantere Aussagen eignen sich Beobachtungsstudien besser und sind auch mit erheblich weniger Aufwand, sowohl finanziell als auch organisatorisch, durchzuführen (11). Mit dieser Arbeit soll ein Anstoß gegeben werden, bei Fragen, die bis *dato* primär mit RCTs beantwortet wurden, in der Studienplanung vermehrt auch an Beobachtungsstudien zu denken und den Weg der nachträglichen „Randomisierung“ durch PS-*Matching* zu erwägen. Somit können Beobachtungsstudien umso präzisere Aussagen und vor allem sowohl intern als auch extern völlig valide Ergebnisse liefern, welche auf den Praxisalltag besser zu übertragen sind als bei RCTs. Durch den häufigeren Einsatz von Beobachtungsstudien können immense Kosten, Zeit und Aufwand gespart werden und z.T. ethische Schwierigkeiten vermieden werden. Dies kann in Zukunft zu einer steigenden Relevanz von und einem steigenden Interesse an Beobachtungsstudien führen.

5 Literatur- und Quellenverzeichnis

1. Weiner BK. On confounders: known and unknown. The Spine Journal. November 2009;9(11):924–5.

2. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Methodology Review Group, Herausgeber. Cochrane Database of Systematic Reviews [Internet]. 29. April 2014 [zitiert 18. April 2018]; Verfügbar unter: http://doi.wiley.com/10.1002/14651858.MR000034.pub2

3. Stuart EA. Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science. Februar 2010;25(1):1–21.

4. Kuss O, Miller M. Unknown confounders did not bias the treatment effect when improving balance of known confounders in randomized trials. Journal of Clinical Epidemiology. Oktober 2020;126:9–16.

5. Rubin DB, Thomas N. Matching Using Estimated Propensity Scores: Relating Theory to Practice. Biometrics. März 1996;52(1):249–64.

6. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. International Journal of Epidemiology. 3. Juli 2016;45(6):2075–88.

7. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Statistics in Medicine. 15. März 2014;33(6):1057–69.

8. Kuss O, Blettner M, Börgermann J. Propensity score: an alternative method of analyzing treatment effects. Deutsches Aerzteblatt Online [Internet]. 5. September 2016 [zitiert 18. April 2018]; Verfügbar unter: http://www.aerzteblatt.de/10.3238/arztebl.2016.0597

9. Rothwell PM. External validity of randomised controlled trials: “To whom do the results of this trial apply?”. The Lancet. Januar 2005;365(9453):82–93.

10. Frieden TR. Evidence for Health Decision Making — Beyond Randomized, Controlled Trials. Drazen JM, Harrington DP, McMurray JJV, Ware JH, Woodcock J, Herausgeber. N Engl J Med. 3. August 2017;377(5):465–75.

11. Borah BJ, Moriarty JP, Crown WH, Doshi JA. Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go? Journal of Comparative Effectiveness Research. Januar 2014;3(1):63–78.

12. Furlan AD, Tomlinson G, Jadad A (Alex) R, Bombardier C. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. Journal of Clinical Epidemiology. März 2008;61(3):209–31.

13. Röhrig B, Prel J-B du, Wachtlin D, Blettner M. Studientypen in der medizinischen Forschung. Deutsches Aerzteblatt Online. 2009;106(15):262–8.

14. Kabisch M, Ruckes C, Seibert-Grafe M, Blettner M. Randomized Controlled Trials: Part 17 of a Series on Evaluation of Scientific Publications. Deutsches Aerzteblatt Online. 30. September 2011;108(39):663–8.

15. ICH E10. Choice of Control Group and Related Issues in Clinical Trials. London UK: International Conference on Harmonization 2000;

16. Lange S, Sauerland S, Lauterberg J, Windeler J. The range and scientific value of randomized trials. Deutsches Aerzteblatt Online. 22. September 2017;635–40.

17. Herbold M. Nützliche Erkenntnisse über Arzneimittel im Alltag. Dtsch Ärztebl. 1996;93(46):A-3010-3012.

18. Hammer GP, Prel J-B du, Blettner M. Vermeidung verzerrter Ergebnisse in Beobachtungsstudien. Dtsch Ärztebl. 2009;106(41):664–8.

19. Stuart EA, Marcus SM, Horvitz-Lennon MV, Gibbons RD, Normand S-LT, Brown CH. Using Non-Experimental Data to Estimate Treatment Effects. Psychiatric Annals. 1. Juli 2009;39(7):719–28.

20. Morgan SL, Harding DJ. Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. Sociological Methods & Research. August 2006;35(1):3–60.

21. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. The Journal of Thoracic and Cardiovascular Surgery. November 2007;134(5):1128-1135.e3.

22. Smith HL. Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. Sociological Methodology. August 1997;27(1):325–53.

23. Ming K, Rosenbaum PR. A Note on Optimal Matching With Variable Controls Using the Assignment Algorithm. Journal of Computational and Graphical Statistics. September 2001;10(3):455–63.

24. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Statistics. März 2011;10(2):150–61.

25. Pattanayak CW, Rubin DB, Zell ER. Propensity Score Methods for Creating Covariate Balance in Observational Studies. Revista Española de Cardiología (English Edition). Oktober 2011;64(10):897–903.

26. Belitser SV, Martens EP, Pestman WR, Groenwold RHH, Boer A, Klungel OH. Measuring balance and model selection in propensity score methods: BALANCE MEASURE FOR PROPENSITY SCORES METHODS. Pharmacoepidemiology and Drug Safety. November 2011;20(11):1115–29.

27. Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. Journal of Clinical Epidemiology. November 2013;66(11):1302–7.

28. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Statistics Soc A. 2008;171(2):481–502.

29. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine. 10. November 2009;28(25):3083–107.

30. IST [zitiert 24. November 2019] [Internet]. Verfügbar unter: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3104487/

31. FUTURA [zitiert 24. November 2019] [Internet]. Verfügbar unter: https://www.clinicalstudydatarequest.com/Posting.aspx?ID=48

32. NEJM [zitiert 24. November 2019] [Internet]. Verfügbar unter: www.nejm.org

33. JAMA [zitiert 24. November 2019] [Internet]. Verfügbar unter: jamanetwork.com

34. Elsevier [zitiert 24. November 2019] [Internet]. Verfügbar unter: https://www.elsevier.com/catalog?producttype=journal

35. Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, u. a. Mortality and Morbidity in Patients Receiving Encainide, Flecainide, or Placebo: The Cardiac Arrhythmia Suppression Trial. New England Journal of Medicine. 21. März 1991;324(12):781–8.

36. Lipid Research Clinics Program. The Lipid Research Clinics Coronary Primary Prevention Trial Results. I. Reduction in incidence of coronary heart disease. JAMA. 1984;251:351–64.

37. Anderson JL, Platia EV, Hallstrom A, Henthorn RW, Buckingham TA, Carlson MD, u. a. Interaction of baseline characteristics with the hazard of encainide, flecainide, and moricizine therapy in patients with myocardial infarction. A possible explanation for increased mortality in the Cardiac Arrhythmia Suppression Trial (CAST). Circulation. 1. Dezember 1994;90(6):2843–52.

