Vergleich von Propensity Score Matching und Propensity Score Adjustierung in primärdatenbasierten Untersuchungen

Natalie Lamp
natalie.lamp@uni-ulm.de

Annabel Müller-Stierlin
annabel.mueller-stierlin@uni-ulm.de

Reinhold Kilian
reinhold.kilian@uni-ulm.de

Verena Schöning
verena.schoening@campus.lmu.de

Klinik für Psychiatrie und Psychotherapie II, Universität Ulm
Ludwig-Heilmeyer-Str. 2, 89312 Günzburg

Zusammenfassung

Propensity Score (PS) Methoden sind eine Möglichkeit, den Selektionsbias bei quasi-experimentellen Studiendesigns zu kontrollieren. Es wurde gezeigt, dass das Propensity Score-Matching (PSM) der Propensity Score-Adjustierung (PSA) hinsichtlich der Effektschätzung überlegen ist. Allerdings wurden bisherige Studien überwiegend auf der Basis von Sekundärdaten mit großen Fallzahlen durchgeführt. Da in Primärdatenuntersuchungen in der Regel wegen erheblich geringerer Fallzahlen die Zahl potenzieller Matchingpartner begrenzt ist, bleibt unklar, ob die Überlegenheit des PSM auf primärdatenbasierte Studien übertragbar ist. Dieser Beitrag vergleicht anhand simulierter Daten das PSM und die PSA. Die Simulation soll den Ablauf in der Versorgungsforschung mit Primärdaten darstellen, im Sinne einer quasi-experimentellen Studie mit wenigen Hundert Studienteilnehmern, die auf zwei gleichgroße Behandlungsgruppen verteilt sind. Für das Matching wird ein Algorithmus von Lanehart et al. verwendet. Vor und nach der Anwendung des PSM wird die Balance unter Verwendung des Makros %STDDIFF bzw. nach der Adjustierung mit dem Makro %GEN3 überprüft. In dieser Arbeit konnte gezeigt werden, dass in allen simulierten Szenarien PSA die exakteren Effektschätzer liefert. Die Verzerrung ist nach PSA immer geringer als nach dem Matching-Verfahren. Das PSM wird zwar in der Literatur häufig empfohlen, konnte jedoch nicht bei diesen Fallzahlen und gleich großen Behandlungsgruppen überzeugen.

Schlüsselwörter: Simulationsstudie, Propensity Score, Matching, Adjustierung

1 Einleitung

Die Randomisierung der Untersuchungsteilnehmer gilt als Goldstandard für die verzerrungsfreie Schätzung von Behandlungseffekten. Im Bereich der Versorgungsforschung ist eine randomisierte Zuweisung von Untersuchungsteilnehmern jedoch häufig nicht möglich [1]. Da bei einer nichtrandomisierten Zuweisung von Studienteilnehmern die Gefahr eines Selektionsbias und damit der Verzerrung der Untersuchungsergebnisse besteht [2], müssen statistische Verfahren zur Biaskontrolle angewendet werden.
Grundlage der Propensity Score (PS) Methode ist die Schätzung der, mit der Effektmessung konfundierte Drittvariablen, bedingten Wahrscheinlichkeit jedes Studienteilneh-

mers, in eine der Untersuchungsgruppen zu gelangen. Werden bei dieser Schätzung alle relevanten Einflussfaktoren der Stichprobenselektion berücksichtigt, so kann der Einfluss dieser Variablen auf die Effektschätzung durch verschiedene Verfahren neutralisiert werden [2]. Die am häufigsten verwendeten Verfahren sind dabei das Propensity Score-Matching (PSM) und die Propensity Score-Adjustierung (PSA). Beim PSM werden dabei jedem Teilnehmer der Interventionsgruppe ein oder mehrere Teilnehmer der Kontrollgruppe mit gleichen oder ähnlichen PS zugeordnet, so dass bei der Effektschätzung Untersuchungsgruppen mit gleichen Merkmalsverteilungen z.B. mittels t-Test verglichen werden können. Bei der PSA erfolgt eine Konstanthaltung des PS im Rahmen eines multivariaten Analyseverfahrens, z.B. einer multiplen Regressionsanalyse [3].

