ÜBERSICHTSARBEIT

Propensity Score – eine alternative Methode zur Analyse von Therapieeffekten

Teil 23 der Serie zur Bewertung wissenschaftlicher Publikationen

Oliver Kuss, Maria Blettner, Jochen Börgermann

ZUSAMMENFASSUNG

Hintergrund: Nur die Randomisierung garantiert in Therapiestudien eine gleichmäßige Verteilung aller bekannten und unbekannten Patientenmerkmale auf eine Interventions- und eine Kontrollgruppe und erlaubt dadurch kausale Aussagen über Therapieeffekte. Randomisierte kontrollierte Studien werden jedoch auch für ihre fehlende externe Validität kritisiert. Nichtrandomisierte Studien sind eine Alternative, allerdings besteht hier die Gefahr, dass sich die Interventions- und die Kontrollgruppe bezüglich bekannter und unbekannter Patientenmerkmale unterscheiden. Zur Analyse von nichtrandomisierten Studien werden in der Regel multiple Regressionsmodelle verwendet, immer häufiger wird aber auch auf die sogenannte Propensity-Score-Methode zurückgegriffen.

Methode: Auf Basis einer selektiven Literaturrecherche und der wissenschaftlichen Erfahrung der Autoren wird die Propensity-Score-Methode anhand eines Beispiels aus der koronaren Bypass-Chirurgie ausführlich dargestellt und erklärt.

Ergebnis: Der Propensity Score (PS) ist definiert als die Wahrscheinlichkeit, mit der ein Patient die zu prüfende Therapie erhält. Der PS wird in einem ersten Schritt aus den vorhandenen Daten geschätzt, beispielsweise in einem logistischen Regressionsmodell. Im zweiten Schritt erfolgt die Schätzung des eigentlich interessierenden Therapieeffekts unter Zuhilfenahme des PS. Dabei stehen vier Methoden zur Verfügung: PS-Matching, „inverse probability of treatment weighting"(IPTW)-Schätzung, Stratifizierung nach dem PS oder Regressionsadjustierung für den PS.

Schlussfolgerung: Die Propensity-Score-Methode ist eine gute Alternative zur Auswertung von nichtrandomisierten Therapiestudien. Sie hat erkenntnistheoretische Vorteile im Vergleich zur herkömmlichen Regressionsanalyse. Der Propensity Score kann allerdings nur für die bekannten und tatsächlich gemessenen Störgrößen adjustieren. Die gleichmäßige Verteilung von unbekannten Störgrößen bleibt die Domäne randomisierter kontrollierter Studien.

▶ **Zitierweise**
Kuss O, Blettner M, Börgermann J: Propensity score: an alternative method of analyzing treatment effects—part 23 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2016; 113: 597–603.
DOI: 10.3238/arztebl.2016.0597

Deutsches Diabetes-Zentrum (DDZ), Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf, Institut für Biometrie und Epidemiologie und Centre for Health and Society (chs), Medizinische Fakultät, Heinrich-Heine-Universität Düsseldorf: Prof. Dr. sc. hum. Kuss

Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI), Universitätsmedizin der Johannes Gutenberg-Universität Mainz: Prof. Dr. rer. nat. Blettner

Klinik für Thorax- und Kardiovaskularchirurgie, Herz- und Diabeteszentrum Nordrhein-Westfalen, Universitätsklinik der Ruhr-Universität Bochum, Bad Oeynhausen: PD Dr. med. Börgermann

Man ist sich in der medizinischen Forschung weitgehend einig darüber, dass Therapien primär in randomisierten kontrollierten Studien geprüft werden sollten. Nur die Randomisierung garantiert eine gleichmäßige Verteilung aller bekannten und unbekannten Patientenmerkmale auf eine Interventions- und eine Kontrollgruppe und erlaubt dadurch kausale Aussagen über Therapieeffekte. Randomisierte kontrollierte Studien sind jedoch in manchen Fällen „unnötig, ungeeignet, unmöglich oder ungenügend" (1) und werden darüber hinaus immer wieder für ihre fehlende externe Validität kritisiert: Patienten in randomisierten kontrollierten Studien sind in der Regel jünger und gesünder als der durchschnittliche Patient (2, 3).

Nichtrandomisierte Studien können für die Evaluierung von Therapien eine Alternative sein, allerdings haben diese das Problem der fehlenden internen Validität: Die Therapiezuteilung erfolgt nicht randomisiert und die Interventions- und die Kontrollgruppe können sich systematisch bezüglich bekannter und (schlimmer noch) unbekannter Patientenmerkmale unterscheiden. Mögliche Unterschiede, die sich in den Gruppen im Verlauf der Studie ergeben, können daher nicht notwendigerweise auf die unterschiedliche Behandlung zurückgeführt werden. Diese Unterschiede könnten auch durch die systematischen Differenzen zwischen den Gruppen zustande gekommen sein.

