Gütemaße und Kriterien bei der Anwendung von Propensity Scores

Quality measures and criteria for the application of propensity scores

Abstract

Propensity scores (PS) have been established as a valid alternative to conventional regression models when evaluating non-randomized treatment studies. The PS describes the probability for an individual to receive a treatment, conditional on a set of observed covariates. PS analyses are performed in two steps. In the first step, the PS is generally estimated via logistic regression. In the second step, the actual treatment effect is estimated. The quality of a PS analysis depends on whether it is possible to achieve a sufficient balance of the patient characteristics in the treatment groups in the first step. This is the only way to ensure that these patient characteristics do not bias the estimate of the treatment effect. Various measures have been proposed to measure this balance, e.g. the standardized difference or the z-difference. Closely related to the balance of patient characteristics (and thus also a measure of the quality for a PS model) is the overlap, i.e. the similarity in the distribution of the estimated propensity scores in the two treatment groups. A valid comparison of the treatments is only possible in regions of sufficient overlap. In this article, the two concepts of balance and overlap are presented and discussed using an example from cardiac surgery.

Zusammenfassung

Propensity Scores (PS) haben sich in den letzten Jahren als eine valide Alternative zu herkömmlichen Regressionsmodellen bei der Auswertung von nichtrandomisierten Behandlungsstudien etabliert. PS-Analysen werden in zwei Schritten durchgeführt. Im ersten Schritt wird der PS, also die Wahrscheinlichkeit, mit der ein Individuum die zu prüfende Behandlung erhält, geschätzt. Im zweiten Schritt erfolgt die Schätzung des eigentlich interessierenden Behandlungseffekts unter Zuhilfenahme des PS. Die Güte einer PS-Analyse ist im Wesentlichen davon abhängig, ob es im ersten Schritt gelingt, eine hinreichende Balanciertheit der PatientInnenmerkmale in den Behandlungsgruppen zu erreichen. Nur dann ist gewährleistet, dass diese PatientInnenmerkmale nicht die Schätzung des Behandlungseffekts verzerren. Zur Messung dieser Balanciertheit wurden verschiedene Maße vorgeschlagen, z.B. die standardisierte Differenz oder die z-Differenz. Eng verwandt mit der Balanciertheit der PatientInnenmerkmale und damit auch ein Maß für die Güte eines PS-Modells ist die Überlappung („overlap“), also die Ähnlichkeit der Verteilung der geschätzten PS in den beiden Behandlungsgruppen. In Wertebereichen des PS ohne Overlap, in denen sich unter Umständen also nur Beobachtungen aus einer der beiden Behandlungsgruppen finden, ist streng genommen ein Vergleich der Behandlungen gar nicht möglich. In diesem Beitrag werden die beiden Konzepte anhand eines Beispiels aus der Herzchirurgie vorgestellt und diskutiert.

Oliver Kuß[1,2]
Alexandra Strobel[3]

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

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

3 Institut für Medizinische Epidemiologie, Biometrie und Informatik, Medizinische Fakultät der Martin-Luther-Universität Halle-Wittenberg, Halle, Deutschland

Einleitung

Propensity Scores (PS) haben sich in den letzten Jahren als eine valide Alternative zu herkömmlichen Regressionsmodellen bei der Auswertung von nichtrandomisierten Behandlungsstudien herausgestellt. Gründe für diese Entwicklung sind sowohl im statistischen als auch im erkenntnistheoretischen Bereich zu suchen, wobei in letzterem die explizite Einbettung von PS in die Theorie der kausalen Inferenz dessen Annahmen und Voraussetzungen klarer und transparenter macht. Der PS wurde 1983 von Rosenbaum & Rubin [1] eingeführt und beschreibt die Wahrscheinlichkeit, mit der ein Individuum eine zu prüfende Behandlung erhält.
PS-Analysen werden in zwei Schritten durchgeführt. Im ersten Schritt wird der PS für jedes Individuum in der Regel mit einem logistischen Regressionsmodell geschätzt. Im zweiten Schritt erfolgt dann die Schätzung des eigentlich interessierenden Behandlungseffekts unter Zuhilfenahme des PS, z.B. durch PS-Matching oder durch Gewichtung für den PS [2].

