Alternative Ansätze für Confounder-Adjustierung durch Gewichtung auf der Grundlage des PropensityScores

Alternative approaches for confounding adjustment using weighting based on the propensity score

Abstract

Propensity score methods (PSMs) such as matching and weighting have become a mainstay for the estimation of treatment effects from observational data. PSMs require the choice of a treatment effect of interest given a specific target population or subpopulation. The most commonly used treatment effects in observational studies are the average treatment effect (ATE) and the average treatment effect among the treated population (ATT). While the traditional matching approach usually discards observations that cannot be matched and is therefore mainly used for estimation of the ATT, the weighting approach utilizes all observations in most cases and can be used for estimation of both, the ATE and ATT. One concern when using the weighting approach is the occurrence of large weights which may disproportionately influence the results and yield estimates with high variance. Recently several newer approaches, including matching weights and overlap weights have been proposed to overcome these limitations. However, this may complicate the interpretation of the results as the population to be inferred depends on the PS estimate itself.

Edin Basic[1]

1 Takeda Pharma Vertrieb GmbH & Co. KG, Berlin, Deutschland

Zusammenfassung

Propensity-Score-Methoden (PSMs) wie Matching und Gewichtung sind zu einem Eckpfeiler der Schätzung von Behandlungseffekten aus Beobachtungsdaten geworden. PSMs erfordern die Wahl eines interessierenden Behandlungseffekts unter Berücksichtigung einer bestimmten Zielpopulation oder Subpopulation. Die am häufigsten verwendeten Behandlungseffekte in Beobachtungsstudien sind der durchschnittliche Behandlungseffekt (average treatment effect (ATE)) und der durchschnittliche Behandlungseffekt für die Gruppe der Personen, die die Behandlung erhalten haben (average treatment effect among the treated population (ATT)). Während der traditionelle Matching-Ansatz in der Regel Beobachtungen, die nicht gematcht werden können, ausschließt und daher hauptsächlich für die Schätzung des ATT verwendet wird, verwendet der Gewichtungsansatz in den meisten Fällen alle Beobachtungen und kann sowohl für die Schätzung des ATE als auch des ATT verwendet werden. Ein Problem bei der Verwendung des Gewichtungsansatzes ist das Auftreten von großen Gewichten, die die Ergebnisse unverhältnismäßig stark beeinflussen und zu Schätzungen mit hoher Varianz führen können. In letzter Zeit wurden jedoch mehrere neuere Ansätze vorgeschlagen, darunter Matching- und Overlap-Gewichte, um diese Einschränkungen zu überwinden. Allerdings kann dies eine erschwerte Interpretation der Ergebnisse zur Folge haben, da die Population, über die eine Aussage getroffen werden soll, von der Propensity-Score-Schätzung selbst abhängt.

Einleitung

Randomisierte kontrollierte Studien (RCT) gelten in der klinischen Forschung als der Goldstandard für die Überprüfung der Wirksamkeit und Sicherheit von neuen Interventionen oder Behandlungen. Sie spielen auch eine entscheidende Rolle für die Bewertung von Gesundheitstechnologien (Health Technology Assessment, HTA). Der Grund hierfür liegt in der Tatsache, dass erwartungsgemäß die Randomisierung die Vergleichbarkeit der zu vergleichenden Gruppen in Bezug auf die bekannten als auch unbekannten Patientencharakteristika gewährleistet. Seit der Verabschiedung des Gesetzes für mehr Sicherheit in der Arzneimittelversorgung (GSAV) im Jahre 2020 können allerdings in Deutschland unter gewissen Voraussetzungen auch versorgungsnahe Daten (Real World Data (RWD)) für die Nutzenbewertung herangezogen werden. Die Verwendung von RWD für die Nutzenbewertung ist jedoch mit einigen methodischen Herausforderungen verbunden. Beispielsweise kann aufgrund der fehlenden Randomisierung nicht gewährleistet werden, dass keine systematischen Unterschiede zwischen den zu vergleichenden Gruppen bestehen. Bei systematischer Ungleichheit würde ein direkter Vergleich der Outcomes zwischen den Behandlungsgruppen zu einer verzerrten Schätzung des Behandlungseffekts führen. Propensity-Score-Methoden (PSM) wie Matching und Gewichtung sind zu einem Eckpfeiler der Schätzung von Behandlungseffekten aus RWD geworden. In diesem Aufsatz beschreiben wir die Implementierung der Propensity-Score-Gewichtungsansätze zusammen mit den grundlegenden Eigenschaften der einzelnen Ansätze.