Yang et al. berichten in ihrem Review über zahlreiche Studien, die belegen, dass die Matching-Methode genauere Effektschätzer liefert als die Stratifizierung, die Gewichtung oder die Adjustierung mit dem Propensity Score [4]. Doch ein direkter Vergleich zwischen dem Matching-Verfahren und der Adjustierung bei wenigen Beobachtungen wurde bisher an simulierten Daten nicht durchgeführt. Bisherige Studien wurden auf Basis von Sekundärdaten mit großen Fallzahlen realisiert. Die große Fallzahl in Sekundärdatenanalysen bietet eine ideale Voraussetzung für die Anwendung des PSM, weil die Zahl potenzieller Matchingpartner ebenfalls hoch ist. Im Gegensatz zu Sekundärdatenanalysen ist in quasi-experimentellen Primärdatenuntersuchungen wegen der deutlich kleineren Fallzahlen auch die Zahl potenzieller Matchingpartner deutlich kleiner. Bislang ist unklar, ob unter diesen Bedingungen die Überlegenheit des PSM gegenüber einer PSA bestehen bleibt.

In dem vorliegenden Beitrag sollen beide Methoden der PS-basierten Biaskontrolle im Hinblick auf eine verzerrungsfreie Effektschätzung zur Verwendung in Primärdatenstudien verglichen werden. Für das PSM wird das 1:1 - Matching ohne Zurücklegen verwendet, basierend auf dem nearest-neighbour Matching mit festem Caliper [5].

Der Treatment-Effekt wird über den Regressionskoeffizienten der linearen Regression geschätzt. Die Schätzung des Bias erfolgt über den mean squared error (MSE). Im Anschluss wird die Balance der interessierenden Variablen nach der Anwendung des PS überprüfen. Zu diesem Zweck wird die standardisierte Differenz verwendet. Die Analyse baut auf zwei Abschlussarbeiten zur PS-Methodik auf [6,7].

2 Methoden

2.1 Datensimulation

Es werden fünf unterschiedliche Szenarien (A-E) mit je 3000 Datensätzen, jeweils mit 500 Beobachtungen, mit der Statistik-Software SAS simuliert. Jedes dieser Szenarien enthält eine dichotome Variable „Treatment“ = t (t = 0 für Kontrollgruppe, t = 1 für Interventionsgruppe). Eine standardnormalverteilte Variable „Outcome“ = y sowie fünf Confounder werden erzeugt. Diese Confounder sind in Szenario A, B und E kontinuierlich und in Szenario C und D kategorial gewählt. Zusätzlich haben sie unterschiedliche Einflussstärken auf die Zielvariable (Tabelle 1). Zur Analyse unbekannter Confounder wird Szenario E mit zwei weiteren Störfaktoren (u_1, u_2) ergänzt. So werden diese in Szenario E1 in die Propensity Score Schätzung aufgenommen, nicht aber in Szenario E2.

Die Gruppenzuteilung (t) wird mit dem logistischen Regressionsmodel geschätzt:

$$p(t) = \frac{e^{\alpha_0+\alpha_1 x_1+\cdots+\alpha_n x_n}}{1+e^{\alpha_0+\alpha_1 x_1+\cdots+\alpha_n x_n}}$$

Damit in den beiden Behandlungsgruppen etwa gleichviele Beobachtungen enthalten sind, wird in den Szenarien A, B und E α_0= 0 definiert, bei den Szenarien mit kategorialen Variablen ist α_0= -1. Zur Erzeugung des Selektionsbias werden $\alpha_1\ bis\ \alpha_5$ in allen Szenarien auf 0,4 gesetzt (Tabelle 1).

Die Zielvariable wird mit Hilfe der linearen Regression simuliert:

$$y = \beta_0 + \beta_T t + \beta_1 x_1 + \cdots + \beta_n x_n$$

Der Einfluss der Behandlung beträgt $\beta_T = 0{,}8$ und entspricht einem starken Einfluss auf die Zielvariable. Die Kovariablen haben abhängig vom Szenarium einen starken Einfluss auf die Zielvariable und werden mit $\beta_1\ bis\ \beta_5 = 0{,}8$ ausgedrückt oder einen schwachen Einfluss. In diesem Fall beträgt $\beta_1\ bis\ \beta_5$ 0,4 (Tabelle 1) [8].

