Wissenschaftliche Ausarbeitung ausgewählter Aspekte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V

VORLÄUFIGER RAPID REPORT

Projekt: A25-13 Version: 1.0 Stand: 28.07.2025

Impressum

Herausgeber

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema

Wissenschaftliche Ausarbeitung ausgewählter Aspekte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V

Auftraggeber

Gemeinsamer Bundesausschuss

Datum des Auftrags

28.01.2025

Interne Projektnummer

A25-13

Anschrift des Herausgebers

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Siegburger Str. 237
50679 Köln

Tel.: +49 221 35685-0
Fax: +49 221 35685-1
E-Mail: berichte@iqwig.de
Internet: www.iqwig.de

Schlagwörter

Produktüberwachung nach Markteinführung, Register, Dateninterpretation – statistische, Forschungsdesign, Modelle – statistische, Konzept

Keywords

Product Surveillance – Postmarketing, Registries, Data Interpretation – Statistical, Research Design, Models – Statistical, Concept

Dieser Bericht wurde unter Beteiligung externer Sachverständiger erstellt. Externe Sachverständige, die wissenschaftliche Forschungsaufträge für das Institut bearbeiten, haben gemäß § 139b Abs. 3 Satz 2 Sozialgesetzbuch – Fünftes Buch – Gesetzliche Krankenversicherung „alle Beziehungen zu Interessenverbänden, Auftragsinstituten, insbesondere der pharmazeutischen Industrie und der Medizinprodukteindustrie, einschließlich Art und Höhe von Zuwendungen“ offenzulegen. Das Institut hat von der Sachverständigen ein ausgefülltes „Formblatt zur Offenlegung von Beziehungen“ erhalten. Die Angaben wurden von dem speziell für die Beurteilung der Interessenkonflikte eingerichteten Gremium des Instituts bewertet. Es wurden keine Interessenkonflikte festgestellt, die die fachliche Unabhängigkeit im Hinblick auf eine Bearbeitung des vorliegenden Auftrags gefährden.

Die beratende Funktion der externen Sachverständigen beinhaltet die Beantwortung von Fragen von Mitarbeiterinnen und Mitarbeitern des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) zu ihren Fachgebieten im Rahmen von Videomeetings.

Für die Inhalte des Berichts ist allein das IQWiG verantwortlich.

Dieser Bericht ist ein vorläufiger Rapid Report. Er wird zur Anhörung gestellt und es können schriftliche Stellungnahmen eingereicht werden. Das Ende der Stellungnahmefrist wird auf der Website des IQWiG (www.iqwig.de) bekannt gegeben. Dort sind auch die notwendigen Formblätter zu finden. Stellungnahmen können alle interessierten Personen, Institutionen und Gesellschaften abgeben. Gegebenenfalls wird eine wissenschaftliche Erörterung zur Klärung unklarer Aspekte aus den schriftlichen Stellungnahmen durchgeführt. Die Anhörung kann zu Änderungen und / oder Ergänzungen des Berichts führen.

Inhaltsverzeichnis

Seite

Tabellenverzeichnis vii
Abbildungsverzeichnis viii
Abkürzungsverzeichnis ix
Glossar xi
1 Hintergrund 1
2 Fragestellung 2
3 Projektverlauf 3
4 Methoden 4
4.1 Sichtung der bisherigen Prüfungen von Studienprotokoll und statistischem Analyseplan 4
4.2 Vorschlag für die Anforderung an eine systematische Confounderidentifikation ... 4
4.3 Informationsbeschaffung 4
5 Bewertung ausgewählter Aspekte zur Generierung versorgungsnaher Daten im Kontext nicht randomisierter vergleichender Studien 5
5.1 Einleitende Anmerkungen zu den Anforderungen an vergleichende Studien ohne Randomisierung zum Zwecke der Nutzenbewertung 5
5.2 Confounderidentifikation und -auswahl 9
5.3 Schätzung notwendiger Fallzahlen vor dem Hintergrund unzureichender Vorabinformationen 12
5.3.1 Grundlagen der Fallzahlschätzung 13
5.3.2 Keine oder unzureichende Vorabinformationen für orientierende Schätzungen zum Umfang der anwendungsbegleitenden Datenerhebung 18
5.3.3 Entscheidungsfindung für oder gegen eine orientierende Fallzahlschätzung 23
5.4 Beobachtungsstart und Behandlungswechsel 25
5.4.1 Beobachtungsstart 25
5.4.2 Behandlungswechsel 30
5.5 Patientenberichtete Endpunkte 33
5.6 Fehlende Werte 37
5.6.1 Maßnahmen zur Vermeidung fehlender Werte 37
5.6.2 Methodischer Umgang mit fehlenden Werten im Kontext von Propensity-Score-Analysen 41
5.6.2.1 Identifikation und Diskussion von geeigneten statistischen Methoden 41