38. Rifkind BM. Lipid research clinics coronary primary prevention trial: Results and implications. The American Journal of Cardiology. August 1984;54(5):30–4.

39. The ACCORD Study Group. Effects of Intensive Blood-Pressure Control in Type 2 Diabetes Mellitus. New England Journal of Medicine. 29. April 2010;362(17):1575–85.

40. The ACCORD Study Group. Effects of Intensive Glucose Lowering in Type 2 Diabetes. New England Journal of Medicine. 12. Juni 2008;358(24):2545–59.

41. The ACCORD Study Group. Effects of Combination Lipid Therapy in Type 2 Diabetes Mellitus. New England Journal of Medicine. 29. April 2010;362(17):1563–74.

42. The DCCT Research Group. The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus. New England Journal of Medicine. 30. September 1993;329(14):977–86.

43. Diabetes Care. Adverse events and their association with treatment regimens in the diabetes control and complications trial. Diabetes Care. November 1995;18(11):1415–27.

44. AFFIRM Investigators. A Comparison of Rate Control and Rhythm Control in Patients with Atrial Fibrillation. New England Journal of Medicine. 5. Dezember 2002;347(23):1825–33.

45. The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major Outcomes in High-Risk Hypertensive Patients Randomized to Angiotensin-Converting Enzyme Inhibitor or Calcium Channel Blocker vs Diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA: The Journal of the American Medical Association. 18. Dezember 2002;288(23):2981–97.

46. The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major Outcomes in Moderately Hypercholesterolemic, Hypertensive Patients Randomized to Pravastatin vs Usual Care: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT-LLT). JAMA: The Journal of the American Medical Association. 18. Dezember 2002;288(23):2998–3007.

47. AMIS Research Group. A Randomized, Controlled Trial of Aspirin in Persons Recovered From Myocardial Infarction. JAMA: The Journal of the American Medical Association. 15. Februar 1980;243(7):661–9.

48. The VA/NIH Acute Renal Failure Trial Network. Intensity of Renal Support in Critically Ill Patients with Acute Kidney Injury. New England Journal of Medicine. 3. Juli 2008;359(1):7–20.

49. The Beta-Blocker Evaluation of Survival Trial Investigators. A Trial of the Beta-Blocker Bucindolol in Patients with Advanced Chronic Heart Failure. New England Journal of Medicine. 31. Mai 2001;344(22):1659–67.

50. A Randomized Trial of Propranolol in Patients With Acute Myocardial Infarction: I. Mortality Results. JAMA. 26. März 1982;247(12):1707.

51. The Digitalis Investigation Group. The Effect of Digoxin on Mortality and Morbidity in Patients with Heart Failure. New England Journal of Medicine. 20. Februar 1997;336(8):525–33.

52. Diabetes Preventing Program Research Group. Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. New England Journal of Medicine. 7. Februar 2002;346(6):393–403.

53. Writing Committee for the ENRICHD Investigators. Effects of Treating Depression and Low Perceived Social Support on Clinical Events After Myocardial Infarction: The Enhancing Recovery in Coronary Heart Disease Patients (ENRICHD) Randomized Trial. JAMA. 18. Juni 2003;289(23):3106.

54. Bostom AG, Carpenter MA, Kusek JW, Levey AS, Hunsicker L, Pfeffer MA, u. a. Homocysteine-Lowering and Cardiovascular Disease Outcomes in Kidney Transplant Recipients: Primary Results From the Folic Acid for Vascular Outcome Reduction in Transplantation Trial. Circulation. 26. April 2011;123(16):1763–70.

55. Di Bisceglie AM, Shiffman ML, Everson GT, Lindsay KL, Everhart JE, Wright EC, u. a. Prolonged Therapy of Advanced Chronic Hepatitis C with Low-Dose Peginterferon. New England Journal of Medicine. 4. Dezember 2008;359(23):2429–41.

56. Eknoyan G, Beck GJ, Cheung AK, Daugirdas JT, Greene T, Kusek JW, u. a. Effect of Dialysis Dose and Membrane Flux in Maintenance Hemodialysis. New England Journal of Medicine. 19. Dezember 2002;347(25):2010–9.

57. Sandercock P, Niewada M, Czlonkowska A, International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. The Lancet. Mai 1997;349(9065):1569–81.

58. Antman EM, The Magnesium in Coronaries (MAGIC) Trial Investigators. Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. The Lancet. Oktober 2002;360(9341):1189–96.

59. Multiple Risk Factor Intervention Trial Research Group. Multiple Risk Factor Intervention Trial: Risk Factor Changes and Mortality Results. JAMA. 24. September 1982;248(12):1465.

60. McConnell JD, Roehrborn CG, Bautista OM, Andriole GL, Dixon CM, Kusek JW, u. a. The Long-Term Effect of Doxazosin, Finasteride, and Combination Therapy on the Clinical Progression of Benign Prostatic Hyperplasia. New England Journal of Medicine. 18. Dezember 2003;349(25):2387–98.

61. Hochman JS, Lamas GA, Buller CE, Dzavik V, Reynolds HR, Abramsky SJ, u. a. Coronary Intervention for Persistent Occlusion after Myocardial Infarction. New England Journal of Medicine. 7. Dezember 2006;355(23):2395–407.

62. The PEACE Trial Investigators. Angiotensin-Converting–Enzyme Inhibition in Stable Coronary Artery Disease. New England Journal of Medicine. 11. November 2004;351(20):2058–68.

63. Bulger EM, May S, Brasel KJ, Schreiber M, Kerby JD, Tisherman SA, u. a. Out-of-Hospital Hypertonic Resuscitation Following Severe Traumatic Brain Injury: A Randomized Controlled Trial. JAMA. 6. Oktober 2010;304(13):1455–64.

64. Bulger EM, May S, Kerby JD, Emerson S, Stiell IG, Schreiber MA, u. a. Out-of-hospital Hypertonic Resuscitation After Traumatic Hypovolemic Shock: A Randomized, Placebo Controlled Trial. Annals of Surgery. März 2011;253(3):431–41.

65. SHEP Coorperative Research Group. Prevention of Stroke by Antihypertensive Drug Treatment in Older Persons With Isolated Systolic Hypertension: Final Results of the Systolic Hypertension in the Elderly Program (SHEP). JAMA. 26. Juni 1991;265(24):3255–64.

66. The SOLVD Investigators. Effect of Enalapril on Survival in Patients with Reduced Left Ventricular Ejection Fractions and Congestive Heart Failure. New England Journal of Medicine. August 1991;325(5):293–302.

67. The SOLVD Investigators. Effect of Enalapril on Mortality and the Development of Heart Failure in Asymptomatic Patients with Reduced Left Ventricular Ejection Fractions. New England Journal of Medicine. 3. September 1992;327(10):685–91.

68. The TIMI Study Group*. Comparison of Invasive and Conservative Strategies after Treatment with Intravenous Tissue Plasminogen Activator in Acute Myocardial Infarction. New England Journal of Medicine. 9. März 1989;320(10):618–27.

69. Coca-Perraillon M. Local and Global Optimal Propensity Score Matching. SASGlobalForum2007.

70. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statistics in Medicine. 30. Mai 2008;27(12):2037–49.

71. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin. 1979;86(2):420–8.

72. Greve W, Wentura D, Gräser H, Schmitz U. Wissenschaftliche Beobachtung: eine Einführung. 2. Aufl. Weinheim: Beltz; 1997. 182 S.

73. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. März 1977;33(1):S. 165.

74. Rasch B, Friese M, Hofmann W, Naumann E, Herausgeber. Deskriptive Statistik, Inferenzstatistik, t-Test, Korrelationstechniken, Regressionsanalyse, Formelsammlung, Glossar, Verteilungstabellen: mit 25 Tabellen. 3., erw. Aufl. Berlin: Springer; 2010. S. 59-60. (Quantitative Methoden).

75. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. Polit Anal. 7. Mai 2019;1–20.

76. Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, u. a. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. European Heart Journal. August 2012;33(15):1893–901.

77. Bolland MJ, Grey A, Gamble GD, Reid IR. Concordance of Results from Randomized and Observational Analyses within the Same Study: A Re-Analysis of the Women's Health Initiative Limited-Access Dataset. Coleman WB, Herausgeber. PLOS ONE. 6. Oktober 2015;10(10):e0139975.

78. Stuart BL, Grebel LE, Butler CC, Hood K, Verheij TJM, Little P. Comparison between treatment effects in a randomised controlled trial and an observational study using propensity scores in primary care. British Journal of General Practice. September 2017;67(662):e643–9.

79. Concato J, Shah N, Horwitz RI. Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. New England Journal of Medicine. 22. Juni 2000;342(25):1887–92.

80. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ. 8. Februar 2016;i493.

81. Franklin JM, Dejene S, Huybrechts KF, Wang SV, Kulldorff M, Rothman KJ. A Bias in the Evaluation of Bias Comparing Randomized Trials with Nonexperimental Studies. Epidemiologic Methods [Internet]. 27. November 2017 [zitiert 1. September 2019];6(1). Verfügbar unter: http://www.degruyter.com/view/j/em.2017.6.issue-1/em-2016-0018/em-2016-0018.xml

82. Nguyen T-L, Collins GS, Spence J, Fontaine C, Daurès J-P, Devereaux PJ, u. a. Magnitude and direction of missing confounders had different consequences on treatment effect estimation in propensity score analysis. Journal of Clinical Epidemiology. Juli 2017;87:87–97.

83. Snedecor GW, Cochran WG. Statistical methods. 7th ed. Ames, Iowa: Iowa State University Press; 1980. 507 S.

84. Smith HL. 6. Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. Sociological Methodology. August 1997;27(1):325–53.

85. Nguyen T-L, Collins GS, Lamy A, Devereaux PJ, Daurès J-P, Landais P, u. a. Simple randomization did not protect against bias in smaller trials. Journal of Clinical Epidemiology. April 2017;84:105–13.