Aufgrund der 3000 Simulationen, werden alle Prozeduren als Makros ausgeführt. Am Anfang steht der Makroname `%macro sample`. Für die Durchführung wird mit dem `%DO` Befehl die Anzahl der Durchläufe, von 1 bis 3000, festgelegt. Diese Anweisung muss mit `%END` beendet werden. Am Ende jedes Makros steht erneut der festgelegte Makroname `%mend sample; %sample;`.

```
%macro sample;
%DO x=1 %TO 3000;
… … … …
%END;  %mend sample;
%sample;
```

Tabelle 1: Überblick Szenario A-E

	Szenario A	**Szenario B**	**Szenario C**	**Szenario D**	**Szenario E1 und E2**
Beobachtungen	500	500	500	500	500
Simulationen	3000	3000	3000	3000	3000
Skalenniveau	kontinuier.	kontinuier.	kategorial	kategorial	kontinuier.
Einflussstärke	stark	schwach	stark	schwach	stark
Confounder	bekannt	bekannt	bekannt	bekannt	bekannt + unbekannt
α_0	0	0	-1	-1	0
$\alpha_1 - \alpha_5$	0,4	0,4	0,4	0,4	0,4
β_0	0	0	0	0	0
β_t	0,8	0,8	0,8	0,8	0,8
$\beta_1 - \beta_5$	0,8	0,4	0,8	0,4	0,4
β_6, β_7	-	-	-	-	0,8

Das Zusammenführen der 3000 Tabellen wird über den Befehl `MERGE BY` ermöglicht.

```
DATA merging;
MERGE simulation1-simulation3000;
BY ID;
RUN;
```

2.2 Propensity Score Schätzung

Der Propensity Score wurde von Rosenbaum und Rubin als die bedingte Wahrscheinlichkeit, mit der ein Patient eine Intervention erhalten würde, unter Betrachtung ausgewählter Variablen definiert [9]. Er wird für jeden Teilnehmer berechnet. Es werden alle Variablen in das Modell aufgenommen, die mit dem Outcome (y) in Verbindung stehen [10]. In SAS kann die Schätzung mit `PROC LOGISTIC` umgesetzt werden, wobei für jedes Szenarium eine separate Modellierung und Propensity Score Schätzung erfolgt.

```
PROC LOGISTIC DESCENDING DATA=data&x;
MODEL t=a1 a2 a3 a4 a5;
OUTPUT OUT=propen&x PROB=prob;
RUN;
```

Mit der Option `PROB` im `OUTPUT` Statement wird für jede Beobachtung die bedingte Wahrscheinlichkeit für die Behandlungsgruppe = 1, unter Berücksichtigung der Kovariablen, ausgegeben. Diese Wahrscheinlichkeit entspricht dem Propensity Score.

In Szenario E1 werden zusätzlich die Variablen u1 und u2 in das Modell aufgenommen. In E2 werden diese, trotz ihres starken Einflusses, ignoriert.

2.3 Common-Support-Region

Voraussetzung für die Nutzung des Propensity Scores ist eine ausreichende Überschneidung der Verteilungen des PS der beiden Gruppen. Das wird auch als die Common-Support-Bedingung bezeichnet. Es ist wichtig, dass dieser Bereich möglichst groß ist [4]. Die Individuen, die außerhalb dieser Region liegen, haben kaum Möglichkeit einen Matchingpartner zu finden, da die Propensity Scores zu unterschiedlich sind [4].
In dieser Arbeit wird die Common-Support-Region mit der Minima-Maxima-Regel dargestellt [4]. In diesem Fall vergleicht man das Minimum und das Maximum des Propensity Scores in den beiden Gruppen miteinander. Die Common-Support-Region ist dann die Schnittmenge der beiden Intervalle. Dafür kann die PROC MEANS Prozedur verwendet werden.

2.4 Propensity Score Matching

In dieser Arbeit wird das 1:1-Matching durchgeführt, da es in der Praxis am häufigsten verwendet wird und die Stichprobe gleichverteilt auf die beiden Gruppen ist [5]. Es wird ein Ziehen ohne Zurücklegen angewendet, das bedeutet, dass Beobachtungen, die einem Partner zugeordnet werden konnten, nicht mehr für andere Beobachtungen zur Verfügung stehen. Die meist verwendete Methode für das PSM basiert auf dem nearest-neighbour matching mit festem Caliper [5].
Beim nearest-neighbour matching wird dem Interventionsteilnehmer der Partner zugewiesen, welcher die geringste Distanz des Propensity Scores aufweist. Ein vorher festgelegter Caliper definiert die maximale Distanz, die zwischen den beiden PS liegen darf [2]. Cochran und Rubin demonstrierten, dass die Verwendung des Calipers, der das 0.2-fache der Standardabweichung des Logit des PS beträgt, respektable Ergebnisse liefert [11]. Für die Umsetzung in SAS wird ein Matchingalgorithmus von Lanehart et al. verwendet [12].