5.6.2.2 Interpretierbarkeit von Analysen mit fehlenden und / oder imputierten Werten 47
5.6.2.3 Anforderungen an die Darstellung 49
5.7 Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven 50
6 Fazit 56
7 Literatur 59
Anhang A Beispiele für digitale PRO-Erhebungen in Registern 75

Tabellenverzeichnis

Seite

Tabelle 1: Vergleich eines Studienprotokolls einer RCT und einer Studie auf Basis versorgungsnaher Daten mittels Target-Trial-Emulation 7
Tabelle 2: Muster für beobachtete Werte für N Personen im Endpunkt und in den Confoundern L1 bis L7 42

Abbildungsverzeichnis

Seite

Abbildung 1: Überlappung der Populationen (gemessen am Propensity Score) xvii
Abbildung 2: Zeitliche Ausrichtung zentraler Elemente des Studiendesigns am Startzeitpunkt der Beobachtung (t_0) 9
Abbildung 3: Vorgeschlagenes 3-schrittiges Verfahren zur systematischen Confounderidentifikation 10
Abbildung 4: Notwendige Fallzahl (N) bei Festlegung der Parameter erwartete Effektgröße, Signifikanzniveau (Fehler 1. Art) und Power (1−Fehler 2. Art) 14
Abbildung 5: Szenarien bei der orientierenden Fallzahlschätzung mit verschobenem Nullhypothesentest ($RR_0 = 2$) 17
Abbildung 6: Aufdeckbare Effektgröße bei Festlegung der Parameter Fallzahl (N), Signifikanzniveau (Fehler 1. Art) und Power (1−Fehler 2. Art) 21
Abbildung 7: Aufdeckbarer Effekt (Effektmaß: Hazard Ratio) in Abhängigkeit von der Stichprobengröße und den Ereignisanteilen in beiden Behandlungsgruppen (Verhältnis von Intervention zu Komparatoren von 1:3) 22
Abbildung 8: Herangehensweisen zur Bewertung der Machbarkeit einer AbD: (A) orientierende Fallzahlschätzung und (B) orientierende Betrachtung von Fallzahlszenarien 23

Abkürzungsverzeichnis

Abkürzung	**Bedeutung**
AbD	anwendungsbegleitende Datenerhebung
ACNU	Active Comparator New User
AIHTA	Austrian Institute for Health Technology Assessment
AHRQ	Agency for Healthcare Research and Quality
ATE	Average Treatment Effect in the Population
ATMP	Advanced Therapy Medicinal Products (Arzneimittel für neuartige Therapien)
ATT	Average Treatment Effect on the Treated
CAR	chimärer Antigenrezeptor
CBDR	Canadian Bleeding Disorders Registry
CDA	Canada's Drug Agency
DHR	Deutsches Hämophilie Register
DKG	Deutsche Krebsgesellschaft
DLBCL	Diffuse large B-Cell Lymphoma (diffus großzelliges B-Zell-Lymphom)
DRS	Disease Risk Score
EMA	European Medicines Agency (Europäische Arzneimittel-Agentur)
EMCL	European Mantle Cell Lymphoma
FCS	Fully Conditional Specification
FDA	Food and Drug Administration (behördliche Lebensmittelüberwachungs- und Arzneimittelzulassungsbehörde der USA)
G-BA	Gemeinsamer Bundesausschuss
GKV	gesetzliche Krankenversicherung
GTR	Gene Therapy Registry
HAS	Haute Autorité de Santé
HTA	Health Technology Assessment
IPTW	Inverse Probability of Treatment Weighting
IMBEI	Institut für Medizinische Biometrie, Epidemiologie und Informatik
IPCW	Inverse Probability of Censoring Weighting
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ITT	Intention to treat
KIS	Krankenhausinformationssystem