Balanciertheit und Overlap als Maße für die Güte eines Propensity-Score-Modells

Zur Messung der Güte eines PS-Modells werden im Allgemeinen zwei Größen herangezogen: Balanciertheit und Overlap. Während die Balanciertheit die Ähnlichkeit der Verteilung *der PatientInnenmerkmale* in beiden Behandlungsgruppen beschreibt, wird durch den Overlap die Ähnlichkeit der Verteilung *des Propensity Scores* in beiden Behandlungsgruppen charakterisiert.
Balanciertheit und Overlap sind wichtige Gütemaße, da nur bei deren Vorliegen valide Aussagen über den Behandlungseffekt im Sinne der Theorie der kausalen Inferenz gemacht werden können [3], [4]. Die Wichtigkeit dieser beiden Maße wird heuristisch klar, wenn man eine PS-Analyse mit einer 1:1-randomisierten klinischen Studie („randomised controlled trial"=RCT) vergleicht. In einem RCT ist die Balanciertheit nahezu perfekt, da die Verteilung der PatientInnenmerkmale durch Randomisierung in den Behandlungsgruppen nicht nur ähnlich, sondern sogar identisch ist, zumindest für hinreichend große Fallzahlen.
Auch bezüglich des Overlaps ist ein RCT optimal, da der PS für jedes Individuum bekannt und insbesondere gleich ist (d.h. PS=½). Die Verteilung des PS in den beiden Behandlungsgruppen ist daher ebenfalls identisch.
Auf eine einfache Formel gebracht: Je besser die Balanciertheit und der Overlap, umso ähnlicher ist eine PS-Analyse einem RCT und umso geringer ist das Risiko, dass die Ergebnisse durch die beobachteten Kovariablen verzerrt sind.
Andere Gütemaße für PS-Modelle, wie z.B. der Hosmer-Lemeshow-Test oder die c-Statistik, sind dagegen weniger geeignet [5]. Ein hoher Wert der c-Statistik ist z.B. weder notwendig noch hinreichend für eine gute Confounderadjustierung [6]. Statistische Tests sollten ebenso mit Vorsicht betrachtet werden, weil deren Ergebnisse im Wesentlichen von der Stichprobengröße abhängig sind (Imai et al. [7] nennen die Verwendung dieser die „balance test fallacy"): In großen Stichproben werden auch irrelevante Abweichungen bei der Balanciertheit statistisch signifikant sein; in kleinen Stichproben werden relevante Imbalancen nicht entdeckt.
Der Vollständigkeit halber sei darauf hingewiesen, dass in einer PS-Analyse nicht nur Balanciertheit und Overlap hinreichend gut sein müssen, um valide kausale Aussagen bezüglich Behandlungen machen zu können. Es gibt darüber hinaus noch eine Reihe von weiteren Annahmen in der Theorie der kausalen Inferenz (Positivität, Abwesenheit von unbekannten Confoundern, keine Interferenz zwischen PatientInnen, s. z.B. [1]), die dafür erfüllt sein müssen.

Maße zur kovariablenspezifischen und zur globalen Balanciertheitsmessung: Die z-Differenz und die Summe der quadrierten z-Differenzen

Um die Balanciertheit der einzelnen PatientInnenmerkmale zu beurteilen, wird häufig empfohlen, die standardisierte Differenz zu berechnen [8]. Diese ist definiert als die Differenz der Mittelwerte oder Anteile in beiden Gruppen, dividiert durch eine gemeinsame Standardabweichung. In der Regel wird ein Wert von 10% oder weniger vorgeschlagen, um eine zufriedenstellende Balanciertheit anzuzeigen [9].
Die standardisierte Differenz hat jedoch mindestens zwei Nachteile. Zum einen hängt deren Verteilung von der Stichprobengröße ab [8]. Zum anderen ist es nicht möglich, standardisierte Unterschiede für PatientInnenmerkmale auf verschiedenen Skalen zu vergleichen. Austin [10] verwendet zum Beispiel den phi-Koeffizienten für binäre Kovariablen und findet, dass eine standardisierte Differenz von 10% bei einer stetigen Kovariablen ungefähr einem phi-Koeffizienten von 5% bei einer binären Kovariablen entspricht. Des Weiteren existieren bisher für ordinale oder nominale Kovariablen keine standardisierten Differenzen.
Ein Maß, das demgegenüber für metrische, binäre und ordinale Merkmale definiert und auf derselben Skala vergleichbar ist, ist die z-Differenz [11]. Für diese wird das jeweilige Unterschiedsmaß (Mittelwertdifferenz, Risikodifferenz, Wilcoxon-Statistik) durch seinen Standardfehler geteilt (z-Standardisierung). Ein Vorteil der z-Differenz ist, dass deren Wert in einer gematchten PS-Analyse mit zwei Referenzpunkten verglichen werden kann. In einem RCT sind die z-Differenzen standard-normalverteilt (N(0,1)) und in einer (im Sinne von Rubin & Thomas [12], [13]) perfekt gematchten Studie N(0,½)-verteilt. Inzwischen liegt auch eine Weiterentwicklung der z-Differenzen

für gewichtete PS-Analysen und eine z-Differenz für nominale Merkmale vor [14].
Die Summe der quadrierten z-Differenzen (SSQ_{zDiff}) kann zudem als globales (d.h. über alle Kovariablen aggregiertes) Maß zur Balanciertheitsmessung verwendet werden: Wenn die z-Differenzen von k Merkmalen standardnormalverteilt sind, dann ist die Summe der quadrierten z-Differenzen, SSQ_{zDiff}, Chi-quadrat-verteilt mit k Freiheitsgraden. Dieser Zusammenhang gilt allerdings nur approximativ, da für eine exakte Gültigkeit die z-Differenzen der einzelnen Kovariablen unabhängig sein müssten, was im Allgemeinen nicht gegeben sein wird. Durch diese Definition erhält man für die SSQ_{zDiff} zwei Referenzwerte, die zur Optimierung eines PS-Modells bzgl. der Balanciertheit herangezogen werden können: In einem RCT ist der Erwartungswert der SSQ_{zDiff} gleich k, in einer perfekt gematchten PS-Studie gleich k/2.