2.5 Effektschätzung

Die Effektschätzung wird mit dem gemischten linearen Modell durchgeführt. Dazu wird in dieser Arbeit die `PROC MIXED` Prozedur verwendet. Es wird der Effekt ohne Adjustierung, mit multivariater Adjustierung, mit PSA und nach PSM untersucht. Zusätzlich wird der Bias und der MSE berechnet. Dazu werden die Formeln $Bias = \bar{x} - \mu$ und $MSE = var + bias^2$ verwendet.
Im ersten Modell erfolgt keine Korrektur, es dient zum besseren Vergleich nach der Adjustierungs- oder Matching-Methode.

```
PROC MIXED DATA=daten&x;
     CLASS t;
     MODEL y = t/SOLUTION;
     LSMEANS t/ PDIFF;
QUIT;
RUN;
```

Zum Vergleich erfolgt eine Auswertung mit der Adjustierung für alle Kovariablen.

```
PROC MIXED DATA=daten&x;
     CLASS t;
     MODEL y = t a1 a2 a3 a4 a5/SOLUTION;
     LSMEANS t/ PDIFF;
QUIT;
RUN;
```

Da der PS den Einfluss mehrere Variablen vereinigt, wird dieser bei der PSA als Kovariate verwendet. Der große Vorteil gegenüber der herkömmlichen Regressionsadjustierung ist, dass nur eine Kovariable verwendet wird, aber für alle Variablen adjustiert werden kann, die in den Propensity Score eingeschlossen wurden [13].

```
PROC MIXED DATA=propen&x;
     CLASS t;
     MODEL y = t prob/SOLUTION;
     LSMEANS t/ PDIFF;
QUIT;
RUN;
```

Die Effektschätzung nach PSM erfolgt mit dem Datensatz, der nur die Beobachtungen mit einem Matchingpartner enthält. Bei der Analyse muss darauf geachtet werden, dass die Unabhängigkeit der Stichproben, aufgrund der Matchingpaare nicht mehr erfüllt ist. Deshalb wird der *match* in das `RANDOM` Statement aufgenommen.

```
PROC MIXED DATA=match_final&x;
CLASS t match;
MODEL y = t /SOLUTION;
RANDOM match;
LSMEANS t/ PDIFF;
QUIT;
RUN;
```

2.6 Balance Tests

Für die Überprüfung der Balance wird die standardisierte Differenz verwendet. Sie hat den Vorteil unabhängig von der Stichprobengröße zu sein. Zusätzlich erlaubt dieses Verfahren eine globale Aussage über die Balance insgesamt, trotz der unterschiedlichen Skalenniveaus der Variablen [2]. Die Grenze für ein Ungleichgewicht wird in der Literatur oft unterschiedlich definiert. In der Regel sollte der absolute Wert nicht größer als 0,10 sein, denn das spricht für ein Ungleichgewicht zwischen den Gruppen [2].

Wird der PS für das Matching verwendet, so wird die Balance in den Gruppen vor und nach dem Matching berechnet. Die standardisierte Differenz vergleicht Mittelwertsunterschiede der Baseline Kovariaten zwischen den gematchten behandelten und nicht be-

handelten Beobachtungen [14]. In dieser Arbeit wird dafür das Makro %STDDIFF von Yang und Dalton verwendet [15].

Douglas et al. entwickelten ein Makro %GEN3 für die Berechnung der gewichteten standardisierten Differenz. Dieses wird in diesem Beitrag nach der Methode der Adjustierung für den Propensity Score verwendet [16].

3 Ergebnisse

Zur Darstellung der Ergebnisse wurden die Kenngröße über die 3000 Simulationen hinweg gemittelt. Bei den Intervallen wurden z.B. die Minima und die Maxima der Datensätze von allen Simulationen gemittelt.