Abkürzung	Bedeutung
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MNAR	Missing not at Random
MICE	multiple Imputation mittels Chained Equations
MPR	Medication Possession Ratio
NICE	National Institute for Health and Care Excellence
PICO	Population, Intervention, Comparison, Outcome
PNU	Prevalent New User
PRO	Patient-reported Outcome (patientenberichteter Endpunkt)
pU	pharmazeutischer Unternehmer
RCT	Randomized controlled Trial (randomisierte kontrollierte Studie)
RR	relatives Risiko
RRMS	Relapsing-remitting multiple Sclerosis (schubförmige remittierende multiple Sklerose)
SAP	statistischer Analyseplan
SCQM	Swiss Clinical Quality Management in Rheumatic Diseases
SGB	Sozialgesetzbuch
SQR	Swedish Rheumatology Quality Register
SMRW	Standardized Morbidity (oder Mortality) Ratio Weighting
WFH	World Federation of Hemophilia

Glossar

Begriffe, die im Glossar erklärt werden, sind im Hauptteil des Berichts jeweils bei ihrer ersten Nennung in einem Abschnitt *kursiv* formatiert und sind mit einem Querverweis zum Glossar versehen.

Active-Comparator-New-User-Design

Für die Nachbildung vergleichender Studien mit Randomisierung auf Basis von Beobachtungsdaten kann ein Active-Comparator-New-User(ACNU)-Design gewählt werden. Darunter wird ein Design verstanden, unter dem lediglich Patientinnen und Patienten in den Vergleich zweier Behandlungsgruppen eingehen, die eine Indikation für eine Umstellung ihrer Therapie haben und demnach eine neue Therapie (Intervention oder Komparator) beginnen [1]. In diesem Fall wird als Beobachtungsstart (t_0) idealerweise der Tag der Therapieentscheidung festgesetzt.

Average Treatment Effect in the Population

Der Estimand Average Treatment Effect in the Population (ATE) beschreibt den durchschnittlichen Behandlungseffekt in der Gesamtpopulation. Betrachtet wird die Frage, wie sich die Ergebnisse unterscheiden würden, wäre die Gesamtpopulation entweder mit der Intervention oder mit dem Komparator behandelt worden [2].

Average Treatment Effect on the Treated

Der Estimand Average Treatment Effect on the Treated (ATT) beschreibt den durchschnittlichen Behandlungseffekt in der Population, die die Intervention erhalten hat. Betrachtet wird die Frage, wie sich die Ergebnisse in dieser Population unterscheiden würden, wären sie nicht mit der Intervention behandelt worden [2].

Confounder

Ein Confounder ist eine Variable, die sowohl die Therapieentscheidung als auch den Endpunkt beeinflusst und somit einen Behandlungseffekt verfälschen kann. In vergleichenden Studien ohne Randomisierung soll mittels Adjustierung für Confounder die fehlende Strukturgleichheit zwischen den Behandlungsgruppen ausgeglichen werden [3]. In der statistischen Analyse kann jedoch nur für bekannte und gemessene Confounder adjustiert werden, die Verzerrung durch unbekannte oder nicht gemessene bzw. nicht messbare Confounder bleibt bestehen (sog. residuelles Confounding).

Disease Risk Score

Disease Risk Score (DRS) – auch Prognosescore genannt – ist definiert als Wahrscheinlichkeit, dass bei einer Person eine Erkrankung auftritt. Die DRS-Methode kann dadurch (als Alternative

zu Propensity-Score-Verfahren) zur Confounderkontrolle eingesetzt werden, mit dem Ziel der Annäherung an eine Strukturgleichheit der Behandlungsgruppen [4].