Propensity Score und Definition von Behandlungseffekten

Der Propensity Score (PS) ist ein Balancing-Score und wird definiert als die auf beobachteten Kovariaten beruhende bedingte Wahrscheinlichkeit, mit der ein Patient die interessierende Behandlung erhält. Auf Basis des PS kann gleichzeitig eine große Anzahl an Kovariaten zwischen der Behandlungs- und der Kontrollpopulation angeglichen werden. Beim Matching wird die Ausbalanciertheit der Kovariaten dadurch sichergestellt, dass die zu vergleichenden Gruppen im Durchschnitt vergleichbare PS haben, d.h. der PS ist innerhalb der gematchten Paare ausbalanciert [1]. Im Gegensatz dazu wird bei Gewichtungsmethoden auf Basis einer Funktion des PS eine Neu-Gewichtung der zu vergleichenden Gruppen vorgenommen, so dass in dieser gewichteten Population die Gruppenzugehörigkeit unabhängig von den Kovariaten ist, die für die Schätzung von PS verwendet wurden [2].
Es lassen sich verschiedene Behandlungseffekte in Beobachtungsstudien definieren, je nachdem, wie die Behandlung definiert und welche Population betrachtet wird. Am häufigsten werden der durchschnittliche Behandlungseffekt (average treatment effect (ATE)) und der durchschnittliche Behandlungseffekt für die Gruppe der Personen, die die Behandlung erhalten haben (average treatment effect among the treated population (ATT)) verwendet. Der ATE beschreibt den Fall, dass alle Patienten der Population behandelt versus nicht oder alternativ behandelt würden, während der ATT den Fall beschreibt, dass alle tatsächlich behandelten Patienten behandelt versus nicht oder alternativ behandelt würden. Bei fehlender Heterogenität des Behandlungseffekts durch Kovariaten (d.h. keine Interaktionseffekte zwischen Behandlung und Kovariaten) sind ATE und ATT identisch, wenn die Effekte additiv sind. Besteht jedoch Behandlungsheterogenität, unterscheiden sich ATE und ATT [3]. Beispielsweise dürfte der Effekt von langwirksamen Beta-Agonisten (LABA) auf die Asthma-Mortalität bei Patienten mit schwerem Asthma stärker ausgeprägt sein als bei Patienten mit nur leichtem Asthma. Bei Vorliegen einer solchen Behandlungsheterogenität wird der ATT sehr unterschiedlich zum ATE ausfallen, auch wenn kein Unterschied in der Behandlung (LABA) besteht.
Es kann auch Situationen geben, in denen das Interesse in der Schätzung von ATE in der Subgruppe der Patienten mit sehr ähnlichen Charakteristika, sog. klinischem Equipoise besteht. Darüber hinaus lassen sich noch zahlreiche andere Behandlungseffekte definieren. Welcher Behandlungseffekt schließlich von Interesse ist, hängt von der jeweiligen Fragestellung und den vorhandenen Daten ab.
Um verschiedene Behandlungseffekte in Beobachtungsstudien schätzen zu können, müssen einige Annahmen getroffen werden. Die erste Annahme bezieht sich auf die bedingte Austauschbarkeit oder die Annahme, dass es kein unbeobachtetes Confounding gibt. Diese Annahme impliziert, dass nur die beobachtbaren Kovariaten einen Einfluss darauf haben, ob ein Patient behandelt wurde oder nicht. Darüber hinaus ist es nötig, Collider und Mediatoren (und deren Nachfahren) aus der Analyse auszuschließen. Hierzu kann das Pearl's Back-Door-Kriterium verwendet werden, welches die grafische Repräsentation kausaler Beziehungsgeflechte ermöglicht [4]. Die zweite Annahme ist als Konsistenz-Annahme bekannt und besagt, dass keine Zusammenhänge zwischen den verschiedenen Patienten und auch keine allgemeinen Gleichgewichtseffekte existieren, da das Outcome eines bestimmten Patienten nicht vom Behandlungsstatus anderer Patienten abhängt. Eine weitere Annahme ist die Overlap-Annahme, welche besagt, dass Patienten mit denselben Kovariatenwerten (sowohl alle Werte einzeln als auch in jeder vorkommenden Kombination der Werte) eine positive Wahrscheinlichkeit haben, sowohl behandelt als auch nicht behandelt zu werden. Eine detaillierte Beschreibung dieser Annahmen findet sich bei Hernan und Robins ([5], S. 25-36).