3.1 Propensity Score

Die Mittelwerte der Propensity Scores sind in den Szenarien relativ ähnlich. Sie bewegen sich in breiten Intervallen in den beiden Behandlungsgruppen (Tabelle 2). In den Szenarien A und B liegt der PS in der Interventionsgruppe im Mittel bei 0,5781 im Intervall von [0,1198 ; 0,9376] und in der Kontrollgruppe bei einem Mittelwert von 0,4219 im Intervall von [0,0621 ; 0,8803]. Szenarien mit den kategorialen Variablen haben etwas schmalere Intervalle. In der Interventionsgruppe liegt dieses bei [0,2677 ; 0,7333] mit einem Mittelwert von 0,5281, während in der Kontrollgruppe das Intervall bei [0,2667 ; 0,7327] mit einem Mittelwert von 0,4717 liegt. Wird in Szenario E mit unbekannten Confoundern der PS geschätzt (E2), so liegt der PS im Intervall von [0,1518 ; 0,9010] mit einem Mittelwert von 0,5567 in der Interventionsgruppe und [0,0999 ; 0,8471] in der Kontrollgruppe mit einem Mittelwert von 0,4431. Werden die Störvariablen u_1 und u_2 in die PS Schätzung aufgenommen (E1), so unterscheidet sich der Mittelwert in den beiden Gruppen deutlicher, mit 0,6457 in der Interventionsgruppe und 0,3542 in der Kontrollgruppe.

Tabelle 2: Propensity Scores in Szenario A bis E, Mittelwerte der Kenngrößen über die 3000 Simulationen

	Intervention				**Kontrolle**			
	min	max	mean	sd	min	max	mean	sd
Szenario A oder B	0,1198	0,9376	0,5781	0,1806	0,0621	0,8803	0,4219	0,1806
Szenario C oder D	0,2677	0,7333	0,5281	0,1135	0,2667	0,7327	0,4717	0,1134
Szenario E1	0,0660	0,9868	0,6457	0,2268	0,0132	0,9348	0,3542	0,2268
Szenario E2	0,1518	0,9010	0,5567	0,1575	0,0999	0,8471	0,4431	0,1574

3.2 Common-Support-Region

Die Common-Support-Region kann aus Tabelle 2 abgelesen werden. Die PS Intervalle sind in beiden Behandlungsgruppen relativ breit. So entstehen in allen Szenarien große Common-Support-Regionen, was für das PSM notwendig ist. Für die Szenarien A und B liegt die Region im Intervall von [0,1198 ; 0,8803]. Die Szenarien mit kategorialen Variablen haben eine Common-Support-Region von 0,2677 bis 0,7327. E1 hat eine Common-Support-Region im Intervall von [0,0660 ; 0,9348], während E2 eine Region von 0,1518 bis 0,8471 hat. Die Anzahl der Beobachtungen aus den jeweiligen Gruppen in der Common-Support-Region kann aus Tabelle 3 entnommen werden.

Tabelle 3: Mittlere Anzahl der Beobachtungen in der Common-Support-Region der 3000 Simulationen in Szenario A bis E

Szenario A oder B	**min**	**max**	**mean**	**sd**
Kontrolle	201	279	240,6	12,50
Intervention	199	286	240,6	12,29
Szenario C oder D				
Kontrolle	198	276	234,1	11,46
Intervention	195	270	234,1	11,38
Szenario E1				
Kontrolle	160	277	230,1	15,58
Intervention	167	280	230,5	15,88
Szenario E2				
Kontrolle	204	282	243,0	11,90
Intervention	195	281	242,8	11,91

3.3 Matchingpaare

Trotz der ausreichenden Common-Support-Regionen können in allen Szenarien nur wenige Matchingpaare gefunden werden. Die meisten Matches konnten bei kategorialen Variablen erzielt werden. Verfügt der Datensatz über unbekannte Confounder, so hat dies keine Auswirkungen auf die Anzahl der Matchingpaare. Bei der Simulation konnten in diesem Fall im Mittel 145 Paare gebildet werden, was keinen auffälligen Unterschied zu anderen Szenarien darstellt (Tabelle 4). Die wenigsten Paare konnten im Szenario E1 gefunden werden, in welchem alle Variablen zur PS-Schätzung einbezogen wurden.

Tabelle 4: mittlere Anzahl an Matchingpaaren in den 3000 Simulationen in Szenarien A bis E

	min	**max**	**mean**	**sd**
Szenario A / B	139	187	164,6	6,81
Szenario C / D	160	200	183,9	6,83
Szenario E1	124	165	144,9	6,29
Szenario E2	149	193	172,5	6,81

3.4 Effektschätzung

3.4.1 Effektschätzung ohne Adjustierung

Das Modell ohne jegliche Adjustierung hat in allen Szenarien stark verzerrte Effektschätzer. In Szenarien mit kontinuierlichen Variablen ist der überschätzte Wert höher als bei kategorialen Variablen. Ebenso machen sich die unterschiedlichen Einflussstärken auf die Effektschätzung bemerkbar. Umso stärker der Einfluss des Confounders auf die Zielvariable, umso höher ist die Überschätzung des Effektes. Die auffälligsten Werte sind in Szenario E1, mit den unbekannten Confoundern zu sehen. Der Effektschätzer erreicht einen Wert von 2,3, mit einem ebenso hohen MSE (Tabelle 5).