6 Anhang: PSinRCT-Makro

```
* PS_IN_RCT.SAS;
 options pagesize=68 linesize=160;
 *options MLOGIC MPRINT SYMBOLGEN;
 options NOMLOGIC NOMPRINT NOSYMBOLGEN;
 %macro PSinRCT(_data=,
            _patid=,
            _treatment=,
            _treatment0=,
            _treatment1=,
            _BinaryBaselineVars=, _OrdinalBaselineVars=, _NominalBaselineVars=, _ContBaselineVars=,
                                  _Outcome=,
                                        _cwidths=10000 9800 9600 9400 9200 9000 8800 8600 8400 8200
8000 7800 7600 7400 7200 7000 6800 6600 6400 6200 6000 5800 5600
                                        5400 5200 5000 4800 4600 4400 4200 4000 3800 3600 3400 3200
3000 2800 2600 2400 2200 2000
                                       1950 1900 1850 1800 1750 1700 1650 1600 1550 1500 1450 1400
1350 1300 1250 1200 1150 1100 1050 1000 950 900 850 800 750 700
                                       650 600 550 500 450 400 350 300 250 200 150 100 95 90 85 80 75 70
65 60 55 50 45 40 35 30 25 20 15 10 9 8 7 6 5 4 3 2 1,
            _Binary=0, _Ordinal=0, _Nominal=0, _Continuous=0,
_Survival=0,_SurvivalCensVar=,_CensValue=,
                                  _seed=,
            _matchnumberofcontrols=1
  );
********************* PRELIMINARIES ***********************************************;
* Define a binary format for the treatment and the binary baseline variables and outcomes;
 proc format; value dichotom 0="No" 1=" Yes"; run;
* Scan the lists of baseline variables and outcomes, count them and write in them in single variables.
 The WORDSCAN macro of Saradha (Saradha P. Maximize the power of %SCAN using WORDSCAN utility.
PharmaSUG 2012,
 AD08, http://www.pharmasug.org/proceedings/2012/AD/PharmaSUG-2012-AD08.pdf) is used here;
 %macro wordscan(string=,prefix=,delim=%str( ));
    %global &prefix.cwordx;
       %let wordx=1;
       %let word=%scan(&string,%eval(&wordx),&delim);
```

```
        %global &prefix&wordx;
    %let &prefix&wordx=&word;
        %do %while(&word ne %str());
          %let wordx=%eval(&wordx+1);
          %let word=%scan(&string,%eval(&wordx),&delim);
          %global &prefix&wordx;
          %let &prefix&wordx=&word;
        %end;
        %let &prefix.c = %eval(&wordx-1);
 %mend wordscan;
 %wordscan(string=&_BinaryBaselineVars,  prefix=_BinaryBaselineVar,  delim=%str( ));
 %wordscan(string=&_ContBaselineVars,    prefix=_ContBaselineVar,    delim=%str( ));
 %wordscan(string=&_OrdinalBaselineVars, prefix=_OrdinalBaselineVar, delim=%str( ));
 %wordscan(string=&_NominalBaselineVars, prefix=_NominalBaselineVar, delim=%str( ));
 %wordscan(string=&_cwidths,             prefix=_cwidth,     delim=%str( ));
* Define the ZDIFF macro to compute the z-differences;
 %macro ZDiff(_datazdiff=,_scalezdiff=,_varzdiff=,_classzdiff=);
      %if &_scalezdiff=continuous %then %do;
                  ods listing close;
         proc sort data=&_datazdiff;by descending &_classzdiff;run;
         proc ttest data=&_datazdiff order=data;
                              class &_classzdiff; var &_varzdiff;
ods output Statistics=_ttestout(keep=Class N Mean StdDev where=(Class ne "Diff (1-2)"));
         run;
                    data _ttestout1 _ttestout2;
                       set _ttestout;
                              if _N_=1 then output _ttestout1; if _N_=2 then output _ttestout2;
                    run;
        data _resultttest(drop=Class _StdErr_Var1 _StdErr_Var2 _Stderr_Vardiff _CV1 _CV2 _M1 _M2
_CV_Common);
            merge _ttestout1(rename=(N=_N1 Mean=_Mean1 StdDev=_StdDev1))
_ttestout2(rename=(N=_N2 Mean=_Mean2 StdDev=_StdDev2));
                                    * Compute the z-difference from the means;
                                _ZDiffContinuous_Mean=(_Mean1-_Mean2)/sqrt(_StdDev1**2/_N1 +
_StdDev2**2/_N2);
          * Compute the z-difference from the variances;
                                    _StdErr_Var1=_StdDev1**2*sqrt(2/(_N1-1));
                                     _StdErr_Var2=_StdDev2**2*sqrt(2/(_N2-1));
```

```
                                    _Stderr_Vardiff=sqrt(_StdErr_Var1**2+_StdErr_Var2**2);
                            _ZDiffContinuous_Variance= (_StdDev1**2-_StdDev2**2)/_Stderr_Vardiff;
                                                * Compute the z-difference from the coefficients of variation;
                                                 _CV1=_StdDev1/_Mean1;
                                                 _CV2=_StdDev2/_Mean2;
                                        _M1=_N1-1;_M2=_N2-1;
                                        _CV_Common=(_M1*_CV1 + _M2*_CV2)/(_M1+_M2);
                    _ZDiffContinuous_CV= (_CV1-_CV2)/
sqrt((1/_M1+1/_M2)*_CV_Common**2*(0.5+_CV_Common**2));
                run;
                data out_&_varzdiff;
                    set _resultttest;
                    length _Variable $32;
                    _ZDiff=_ZDiffContinuous_Mean;
                    _Variable="&_varzdiff";
                    keep _Variable _ZDiff;
                run;
                proc print data=out_&_varzdiff label noobs; title"Z-difference, data=&_datazdiff,
scale=&_scalezdiff, var=&_varzdiff, class=&_classzdiff"; run;
                                        ods listing;
            %end;
                                %if &_scalezdiff=ordinal %then %do;
                                ods listing close;
                                    proc npar1way data=&_datazdiffwilcoxon correct=no;
                    class &_classzdiff; var &_varzdiff;
                  ods output WilcoxonTest=_WilcoxonOut(where=(Name1="Z_WIL")
rename=(nValue1=_ZDiffOrdinal) drop=Variable Label1 cValue1);
                run;
                data out_&_varzdiff;
                    set _WilcoxonOut;
                    length _Variable $32;
                    _ZDiff=_ZDiffOrdinal;
                    _Variable="&_varzdiff";
                    keep _Variable _ZDiff;
                run;
                proc print data=out_&_varzdiff label noobs; title"Z-difference, data=&_datazdiff,
scale=&_scalezdiff, var=&_varzdiff, class=&_classzdiff"; run;
                                        ods listing;
```

```
      %end;
                    %if &_scalezdiff=binary %then %do;
                       proc sort data=&_datazdiff;by descending &_classzdiff descending &_varzdiff;run;
                    ods listing close;
                       proc freq data=&_datazdiff order=data;
           tables &_varzdiff*&_classzdiff / riskdiffnorownopercent out=Frequencies(drop=PERCENT);
          ods output RiskDiffCol1=_RiskDiffOut(where=(Row="Difference") keep=Risk ASE Row);
        run;
                              data _RiskDiffOut2;
                                 set _RiskDiffOut;
                                 _ZDiffBinary=Risk/ASE;
        run;
        data out_&_varzdiff;
            set _RiskDiffOut2;
            length _Variable $32;
            _ZDiff=_ZDiffBinary;
            _Variable="&_varzdiff";
            keep _Variable _ZDiff;
         run;
         proc print data=out_&_varzdiff label noobs; title"Z-difference, data=&_datazdiff,