Ein Beispiel aus der Herzchirurgie

Daten

Zur Darstellung von Balanciertheit und Overlap verwenden wir ein Beispiel aus einer publizierten PS-Analyse in der Aortenklappenchirurgie [15]. Grundlage der Studie waren PatientInnen, denen zwischen Juli 2009 und Juli 2017 am Herz- und Diabeteszentrum NRW in Bad Oeynhausen eine neue Aortenklappe eingesetzt wurde. In der Originalpublikation wurde die konventionelle offene Operation (Ministernotomie, MIC, N=1.929) mit zwei katheterbasierten Behandlungen (transapikal, TA, N=607 und transfemoral, TF, N=1.273) verglichen. Aus Gründen der Übersichtlichkeit beschränken wir uns hier auf den Vergleich von MIC und TA, sodass der Analyse 2.536 Beobachtungen zugrunde liegen. Die Entscheidung bzgl. der Auswahl zwischen MIC und TA wurde nichtrandomisiert durch Konsens des TAVI-Teams (unter Beteiligung von Kardiochirurgie, Kardiologie und Anästhesiologie) getroffen. Als primärer klinischer Outcome wurde die Zeit bis zum Tod der PatientInnen im Follow-up gewählt; die mediane Beobachtungszeit betrug dabei 36,1 Monate.
Dieses Beispiel ist dahingehend extrem, dass es zwei Behandlungen vergleicht, die in sehr unterschiedlichen Gruppen von PatientInnen durchgeführt werden. Die katheterbasierte Implantation der Aortenklappe verzichtet im Gegensatz zur Ministernotomie auf eine Öffnung der Brust (Sternotomie), wodurch ein wesentlich kränkeres Kollektiv von PatientInnen („high risk patients“) diese Behandlung erhalten kann.

Methoden

Für die Auswertung wurde im ersten Schritt zur Schätzung des PS-Modells ein logistisches Regressionsmodell mit insgesamt 23 präspezifizierten Kovariablen berechnet. Im zweiten Schritt und zur Schätzung des Behandlungseffekts für den klinischen Outcome wurde zunächst ein PS-Matching mit Matching-Ratio 1:1 unter Verwendung eines „optimal matching algorithm“ für den logit-transformierten PS [16] durchgeführt. Die Caliperweite wurde (verblindet für den klinischen Outcome) so festgelegt, dass die SSQ_{zDiff} über alle 23 Kovariablen minimal war.

Ergebnisse

Einen Eindruck über die Unterschiede der PatientInnenmerkmale in beiden Gruppen im vollen Datensatz, d.h. vor dem PS-Matching, erhält man aus Tabelle 1, in der exemplarisch die Merkmale Alter, Nierenfunktion (gemessen als „estimated Glomerular Filtration Rate“ (eGFR)) und Vorliegen von Diabetes dargestellt sind. Die PatientInnen in der TA-Gruppe waren wesentlich älter, hatten eine schlechtere Nierenfunktion (niedrigere eGFR) und eine höhere Diabetesprävalenz. Die immensen Unterschiede zwischen den beiden PatientInnengruppen über alle 23 Kovariablen zusammengefasst zeigen sich im Wert der SSQ_{zDiff}. Hier wird vor dem PS-Matching ein Wert von 6.460,37 beobachtet, der um Größenordnungen höher ist, als man diesen aus einem RCT (erwartete SSQ_{zDiff}=23) oder gar aus einer perfekt gematchten PS-Studie (erwartete SSQ_{zDiff}=23/2=11,5) erwarten würde. Nach dem PS-Matching mit optimaler Caliperweite wird die Balanciertheit einer perfekt PS-gematchten Studie mit dem Wert von SSQ_{zDiff}=12,01 nahezu erreicht. Auch die Unterschiede zwischen beiden PatientInnengruppen bzgl. Alter, Nierenfunktion und Diabetesprävalenz sind dann klinisch irrelevant. In Abbildung 1 ist der Verlauf des Optimierungsprozesses für die Caliperweite graphisch dargestellt. Es zeigt sich, dass mit einer sorgfältigen, datengestützten Auswahl der Caliperweite die Balanciertheit der Kovariablen relevant verbessert bzw. optimiert werden kann.
Parallel zur Balanciertheit der Kovariablen ist durch das PS-Matching auch ein guter Overlap entstanden (Abbildung 2). Die Verteilungen des Logit(PS), die vor dem PS-Matching (Abbildung 2a) deutlich separiert sind, sind nach dem PS-Matching nicht mehr zu unterscheiden (Abbildung 2b).
Mit der Optimierung von Balanciertheit und Overlap geht notwendigerweise eine Veränderung der PatientInnenpopulation in der PS-gematchten Stichprobe einher. Dies betrifft zum einen die Fallzahl, die von 2.536 auf 308 zurückgeht, zum anderen die klinischen Eigenschaften der Population. Das mittlere Alter in der PS-gematchten Population liegt nun nahe an dem der TA-Ausgangspopulation, die eGFR zwischen der der beiden Ausgangspopulationen. Die Diabetesprävalenz ist in beiden Gruppen sogar höher als vor dem PS-Matching. Dieser Rückgang in der Fallzahl samt Veränderung der Population wird häufig als eine *Schwäche* des PS-Matchings aufgefasst. Die Autoren sehen diese beiden Veränderungen jedoch als *Stärke* des PS-Matchings, weil damit explizit transparent gemacht wird, welche PatientInnen in den beiden Behandlungsgruppen überhaupt vergleichbar sind und für welche Population Aussagen bezüglich des Behandlungseffekts getroffen werden können.
Die Ergebnisse bezüglich des Behandlungseffekts auf die Zeit bis zum Tod im Follow-up sind in Tabelle 2 als