Estimand

Bei Propensity-Score-Analysen wird das Konstrukt der Estimands genutzt, um die Population, für die der Effekt einer Behandlung geschätzt werden soll, zu beschreiben. Regelhaft werden folgende Estimands unterschieden: ATE, ATT, Average Treatment Effect on the Untreated und Average Treatment Effect in the Overlap [2]. Die Auswahl eines konkreten Propensity-Score-Verfahrens orientiert sich am interessierenden Estimand [5]. Je nach gewähltem Propensity-Score-Verfahren und Überlappung der Propensity-Score-Verteilungen kann sich ergeben, dass die Estimands in einer gegebenen Datensituation fließend ineinander übergehen (zu den Estimands ATE und ATT, siehe Glossar oben).

Der Begriff ist abzugrenzen von dem Konstrukt der Estimands im Rahmen klinischer Studien. Hierzu beschreibt ein Papier der europäischen Zulassungsbehörde (Addendum zur Guideline ICH E9 zu statistischen Prinzipien für klinische Studien [6]) 5 Strategien, die zu verschiedenen Estimands führen (darunter der Estimand Treatment-Policy-Strategie). Die darin getroffenen Festlegungen fokussieren auf den Umgang mit interkurrenten Ereignissen, d. h. Ereignissen, die im Verlauf der Beobachtung einer Patientin bzw. eines Patienten auftreten (z. B. Wechsel der Behandlung).

Immortal-Time-Bias

Immortal-Time-Bias kann vorliegen, insofern es aufgrund des Studiendesigns einen Zeitraum (Wartezeit) gibt, in dem ein Ereignis (z. B. Tod) nicht auftreten kann [7,8]. Klassisches Beispiel hierfür ist, wenn der Beobachtungsstart (t_0) auf den Zeitpunkt der Erfüllung der Einschlusskriterien, aber vor den Zeitpunkt der Behandlungszuteilung gesetzt wird. Dies hat zur Folge, dass die Behandlungszuteilung auf Basis von Informationen aus dem Follow-up erfolgt [9]. Dies trifft z. B. auf solche Situationen zu, in denen Patientinnen und Patienten erst ab Erhalt einer CAR-T-Zelltherapie (und somit nach t_0) der Interventionsgruppe zugeordnet werden. In diesen Fällen müssen die Patientinnen und Patienten bis zum Zeitpunkt des Beginns der CAR-T-Zelltherapie (unter Behandlung mit einer Brückentherapie) überlebt haben.

Inverse Probability of Treatment Weighting

Inverse Probability of Treatment Weighting (IPTW) ist ein Gewichtungsansatz auf Basis von Propensity Scores. Das Gewicht, das jede Person erhält, entspricht der inversen Wahrscheinlichkeit, die Intervention erhalten zu haben. Patientinnen und Patienten aus der Interventionsgruppe erhalten das Gewicht 1/Propensity Score und Patientinnen und Patienten aus der Kontrollgruppe das Gewicht 1/(1–Propensity Score). Dies bedeutet, dass Patientinnen und Patienten aus der Interventionsgruppe mit einem niedrigen Propensity Score (d. h. mit einer geringen Wahrscheinlichkeit, die Intervention zu erhalten) ein hohes Gewicht erhalten, weil

sie den Patientinnen und Patienten aus der Kontrollgruppe hinsichtlich der beobachteten Charakteristika ähneln (ausgedrückt durch deren niedrigen Propensity Score). Demgegenüber werden Patientinnen und Patienten aus der Kontrollgruppe mit einem hohen Propensity Score (d. h. mit einer hohen Wahrscheinlichkeit, die Intervention zu erhalten) stärker gewichtet, da sie den Patientinnen und Patienten aus der Interventionsgruppe hinsichtlich beobachteter Charakteristika ähnlicher sind [10].