Eine Reihe von statistischen Verfahren wurde entwickelt, um diese Unterschiede bei der Auswertung zu berücksichtigen. Standardverfahren sind dabei die multiplen Regressionsmodelle. Immer häufiger wird jedoch auch die sogenannte Propensity-Score-Methode angewendet (4). Im Folgenden wird der Propensity Score eingeführt und zuerst allgemein, dann anhand eines Beispiels aus der koronaren Bypass-Chirurgie ausführlich dargestellt und erklärt. In einem weiteren Abschnitt wird die Methode gegenüber den herkömmlichen Regressionsmodellen abgegrenzt. Der Artikel schließt mit einigen grundsätzlichen Bemerkungen zum Erkenntnisgewinn in der medizinischen Forschung.

Propensity-Score-Methode

Der Propensity Score (PS) ist die Wahrscheinlichkeit, mit der ein Patient die zu prüfende Therapie erhält. In einer 1:1-randomisierten Studie ist diese gerade 0,5. In einer nichtrandomisierten Studie ist diese Wahrschein-

TABELLE 1

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

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

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

lichkeit für jeden einzelnen Patienten unbekannt und hängt von dessen Merkmalen ab. Der PS muss also in einem ersten Schritt aus den vorliegenden Daten geschätzt werden. Hierzu kann ein logistisches Regressionsmodell eingesetzt werden, in dem die zugeteilte Therapie die abhängige Variable ist und die bei Therapiebeginn bestehenden Patientenmerkmale als unabhängige Variablen verwendet werden. Aus den geschätzten Parametern dieses PS-Modells kann dann der Propensity Score für jeden einzelnen Patienten berechnet werden. Bei der Auswahl der unabhängigen Variablen für das PS-Modell sollte man darauf achten, Merkmale heranzuziehen, die den späteren Therapieerfolg (und nicht etwa die Therapiezuteilung) vorhersagen, da diese die Varianz des Behandlungseffekts verringern, ohne einen zusätzlichen Bias zu erzeugen (5). Für die Therapieentscheidung unbekannte oder nicht gemessene Faktoren können im PS-Modell natürlich nicht berücksichtigt werden.

In einem zweiten Schritt wird der eigentliche interessierende Therapieeffekt unter Zuhilfenahme des Propensity Score geschätzt. Dabei stehen vier Methoden zur Berücksichtigung des Propensity Score zur Verfügung (6):

- PS-Matching
- „inverse probability of treatment weighting"(IPTW)-Schätzung
- Stratifizierung
- Regressionsadjustierung für den PS.

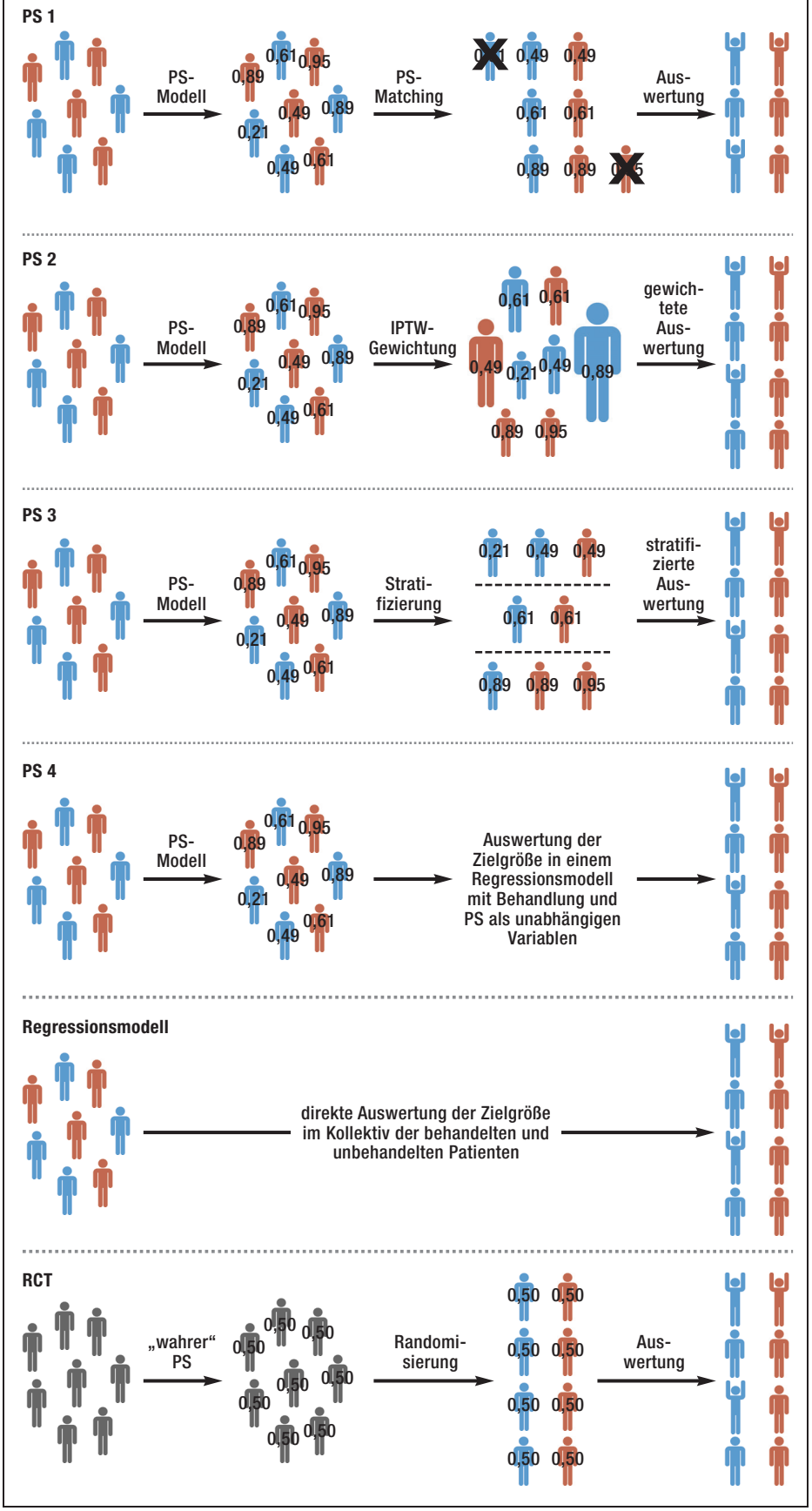
PS-Matching: Beim PS-Matching wird jedem behandelten Patienten ein („1:1-Matching") unbehandelter Patient oder es werden ihm mehrere („1:n-Matching") unbehandelte Patienten mit demselben PS oder mit einem nur minimal innerhalb vorher definierter Grenzen abweichenden PS zugeteilt. Im gematchten Kollektiv wird dann der Therapieeffekt unter Berücksichtigung des Matchings (7) geschätzt.