Tabelle 1: Deskription der Merkmale Alter, Nierenfunktion (eGFR) und Vorliegen eines Diabetes und der Balanciertheit (z-Differenz für das jeweilige Merkmal, Summe der quadrierten z-Differenzen [SSQ_{zDiff}] über alle 23 Merkmale im PS-Modell) vor und nach dem PS-Matching in der Beispielstudie

	Vor dem PS-Matching (n=2.536)			Nach dem PS-Matching (n=308)		
	MIC (n=1.929)	TA (n=607)	z-Differenz	MIC (n=154)	TA (n=154)	z-Differenz
Alter [Jahre, MW±SD]	67±11	81±6	–38,2	79±6	78±7	1,14
eGFR [ml/min, MW±SD]	79±20	56±23	22,2	66±23	65±21	0,38
Diabetes [%]	19	35	–7,75	46	54	–0,87
SSQ_{zDiff} (über alle 23 Kovariablen im PS-Modell)	6.460,37			12,01		

MIC: konventionelle Aortenklappenoperation per Ministernotomie; TA: transapikale, katheterbasierte Aortenklappenimplantation; MW: Mittelwert; SD: Standardabweichung; eGFR: geschätzte glomeruläre Filtrationsrate [estimated glomerular filtration rate]; SSQ_{zDiff}: Summe der quadrierten z-Differenzen

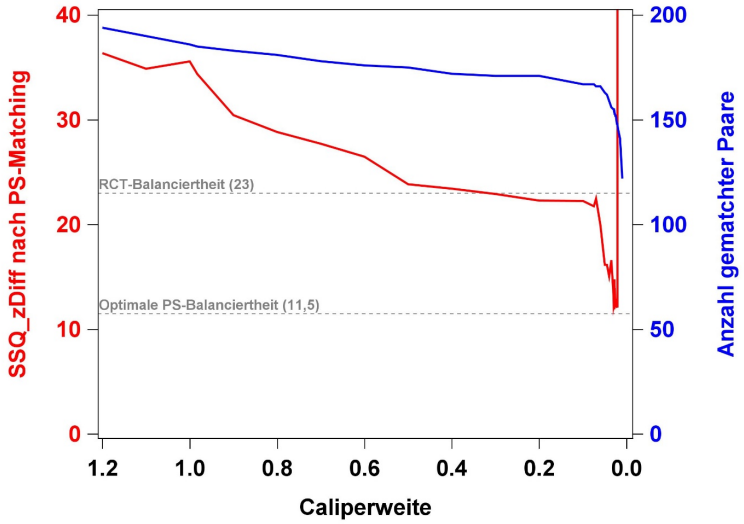

Abbildung 1: Balanciertheit (gemessen als Summe der quadrierten z-Differenzen, rot, linke y-Achse) und Anzahl der PS-gematchten Paare (blau, rechte y-Achse) in Abhängigkeit von der gewählten Caliperweite im PS-Matching-Algorithmus. Die optimale Balanciertheit (minimale SSQ_{zDiff}) von 12,01 wird für die Caliperweite 0,0284 erreicht.