scale=&_scalezdiff, var=&_varzdiff, class=&_classzdiff"; run;
ods listing;
      %end;
  %mend;
********************* DESCRIPTIVE ANALYSES ****************************************;
ods select none;
* Categorical baseline variables;
  proc freq data=&_data order=formatted;
     %if &_BinaryBaselineVars ne %then %do; tables &_treatment * (&_BinaryBaselineVars)
/nopercentnorow; %end;
     %if &_OrdinalBaselineVars ne %then %do; tables &_treatment * (&_OrdinalBaselineVars)
/nopercentnorow; %end;
     %if &_NominalBaselineVars ne %then %do; tables &_treatment * (&_NominalBaselineVars)
/nopercentnorow; %end;
           format  &&_treatment dichotom.;
             %if &_BinaryBaselineVars ne %then %do;
         %do i=1 %to &_BinaryBaselineVarc;
           format  &&_BinaryBaselineVar&idichotom.;
```

```
        %end;
            %end;
          title"Description of categorical baseline variables before matching";
 run;
* Continuous baseline variables;
 %if &_ContBaselineVars ne %then %do;
    proc means data=&_data median min q1 q3 max mean maxdec=2;
       var &_ContBaselineVars;
             class &_treatment;
          format  &&_treatment dichotom.;
          title"Description of continuous baseline variables  before matching";
    run;
 %end;
* Compute z-differences;
* Define macros %ZDiffContPreMatching, %ZDiffBinaryPreMatching, and %ZDiffOrdinalPreMatching that
loop the computation of the z-differences
 across the respective continuous, binary, and ordinal baseline variables. Note that there is no z-
difference for nominal baseline variables;
 %macro ZDiffContPreMatching;
      %do i=1 %to &_ContBaselineVarc;
        %ZDiff(_datazdiff=&_data, _varzdiff=&&_ContBaselineVar&i,_classzdiff=&_treatment,
_scalezdiff=continuous);
      %end;
 %mend ZDiffContPreMatching;
 %macro ZDiffBinaryPreMatching;
      %do i=1 %to &_BinaryBaselineVarc;
        %ZDiff(_datazdiff=&_data, _varzdiff=&&_BinaryBaselineVar&i,_classzdiff=&_treatment,
_scalezdiff=binary);
      %end;
 %mend ZDiffBinaryPreMatching;
 %macro ZDiffOrdinalPreMatching;
      %do i=1 %to &_OrdinalBaselineVarc;
        %ZDiff(_datazdiff=&_data, _varzdiff=&&_OrdinalBaselineVar&i,_classzdiff=&_treatment,
_scalezdiff=ordinal);
      %end;
 %mend ZDiffOrdinalPreMatching;
* Run the macros;
%ZDiffContPreMatching;%ZDiffBinaryPreMatching;%ZDiffOrdinalPreMatching;
```

```
%if &_OrdinalBaselineVars= %then %do;
data _ZDiffOrdinalPreMatching; run;%end;
%else %do;
data _ZDiffOrdinalPreMatching; set %do i=1 %to &_OrdinalBaselineVarc; out_&&_OrdinalBaselineVar&i
%end;; run;
%end;
%if &_BinaryBaselineVars= %then %do;
data _ZDiffBinaryPreMatching; run;%end;
%else %do;
data _ZDiffBinaryPreMatching; set %do i=1 %to &_BinaryBaselineVarc; out_&&_BinaryBaselineVar&i
%end;; run;
%end;
%if &_ContBaselineVars= %then %do;
data _ZDiffContPreMatching; run;%end;
%else %do;
data _ZDiffContPreMatching; set %do i=1 %to &_ContBaselineVarc; out_&&_ContBaselineVar&i %end;;
run;
%end;
* Collect z-differences in a single data set;
 data _ZDiffPreMatching;
    set _ZDiffContPreMatching _ZDiffBinaryPreMatching _ZDiffOrdinalPreMatching;
          _ZDiffSquared=_ZDiff*_ZDiff;
 run;
 proc print data=_ZDiffPreMatching;
title"Z-differences before matching";
 run;
* Compute the number and the sum of squared z-differences ...;
 proc means data=_ZDiffPreMatching n sum;
    var _ZDiffSquared;
    output out=_tempZDiffPreMatchout n(_ZDiffSquared)=_nZDiffSquared
sum(_ZDiffSquared)=_sumZDiffSquaredBefore;
title"Squared Z-differences before matching";
 run;
* ... and write them to the macro variables _nZDiffSquared and &_sumZDiffSquareBefore;
 data _ZDiffPreMatchout;
    set _tempZDiffPreMatchout;
          call symput('_nZDiffSquared',_nZDiffSquared);
          call symput('_sumZDiffSquaredBefore',_sumZDiffSquaredBefore);
```

```
 run;
ods graphics on; ods exclude none; ods results;
*************** COMPUTE PS MODEL ************************************************;
ods select none;
 proc logistic data=&_data order=formatted;
    class &_BinaryBaselineVars&_OrdinalBaselineVars&_NominalBaselineVars / param=ref;
    model &_treatment=
&_BinaryBaselineVars&_OrdinalBaselineVars&_NominalBaselineVars&_ContBaselineVars;
          format  &&_treatment dichotom.;
    %do i=1 %to &_BinaryBaselineVarc;
          format  &&_BinaryBaselineVar&idichotom.;
    %end;
            output out=_PSOut P=_PS XBETA=_LinPred;
          title"Compute Propensity Score model";
 run;
* In the case of a matched analysis a caliper width is computed from the standard deviation of the linear
predictor;
 proc means data=_PSOut std;
    var _Linpred;
            output out=_TempStdLinpred(where=(_STAT_="STD") rename=(_LinPred=_StdLinPred));
 run;
 data _StdLinpred;
    set _TempStdLinpred(keep=_StdLinPred);
          * Write the standard deviation of the linear predictor to the macro variable &_StdLinPred;
           call symput('_StdLinPred',_StdLinPred);
 run;
 %put &_StdLinPred;
ods graphics on;ods exclude none; ods results;
*************** CALIPER-MAKRO ***************************************************;
 %macro caliper(_tenthousandthSTDcaliper=);
      data caliperwidth;
         _caliperwidth=(&_tenthousandthSTDcaliper*&_StdLinPred)/10000;
                call symput('_caliperwidth',_caliperwidth);
run;
      %put&_caliperwidth;
           * Im Verlauf des Caliper-Makros wird der Datensatz sortiert, das muss hier wieder
r�ckg�ngig gemacht werden,
                    sonst wird das Matchinggest�rt;
```

```
                    proc sort data=_psout;by&_PatID;run;
********** MATCHING **********************************************************;
    * The matching is performed by using a restricted version of the %PSMatching-macro of Coca-
Perraillon.
     Following recommendations by Austin, 2014, (see below) we set the following defaults:
     Method     = Caliper, and Replacement = No,
     In conclusion, we would recommend that, in most situations, nearest neighbor caliper matching
without
     replacement (random order or closest distance) be used when forming pairs of treated and
untreated
     subjects with similar values of the propensity score.
     Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014
Mar 15,33(6):1057-69.;
    * Prepare the data set from the PS model;
    data _PSOut_&_treatment0 _PSOut_&_treatment1;
       set _PSOut;
       if &_treatment=0 then output _PSOut_&_treatment0;
         if &_treatment=1 then output _PSOut_&_treatment1;
    run;
****** "Demacrofied" and shortened version of the Coca-Perraillon-Code **********************;
    * Create copies of the treated units if N > 1 */;
    data _PSOut_Match_&_treatment1(rename=(&_PatID=_PatID1 _LinPred=_LinPred1) drop=_i);
       * Initialize the random generator by a specified value for &_seed;
            call streaminit(&_seed);
        set _PSOut_&_treatment1;
            do _i=1 to &_matchnumberofcontrols;
          call ranuni(seed,_RandomNumber);
        output;
        end;
    run;
    proc sort data=_PSOut_Match_&_treatment1 out=_Treatment1(drop=_RandomNumber);by
_RandomNumber;run;
    data _PSOut_Match_&_treatment0(rename=(&_PatID=_PatID0 _LinPred=_LinPred0));