Korrekte Spezifikation des PS-Modells

In der Praxis ist der PS unbekannt und muss daher anhand der verfügbaren Daten geschätzt werden. Bei der

Schätzung des PS muss ein statistisches Modell festgelegt sowie eine Auswahl der beobachteten Kovariaten, die in das Modell aufgenommen werden, getroffen werden. Die logistische Regression ist das am häufigsten verwendete Modell zur Schätzung des PS. Es können jedoch auch flexiblere Modelle wie Boosting, Random Forests oder verallgemeinerte additive Modelle für die PS-Schätzung verwendet werden [6], [7]. Die Schätzung des ATE und ATT beruht auf der Annahme der bedingten Unabhängigkeit, die voraussetzt, dass das Outcome unabhängig von der Behandlung ist, gegeben der PS. Die Schätzung des PS erfordert daher die Auswahl von Kovariaten, die diese Bedingung erfüllen. Daher müssen alle beobachteten Confounder, d.h. Kovariaten, die gleichzeitig die Behandlung und das Outcome beeinflussen, in das PS-Modell aufgenommen werden. Andererseits gibt es keinen einheitlichen Konsens über die Berücksichtigung von Kovariaten, die das Outcome, aber nicht die Behandlung beeinflussen, oder von Kovariaten, die nur die Behandlung beeinflussen. Beispielsweise zeigen Brookhart et al., dass die Berücksichtigung von Kovariaten, die mit dem Outcome, aber nicht mit der Behandlung zusammenhängen, die Varianz des geschätzten Behandlungseffekts verringern können, ohne die Verzerrung zu erhöhen [8]. Im Gegensatz dazu kann die Berücksichtigung von Kovariaten, die mit der Behandlung, aber nicht mit dem Outcome zusammenhängen, die Varianz des geschätzten Behandlungseffekts erhöhen und bei Vorliegen von ungemessenem Confounding die Verzerrung verstärken [9], [10]. In der Praxis kann die Unterscheidung zwischen den obigen Variablen eine Herausforderung darstellen. Eine Abhilfe hierzu bieten die sog. kausalen gerichteten azyklischen Graphen (DAG). Sie sind besonders hilfreich bei der Abgrenzung und dem Verständnis von Störgrößen und potenziellen Ursachen für Verzerrungen in Behandlungs-Outcome-Beziehungen [11]. Weiterhin schlagen Schneeweiss et al. vor, das hochdimensionale PS-Modell zu verwenden, um das Risiko von ungemessenem Confounding weiter zu verringern [12]. Der hochdimensionale PS bezieht sich auf die Verwendung einer großen Anzahl von Kovariaten, die als Proxies für unbeobachtete Confounder dienen. Es sollte jedoch bedacht werden, dass die besten Prädiktionsmodelle nicht unbedingt hilfreich sind, da dies zu einer Verletzung der Overlap-Bedingung führen könnte. Außerdem ist bei seltenen Krankheiten mit einer geringen Anzahl behandelter Patienten die Anzahl der Kovariaten, die für die PS-Schätzung verwendet werden können, ebenfalls begrenzt. Die aufgeführten Punkte implizieren, dass kein allumfassendes Konzept verfügbar ist, sondern wissenschaftliches Verständnis, frühere empirische Erkenntnisse und die vorliegende Datenstruktur eine entscheidende Rolle bei der Auswahl der Variablen für das PS-Modell spielen können.

Evaluation der Überlappung der Propensity-Score-Verteilung zwischen Behandlungs- und Referenzgruppe

Nach der Schätzung des PS sollte die Überlappung der Verteilungen der geschätzten PS für die Behandlungs- und die Kontrollgruppe grafisch untersucht werden. Hierzu gibt es drei mögliche Szenarien:

1. Die PS-Verteilungen können sich sowohl in Bezug auf den Wertebereich als auch auf die Dichte sehr gut überlappen, was bedeutet, dass beide Gruppen in Bezug auf die gewählten Kovariaten-Verteilungen sehr ähnlich sind.
2. Es existiert kaum Überlappung zwischen den PS-Verteilungen.
3. Die PS-Verteilungen überschneiden sich bis zu einem gewissen Grad.