IPTW-Schätzung: Bei der IPTW-Schätzung wird jedem Patienten der Kehrwert der Behandlungswahrscheinlichkeit, die zu seiner tatsächlichen Behandlung gehört, als statistisches Gewicht zugeteilt: Ein behandelter Patient erhält das Gewicht 1/PS, ein unbehandelter Patient das Gewicht 1/(1-PS). Diese Definition hat mathematische Gründe, kann aber auch intuitiv interpretiert werden (8): Ein behandelter Patient mit einem niedrigen PS (für die Behandlung) bekommt ein hohes Gewicht deshalb, weil er den unbehandelten Patienten bezüglich ihrer Merkmale ähnelt (ausgedrückt durch deren niedrigen PS) und daher einen validen Vergleich mit diesen ermöglicht. In die Auswertung für den Therapieeffekt gehen die Patienten dann entsprechend ihres statistischen Gewichts ein.

Stratifizierung: Die Stratifizierung für den PS entspricht einem vergröberten PS-Matching. Hier wird der gesamte Datensatz in gleich große Teile (zum Beispiel Quintile) bezüglich des geschätzten PS eingeteilt. In jedem dieser Teile wird mit herkömmlichen Methoden ein Therapieeffekt geschätzt, die so erhaltenen Therapieeffekte werden dann mit metaanalytischen Methoden zusammengefasst.

Regressionsadjustierung: Bei der Regressionsadjustierung für den PS wird ein herkömmliches Regressionsmodell mit der interessierenden Zielgröße als abhängiger Variable und dem Therapieeffekt und dem PS

GRAFIK



Durchführung einer PS-Analyse im Vergleich zu einer herkömmlichen Regressionsanalyse und einer randomisierten kontrollierten Studie

Die Abkürzungen PS 1–PS 4 stehen für die vier Methoden zur Berücksichtigung des Propensity Scores (PS):
PS 1 = PS-Matching,
PS 2 = „inverse probability of treatment weighting“(IPTW)-Schätzung,
PS 3 = Stratifizierung und
PS 4 = Regressionsadjustierung für den PS.

Am Beginn jeder PS-Analyse steht eine Gruppe von Patienten, die mit der interessierenden Intervention behandelt (rot) oder nicht behandelt (blau) wurden. Mit Hilfe der vorliegenden Patientenmerkmale wird ein PS-Modell geschätzt und für jeden Patienten der Propensity Score berechnet (in der Grafik als Zahlenwerte bei den Piktogrammen). Entsprechend der jeweiligen PS-Methode werden Patienten dann gematcht (PS 1, in der Regel werden Patienten ausgeschlossen, für die kein Matching-Partner gefunden worden ist; sie sind mit einem X gekennzeichnet), in Abhängigkeit von ihrem PS gewichtet (PS 2, Patienten mit höherem IPTW-Gewicht sind in der Grafik größer dargestellt), stratifiziert (PS 3, hier in Terzilen), oder es wird (PS 4) eine Regressionsanalyse unter Berücksichtigung des PS durchgeführt. Entsprechend der PS-Methode werden die klinischen Zielgrößen ausgewertet. (In der Grafik sind geheilte Patienten der Einfachheit halber in einer Jubelpose dargestellt.)

In einem herkömmlichen Regressionsmodell wird dagegen ein einziges statistisches Modell berechnet, in das die klinische Zielgröße als abhängige Variable, die Therapie und die anderen Patientenmerkmale als unabhängige Variablen eingehen.

