Workshop der Arbeitsgruppe *"Therapeutische Forschung"* auf der GMDS 2022

Donnerstag, 17.11.2022

# Praktische Aspekte bei der Anwendung von Propensity Scores

Ralf Bender, Anika Großhennig, Friedhelm Leverkus

Bei einigen neuen Arzneimitteln sind die Studiendaten zum Zeitpunkt der Zulassung für die Nutzenbewertung nur wenig aussagekräftig. Nach dem Gesetz für mehr Sicherheit in der Arzneimittelversorgung (GSAV) von 2019 kann der Gemeinsame Bundesausschuss (G-BA) vom pharmazeutischen Unternehmer verlangen, dass während der Behandlung mit diesem Arzneimittel in der täglichen Praxis Daten gesammelt und ausgewertet werden. Diese sogenannte "anwendungsbegleitende Datenerhebung" (abD) soll dazu beitragen, mehr Informationen über den Nutzen und Schaden des neuen Arzneimittels zu gewinnen um eine Nutzenbewertung zu ermöglichen. Die abD soll damit Informationen liefern, die über die bereits bekannten Studiendaten aus der Zulassung hinausgehen.

Im Gegensatz zu den üblicherweise randomisierten Zulassungsstudien, sind für eine abD insbesondere nicht randomisierte Studien vorgesehen. Bei nicht randomisierten Studien ergibt sich die zwingende Notwendigkeit einer adäquaten Adjustierung aller potenziellen Confounder. Daher haben für abDs die für Beobachtungsstudien entwickelten Methoden zur Confounder-kontrolle eine grundlegende Bedeutung, und zwar insbesondere die Verfahren auf Basis von Propensity Scores. Bei der Anwendung von Methoden auf Basis von Propensity Scores spielen eine ausreichende Positivität, Überlappung und Balanciertheit eine zentrale Rolle.

In diesem Workshop sollen relevante praktische Aspekte bei der Anwendung von Propensity Scores vorgestellt und erläutert werden. Insbesondere soll dabei auf verfügbare Methoden zur Überprüfung von Positivität, Überlappung und Balanciertheit eingegangen werden. Im Anschluss an die Vorträge werden dann in einer Podiumsdiskussion diese Verfahren hinsichtlich Ihrer konkreten Anwendung in der Nutzenbewertung diskutiert.

Programm:

| | |
|---|---|
| 10:00 – 10:10 | Begrüßung |
| 10:10 – 10:25 | Thomas Kaiser (IQWiG, Köln): *„Anwendungsbegleitende Datenerhebungen (AbD): Warum, wann und wie?*“ |
| 10:25 – 10:45 | Oliver Kuß (DDZ, Düsseldorf): *„Einführung in Propensity Scores“* |
| 10:45 – 11:15 | Tim Mathes (UMG, Göttingen): *„Anforderungen an die Daten: Eine Diskussion anhand von Patientenregistern“* |
| 11:15 – 11:30 | Kaffeepause |
| 11:30 – 12:00 | Edin Basic (Takeda, Berlin): *„Alternative Ansätze für Confounder-Adjustierung durch Gewichtung auf der Grundlage des Propensity-Score“* |
| 12:00 – 12:30 | Oliver Kuß (DDZ, Düsseldorf): *„Gütemaße und Kriterien bei der Anwendung von Propensity Scores“* |
| 12:30 – 13:00 | Diskussion |
| 13:00 – 13:30 | AG-Sitzung |

# Abstracts

## Anwendungsbegleitende Datenerhebungen (AbD): Warum, wann und wie?

Thomas Kaiser
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln

Mit dem Gesetz für mehr Sicherheit in der Arzneimittelversorgung (GSAV) wurde 2019 die anwendungsbegleitende Datenerhebung (AbD) in das Gesamtverfahren der frühen Nutzenbewertung nach §35a SGB V integriert. Damit wurde dem Gemeinsamen Bundesausschuss (G-BA) die Möglichkeit eröffnet, eine solche AbD für Arzneimittel für seltene Erkrankungen (Orphan Drugs), Arzneimittel mit bedingter Zulassung sowie unter besonderen Umständen zugelassenen Arzneimitteln zu fordern. Adressat dieser Forderung ist der zuständige pharmazeutische Unternehmer, der auch die Kosten tragen muss.

Auch vor Einführung der AbD konnte der G-BA bereits weitere Daten im Rahmen einer Befristungsauflage fordern. Mit der AbD sind jedoch einige Besonderheiten verbunden, insbesondere:

- die Beschränkung auf die oben genannten Arzneimittel,
- die Beschränkung auf Studien ohne Randomisierung,
- die Möglichkeit, die Verordnungsbefugnis auf diejenigen Ärztinnen und Ärzte zu beschränken, die an der AbD teilnehmen, sowie
- die Sanktionierung bei Nichtdurchführung einer AbD (Abschläge vom Erstattungsbetrag).

Im Vortrag wird dargelegt, warum die AbD in Deutschland eingeführt wurde, wann eine solche AbD prinzipiell beauflagt werden kann und wie diese dann durchzuführen ist.

## Einführung in Propensity Scores

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

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 häufig und zurecht für ihre fehlende externe Validität kritisiert. Nichtrandomisierte Studien sind eine Alternative, allerdings besteht hier die Gefahr, dass sich Interventions- und die Kontrollgruppe bezüglich bekannter und (schlimmer noch) unbekannter Patientenmerkmale unterscheiden.