3.4.2 Effektschätzung mit multivariater Adjustierung

Die Effektschätzer nach der multivariaten Adjustierung konnten sehr exakt geschätzt werden. Die unterschiedlich starken Einflüsse der Confounder bewirken keine Veränderung in den Effektschätzern. Ein geringer Unterschied ist zwischen den kontinuierlichen und kategorialen Störfaktoren zu sehen. Zu einer Verzerrung führt allerdings das Modell mit den unbekannten Confoundern. Werden nicht alle Störvariablen im Modell aufgenommen, führt das zur Überschätzung des Effektes (Tabelle 5).

3.4.3 Effektschätzung mit PS Adjustierung

Die Effektschätzer nach Propensity Score Adjustierung weichen nur minimal von den Ergebnissen nach der multivariaten Adjustierung ab. Die PSA Methode konnte ebenso genaue Effektschätzer mit einem niedrigen MSE liefern. Aber auch hier wurde der Effekt nicht genau geschätzt, wenn nicht alle Confounder in das Modell aufgenommen werden (Tabelle 5).

3.4.4 Effektschätzung mit PS Matching

Die Effektschätzer nach der PSM Methode variieren in den Szenarien. Bei kontinuierlichen Variablen ist der Schätzer zwar nicht so hoch wie in dem Roh-Modell, ist aber auch deutlich überschätzt mit einem hohen MSE Wert. Auch hier verbessert sich die Schätzung, wenn die Einflussstärke abnimmt. Das Modell mit kategorialen Variablen hat eine deutlich bessere Effektschätzung mit minimaler Verzerrung erreicht dennoch nicht die Genauigkeit der Adjustierungsmethoden. Problematisch wird es auch hier beim Ignorieren von Confoundern. Szenario E2 zeigt auch hier eine deutliche Verzerrung der Effektschätzer (Tabelle 5).

Tabelle 5: Übersicht der mittleren Effektschätzer der 3000 Simulationen in Szenario A bis E

		Roh-Modell	multivariate Adjustierung	PS Adjustierung	PS Matching
Szenario A	Schätzer	2,1670	0,8013	0,8020	1,0369
	MSE	1,898	0,009	0,009	0,072
Szenario B	Schätzer	1,4838	0,8013	0,8017	0,9198
	MSE	0,481	0,009	0,009	0,027
Szenario C	Schätzer	1,1847	0,7984	0,7983	0,8431
	MSE	0,162	0,009	0,009	0,014
Szenario D	Schätzer	0,9916	0,7984	0,7983	0,8205
	MSE	0,05	0,009	0,009	0,012
Szenario E1	Schätzer	1,4923	0,8015	0,8026	1,1491
	MSE	0,500	0,002	0,012	0,140
Szenario E2	Schätzer	2,2997	1,8300	1,8309	1,8998
	MSE	2,269	1,079	1,080	1,234

3.5 Balancetests

Tabelle 6 zeigt den Anteil der Simulationen in welchen die Variablen ausbalanciert sind, gemessen an einer standardisierten Differenz kleiner 0.1. Nach PSA und PSM waren die Variablen in ca. 75% der Simulationen ausbalanciert. In dem unadjustierten Modell kann kaum eine Balance nachgewiesen werden.

Tabelle 6: Anteil der Simulationen in welchen die Variablen ausbalanciert sind in Szenarien A bis E

	Roh-Modell	nach PS Adjustierung	nach PS Matching
Szenario A und B: x1 bis x5	0.1% bis 0.2%	76.4% bis 78.5%	75.7% bis 76.7%
Szenario C und D: x1 bis x5	13.9% bis 16.7%	81.0% bis 82.5%	92.5% bis 93.3%
Szenario E1: x1 bis x5	1.5% bis 2.2%	70.9% bis 72.6%	67,6% bis 69.0%
Szenario E2: x1 bis x5	1.5% bis 2.2%	78.7% bis 79.9%	83.6% bis 86.0%