Das erste Szenario, das im Rahmen einer randomisierten kontrollierten Studie zu erwarten ist, ist für die meisten Beobachtungsstudien sehr unwahrscheinlich. Das zweite Szenario kann darauf hindeuten, dass zu viele Unterschiede zwischen den behandelten Personen und den Vergleichspersonen bestehen, um eine kausale Aussage treffen zu können. Das dritte Szenario ist das in der Praxis am häufigsten anzutreffende Szenario. In der praktischen Anwendung können geschätzte PS nahe 0 und nahe 1 ein Hinweis auf eine unzureichende Überlappung bzw. auf die Verletzung der Overlap-Annahme sein.

Überlegungen bei der Auswahl einer Gewichtungsmethode

Basierend auf der Wahl des interessierenden Behandlungseffekts existieren verschiedene Gewichtungsansätze, um die zu vergleichenden Behandlungsgruppen in Bezug auf die beobachteten Confounder anzugleichen.

Gewichtungsansätze zur Schätzung von ATE

Es existieren zwei Gewichtungsansätze, *inverse probability treatment weighting* (IPTW) und *fine stratifcation weights* [13], für die Schätzung von ATE in Beobachtungsstudien. Beiden Ansätzen ist gemeinsam, dass sie versuchen, die Verteilung der Kovariaten in der Behandlungs- und der Kontrollgruppe anzugleichen, so dass sie der Verteilung in der Gesamtstichprobe entsprechen.

Inverse probability treatment weighting (IPTW)

Bei diesem Ansatz entsprechen die Gewichte der inversen Wahrscheinlichkeit, die tatsächliche Behandlung erhalten zu haben. Demzufolge erhalten Patienten aus der Behandlungsgruppe das Gewicht 1/PS und die aus der Referenz-

gruppe das Gewicht 1/(1–PS). Dies bedeutet, dass Patienten, die der Behandlungsgruppe zugeordnet wurden, obwohl sie nach ihren Charakteristika viel wahrscheinlicher der Kontrollgruppe hätten zugeordnet werden müssen, ein hohes Gewicht bekommen. Auf der anderen Seite erhalten Patienten, die der erwarteten Behandlungsgruppe zugeordnet wurden, ein kleines Gewicht.
Aufgrund der direkten Verwendung des PS zur Berechnung der Gewichte können extreme Gewichte vorkommen. Dies geschieht, wenn der geschätzte PS nahe 0 oder 1 ist. In einem solchen Fall können einzelne Patienten einen sehr hohen Einfluss auf die Schätzung haben. Hierbei sollte zuerst überprüft werden, ob diese Patienten überhaupt zur Zielpopulation gehören oder stattdessen als Ausreißer betrachtet und aus der Stichprobe entfernt werden sollten. Darüber hinaus können extreme Gewichte entweder durch Stabilisierung, Trunkierung oder Trimming behandelt werden. Eine Gewicht-Stabilisierung kann erreicht werden, indem der Zähler – der bei den nicht stabilisierten Gewichten 1 ist – durch die marginale Behandlungswahrscheinlichkeit ersetzt wird [14]. Als weitere Maßnahme können extreme Gewichte auch durch Trunkierung behandelt werden [15]. Hierzu werden besonders kleine PS (z.B. <0.01 oder <0.001) durch den Trunkierungswert ersetzt, um hohe Gewichte zu verhindern. Dies erzeugt eine verzerrte Schätzung des Behandlungseffekts, kann aufgrund potentiell geringerer Varianz aber zu einem besseren mittleren quadratischen Fehler (MSE) führen. Schließlich besteht noch die Möglichkeit, ein Trimming durchzuführen, d.h. alle Beobachtungen mit einem PS unter dem unteren und über dem oberen PS-Grenzwert auszuschließen. Dabei existieren mehrere Methoden, um die Grenzwerte zu definieren:

1. Bei der Common-Range-Methode wird als unterer Grenzwert der niedrigste PS bei den Behandelten und als oberer Grenzwert der höchste PS bei den Patienten aus der Kontrollgruppe genommen.
2. Bei der Stürmer-Methode entspricht der untere Grenzwert dem 5%-Perzentil der PS-Verteilung der Behandelten und der obere Grenzwert dem 95%-Perzentil der Nicht-Behandelten.
3. Die Walker-Methode geht von einem unteren Grenzwert von 0.3 des sog. Präferenzwertes und einem oberen von 0.7 aus, wobei höhere Werte eine höhere Präferenz für die Behandlung bei gemessenen Kovariaten widerspiegeln und der Logit des Präferenzwertes als der Logit des PS minus der Logit der Behandlungsprävalenz definiert ist.
4. Bei der Crump-Methode wird der untere Grenzwert auf 0.1 und der obere auf 0.9 festgelegt [15], [16], [17], [18].