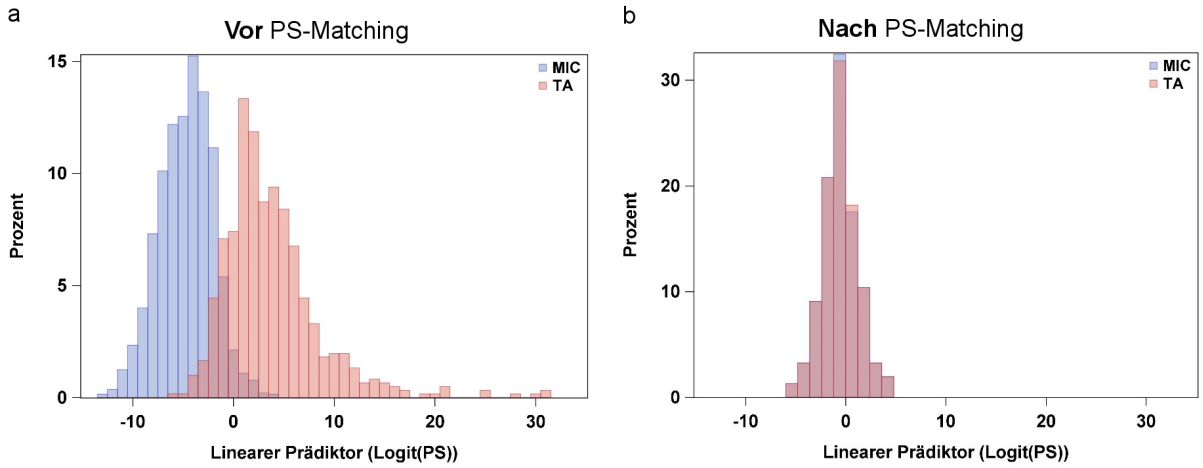

Abbildung 2: Histogramme des logit-transformierten PS („Linearer Prädiktor") vor (a) und nach (b) dem PS-Matching in den beiden Behandlungsgruppen in der Beispielstudie
(MIC: konventionelle Aortenklappenoperation per Ministernotomie; TA: transapikale, katheterbasierte Aortenklappenimplantation)

Tabelle 2: Ergebnisse verschiedener Auswertungsmethoden für den klinischen Outcome Zeit bis zum Tod im Follow-up in der Beispielstudie. Angegeben sind (Spalte 2) die (originale oder gewichtete) Fallzahl und die Anzahl der beobachteten Todesfälle und (Spalte 3) das Hazard Ratio mit 95%-Konfidenzintervall mit der Referenzkategorie MIC. Berechnet werden jeweils Cox-Modelle in verschiedenen Varianten bzgl. der eingeschlossenen PatientInnen, Kovariablen bzw. Gewichtungen.

	N/ Anzahl Todesfälle	Hazard Ratio [95%-KI, Ref.: MIC]
Cox-Modell mit Behandlung als einzige Kovariable im vollen Datensatz	2.536/479	6,40 [5,33; 7,69]
Cox-Modell mit Behandlung und den 23 Kovariablen aus dem PS-Modell als Kovariablen im vollen Datensatz	2.536/479	1,64 [1,23; 2,19]
Stratifiziertes Cox-Modell mit Behandlung als einzige Kovariable in der PS-gematchten Population	308/108	1,25 [0,79; 1,99]
Gewichtetes Cox-Modell mit Behandlung als einzige Kovariable und IPTW-Gewichten (max. Gewichte: 403,9; 237,5; 89,6)	4.737,7/479	1,57 [1,33; 1,85]
Gewichtetes Cox-Modell mit Behandlung als einzige Kovariable und Matching-Gewichten	341,8/479	1,27 [1,11; 1,45]
Gewichtetes Cox-Modell mit Behandlung als einzige Kovariable und Overlap-Gewichten	257,3/479	1,25 [1,09; 1,43]

IPTW: Inverse probability of treatment weighting; MIC: konventionelle Aortenklappenoperation per Ministernotomie; PS: Propensity Score

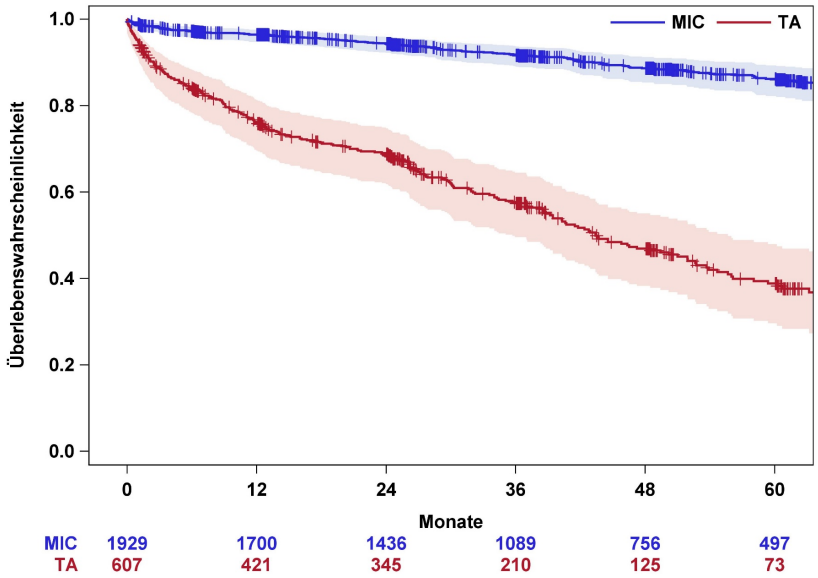

Abbildung 3: Kaplan-Meier-Schätzer für den klinischen Outcome Zeit bis zum Tod im Follow-up in der ursprünglichen PatientInnenpopulation (vor dem PS-Matching) in der Beispielstudie. Das zugehörige Hazard Ratio [95%-KI] aus einem Cox-Modell ist 6,40 [5,33; 7,69].
(MIC: konventionelle Aortenklappenoperation per Ministernotomie; TA: transapikale, katheterbasierte Aortenklappenimplantation)