4 Diskussion

PSM und PSA werden häufig in der Praxis verwendet und sind ein wichtiges Werkzeug bei nicht-randomisierten Studien [4]. In dieser Arbeit wurde anhand einer Simulation untersucht, welches dieser PS-Verfahren die genaueren Effektschätzer liefert. Der Unterschied zu vorherigen Untersuchungen aus der Literatur liegt darin, dass die Stichprobe viel kleiner ist. Ein zusätzlicher Unterschied ist, dass auch die Stichproben in dieser Arbeit gleich groß sind. Mit dem gemischten linearen Modell wurden die Effektschätzer nach Anwendung der jeweiligen PS-Verfahren ermittelt. MSE sowie Bias wurden zusätzlich berechnet. Hier konnte ebenso die Frage geklärt werden, welchen Einfluss nicht-messbare Confounder auf die Auswertung haben können. Im Anschluss wurde die Balance in den beiden Behandlungsgruppen untersucht.

Alle fünf Szenarien haben gezeigt, dass das Verfahren der PSA ebenso wie die multivariate Adjustierung zufriedenstellende Schätzer liefern konnte. PSA ist mit kleinen Stichproben gut umzusetzen. Die Effektschätzung der Szenarien mit kontinuierlichen Variablen unterscheidet sich nur sehr gering von den Szenarien mit kategorialen Variablen. Auch die unterschiedlichen Einflussstärken der Confounder hatten nur einen geringen Einfluss auf die Ergebnisse. Die Effektschätzung ist relativ genau mit einer geringen Verzerrung und niedrigem MSE. Enthält die Stichprobe allerdings Confounder, die bei der Analyse nicht berücksichtigt werden, weil sie beispielsweise unbekannt sind, so können die Effekte nicht mehr richtig geschätzt werden. Die Verzerrung ist in diesem Fall sehr hoch.

Im Vergleich dazu variierten die Effektschätzer des PSM von Szenario zu Szenario. Bei kontinuierlichen Variablen waren die Effektschätzer deutlich überschätzt, mit relativ hohen MSE und Bias. Bei kategorialen Variablen waren die Schätzer zwar niedriger, konnten allerdings die Genauigkeit der PSA nicht erreichen. Wie in der Literatur bisher beschrieben, kann die Simulation zwar zeigen, dass das PSM die Verzerrung verringert und die Effektschätzung verbessert [5], die Ergebnisse sind dennoch nicht zufriedenstellend. Szenarien mit kategorialen Variablen liefern genauere Schätzer als die Szenarien mit kontinuierlichen Variablen, ebenso ist auch der Unterschied in den verschiedenen Einflussstärken zu erkennen. Bei einem schwächeren Einfluss der Confounder kann der Effektschätzer genauer berechnet werden. Auch bei dieser Methode haben die unbekannten Confounder zu deutlichen Verzerrungen geführt.

In allen Szenarien werden nur wenige Paare gebildet, obwohl die Common-Support-Region beinahe über das gesamte Intervall verläuft. Aufgrund der geringen Fallzahl können nur wenige Matchingpaare gefunden werden und dies führt zu den ungenaueren Effektschätzungen des Matching-Verfahrens. Stärkere Einflüsse der Confounder verursachen größere Verzerrungen der Effektschätzer und einen größeren Bias. Modelle mit kontinuierlichen Variablen sind für die Verzerrungen anfälliger als jene mit kategorialen Variablen. Bei der Adjustierung dagegen war die Analyse nicht so empfindlich.
Die in der Literatur beschriebene standardisierte Differenz konnte in dieser Arbeit für die Beurteilung der Balance verwendet werden [2]. Sowohl die PSA als auch das PSM haben die Balance in den beiden Behandlungsgruppen herstellen können, trotz der geringen Fallzahl.

Diese Arbeit hat gezeigt, dass bei Verwendung des PSM die Effektschätzer vermutlich überschätzt werden. Die alternative PSA ist eher zu empfehlen. Dieses Verfahren hat in jeder der simulierten Situationen überzeugen können. Es wurde vermutet, dass die Adjustierung aufgrund der Stichprobengröße tatsächlich genauer ist als das PSM [17]. Diese Vermutung konnte mit dieser Simulation bestätigt werden. Tauchen allerdings in der Stichprobe Confounder auf, die in die PS-Schätzung nicht eingeschlossen werden, so kann man sich in beiden Verfahren nicht mehr auf die Ergebnisse verlassen.

Möglicherweise hätte das PSM bessere Effektschätzer erzeugen können, unter Verwendung anderer Matching-Verfahren, wie beispielsweise dem Matching mit Zurücklegen oder optimal matching [18]. Auf diese Weise können zwar nicht alle Beobachtungen für die Analyse verwendet werden, dafür aber mehr Matchingpartner bzw. ähnlichere Paare gebildet werden.