Hierbei ist zu beachten, dass bei Ausschluss von Beobachtungen generell die Gefahr von Selektionsbias besteht. Nach der Berechnung der Gewichte werden für die Schätzung des Behandlungseffekts Patienten gemäß ihres Gewichts verwendet. Dabei wird entsprechend dem interessierenden Outcome ein gewichtetes statistisches Modell spezifiziert, z.B. eine gewichtete logistische Regression bei binären Outcomes bzw. ein gewichtetes Cox-Regressionsmodell bei Time to Event Outcomes. Hierbei ist zu beachten, dass durch die Gewichtung eine „Pseudopopulation“ generiert wird, die „Replikationen“ von Patienten enthält, wodurch eine Korrelation innerhalb der Patienten induziert wird. Dieser Mangel an Unabhängigkeit muss berücksichtigt werden, um die Varianz korrekt zu schätzen. Zusätzlich muss auch berücksichtigt werden, dass die Gewichte eine Variation aufweisen, da sie auf dem geschätzten PS basieren. Dies kann bei der Varianzschätzung entweder durch die Verwendung eines robusten „Sandwich“-Varianzschätzers oder durch Bootstrap-basierte Methoden berücksichtigt werden [14].

Fine stratification weights

Bei diesem Ansatz werden die geschätzten PS nicht zur Berechnung der Gewichte, sondern zur Bildung der sog. Feinschichten verwendet. Es gibt mehrere Möglichkeiten, die Schichten auf Basis von PS zu bilden:

1. Die PS-Verteilung der gesamten Kohorte wird verwendet.
2. Die PS-Verteilung der kleineren der beiden zu vergleichenden Gruppen wird verwendet.
3. Eine feste Breite der PS (z.B. 0–0.1 Schicht 1, >0.1–0.2 Schicht 2, ... >0.9–1) wird verwendet.

Bei einer niedrigen Behandlungsprävalenz gewährleistet die Bildung von Schichten auf Basis der PS-Verteilung der behandelten Patienten die Zuordnung aller Patienten zu den jeweiligen Schichten. Nach der Bildung von Schichten werden die Gewichte für beide Gruppen in allen Schichten berechnet. Hierzu wird in einem ersten Schritt innerhalb der Schichten ein neuer, vereinfachter PS berechnet. Dieser neue PS ergibt sich aus der Anzahl der behandelten Patienten dividiert durch die Gesamtanzahl der Patienten in einer bestimmten Schicht. In einem zweiten Schritt wird der Kehrwert dieses neuen PS gebildet. Anschließend wird dieser Kehrwert mit der marginalen Wahrscheinlichkeit der Gesamtstichprobe behandelt zu werden (Stabilisierungsfaktor bei dem IPTW-Ansatz) multipliziert. Analog hierzu werden die Gewichte für nichtbehandelte Patienten gebildet.
Anschließend erhalten innerhalb jeder Schicht alle behandelten bzw. nicht behandelten Patienten das gleiche Gewicht. Solange keine schwach besetzten Schichten vorkommen, sind extreme Gewichte bei diesem Ansatz unwahrscheinlich [19]. Bezüglich der Varianzschätzung gelten hier die gleichen Überlegungen wie bei IPTW.

Gewichtungsansätze zur Schätzung von ATT

Standardised mortality ratio weighting (SMRW) und *fine stratification weights* werden für die Schätzung von ATT in Beobachtungsstudien herangezogen. Beide Ansätze versuchen die Verteilung der Kovariaten in der Kontrollgruppe an die Kovariaten-Verteilung der Behandlungsgruppe anzugleichen.

Standardised mortality ratio weighting (SMRW)

Bei diesem Ansatz gehen die behandelten Patienten unverändert in die Analyse ein, während die nicht behandelten Patienten so umgewichtet werden, dass sie den Behandelten sehr ähnlich sind. Hierfür wird den Patienten aus der Behandlungsgruppe das Gewicht 1 und denen aus der Kontrollgruppe das Gewicht PS/(1–PS) zugewiesen. Trotz der Verwendung einer Art der Stabilisierung bei der Berechnung der Gewichte für die Kontrollgruppe ist das Auftreten extremer Gewichte möglich. In einem solchen Fall kann Trunkierung oder Trimming verwendet werden. Bezüglich der Varianzschätzung gelten auch hier die gleichen Überlegungen wie bei IPTW.