Hazard Ratios dargestellt. Berechnet wurden jeweils Cox-Modelle in verschiedenen Varianten bezüglich der eingeschlossenen PatientInnen, Kovariablen und Gewichtungen. In der unadjustierten Analyse, d.h. aus einem Cox-Modell mit Behandlung als einziger Kovariable im vollen Datensatz (vgl. auch die Kaplan-Meier-Schätzer in Abbildung 3), finden wir einen extrem hohen Wert des Hazard Ratio von 6,40 (95%-KI: [5,33; 7,69]) mit einer dramatisch höheren Sterblichkeit in der TA-Gruppe. Dieser ist selbstverständlich nicht kausal, sondern auf die großen strukturellen Unterschiede der beiden PatientInnenpopulationen zurückzuführen. Eine herkömmliche Regressionsadjustierung für die initial festgelegten 23 Kovariablen reduziert das Hazard Ratio bereits beträchtlich auf 1,64 [95%-KI: 1,23; 2,19].

Ein weiterer Rückgang findet sich dann in der primär spezifizierten und aus unserer Sicht validen Analyse eines (für das Matchingstratum) stratifizierten Cox-Modells in der PS-gematchten Population (vgl. auch Abbildung 4 für die Kaplan-Meier-Schätzer), in welcher das Hazard zu versterben in der TA-Gruppe um 25% erhöht ist (Hazard Ratio: 1,25 [95%-KI: 0,79; 1,99]).

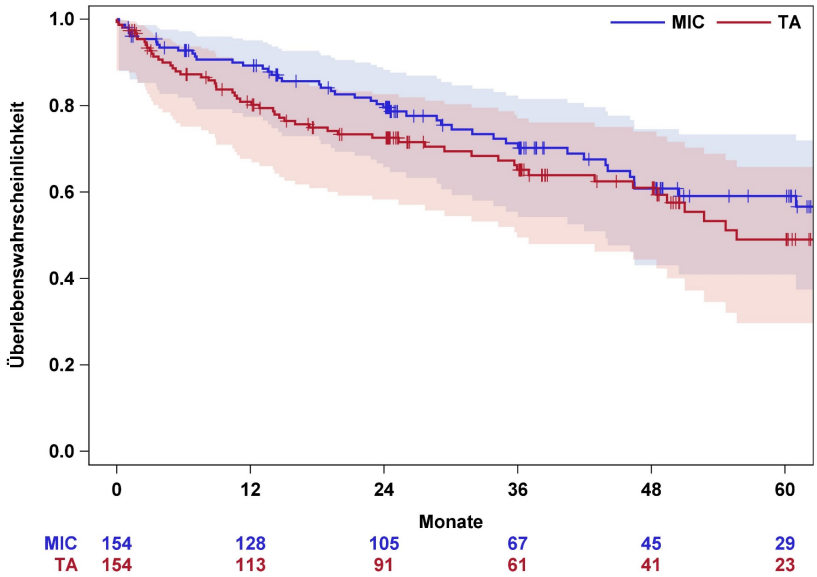

Abbildung 4: Kaplan-Meier-Schätzer für den klinischen Outcome Zeit bis zum Tod im Follow-up in der PS-gematchten Population (nach dem PS-Matching) in der Beispielstudie. Das zugehörige Hazard Ratio [95%-KI] aus einem für das Matchingstratum stratifizierten Cox-Modell ist 1,25 [0,79; 1,99].
(MIC: konventionelle Aortenklappenoperation per Ministernotomie; TA: transapikale, katheterbasierte Aortenklappenimplantation)

Schlussfolgerung und Ausblick

Eine hinreichende/optimale Balanciertheit der relevanten Kovariablen in den Behandlungsgruppen ist das zentrale Qualitätskriterium einer PS-Analyse. In der Regel folgt aus einer guten Balanciertheit der Kovariablen auch ein guter Overlap des PS.

Bei a priori stark unterschiedlichen Behandlungsgruppen, d.h. initial schlechtem Overlap, hat ein PS-Matching den Vorteil, dass es transparent macht, für welche Populationen überhaupt Aussagen bezüglich des Behandlungseffekts gemacht werden dürfen. Dies ist trotz der damit verbundenen Reduktion in der Größe der Stichprobe und des damit einhergehenden Powerverlusts eine Stärke und nicht etwa eine Schwäche des PS-Matchings.