       set _PSOut_&_treatment0;
            call ranuni(seed,_RandomNumber);
    run;
    proc sort data=_PSOut_Match_&_treatment0 out=_Treatment0(drop=_RandomNumber); by
_RandomNumber; run;
```

```
      * Randomly sort both datasets;
       data _output(keep = _treatment0_SelectedID _treatment1_MatchedID);
         * Load Control dataset into the hash object;
          if _N_= 1 then do;
            declare hash _h(dataset: "_PSOut_Match_&_treatment0", ordered: 'no'); declare
hiteriter('_h');
            _h.defineKey("_PatID0"); _h.defineData("_LinPred0", "_PatID0");  _h.defineDone();
            call missing(_PatID0, _LinPred0);
          end;
       * Open the treatment;
        set _Treatment1;
      retain _BestDistance 99;
       * Iterate over the hash;
        _rc= iter.first();
       if (_rc=0) then _BestDistance= 99;
         do while (_rc = 0);
           if (_LinPred1 - &_caliperwidth) <= _LinPred0 <= (_LinPred1 + &_caliperwidth) then do;
             _ScoreDistance = abs(_LinPred1 - _LinPred0);
             if _ScoreDistance< _BestDistance then do;
               _BestDistance = _ScoreDistance;
                                              _treatment0_SelectedID = _PatID0;
               _treatment1_MatchedID = _PatID1;
             end;
           end;
          _rc = iter.next();
         * Output the best control and remove it;
          if (_rc ~= 0) and _BestDistance ~=99 then do;
            output;
            _rc1 = _h.remove(key: _treatment0_SelectedID);
          end;
        end;
        run;
        proc sort data=_output;by _treatment1_MatchedID;run;
        proc sql; create table _matched_treatment1 as select distinct _treatment1_MatchedID  as&_PatID
from _output; quit;
        proc sql; create table _matched_treatment0   as select distinct _treatment0_SelectedID as &_PatID
from _output;  quit;
        data _matched_distinct;
```

```
        set _matched_treatment0 _matched_treatment1;
        by &_PatID;
    run;
    proc sort data=_psout;by&_PatID;run;
    data _tempmatched;
        merge _psout _matched_distinct (in=_in_matched_distinct);
        by &_PatID;
        if _in_matched_distinct then output;
    run;
    data _outmatched;
      set _output;
    run;
   * Define Matching-Strata;
    data _tempoutmatched  (rename=(_treatment0_SelectedID =_Matched_&_treatment0
_treatment1_MatchedID = _Matched_&_treatment1)
                    keep=_treatment0_SelectedID  _treatment1_MatchedID  _MatchingStratum);
        set _outmatched;
                retain _MatchingStratum 0;
                if _treatment1_MatchedID = LAG(_treatment1_MatchedID) then _MatchingStratum
=_MatchingStratum;
                else if _treatment1_MatchedID ^= LAG(_treatment1_MatchedID) then _MatchingStratum
= _MatchingStratum+1;
    run;
    data _outmatched_&_treatment0 ;
        set _tempoutmatched (rename=(_Matched_&_treatment0 = &_patid)
keep=_Matched_&_treatment0 _MatchingStratum);
    run;
    data _outmatched_&_treatment1 (rename=(_Matched_&_treatment1 = &_patid)
keep=_Matched_&_treatment1  _MatchingStratum);
set  _tempoutmatched;
    run;
    data _matchingstrata;
        set _outmatched_&_treatment1 _outmatched_&_treatment0;
    run;
   * Merging the data set of matched observations with the matching strata;
    proc sort data=_matchingstrata; by &_patid; run;
    proc sort data=_tempmatched; by &_patid; run;
    data _matched;
```

```
            merge _matchingstrata _tempmatched;
            by &_patid;
                    _int=1;
        run;
        proc sort data=_matched nodup; by _MatchingStratum; run;
       * Count the number of observations in the matched data set;
ods select none;
        proc means data=_matched n;
            var _int;
                    output out=_tempnmatchedout n(_int)=_nmatchedobs;
        run;
ods graphics on;ods exclude none; ods results;
        data _nmatchedout;
            set _tempnmatchedout;
                    call symput('_nmatchedobs',_nmatchedobs);
        run;
********** DESCRIPTIVE ANALYSES AFTER MATCHING **********;
       * Categorical baseline variables;
ods select none;
        proc freq data=_matched order=formatted;
            %if &_BinaryBaselineVars ne %then %do; tables &_treatment * (&_BinaryBaselineVars)
/nopercentnorow; %end;
            %if &_OrdinalBaselineVars ne %then %do; tables &_treatment * (&_OrdinalBaselineVars)
/nopercentnorow; %end;
            %if &_NominalBaselineVars ne %then %do; tables &_treatment * (&_NominalBaselineVars)
/nopercentnorow; %end;
             format  &&_treatment dichotom.;
                    %if &_BinaryBaselineVars ne %then %do;
              %do i=1 %to &_BinaryBaselineVarc;format&&_BinaryBaselineVar&idichotom.;%end;
                    %end;
             title"Description of categorical baseline variables after matching, &_tenthousandthSTDcaliper";
        run;
       * Continuous baseline variables;
        %if &_ContBaselineVars ne %then %do;
           proc means data=_matched median min q1 q3 max mean maxdec=2;
              var &_ContBaselineVars;
                      class &_treatment;
             format  &&_treatment dichotom.;
```

```
            title"Description of continuous baseline variables after matching, &_tenthousandthSTDcaliper";
          run;
       %end;
     * Compute z-differences;
     * Define macros %ZDiffContPostMatching, %ZDiffBinaryPostMatching, and
%ZDiffOrdinalPostMatching that loop the computation of the
       z-differences across the respective continuous, binary, and ordinal baseline variables.
       Note that there is no z-difference for nominal baseline variables;
       %macro ZDiffContPostMatching;
            %do i=1 %to &_ContBaselineVarc;
              %ZDiff(_datazdiff=_matched, _varzdiff=&&_ContBaselineVar&i,_classzdiff=&_treatment,
_scalezdiff=continuous);
            %end;
       %mend ZDiffContPostMatching;
       %macro ZDiffBinaryPostMatching;
            %do i=1 %to &_BinaryBaselineVarc;
              %ZDiff(_datazdiff=_matched, _varzdiff=&&_BinaryBaselineVar&i,_classzdiff=&_treatment,
_scalezdiff=binary);
            %end;
       %mend ZDiffBinaryPostMatching;
       %macro ZDiffOrdinalPostMatching;
            %do i=1 %to &_OrdinalBaselineVarc;
              %ZDiff(_datazdiff=_matched, _varzdiff=&&_OrdinalBaselineVar&i,_classzdiff=&_treatment,
_scalezdiff=ordinal);
            %end;
       %mend ZDiffOrdinalPostMatching;

%ZDiffContPostMatching;%ZDiffBinaryPostMatching;%ZDiffOrdinalPostMatching;
%if &_OrdinalBaselineVars= %then %do;
data _ZDiffOrdinalPostMatching; run;%end;
%else %do;
data _ZDiffOrdinalPostMatching; set %do i=1 %to &_OrdinalBaselineVarc;
out_&&_OrdinalBaselineVar&i %end;; run;
%end;
%if &_BinaryBaselineVars= %then %do;
data _ZDiffBinaryPostMatching; run;%end;
%else %do;
```

```
data _ZDiffBinaryPostMatching; set %do i=1 %to &_BinaryBaselineVarc; out_&&_BinaryBaselineVar&i
%end;; run;
%end;
%if &_ContBaselineVars= %then %do;
data _ZDiffContPostMatching; run;%end;
%else %do;
data _ZDiffContPostMatching; set %do i=1 %to &_ContBaselineVarc; out_&&_ContBaselineVar&i %end;;
run;