In der Literatur wird empfohlen, das PSM in nicht-randomisierten Studiendesigns zu verwenden [11]. Diese Arbeit konnte zeigen, dass diese Empfehlung bei kleinen Stichproben nicht zutrifft. Die Vermutung wurde hier bestätigt, dass das Matching-Verfahren aufgrund der kleinen Stichprobe und der gleich großen Behandlungsgruppen keine exakten Effektschätzer liefern kann. Die Methode der PSA kann dagegen bei kleinen Fallzahlen mit gleichgroßen Behandlungsgruppen angewendet werden.

Die Durchführung von PS-Verfahren ist bei nicht-randomisierten Studien vorteilhaft, denn beide Verfahren reduzierten den Bias und konnten die Verzerrung der Effektschätzer minimieren. Auch konnten beide Verfahren die Balance zwischen den beiden Behandlungsgruppen herstellen. PS-Methoden haben sich in den letzten Jahren schnell weiter entwickelt, auch wurde bereits viel geforscht auf diesem Gebiet [4]. Es sind aber weitere Untersuchungen mit kleinen Stichproben notwendig, denn es gibt so viele verschiedene Möglichkeiten ein Matching durchzuführen, allerdings kaum Empfehlungen für kleine Stichproben. Standardisierte Vorgehensweisen sind dringend nötig, weshalb es vorteilhaft wäre, eine Art Leitfaden für kleine Fallzahlen zu entwickeln.

Literatur

[1] Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ. 1996:1215–8.

[2] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011;46:399–424. doi:10.1080/00273171.2011.568786.

[3] Holmes WM. Using Propensity Scores in Quasi-Experimental Designs: SAGE Publicatins; 2014.

[4] Yang G, Stemkowski S, Saunders W. A Review of Propensity Score Application in Healthcare Outcome and Epidemiology. Denver: PharmaSug Konferenz, Premier Inc. Working Paper PR02; 2007.

[5] Austin PC. A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality. Multivariate Behav Res. 2011;46:119–51. doi:10.1080/00273171.2011.540480.

[6] Lamp N, Effektschätzung mit Hilfe des Propensity Scores: Vergleich von Propensity Score Matching und Propensity Score Adjustierung bei kleinen Stichproben in

einer Simulationsstudie. [Masterthesis]. Heidelberg: Ruprecht-Karls-Univesität; März 2017.

[7] Schöning VM. Propensity Score-Methoden zur Kontrolle des Selektionsbias bei nicht-randomisierten Studien. [Bachelorthesis]. Ulm: Hochschule Ulm, Juli 2015.

[8] Stierlin AS. Assessment of stability and validity of logistic regression models using bootstrap resampling in simulated data as well as in real data: To what extent does the approach of missing data imputation and variable selection affect the composition and performance of logistic regression models? [Masterthesis]. Heidelberg: Ruprecht-Karls-Univesität; June 2014.

[9] Rosenbaum PR, Rubin DR. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983:41–55. doi:10.1093/biomet/70.1.41.

[10] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163:1149–56. doi:10.1093/aje/kwj149.

[11] Austin PC. Some Methods of Propensity-Score Matching had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo simulations. Biometrical Journal. 2009:171–84.

[12] Lanehart RE, Rodriguez de Gil P, Kim ES, Bellara AP, Kromrey J, Lee R. Propensity Score Analysis and Assessment of Propensity Score Approaches Using SAS® Procedures. SAS Global Forum 2012.

[13] Kuss O, Blettner M, Borgermann J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. Dtsch Arztebl Int. 2016;113:597–603. doi:10.3238/arztebl.2016.0597.

[14] Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. Pharmacoepidemiol Drug Saf. 2008:1202–17. doi:10.1002/pds.1673.

[15] Yang D, Dalton JE. Standardized Difference: An Index to Measure the Effect Size between Two Groups. SAS Global Forum. 2012.

[16] Faries DE, Obenchain R, Haro JM, Leon AC. Analysis of Observational Health Care Data Using SAS: SAS Institute; 2010.

[17] King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. [Working Paper]. Cambridge: Harvard University, December 2016.

[18] Stone CA, Tang Y. Comparing Propensity Score Methods in Balancing Covariates and Recovering Impact in Small Sample Educational Program Evaluations. Practical Assessment, Research & Evaluation. 2013;18:1–12.