Zur Analyse von nichtrandomisierten Studien werden in der Regel multiple Regressionsmodelle verwendet, immer häufiger wird aber auch auf Propensity Scores zurückgegriffen. 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 mit einem logistischen Regressionsmodell. Im zweiten Schritt erfolgt die Schätzung des eigentlich interessierenden Therapieeffekts unter Zuhilfenahme des PS. Im Vortrag wird eine kurze, nichttechnische Einführung in Propensity Scores gegeben.

## Anforderungen an die Daten: Eine Diskussion anhand von Patientenregistern

Tim Mathes
Universitätsmedizin Göttingen, Institut für Medizinische Statistik, Göttingen

In dem Beitrag wird zunächst auf grundlegende Anforderungen (z. B. Erfassung von relevanten Confoundern) und häufige Biasquellen (z. B. immortal Time-Bias) bei der Nutzung von Registern und anderen versorgungsnahen Daten, als Grundlage für die vergleichende Analyse von Therapien, eingegangen. Es werden die möglichen zu schätzenden Estimands diskutiert. Weiterhin werden, neben der Datenqualität, zusätzliche relevante Anforderungen an die Daten hinsichtlich Machbarkeit (z. B. notwendige Fallzahl) und Validität (z. B. Positivität, Überlappung und Balanciertheit), insbesondere zur Durchführung von Propensity-Score basierten Analysen, erläutert. Dabei wird auch die potentielle Anwendbarkeit der Analyse-methoden vor dem Hintergrund der deutschen Registerlandschaft betrachtet. Abschließend werden die Erfahrungen von Projekten, die versucht haben RCT-Effekte mittels versorgungsnaher Daten zu emulieren aufgezeigt.

## Alternative Ansätze für Confounder-Adjustierung durch Gewichtung auf der Grundlage des Propensity-Score

Edin Basic[1], Friedhelm Leverkus[2], Sarah Böhme[2], Jens-Otto Andreas[3], Dietrich Knoerzer[4], Katrin Kupas[5], Tobias Bluhmki[5;]
[1]Takeda Pharma Vertrieb GmbH & Co. KG, Berlin; [2]Pfizer Pharma GmbH, Berlin; [3]UCB Biosciences GmbH, Monheim; [4]Roche Pharma AG, Grenzach-Wyhlen; [5]BMS GmbH, München

Propensity-Score-Methoden (PSMs) wie Matching und Gewichtung sind zu einem Eckpfeiler der Schätzung von Behandlungseffekten aus Beobachtungsdaten geworden. PSMs erfordern die Wahl eines „Estimands“, des interessierenden Effekts unter Berücksichtigung einer bestimmten Zielpopulation oder Subpopulation. Die am häufigsten verwendeten „Estimands“ 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. In dieser Präsentation beschreiben wir die Implementierung dieser alternativen Propensity-Score-Gewichtungsmethoden zusammen mit den Haupteigenschaften der einzelnen Ansätze.

# Gütemaße und Kriterien bei der Anwendung von Propensity Scores

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

Propensity Score-Analysen werden in zwei Schritten durchgeführt. Im ersten Schritt wird der Propensity Score (PS), also die Wahrscheinlichkeit, mit der ein Patient die zu prüfende Therapie erhält, geschätzt, in der Regel mit einem logistischen Regressionsmodell. Im zweiten Schritt erfolgt dann die Schätzung des eigentlich interessierenden Therapieeffekts unter Zuhilfenahme des PS.

Die Validität einer PS-Analyse ist im Wesentlichen davon abhängig, ob es im ersten Schritt gelingt, eine hinreicheichende Balanciertheit der Confounder in den Therapiegruppen zu erreichen. Nur dann ist gewährleistet, dass diese Confounder nicht die Schätzung des Therapieeffekts verzerren. Zur Messung dieser Balanciertheit wurden verschiedene Maße vorgeschlagen, z. B. die standardisierte Differenz oder die *z*-Differenz. Häufig übersehen wird dabei, dass die Standardmaßnahmen zur Beurteilung der Modellgüte im logistischen Regressionsmodell (Hosmer-Lemeshow-Test, *c*-Statistik, Analyse der Residuen, …) hier nicht angewendet werden sollten, da diese eine ungenügende Adjustierung für Confounder nicht entdecken können. Darüber hinaus maximieren logistische Regressionsmodelle die Prädiktionsgüte (unabhängig von der Balancierung der Kovariablen) und es ist zu überlegen, ob man im ersten Schritt einer PS-Analyse nicht statistische Modelle verwenden sollte, die explizit die Balanciertheit der Kovariablen optimieren.

Eng verwandt mit der Balanciertheit der Confounder und damit auch ein Maß für die Güte eines PS-Modells ist die Überlappung (“overlap”), also die Ähnlichkeit der Verteilung der geschätzten Propensity Scores in den beiden Therapiegruppen. In Wertebereichen des PS ohne Overlap, wo sich also nur Patienten aus einer der beiden Therapiegruppen finden, ist streng genommen ein Vergleich der Therapien nicht möglich. Im Vortrag werden die beiden Gütekriterien kurz vorgestellt und diskutiert.