Für die Zukunft erwarten und empfehlen wir die Verwendung von Matching- [17] oder Overlap-Gewichten [18]. Diese haben gegenüber dem PS-Matching und auch anderen Gewichtungsverfahren mathematische Vorteile bezüglich der Effizienz der Parameterschätzung. Im Vergleich zum herkömmlichen PS-Matching haben sie zudem den Vorteil, dass keine Beobachtungen gelöscht werden. Im Vergleich zur Standardgewichtung mit inverse probability of treatment (IPTW)-Gewichten werden bei diesen neuen Gewichtungsvarianten Beobachtungen mit „ungewöhnlichen“ Werten des PS (hier: PatientInnen in der TA-Gruppe, die aber eine hohe Wahrscheinlichkeit für eine MIC haben, und umgekehrt) nicht *herauf*-, sondern *herunter*gewichtet. Damit vermeidet man in Situationen mit initial schlechtem Overlap extreme Gewichte, wie sie bei der Standard-IPTW-Gewichtung häufig vorkommen [17] und dann sogar zum kompletten nummerischen Zusammenbruch der Schätzverfahren führen können. In unserem Datenbeispiel (vgl. Tabelle 2) wird zwar ein IPTW-Schätzer berechnet, es finden sich aber maximale Gewichte von einzelnen Beobachtungen von über 200. Das heißt, es gibt Beobachtungen, die mit mehr als dem 200-fachen statistischen Gewicht in die Analyse eingehen. Das sind aber gerade solche Fälle, bei denen eine gänzlich unerwartete Behandlung durchgeführt worden ist, die also auch mit einem gewissen Risiko fehlklassifiziert sein könnten. In Situationen mit extremen Gewichten wird häufig eine Trunkierung der Gewichte vorgeschlagen (s. z.B. [19]), d.h. es werden Beobachtungen mit extremen Gewichten ausgeschlossen. Dies ist prinzipiell möglich, bringt aber z.B. das Problem mit sich, dass nicht offensichtlich klar ist, ab welchem Perzentil oder ab welcher Größe des Gewichts Beobachtungen ausgeschlossen werden sollen.

Da eine Analyse mit Matching-Gewichten asymptotisch äquivalent zum herkömmlichen PS-Matching ist und auch Matching- und Overlap-Gewichte sehr ähnlich sind, überrascht es nicht, dass alle drei Verfahren zu sehr ähnlichen Schätzern des Hazard Ratio führen. Der Effizienzgewinn der beiden modernen Gewichtungsverfahren ist allerdings beträchtlich; die jeweiligen Konfidenzintervalle sind bedeutend schmaler als beim PS-Matching.

Ein wesentlicher Teil der hier gemachten Empfehlungen basiert auf der Verwendung der SSQ_{zDiff} als globales Balanciertheitsmaß. Bisher liegen keine Erkenntnisse zur Validität der SSQ_{zDiff} vor; unsere Empfehlungen gründen sich im Wesentlichen auf die positive Erfahrung, die wir mit dieser in der praktischen Anwendung gemacht haben. Es existieren aber erste, bisher noch unveröffentlichte Erkenntnisse, dass die genannte Chi-Quadrat-Verteilung auch für mittlere Korrelationen zumindest bezüglich des Erwartungswerts noch gültig ist. Wir arbeiten des Weiteren an der Herleitung der Verteilung der SSQ_{zDiff} im allgemeineren Fall von korrelierten z-Differenzen und planen auch, die empirischen Korrelationen der z-Differenzen in den

Individualdaten großer RCTs zu prüfen [20]. Bei Vorliegen einer Verteilung für die SSQ_{zDiff} wird es in Zukunft auch möglich sein, Konfidenzintervalle für diese anzugeben, um die Abweichungen der geschätzten SSQ_{zDiff} von den Referenzpunkten eines RCT (SSQ_{zDiff}=k) und einer optimal gematchten PS-Analyse (SSQ_{zDiff}=k/2) noch besser einordnen zu können.

Das vorgeschlagene Vorgehen zur Minimierung der SSQ_{zDiff} durch eine datengestützte Minimierung der Caliperweite hat einen gewissen Ad-hoc-Charakter. Es ist nicht automatisch gewährleistet, dass dieses Vorgehen zu einem Effektschätzer führt, der einen guten Kompromiss zwischen guter Balanciertheit und hinreichend großer Fallzahl darstellt. Im Gegenteil, es wäre auch denkbar, dass durch die Minimierung der SSQ_{zDiff} ein PatientInnenkollektiv zur Modellierung benutzt wird, das zu wenig extern valide ist. Hier wären in der Zukunft theoretische Überlegungen oder Simulationsuntersuchungen hilfreich.

Eine bekannte Limitation des PS-Matchings ist, dass damit nur der „Average treatment effect in the treated“ (ATT) geschätzt werden kann. Dieser entspricht dem Effekt in einer Population, deren Kovariablenverteilung gleich dem der behandelten Population entspricht. Im vorliegenden Beispiel aus der Herzchirurgie könnte aber auch der Behandlungseffekt in der gesamten Gruppe der Menschen, die eine neue Aortenklappe benötigen, von Interesse sein. Dieser Effekt wäre dann der „Average treatment effect“ (ATE) und könnte für PatientInnen relevant sein, die sich autonom für eine der beiden Behandlungsmöglichkeiten entscheiden möchten. Eine Schätzung des ATE kann z.B. mit IPTW-Gewichten erfolgen [21].