Fine stratification weights

Ähnlich wie bei den *fine stratification weights* für die Schätzung von ATE werden auch hier PS zur Bildung von Schichten verwendet. Allerdings erhalten hier alle behandelten Patienten das Gewicht 1, während die Patienten aus der Kontrollgruppe ein Gewicht, das im Verhältnis zur Verteilung der behandelten zu den nicht behandelten Patienten in der jeweiligen Schicht steht, bekommen. Durch diese Gewichtung wird eine „Pseudopopulation" generiert, in der die Kovariaten-Verteilung zumindest in jeder Schicht zwischen der Behandlungs- und Kontrollgruppe annähernd übereinstimmt. Bezüglich der Varianzschätzung gelten auch hier die gleichen Überlegungen wie bei IPTW.

Gewichtungsansätze zur Schätzung von ATE in der Subgruppe der Patienten mit sehr ähnlichen Charakteristika, sog. klinischem Equipoise

Die nächsten beiden Gewichtungsansätze, Matching- und Overlap-Gewichte, haben ein variables Ziel der Inferenz, das stark von der Überlappung der PS-Verteilungen der zu vergleichenden Gruppen abhängt [20], [21]. Vereinfacht gesagt wird mit den beiden Ansätzen der ATE in einer Subgruppe der Patienten mit einem gewissen klinischem Equipoise geschätzt. Demzufolge versuchen beide Ansätze die Verteilung der Kovariaten in der Behandlungs- und der Kontrollgruppe so anzugleichen, dass sie der Verteilung der Kovariaten in der Teilmenge der Gesamtstichprobe entsprechen. In dieser haben Patienten aus beiden Gruppen relativ hohe Wahrscheinlichkeiten sowohl behandelt als auch nicht behandelt zu sein.

Matching-Gewichte

Bei diesem Ansatz entsprechen die Gewichte dem Verhältnis der niedrigeren der beiden vorhergesagten Wahrscheinlichkeiten (PS bzw. 1–PS) zur vorhergesagten Wahrscheinlichkeit der tatsächlich erhaltenen Behandlung. Es handelt sich um eine Modifikation der Gewichte aus dem IPTW Ansatz, indem der Zähler der IPTW-Gewichte durch das Minimum aus PS und 1–PS ersetzt wird. Dies bedeutet, dass alle behandelten Patienten mit einem geschätzten PS<0.5 unverändert in die Analyse eingehen, während Patienten aus der Kontrollgruppe mit dem gleichen PS ein Gewicht <1 bekommen. Für Patienten mit einem geschätzten PS>0.5 ist es genau umgekehrt, d.h. hierbei werden alle behandelten Patienten heruntergewichtet, während Patienten aus der Kontrollgruppe unverändert in die Analyse eingehen.
Bei diesem Ansatz können keine extremen Gewichte auftreten, da sie aufgrund des Designs immer zwischen 0 und 1 liegen. Das Ziel der Inferenz ist nahe am ATE in der Gesamtpopulation, wenn die zu vergleichenden Gruppen gleich groß sind und die PS-Verteilungen eine gute Überlappung aufweisen. Im Falle ungleicher Gruppengrößen und einer guten Überlappung der PS-Verteilungen wird der ATT in der kleineren Gruppe geschätzt. Bei einer begrenzten Überlappung der PS-Verteilungen wird der Behandlungseffekt in einer Teilpopulation geschätzt, die weder die Gruppe der behandelten Patienten noch die Gesamtpopulation repräsentiert. Bei der Varianzschätzung sollten robuste „Sandwich"-Varianzschätzer oder Bootstrap-basierte Methoden verwendet werden, da dadurch eine eventuelle Fehlspezifikation des PS-Modells korrigiert werden kann [20], [21].