Die Ähnlichkeit zwischen einer randomisierten kontrollierten Studie (RCT, „randomised controlled trial“) und einer PS-Analyse wird im unteren Teil der Grafik verdeutlicht: Zu Beginn sind die Patienten in einer RCT noch unbehandelt (grau) und ihr PS (also die Wahrscheinlichkeit, die Intervention zu erhalten) ist bekannt und gleich 0,5. Bei der Randomisierung wird jedem Patienten eine Therapie zugeteilt, so dass wie beim PS je eine Gruppe von behandelten und unbehandelten Patienten gebildet wird. Schließlich folgt als letzter Schritt die Auswertung der klinischen Zielgröße.

Unbehandelt
Intervention
Kontrolle

TABELLE 2

Präoperative Patientenmerkmale vor und nach PS-Matching* (modifiziert nach [16])

	alle Patienten (n = 1,282)			PS-gematchte Patienten (n = 788)		
	Less-OPCAB (n = 395)	cCABG (n = 887)	z-Differenz	Less-OPCAB (n = 394)	cCABG (n = 394)	z-Differenz
Alter (Jahre)	69,3 ± 9,1	67,5 ± 9,4	3,24	69,3 ± 9,1	69,0 ± 8,9	0,46
männlich (%)	78,2	77,9	0,13	78,2	77,9	0,09
BMI (kg/m²)	27,8 ± 4,2	28,3 ± 4,5	−1,83	27,8 ± 4,2	28,0 ± 4,2	−0,60
Hauptstammstenose (%)	25,3	25,5	−0,06	25,1	24,9	0,08
LVEF (%)	56,7 ± 12,3	55,4 ± 14,1	1,64	56,6 ± 12,2	56,9 ± 13,3	−0,28
präoperativer Myokardinfarkt (%)	27,1	35,7	−3,14	27,2	26,7	0,16
Hypertonie (%)	82,3	84,1	−0,80	82,2	82,2	0
Diabetes mellitus (%)	22,8	31,7	−3,39	22,8	19,8	1,05
COPD (%)	5,8	7,1	−0,88	5,8	6,1	−0,15
Niereninsuffizienz (%)	0,8	1,2	−0,86	0,8	0,3	1,16
Schlaganfall (%)	1,0	2,4	−2,03	1,0	1,8	−0,95
pAVK (%)	11,9	11,4	0,26	11,7	14,7	−1,27
Voroperationen (n)	0,05 ± 0,26	0,08 ± 0,39	−1,56	0,05 ± 0,26	0,06 ± 0,27	−0,80
Dringlichkeit (%)			−4,82			0,25
elektiv	91,9	81,0		91,9	92,4	
dringlich	2,5	9,8		2,5	2,3	
Notfall	5,3	8,7		5,3	4,8	
ultima ratio	0,3	0,6		0,3	0,5	
präoperative IABP (%)	1,0	1,5	−0,71	1,0	1,0	0

*angegeben sind für die metrischen Patientenmerkmale Mittelwert ± Standardabweichung und für die kategorialen Patientenmerkmale die relative Häufigkeit in %;
BMI, Body-mass-Index; cCABG, konventionelle CABG; CABG, „coronary artery bypass graft"; COPD, chronisch obstruktive Lungenerkrankung; IABP, intraaortale Ballonpumpe; lessOPCAB, clampless OPCAB, ohne Abklemmen der Aorta; LVEF, linksventrikuläre Ejektionsfraktion; OPCAB, „off-pump coronary artery bypass grafting"; pAVK, periphere arterielle Verschlusskrankheit; PS, Propensity Score

als unabhängigen Variablen berechnet. Der Einfluss der Behandlung auf die Zielgröße ist so für den PS und damit für alle in den PS eingeschlossenen Patientenmerkmale adjustiert.

Jede dieser Methoden hat spezifische Stärken und Schwächen, jedoch wird das PS-Matching immer wieder als das zu bevorzugende Verfahren genannt (9, 10). Der Hauptvorteil des PS-Matching ist die Möglichkeit, die erhobenen Merkmale von behandelten und unbehandelten Patienten, ähnlich wie in der „Table 1" einer randomisierten kontrollierten Studie, explizit darzustellen. Dadurch kann geprüft werden, ob die Verteilung dieser Merkmale bei behandelten und unbehandelten Patienten ähnlich ist. Zusätzlich sollte auch die Verteilung der Patientenmerkmale vor dem PS-Matching dargestellt werden, um deutlich zu machen, inwiefern das PS-Matching ursprünglich vorhandene Unterschiede ausgeglichen hat.

Notwendigerweise werden beim PS-Matching Patienten ausgeschlossen, für die kein Matching-Partner gefunden wurde, während alle anderen PS-Methoden den vollen Datensatz für die Analyse nutzen. Dies kann beim PS-Matching mit einer Reduktion der Fallzahl und damit einem Verlust an statistischer Power einhergehen, es hat aber auch den Vorteil, dass bei der Betrachtung der ausgeschlossenen Patienten klar wird, welche Patienten in der Therapiegruppe über- oder unterrepräsentiert waren. Daher dürfen im Folgenden auch keine Aussagen für diese Teilgruppen gemacht werden.