Durch die Verwendung von Overlap- und Matching-Gewichten ergibt sich auch bezüglich der Populationen, für die die berechneten Effektschätzer gültig sind, eine neue Sichtweise. Mit diesen Gewichten wird der Effekt in der Population geschätzt, für die die Verteilung der Kovariablen in beiden Behandlungsgruppen identisch ist und beide Behandlungen möglich sind („Average treatment effect in the overlap population“ (ATO)). Damit entspricht diese Population der eines RCTs, in dem für alle PatientInnen beide Behandlungen gleich legitim sind und für die Behandlungsempfehlungen eigentlich am notwendigsten sind [18].

Zusammenfassend stellen PS-Analysen eine valide Methode zur Auswertung von nichtrandomisierten Studien dar. Die Validität einer solchen Analyse hängt allerdings wesentlich von der Balanciertheit der PatientInnenmerkmale in der PS-spezifischen Analyse und dem Overlap des PS in beiden Behandlungsgruppen ab. Nur bei Vorliegen von Balanciertheit und hinreichend Overlap können Behandlungseffekte unverzerrt geschätzt und kann die Population identifiziert werden, für die diese Behandlungseffekte gelten.

Anmerkungen

Finanzielle Unterstützung

Diese Arbeit wurde nicht extern gefördert. Das Deutsche Diabetes-Zentrum wird vom Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen und vom Bundesministerium für Gesundheit finanziert.

Interessenkonflikte

Die AutorInnen erklären, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel haben.

Literatur

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41-55.
2. Kuss O, Blettner M, Börgermann J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. Dtsch Arztebl Int. 2016 Sep 5;113(35-36):597-603. DOI: 10.3238/arztebl.2016.0597
3. Oakes JM, Johnson PJ. Propensity score matching methods for social epidemiology. In: Oakes JM, Kaufman JS, editors. Methods in Social Epidemiology. San Francisco: Jossey-Bass/Wiley; 2006. p. 364-86.
4. Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010 Feb 1;25(1):1-21. DOI: 10.1214/09-STS313
5. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf. 2005 Apr;14(4):227-38. DOI: 10.1002/pds.986
6. 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 Mar;20(3):317-20. DOI: 10.1002/pds.2074
7. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. J R Stat Soc Ser A-Statistics Soc. 2008;171:481-502.
8. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009 Nov 10;28(25):3083-107. DOI: 10.1002/sim.3697
9. Cohen J. Statistical power analysis for the behavioral sciences. Toronto: Academic Press, Inc.; 1977. The t Test for Means. p. 19-74.
10. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. Commun Statistics-Simulation Comput. 2009;38(6):1228-34.
11. Kuss O. The z-difference can be used to measure covariate balance in matched propensity score analyses. J Clin Epidemiol. 2013 Nov;66(11):1302-7. DOI: 10.1016/j.jclinepi.2013.06.001
12. Rubin DB, Thomas N. Characterizing the effect of matching using linear propensity score methods with normal distributions. Biometrika. 1992;79(4):797-809.

13. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics. 1996 Mar;52(1):249-64.

14. Filla T, Schwender H, Kuss O. Measuring covariate balance in weighted propensity score analyses by the weighted z-difference [Preprint]. arXiiv. 2022. DOI: 10.48550/arXiv.2212.09490

15. Furukawa N, Kuss O, Emmel E, Scholtz S, Scholtz W, Fujita B, Ensminger S, Gummert JF, Börgermann J. Minimally invasive versus transapical versus transfemoral aortic valve implantation: A one-to-one-to-one propensity score-matched analysis. J Thorac Cardiovasc Surg. 2018 Nov;156(5):1825-34. DOI: 10.1016/j.jtcvs.2018.04.104

16. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007 Jan 15;26(1):20-36. DOI: 10.1002/sim.2739

17. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. Int J Biostat. 2013 Jul 31;9(2):215-34. DOI: 10.1515/ijb-2012-0030

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

19. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I; "on behalf of" the topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. Stat Med. 2020 Dec 30;39(30):4922-48. DOI: 10.1002/sim.8741

20. Kuss O, Miller M. Unknown confounders did not bias the treatment effect when improving balance of known confounders in randomized trials. J Clin Epidemiol. 2020 Oct;126:9-16. DOI: 10.1016/j.jclinepi.2020.06.012

21. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424. DOI: 10.1080/00273171.2011.568786

Korrespondenzadresse:
Dr. Oliver Kuß
Institut für Biometrie und Epidemiologie, Deutsches Diabetes-Zentrum (DDZ), Auf'm Hennekamp 65, 40225 Düsseldorf, Deutschland
oliver.kuss@ddz.de

Bitte zitieren als
Kuß O, Strobel A. Gütemaße und Kriterien bei der Anwendung von Propensity Scores. GMS Med Inform Biom Epidemiol. 2024;20:Doc01. DOI: 10.3205/mibe000257, URN: urn:nbn:de:0183-mibe0002573

Artikel online frei zugänglich unter
https://doi.org/10.3205/mibe000257

Veröffentlicht: *05.01.2024*