Overlap-Gewichte

Bei diesem Ansatz entsprechen die Gewichte der Wahrscheinlichkeit die entgegensätzliche Behandlung bekommen zu haben, d.h. Patienten aus der Behandlungsgruppe bekommen das Gewicht 1–PS und Patienten aus der Referenzgruppe das Gewicht PS. Die Verwendung dieser Art von Gewichten bedeutet, dass alle Patienten heruntergewichtet werden. Wie stark Patienten heruntergewichtet werden, hängt von ihrem geschätzten PS ab. So werden bei einem hohen geschätzten PS Patienten aus der Interventionsgruppe deutlich stärker heruntergewichtet als Patienten aus der Kontrollgruppe. Im Gegensatz dazu werden bei einem niedrigen geschätzten PS Patienten aus der Kontrollgruppe stärker heruntergewichtet, während Patienten aus beiden Gruppen, deren geschätzter PS nahe 0.5 liegt, gleich stark heruntergewichtet werden. Dies bedeutet, dass Patienten mit einem geschätzten PS von 0.5 den größten Beitrag zur Analyse leisten. Je weiter sich der PS von 0.5 entfernt, desto kleiner fällt der Beitrag zur Analyse der einzelnen Patienten aus. Ähnlich wie bei dem Matching-Ansatz sind extreme Gewichte unmöglich, da sie per Definition durch 0 und 1 begrenzt sind. Eine weitere attraktive Eigenschaft dieses Ansatzes besteht darin, dass die Kovariaten im Mittel zwischen den Gruppen exakt ausbalanciert sind [21]. Das Ziel der Inferenz bei diesem Ansatz ist der ATE in der Überlappungspopulation, der sich von dem ATE oder dem ATT in der gesamten Studienpopulation unterscheiden kann. Bezüglich der Varianzschätzung gelten hier die gleichen Überlegungen wie beim Matching-Ansatz.

Diskussion

Gewichtungsansätze auf Basis von PS stellen eine flexible Möglichkeit zur Schätzung von Behandlungseffekten in Beobachtungsstudien dar. Im Unterschied zum Matching-Ansatz, mit dem gewöhnlicherweise der ATT geschätzt wird, lassen sich mit den Gewichtungsansätzen verschiedene Behandlungseffekte relativ einfach schätzen. Allerdings ist zu beachten, dass die Gewichtungsansätze eine korrekte Spezifikation des PS-Modells erfordern. Hierbei ist wiederum zu beachten, dass mit der „korrekten" Spezifikation des PS-Modells nicht das vollständige Wissen über den zugrundeliegenden Datengenerierungsprozess, der der Behandlungszuweisung zugrunde liegt, verlangt wird [22]. Vielmehr geht es darum, alle Confounder und deren Einfluss auf die Behandlungszuweisung adäquat im PS-Modell zu berücksichtigen. Die Simulationsergebnisse zur Performance verschiedener Gewichtungsansätze zeigen jeweils eine relativ hohe Sensitivität bzgl. einer Fehlspezifikation des PS-Modells [16], [21]. In der Publikation von Stürmer et al. findet sich jedoch ein Hinweis, dass der IPTW mit anschließendem Trimming zu einer Reduzierung des *unmeasured confounding* führen kann [16].

Ein weiterer wichtiger Punkt bei der Nutzung der Gewichtungsansätze ist das Vorliegen einer ausreichenden Überlappung. Es herrscht kein einheitlicher Konsens darüber, zu welchem Grad sich die beiden PS-Verteilungen überlappen müssen, um die Gewichtungsansätze anwenden zu können. Vor allem geht es hier um die Frage, ab welchem Grad der Nicht-Überlappung die ursprüngliche Fragestellung mit den vorhandenen Daten nicht mehr beantwortet werden kann [18].

Schließlich muss nach der Anwendung der Gewichtungsansätze die Balanciertheit der einzelnen Confounder überprüft werden. Hierzu können die standardisierten Differenzen oder z-Differenzen herangezogen werden [23], [24]. Es sollten aber auch die ungewichteten Differenzen als eine Art Benchmark dargestellt werden.

All diese Punkte implizieren, dass die Schätzung von Behandlungseffekten in Beobachtungsstudien mittels PS einen mehrstufigen Prozess darstellt. Zunächst muss die korrekte Spezifikation des PS-Modells geprüft werden. Anschließend ist das Ausmaß der Überlappung zwischen den zu vergleichenden Behandlungsgruppen zu beurteilen. Wenn eine ausreichende Überlappung vorliegt, muss die Entscheidung bezüglich des interessierenden Behandlungseffekts getroffen werden. Darauffolgend ist der Gewichtungsansatz auszuwählen und die Überlappung der gewichteten PS-Verteilungen sowie die Balanciertheit der einzelnen Confounder zu prüfen.