Letztendlich ist die Frage von PS-Matching versus andere PS-Methoden immer die nach einem Ausgleich zwischen einer verzerrten (Bias) oder einer ungenauen (Varianz) Schätzung des Therapieeffektes (8). PS-Matching sollte dann verwendet werden, wenn die Gruppen möglichst ähnlich sein sollen (kein Bias). Man muss dann allerdings aufgrund der kleineren Fallzahl eine größere Varianz in Kauf nehmen. Eine Übersicht der Stärken und Schwächen der verschiedenen Methoden gibt *Tabelle 1* wieder. In der *Grafik* ist der Ablauf einer PS-Analyse im Vergleich zu einer randomisierten kontrollierten Studie und einer herkömmlichen Regressionsanalyse schematisch dargestellt.

Die Güte eines PS-Modells sollte allein daran gemessen werden, wie gut die Patientenmerkmale in den beiden Therapiegruppen balanciert ist. Eine Berechnung von Anpassungstests wie dem Hosmer-Lemeshow-Test (11) oder von Diskriminationsmaßen wie der c-Statistik (12) führt hier nicht zum Ziel. Beide Verfahren sind nicht geeignet, unbekannte Störgrößen zu entdecken (13). Schlimmer noch, ein hoher Wert der c-Statistik ist weder notwendig noch hinreichend für ei-

ne gute Adjustierung der Störgrößen. Das lässt sich gut an einer randomisierten kontrollierten Studie veranschaulichen, in der man per Konstruktion eine sehr gute Balanciertheit der Störgrößen erreicht, die c-Statistik aber einen sehr kleinen Wert (um die 0,5) ergeben wird (14). Zur konkreten Messung der Balanciertheit von Patientenmerkmalen sind viele Maße vorgestellt worden (6, 15).

Die methodische Weiterentwicklung der Propensity-Score-Methode schreitet nach wie vor voran. Auf einige wichtige Punkte (zum Beispiel Umgang mit fehlenden Werten, Mindestanforderungen an Fallzahlen, Software, Einfluss von verschiedenen Matching-Algorithmen) kann hier leider nicht näher eingegangen werden.

Ein Beispiel

Im Folgenden wird eine publizierte PS-Analyse aus der koronaren Bypass-Chirurgie (16) dargestellt. Diese haben der Erst- und der Letztautor des vorliegenden Artikels gemeinsam durchgeführt. Grundlage war ein Datensatz von insgesamt 1 282 Patienten, die zwischen Juli 2009 und November 2010 am Herz- und Diabeteszentrum NRW in Bad Oeynhausen isoliert koronarchirurgisch versorgt worden waren. Von diesen Patienten waren 69,2 % (n = 887) konventionell mit einem Koronararterien-Bypass („conventional coronary artery bypass graft“ [cCABG]) unter Einsatz der Herz-Lungen-Maschine im kardioplegischen Herzstillstand versorgt worden, 30,8 % (n = 395) waren mit Hilfe der Clampless-off-pump(less-OPCAB)-Technik ohne Herz-Lungen-Maschine und ohne Abklemmen der Aorta operiert worden. Die Entscheidung für eine Operationsmethode hatten die jeweiligen Operateure getroffen. Zur Schätzung des PS für jeden Patienten wurde ein logistisches Regressionsmodell berechnet. Alle in dieses Modell als unabhängige Variablen eingehenden Patientenmerkmale wurden dafür a priori festgelegt und sind in *Tabelle 2* dargestellt. Es wurde ein 1:1-Matching durchgeführt unter Verwendung eines Optimal-Matching-Algorithmus mit einer Caliper-Weite von 0,2 Standardabweichungen des linearen Prädiktors (17).

Es wurde weiterhin geprüft, ob die präoperativen Patientenmerkmale in den beiden Behandlungsgruppen nach dem PS-Matching hinreichend balanciert waren. Dazu können standardisierte Differenzen (9) oder z-Differenzen (18) herangezogen werden. Die z-Differenzen sind in einer randomisierten kontrollierten Studie standard-normalverteilt (N[0,1]) und in einer perfekt gematchten Studie normalverteilt, ebenfalls mit Erwartungswert 0, allerdings mit Varianz =½ (N[0, ½]) (19). Das bedeutet also, dass ein PS-Matching in der Regel eine bessere Balancierung für die bekannten Variablen erreicht als eine Randomisierung. Zur Beurteilung der Größe der z-Differenzen kann die bekannte 2σ-Regel (20) herangezogen werden: Wenn Daten annähernd normalverteilt $N(\mu,\sigma^2)$ sind, liegen im Bereich zwischen μ-2σ und μ+2σ ungefähr 95 % aller Beobachtungen. Gemäß der N(0, ½)-Verteilung der z-Differenzen wären also absolute z-Differenzen von $\sqrt{2}$ = 1,4142… und größer „auffällig“. Solche auffälligen z-Differenzen sollten demnach bei höchstens 5 % der Patientenmerkmale vorkommen, um noch von einem guten PS-Matching sprechen zu können. In der Tat finden wir im ungematchten Kollektiv einige Patientenmerkmale, die eine beträchtlich größere z-Differenz haben, jedoch keine mehr im gematchten Kollektiv.