%end;
     * Collect z-differences in a single data set;
      data _ZDiffPostMatching;
         set _ZDiffContPostMatching _ZDiffBinaryPostMatching _ZDiffOrdinalPostMatching;
                _ZDiffSquared=_ZDiff*_ZDiff;
      run;
      proc print data=_ZDiffPostMatching;
title"Z-differences after matching, &_tenthousandthSTDcaliper";
      run;
     * Compute the number and the sum of squared z-differences ...;
      proc means data=_ZDiffPostMatching n sum;
         var _ZDiffSquared;
                output out=_tempZDiffPostMatchout n(_ZDiffSquared)=_nZDiffSquaredafter
sum(_ZDiffSquared)=_sumZDiffSquaredafter;
title"Squared Z-differences after matching, &_tenthousandthSTDcaliper";
      run;
     * ... and write them to the macro variables _nZDiffSquaredafter and &_sumZDiffSquaredafter;
      data _ZDiffPostMatchout;
         set _tempZDiffPostMatchout;
                call symput('_sumZDiffSquaredafter',_sumZDiffSquaredafter);
      run;
ods graphics on;ods exclude none; ods results;
 ********** ANALYSIS OF MATCHED OUTCOMES **********;
     * Following recommendations in Austin, 2008 (Austin PC. A critical appraisal of propensity-score
matching in the medical
      literature between 1996 and 2003. Stat Med. 2008 May 30,27(12):2037-49.) the analysis for
assessing the treatment effect
      accounts for the matched design in the matched sample. To be concrete,
      for binary outcomes          a stratified logistic regression (with the matching stratum being the
strata variable), PROC LOGISTIC
```

```
for continuous outcomes        a mixed model with a random intercept for the matching stratum,
PROC MIXED
for ordinal outcomes           a mixed proportional odds model with a random intercept for the
matching stratum, PROC GLIMMIX
for survival outcomes          a stratified proportional hazard model (with the matching stratum being
the strata variable), PROC PHREG
is fitted;
%macro condlogistic(_BinOutcome);
    ods select none;
    proc freq data=_matched order=formatted;
        tables &_treatment * &_BinOutcome / nopercentnorow;
                format &_BinOutcome&&_treatment dichotom.;
        title"Description, Matched sample, Outcome=&_Binoutcome";
        run;
                                    proc logistic data=_matched order=formatted;
        class &_BinOutcome&_treatment _MatchingStratum / param=ref;
        model &_BinOutcome=&_treatment / link=logit clparm=wald MAXITER=1000;
                strata _Matchingstratum;
                        ods output CLparmWald=_temp&_BinOutcome;
                format &_BinOutcome&&_treatment dichotom.;
                        title"Stratified logistic regression, Matched sample,
Outcome=&_Binoutcome";
        run;
                    data &_BinOutcome&_tenthousandthSTDcaliper(drop=Parameter ClassVal0);
                        set _temp&_BinOutcome;
                                Outcome="&_BinOutcome";
                                        Outcometype="Binary";
                            Daten="&_data";
                            Caliperwidth=&_tenthousandthSTDcaliper;
                                        Nmatchedobs=&_nmatchedobs;
nZDiffSquared=&_nZDiffSquared;
        sumZDiffSquaredBefore=&_sumZDiffSquaredBefore;
sumZDiffSquaredAfter=&_sumZDiffSquaredAfter;
                    run;
ods select all;
                        /*proc print data=&_BinOutcome&_tenthousandthSTDcalipernoobs;run;*/
    %mend;
    %if &_Binary=1 %then %do; %condlogistic(&_Outcome); %end;
```

```
%macro condmixed(_ContOutcome);
proc means data=_matched order=formatted;
        var &_ContOutcome;
        class &_treatment;
        output out=means&_ContOutcome;
                format &&_treatment dichotom.;
      title"Description, Matched sample, Outcome=&_ContOutcome";
 run;
      data RCTSTD1_&_ContOutcome; set means&_ContOutcome;
      STD1=&_ContOutcome; Outcome="&_ContOutcome"; Nummer = _N_; RUN;
      data RCTSTD0_&_ContOutcome; set means&_ContOutcome;
      STD0=&_ContOutcome; Outcome="&_ContOutcome"; Nummer = _N_; RUN;
      PROC SQL;
      DELETE FROM RCTSTD1_&_ContOutcome WHERE Nummer NE 10;
      DELETE FROM RCTSTD0_&_ContOutcome WHERE Nummer NE 15;
      RUN; QUIT;
      DATA RCTSTD_&_ContOutcome; MERGE RCTSTD1_&_ContOutcome
RCTSTD0_&_ContOutcome; BY Outcome;
      DROP Nummer _TYPE_ _LEVEL_ _STAT_ _FREQ_ RZGROUP &_ContOutcome; RUN;
      proc mixed data=_matched order=formatted PLOTS(MAXPOINTS=NONE);
        class &_treatment;
        model &_ContOutcome=&_treatment / s cl;
                        ods output SolutionF=_temp&_ContOutcome;
                            format &&_treatment dichotom.;
title"Conditional model for the continuous outcome &_ContOutcome, Matched sample";
      run;

         data pre&_ContOutcome&_tenthousandthSTDcaliper(drop=StdErr Effect DF tValueProbt
Alpha Lower Upper)      ;
                        set _temp&_ContOutcome;
        where Effect="&_treatment";
        where &_treatment=1;
                            Outcome="&_ContOutcome";
                                    Outcometype="Continuous";
                        Daten="&_data";
                        Caliperwidth=&_tenthousandthSTDcaliper;
nZDiffSquared=&_nZDiffSquared;
sumZDiffSquaredBefore=&_sumZDiffSquaredBefore;
```

```
sumZDiffSquaredAfter=&_sumZDiffSquaredAfter;
Nmatchedobs=&_nmatchedobs;
                                                LowerCL=Lower;
UpperCL=Upper;
                        run;
                        data &_ContOutcome&_tenthousandthSTDcaliper; MERGE
pre&_ContOutcome&_tenthousandthSTDcaliper RCTSTD_&_ContOutcome; BY Outcome;
                        run;
                        PROC SQL;
                        UPDATE &_ContOutcome&_tenthousandthSTDcaliper
                        SET Estimate=Estimate/Sqrt((STD1*STD1+STD0*STD0)/2);
                        UPDATE &_ContOutcome&_tenthousandthSTDcaliper
                        SET UpperCL=UpperCL/Sqrt((STD1*STD1+STD0*STD0)/2);
                        UPDATE &_ContOutcome&_tenthousandthSTDcaliper
                        SET LowerCL=LowerCL/Sqrt((STD1*STD1+STD0*STD0)/2);
                        run; QUIT;
ods select all;
                       /*proc print data=&_ContOutcome&_tenthousandthSTDcalipernoobs;run;*/
    %mend;
    %if &_Continuous=1 %then %do; %condmixed(&_Outcome); %end;
    %macro condordinal(_OrdinalOutcome);
                ods select none;
        proc freq data=_matched order=formatted;
            tables &_treatment * &_OrdinalOutcome / nopercentnorow;
                format &&_treatment dichotom.;
        title"Description, Matched sample, Outcome=&_OrdinalOutcome";
        run;
        proc glimmix data=_matched order=formatted MAXOPT=500;
            class &_OrdinalOutcome&_treatment _MatchingStratum;
            model &_OrdinalOutcome=&_treatment / link=clogit s cl;
                random intercept / subject=_Matchingstratum;
                    ods output ParameterEstimates=_temp&_OrdinalOutcome;
                format &&_treatment dichotom.;
                    title"Conditional model for the ordinal outcome &_OrdinalOutcome, Matched
sample";
        run;