Anmerkungen

Danksagung

Der Autor dankt Sarah Böhme, Jens-Otto Andreas, Dietrich Knoerzer, Tobias Bluhmki und Friedhelm Leverkus für deren wertvolle und hilfreiche Anmerkungen.

Interessenkonflikte

Edin Basic ist Mitarbeiter der Takeda Pharma Vertrieb GmbH & Co. KG.

Literatur

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41-55. DOI: 10.1093/biomet/70.1.41
2. Rosenbaum PR. Model-based direct adjustment. J Am Stat Assoc. 1987;82:387-94. DOI: 10.1080/01621459.1987.10478441
3. Moodie EE, Saarela O, Stephens DA. A doubly robust weighting estimator of the average treatment effect on the treated. Stat. 2018;7:e205. DOI: 10.1002/sta4.205
4. Pearl J. The Foundations of Causal Inference. Sociological Methodology. 2010;40:75-149.
5. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
6. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods. 2004 Dec;9(4):403-25. DOI: 10.1037/1082-989X.9.4.403
7. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf. 2008 Jun;17(6):546-55. DOI: 10.1002/pds.1555
8. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006 Jun 15;163(12):1149-56. DOI: 10.1093/aje/kwj149
9. Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care. 2010 Jun;48(6 Suppl):S114-20. DOI: 10.1097/MLR.0b013e3181dbebe3
10. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011 Dec 1;174(11):1213-22. DOI: 10.1093/aje/kwr364
11. Williams TC, Bach CC, Matthiesen NB, Henriksen TB, Gagliardi L. Directed acyclic graphs: a tool for causal studies in paediatrics. Pediatr Res. 2018 Oct;84(4):487-93. DOI: 10.1038/s41390-018-0071-3
12. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009 Jul;20(4):512-22. DOI: 10.1097/EDE.0b013e3181a663cc
13. Desai RJ, Rothman KJ, Bateman BT, Hernandez-Diaz S, Huybrechts KF. A Propensity-score-based Fine Stratification Approach for Confounding Adjustment When Exposure Is Infrequent. Epidemiology. 2017 Mar;28(2):249-57. DOI: 10.1097/EDE.0000000000000595

14. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000 Sep;11(5):561-70. DOI: 10.1097/00001648-200009000-00012

15. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika. 2009;96(1):187-99. DOI: 10.1093/biomet/asn055

16. Stürmer T, Webster-Clark M, Lund JL, Wyss R, Ellis AR, Lunt M, Rothman KJ, Glynn RJ. Propensity Score Weighting and Trimming Strategies for Reducing Variance and Bias of Treatment Effect Estimates: A Simulation Study. Am J Epidemiol. 2021 Aug 1;190(8):1659-70. DOI: 10.1093/aje/kwab041

17. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution – a simulation study. Am J Epidemiol. 2010 Oct 1;172(7):843-54. DOI: 10.1093/aje/kwq198

18. Walker AM, Patrick AR, Lauer MS, Hornbrook MC, Marin MG, Platt R, Roger VL, Stang P, Schneeweiss S. A tool for assessing the feasibility of comparative effectiveness research. Comp Eff Res. 2013;2013(3):11-20. DOI: 10.2147/CER.S40357

19. Franklin JM, Eddings W, Austin PC, Stuart EA, Schneeweiss S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. Stat Med. 2017 May 30;36(12):1946-63. DOI: 10.1002/sim.7250

20. 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

21. Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights. Am J Epidemiol. 2019 Jan 1;188(1):250-7. DOI: 10.1093/aje/kwy201

22. Chakraborty B, Moodie E. Statistical methods for dynamic treatment regimes. New York: Springer; 2013.

23. Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. The American Statistican. 1985 Feb;39(1):33-8. DOI: 10.2307/2683903

24. 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

Korrespondenzadresse:
Dr. Edin Basic
Takeda Pharma Vertrieb GmbH & Co. KG, Potsdamer Str. 125, 10783 Berlin, Deutschland
Edin.Basic@takeda.com

Bitte zitieren als
Basic E. Alternative Ansätze für Confounder-Adjustierung durch Gewichtung auf der Grundlage des PropensityScores. GMS Med Inform Biom Epidemiol. 2024;20:Doc02.
DOI: 10.3205/mibe000258, URN: urn:nbn:de:0183-mibe0002583

Artikel online frei zugänglich unter
https://doi.org/10.3205/mibe000258

Veröffentlicht: *05.01.2024*