Zur Beurteilung des Therapieeffekts wurden in der PS-gematchten Stichprobe drei klinische Zielgrößen betrachtet *(Tabelle 3)*:

- eine dichotome (Tod oder Schlaganfall im Verlauf der Behandlung in der Klinik, ja/nein)
- eine metrische (Operationszeit in Minuten)
- eine Ereigniszeit (Zeit bis zum Tod oder Schlaganfall in der Nachbeobachtung).

Für die Nachbeobachtung der Patienten ist im Herz- und Diabeteszentrum NRW ein standardisiertes Vorgehen etabliert worden, bei dem allen operierten Patienten jährlich ein Fragebogen zugeschickt wird. In den Fragebögen berichtete relevante Ereignisse werden über die behandelnden Institutionen (beispielsweise lokales Krankenhaus, Hausarzt) validiert. Bei der Auswertung ist, wie bereits beschrieben, darauf zu achten, dass zum Beispiel durch eine konditionale Analyse für das Matching-Stratum adjustiert wird (7). Wie aus *Tabelle 3* ersichtlich, ist die Less-OPCAB-Technik der cCABG-Technik bezüglich aller drei Zielgrößen überlegen. Qualitativ sehr ähnliche Ergebnisse ergeben sich für die drei anderen genannten PS-Methoden und das parallele herkömmliche Regressionsmodell.

PS-Analysen versus herkömmliche Regressionsmodelle

Es gibt eine Reihe von Vorteilen der PS-Methode im Vergleich zu den herkömmlichen Regressionsmodellen, die immer noch die Standardmethode zur Adjustierung für Patientenmerkmale in nichtrandomisierten Studien darstellen. Ein erster Vorteil besteht darin, dass PS-Analysen vom Vorgehen her eher einer randomisierten kontrollierten Studie ähneln *(Grafik)*. Insbesondere wird ein PS-Modell so geschätzt, dass keine Information über die eigentlich interessierenden Zielgrößen eingeht, sondern es fließen nur die bei Studienbeginn vorhandenen Patientenmerkmale ein (21). Die Berechnung des PS-Modells gehört also noch zum Design der Studie und nicht zu deren Auswertung.

Randomisierte kontrollierte Studien (RCT, „randomised controlled trial“) und PS-Matching ähneln sich des Weiteren dahingehend, dass sie beide Zwei-Schritt-Verfahren sind: Im ersten Schritt wird darauf Wert gelegt, dass beide Therapiegruppen hinsichtlich der Patientenmerkmale ähnlich sind (beim RCT durch Randomisierung, beim PS durch Berechnung des PS-Modells). Im zweiten Schritt wird dann in der balancierten Stichprobe der eigentlich interessierende Therapieeffekt geschätzt. Demgegenüber ist das herkömmliche Regressionsmodell ein Ein-Schritt-Verfahren: Der Einfluss der Therapie auf die Zielgröße wird gleichzeitig mit den anderen unabhängigen Variablen geschätzt (22).

Ein Problem der herkömmlichen Regressionsmodelle ist, dass diese immer Therapieeffekte schätzen, und zwar selbst dann, wenn sich die beiden Gruppen von Behandelten und Nicht-Behandelten so extrem unterscheiden, dass eine solche Schätzung nicht sinnvoll ist. Regressionsmo-

TABELLE 3

Ergebnisse für die drei klinischen Zielgrößen in der Gruppe der PS-gematchten Patienten (n = 788) (modifiziert nach [16])

dichotome Zielgröße				
	Less-OPCAB (n = 394)	**cCABG (n = 394)**	**Odds Ratio [95-%-KI]**	**p-Wert**
postoperativer Tod oder Schlaganfall [n (%)]	6 (1,5)	22 (5,6)	0,27 [0,11; 0,67]	0,005
metrische Zielgröße				
	Less-OPCAB (n = 394)	**cCABG (n = 394)**	**MWD [95-%-KI]**	**p-Wert**
Operationszeit in Minuten (Mittelwert; SD)	175 (38)	180 (47)	5 [−1; 11]	0,12
Ereigniszeit als Zielgröße				
	Less-OPCAB (n = 394)	**cCABG (n = 394)**	**Hazard Ratio [95-%-KI]**	**p-Wert**
Zeit bis Tod oder Schlaganfall im Follow-up (Ein-Jahres-Wahrscheinlichkeit für Ereignisfreiheit in %)	94,7	89,8	0,60 [0,35; 1,03]	0,06

cCABG, konventionelle CABG; CABG, „coronary artery bypass grafting"; KI, Konfidenzintervall; Less-OPCAB, clampless OPCAB,ohne Abklemmen der Aorta; OPCAB, „off-pump coronary artery bypass grafting"; MWD, Mittelwertdifferenz; PS, Propensity Score; SD, Standardabweichung

delle erlauben Aussagen darüber, was passiert wäre, wenn Behandelte nicht therapiert worden wären. Dabei werden aber die Informationen von Nichtbehandelten benutzt, die unter Umständen vollkommen anders sind als die Behandelten. Die Information über die Nichtbehandelten wird dabei (durch Extrapolation) nur geschätzt, ist aber nicht wirklich beobachtet worden (8). Das heißt konkret: Wenn zum Beispiel der älteste behandelte Patient 30 Jahre alt und männlich war, wird ein herkömmliches Regressionsmodell zur Beurteilung der Intervention auch die Information über eine unbehandelte 80-jährige Dame nutzen (23).