                                data &_OrdinalOutcome&_tenthousandthSTDcaliper(drop= Lower
Upper);
```

```
                        set _temp&_OrdinalOutcome;
            where Effect="&_treatment";
            where &_treatment=1;
                                Outcome="&_OrdinalOutcome";
                                        Outcometype="Ordinal";
                        Daten="&_data";
                        Caliperwidth=&_tenthousandthSTDcaliper;
                                        Nmatchedobs=&_nmatchedobs;
nZDiffSquared=&_nZDiffSquared;
        sumZDiffSquaredBefore=&_sumZDiffSquaredBefore;
sumZDiffSquaredAfter=&_sumZDiffSquaredAfter;
                                        LowerCL=Lower;
UpperCL=Upper;
                    run;
ods select all;
                    /*proc print data=&_OrdinalOutcome&_tenthousandthSTDcalipernoobs;run;*/
    %mend;
    %if &_Ordinal=1 %then %do; %condordinal(&_Outcome); %end;
    %macro condsurvival(_SsurvivalOutcome, _SurvivalCensvar, _CensValue);
                ods select none;
        proc lifetest data=_matched;
                    time &_SurvivalOutcome*&&_SurvivalCensVar(&_CensValue);
                strata &_treatment;
                format &&_treatment dichotom.;
        title"Description, Matched sample, Outcome=&_SurvivalOutcome";
        run;
        proc phreg data=_matched;
            class &_treatment _MatchingStratum;
            model &_SurvivalOutcome*&_SurvivalCensVar(&_CensValue)=&_treatment /
risklimits=wald MAXITER=1000;
                strata _Matchingstratum;
                        ods output ParameterEstimates=_temp&_SurvivalOutcome;
                format &&_treatment dichotom.;
                        title"Stratified Proportional hazard model, Matched sample,
Outcome=&_Survivaloutcome";
        run;
        proc print data=_temp&_SurvivalOutcome;run;
```

```
                    data &_SurvivalOutcome&_tenthousandthSTDcaliper(drop=StdErr Parameter
ClassVal0 DF ChiSqProbChiSqHazardRatioHRLowerCLHRUpperCL Label);
                        set _temp&_SurvivalOutcome;
            Outcome="&_SurvivalOutcome";
                                    Outcometype="Survival";
                        Daten="&_data";
                        Caliperwidth=&_tenthousandthSTDcaliper;
                                    Nmatchedobs=&_nmatchedobs;
nZDiffSquared=&_nZDiffSquared;
          sumZDiffSquaredBefore=&_sumZDiffSquaredBefore;
sumZDiffSquaredAfter=&_sumZDiffSquaredAfter;
                                    LowerCL=Estimate - probit(0.975)*StdErr;
UpperCL=Estimate + probit(0.975)*StdErr;
                        run;
ods select all;
         /* proc print data=&_SurvivalOutcome&_tenthousandthSTDcalipernoobs;run;*/
      %mend;
      %if &_Survival=1 %then %do; %condsurvival(&_Outcome,&_SurvivalCensvar, &_CensValue); %end;
  %mend caliper;
  %macro runcaliper;
       %do j=1 %to &_cwidthc;
         %caliper(_tenthousandthSTDcaliper=&&_cwidth&j);
                        %put &&_cwidth&j;
       %end;
  %mend runcaliper;
  %runcaliper;


********************* COLLECTING RESULTS ******************************************;
LIBNAME Finals "/folders/myfolders/Finale Ergebnisse";
                    data &_data&_Outcome;
                        set %do i=1 %to &_cwidthc; &_Outcome&&_cwidth&i %end;;
              run;
                     proc print data=&_data&_Outcome noobs;
                    title"&_data&_Outcome, Finale Outcome-Datei";
                    run;
 /*    * Aufr□umen: Alle Dateien au□er der zentralen &_data&_Outcome-Datei und den eingelesenen
Daten-Dateien
```

```
                    werden gel�scht;
                procdatasetsmemtype=data;
                save &_data&_Outcome;
                    save accordbpfinal ACCORDGLYFINAL accolipfinalaffirmfinal ALLHATCAFINAL
ALLHATCLFINAL ALLHATLLTFINAL AMISFINAL ATNFINAL BESTFINAL BHATFINAL
            CASTFINAL CPPTFINAL DIGFINAL DCCTFINAL DPPPLFINAL DPPPMFINAL ENRICHDFINAL
FAVORITFINAL HALTCFINAL HEMOKTVFINAL HEMOFLUXFINAL
            ISTASPFINAL ISTHEPFINAL ISTHEPDOSFINAL MAGICFINAL MRFITFINAL MTOPSPDFINAL
MTOPSPFFINAL MTOPSPCFINAL OATFINAL PEACEFINAL ROCHSFINAL
            ROCTBIFINAL SHEPFINAL SOLVDPREFINAL SOLVDINTFINAL TIMIFINAL TIMIBFINAL;
run;quit;*/
 data      Finals.&_data&_Outcome;
         set &_data&_Outcome;
        run;
%mend PSinRCT;
******** Einlesen der Daten ************************************************************;
libname ACCORD "/folders/myfolders/ACCORD BP";
  data accordbpfinal;
     set ACCORD.accordbpfinal;
libname AFFIRM "/folders/myfolders/AFFIRM";
  data affirmfinal;
      set AFFIRM.affirmfinal;
libname  ALLHAT "/folders/myfolders/ALLHATALFINAL ";
  data ALLHATALFINAL;
      set ALLHAT.ALLHATALFINAL;
LIBNAME ATN "/folders/myfolders/ATN";
  data ATNFINAL;
      set ATN.ATNFINAL;
 *options MLOGIC MPRINT SYMBOLGEN;
  options NOMLOGIC NOMPRINT NOSYMBOLGEN;
 %PSinRCT(_data=accordbpfinal,_patid=MaskID,_treatment=arm,_treatment0=0,_treatment1=1,
         _BinaryBaselineVars=female cvd_hx_baselinehartfail,
         _OrdinalBaselineVars=edu,_NominalBaselineVars=cigarett RACECLASS,
         _ContBaselineVars=baseline_agesbpdbpyrsdiabwt_kg
ht_cm BMI b1antihp hba1c chol trig ldlhdlfpg potassium
screatgfruacr,
         _matchnumberofcontrols=1,_seed=59063,
         _Outcome=fuyrs_tm,_Survival=1,_SurvivalCensVar=censor_tm,_CensValue=0);
```

```
%PSinRCT(_data=affirmfinal,_patid=newID,_treatment=ArmMain,_treatment0=0,_treatment1=1,
_BinaryBaselineVars=CHF01 Fail01 First01 Gender Minority,
        _OrdinalBaselineVars=Durat01,_NominalBaselineVars=Cause01,
        _ContBaselineVars=Age,
_matchnumberofcontrols=1,_seed=59065,
_Outcome=Hosp05,_Binary=1);
%PSinRCT(_data=ALLHATALFINAL,_patid=STUDYID,_treatment=RZGROUP,_treatment0=0,_treatment1=
1,
_BinaryBaselineVars=SEX MISTROKE HXCABG STDEPR OASCVD
DIABETES HDLLT35 LVHECG WALL25 LCHD LLT,
    _OrdinalBaselineVars=BLMEDS,_NominalBaselineVars=CURSMOKE
        ASPIRIN ESTROGEN ETHNIC,
        _ContBaselineVars=AGE BLBMI BV2SBP BV2DBP EDUCAT,
_matchnumberofcontrols=1,_seed=5432,
_Outcome=SBP6M12,_Continuous=1);
%PSinRCT(_data=ATNFINAL,_patid=PatientKey,_treatment=TreatmentGp,_treatment0=0,_treatment1=
1,
_BinaryBaselineVars=olig Gender MechVentilation Ischemic
Nephrotoxic Sepsis Multifactorial HemoCRRTsepsisbase,
_OrdinalBaselineVars=,_NominalBaselineVars=PrimaryTreat
Race,
_ContBaselineVars=Form01Premorbidcreat Age Weight BUNPR
daysICUbranddayshospbrandbCoagulationbCardiovascular
bCentralNervebTotApacheII,
_matchnumberofcontrols=1,_seed=59069,
_Outcome=rrfstage,_Ordinal=1
);
```

Danksagung

Als Erstes danke ich Jesus Christus, meinem Herrn und Erlöser, dass er mir die nötige Kraft gab und mich die ganze Zeit begleitet.

Als Zweites danke ich meinen Eltern und meinen Geschwistern, welche mich in allen Phasen der Dissertation unterstützten. Danke, dass es euch gibt.

Als Drittes danke ich Prof. Kuß für die sehr gute Betreuung. Einen besseren Doktorvater hätte ich mir nicht erträumen können.

Als Viertes danke ich allen weiteren Verwandten, Freunden und Bekannten, welche mich in Wort, Tat und Gebet unterstützten.