Schließlich ist die PS-Methode gerade für das Modellieren von seltenen Ereignissen den herkömmlichen Regressionsmodellen besonders überlegen (24). Dies hat folgenden Grund: Wenn die zu vergleichenden Therapien jeweils häufig angewandt werden, das eigentlich interessierende Zielereignis aber selten vorkommt, dann wird es in der Regel so sein, dass nicht genug Information vorhanden ist, um den Zusammenhang zwischen Zielgröße und Patientenmerkmalen (einschließlich der Therapie) in einem herkömmlichen Regressionsmodell gut zu schätzen. Umgekehrt kann das PS-Modell gut geschätzt werden, weil für die Messung des Zusammenhangs von Therapiezuteilung (die abhängige Variable im PS-Modell) und Patientenmerkmalen (die unabhängigen Variablen im PS-Modell) hinreichend Informationen vorhanden sind (25).

Fazit

Die Propensity-Score-Methode kann eine Randomisierung nicht ersetzen, stellt aber eine gute Alternative zur Auswertung von nichtrandomisierten Therapiestudien dar, die erkenntnistheoretische Vorteile im Vergleich zur herkömmlichen Regression hat. Wendet man in der Gruppe der PS-Methoden zusätzlich noch die Matching-Methode an, ergibt sich eine Reihe von Vorteilen. Der wichtigste unter diesen ist die Möglichkeit, explizit die beiden Therapiegruppen bezüglich der Risikofaktoren zu vergleichen.

Eines ist dabei jedoch stets zu bedenken: Der Propensity Score kann, wie die herkömmlichen Regressionsmodelle auch, nur für die bekannten und tatsächlich gemessenen Patientenmerkmale adjustieren. Die gleichmäßige Verteilung auch der unbekannten Störgrößen kann nur in randomisierten konrollierten Studien erreicht werden.

Die randomisierte kontrollierte Studie ist nach wie vor das Design der Wahl, um die Wirksamkeit von Therapien zu prüfen. Die klinische Forschung sollte jedoch darauf achten, dass diese Erkenntnis nicht zum Dogma erstarrt. Zum einen gibt es immer bessere Evidenz, dass randomisierte kontrollierte und nichtrandomisierte Studien in den meisten Fällen zu ähnlichen Ergebnissen führen (26, 27). Beispiele, in denen sich Evidenz aus randomisierten kontrollierten und nichtrandomisierten Studien explizit widerspricht (zum Beispiel im Falle der Women's Health Initiative (WHI)-Studie zur Hormonersatztherapie bei postmenopausalen Frauen [28]) sind aus historischen, pragmatischen oder pädagogischen Gründen wichtig (29), bleiben aber Ausnahmen und können bei genauerer Analyse auch oft erklärt werden (30). Unter Umständen ist die Gefahr durch unbekannte Patientenmerkmale bei der PS-Analyse auch nicht so groß wie befürchtet. Diese unbekannten Patientenmerkmale sind nur dann eine wirkliche Gefahr, wenn diese nicht mit den bekannten Patientenmerkmalen assoziiert sind. Wenn bekannte und unbekannte Patientenmerkmale assoziiert sind, dann wird durch das Adjustieren für bekannte auch für die unbekannten Patientenmerkmale mitadjustiert (31).

Gemeinsam mit Borah und Koautoren (32) erwarten wir, dass die Nachfrage von Patienten, Klinikern und dem gesamten Gesundheitssystem nach Evidenz aus nichtrandomisierten Studien in den nächsten Jahren noch ansteigen wird. Es gibt schlichtweg zu viele Fragestellungen in der medizinischen Versorgung, als dass alle in randomisierten kontrollierten Studien beantwortet werden könnten. Zudem wird sich die Gesellschaft weder die dazu nötigen Mittel noch die dazu nötige Zeit leisten können oder wollen.

KERNAUSSAGEN

- Zur Auswertung von nichtrandomisierten Studien wird immer häufiger die Propensity-Score-Methode herangezogen.
- Die randomisierte kontrollierte Studie ist nach wie vor das Design der Wahl, um die Wirksamkeit von Therapien zu prüfen. Die klinische Forschung sollte jedoch darauf achten, dass diese Erkenntnis nicht zum Dogma erstarrt.
- Die Propensity-Score-Methode kann eine Randomisierung nicht ersetzen, stellt aber eine gute Alternative zur Auswertung von nichtrandomisierten Therapiestudien dar.
- Der Propensity Score kann, wie die herkömmlichen Regressionsmodelle auch, nur für die bekannten und tatsächlich gemessenen Patientenmerkmale adjustieren. Die gleichmäßige Verteilung auch der unbekannten Merkmale kann nur in randomisierten kontrollierten Studien erreicht werden.
- Die Nachfrage von Patienten, Klinikern und dem gesamten Gesundheitssystem nach Evidenz aus nichtrandomisierten Studien wird in den nächsten Jahren noch steigen.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt besteht.

Manuskriptdaten
eingereicht: 9. 5. 2016; revidierte Fassung angenommen: 23. 6. 2016

LITERATUR

1. Black N: Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996; 312:1215–8.
2. McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C: Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. BMJ 1999; 319: 312–5.
3. Rothwell PM: External validity of randomised controlled trials: „to whom do the results of this trial apply?". Lancet 2005; 365: 82–93.
4. Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41–55.
5. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T: Variable selection for propensity score models. Am J Epidemiol 2006; 163: 1149–56.
6. Austin PC: The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making 2009; 29: 661–77.
7. Austin PC: Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. Int J Biostat 2009; 5: Article 13.
8. Stuart EA, Marcus SM, Horvitz-Lennon MV, Gibbons RD, Normand SL: Using non-experimental data to estimate treatment effects. Psychiatr Ann 2009; 39: 719–28.
9. Austin PC: Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. J Thorac Cardiovasc Surg 2007; 134: 1128–35.
10. Morgan SL, Harding DJ: Matching estimators of causal effectsd prospects and pitfalls in theory and practice. Sociological Methods Res 2006; 35: 3–60.
11. Hosmer DW, Lemeshow S: Goodness of fit tests for the multiple logistic regression model. Commun Stat – Theor M 1980; 9: 1043–69.
12. Harrell FE: Regression modeling strategies. New York: Springer 2001; 257.
13. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor VM: Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf 2005; 14: 227–38.
14. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T: The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 2011; 20: 317–20.
15. Belitser SV, Martens EP, Pestman WR, Groenwold RH, de Boer A, Klungel OH: Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf 2011; 20: 1115–29.
16. Börgermann J, Hakim K, Renner A, et al.: Clampless off-pump versus conventional coronary artery revascularization: a propensity score analysis of 788 patients. Circulation 2012; 126 (11 Suppl 1): S176–82.
17. Austin PC: Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat 2011; 10: 150–61.
18. Kuss O: The z-difference can be used to measure covariate balance in matched propensity score analyses. J Clin Epidemiol 2013; 66: 1302–7.
19. Rubin DB, Thomas N: Matching using estimated propensity scores: relating theory to practice. Biometrics 1996; 52: 249–64.
20. Hedderich J, Sachs L: Angewandte Statistik. Berlin, Heidelberg, New York: Springer 2016; 264.
21. Rubin DB: The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007; 26: 20–36.
22. Martens EP, de Boer A, Pestman WR, Belitser SV, Stricker BH, Klungel OH: Comparing treatment effects after adjustment with multivariable cox proportional hazards regression and propensity score methods. Pharmacoepidemiol Drug Saf 2008; 17: 1–8.
23. Pattanayak CW, Rubin DB, Zell ER: [Propensity score methods for creating covariate balance in observational studies]. Rev Esp Cardiol 2011; 64: 897–903.
24. Cepeda MS, Boston R, Farrar JT, Strom BL: Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003; 158: 280–7.
25. Braitman LE, Rosenbaum PR: Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med 2002; 137: 693–5.
26. Anglemyer A, Horvath HT, Bero L: Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014; 4: MR000034.
27. Eichler M, Pokora R, Schwentner L, Blettner M: Evidenzbasierte Medizin – Möglichkeiten und Grenzen. Dtsch Arztebl 2015; 112: A 2190–2.
28. Manson JE, Hsia J, Johnson KC, et al., Women's Health Initiative Investigators: Estrogen plus progestin and the risk of coronary heart disease. N Engl J Med 2003; 349: 523–34.
29. Abel U, Koch A: The role of randomization in clinical studies: myths and beliefs. J Clin Epidemiol 1999; 52:487–97.
30. Hernán MA, Alonso A, Logan R: Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 2008; 19: 766–79.
31. Stuart EA: Matching methods for causal inference: a review and a look forward. Statistical Science 2010; 25: 1–21.
32. 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? J Comp Eff Res 2014; 3: 63–78.

Anschrift für die Verfasser
Prof. Dr. sc. hum Oliver Kuß
Deutsches Diabetes-Zentrum (DDZ)
Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität
Institut für Biometrie und Epidemiologie
Auf'm Hennekamp 65
40225 Düsseldorf
oliver.kuss@ddz.uni-duesseldorf.de

Zitierweise
Kuss O, Blettner M, Börgermann J: Propensity score: an alternative method of analyzing treatment effects—part 23 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2016; 113: 597–603. DOI: 10.3238/arztebl.2016.0597

@ The English version of this article is available online:
www.aerzteblatt-international.de