Konvergenz

Für die Berechnung von Propensity Scores, die multiple Imputation bei fehlenden Werten und für Effektschätzungen werden Maximum-Likelihood-Verfahren eingesetzt. Zu den Verfahren gehören u. a. verallgemeinerte lineare Modelle (z. B. multivariate logistische Regression) und das Cox-Proportional-Hazards-Modell. Die diesen Verfahren zugrunde liegenden Algorithmen beruhen auf iterativen Prozessen und mehreren Konvergenzkriterien, z. B. Konvergenz der Parameterschätzung oder der Likelihood-Schätzung. Eine Konvergenz eines Algorithmus liegt dann vor, wenn in 2 aufeinanderfolgenden Iterationen die Änderungen in den jeweiligen Schätzungen kleiner als vorgegebene Werte sind.

Wenn nach einer vorgegebenen Anzahl von Iterationen für ein oder mehrere Kriterien keine Konvergenz vorliegt, liegen keine Parameterschätzungen vor oder die Parameterschätzungen sind potenziell stark verzerrt. Die Gründe, warum ein Algorithmus nicht konvergiert, sind vielfältig. Dazu zählen eine zu geringe Anzahl von Beobachtungen (Fallzahl) und eine nicht angemessene Modellierung (durch zu viele Kovariaten) [11].

Latency-Time-Bias

Latency-Time-Bias tritt dann auf, wenn die Vortherapie zu einer Erhöhung des Risikos für das Ereignis von Interesse führt, das Ereignis selbst jedoch erst unter Studienintervention auftritt. Ein Beispiel hierfür ist: Sekundäre Primärtumore können mehrere Monate bis Jahre nach einer Chemotherapie auftreten [12]. Da diese Ereignisse nicht in einem kausalen Zusammenhang mit der Studienintervention stehen, sollten Latenzzeitfenster definiert werden. Ereignisse, die in diesem Zeitfenster auftreten, gehen nicht in die Analyse ein [13].

Matching Weights

Matching Weights ist ein Gewichtungsansatz auf Basis von Propensity Scores. Das Gewicht, das jede Person erhält, entspricht dem Verhältnis der niedrigeren der beiden vorhergesagten Wahrscheinlichkeiten (Propensity Score bzw. 1–Propensity Score) zur vorhergesagten Wahrscheinlichkeit der tatsächlich erhaltenen Behandlung [5].

Medication Possession Ratio

Medication Possession Ratio (MPR) ist ein Maß, in das jegliche vorherige Exposition gegenüber 1 oder mehreren Vergleichspräparat(en) eingeht, die im zu bewertenden Anwendungsgebiet

eingesetzt werden. Durch MPR-Matching wird versucht, die Behandlungsgruppen bezüglich der vorherigen Expositionsmuster anzugleichen [14].

Missing at Random

Als Missing at Random (MAR) wird eine Annahme über das Fehlen eines Wertes bezeichnet. Unter dieser Annahme hängt die Wahrscheinlichkeit für das Fehlen eines Wertes von anderen beobachteten Werten einer Person ab, nicht aber vom fehlenden Wert selbst. Solche fehlenden Werte führen zu verzerrten Ergebnissen, insofern die Abhängigkeit in der Analyse unberücksichtigt bleibt. Um solche Abhängigkeiten in der Analyse zu berücksichtigen, muss der Ausfallmechanismus klar ersichtlich mithilfe der beobachteten Daten erklärbar sein. Die sachgerechte Berücksichtigung in der Analyse führt zu einer unverzerrten Schätzung [15]. Ein Beispiel hierfür: Männer neigen eher dazu, einen Fragebogen zur Feststellung der Schwere der Depression nicht auszufüllen als Frauen. Die Beantwortung des Fragebogens hängt vom Geschlecht, nicht aber von der tatsächlichen Schwere der Depression selbst ab [16].

Missing completely at Random

Als Missing completely at Random (MCAR) wird eine Annahme über das Fehlen eines Wertes bezeichnet. Die Wahrscheinlichkeit für das Fehlen eines Wertes ist unabhängig von anderen beobachteten Werten einer Person (z. B. Kovariablen oder Werten zu früheren Zeitpunkten) und dem unbeobachteten, fehlenden Wert selbst. Solche fehlenden Werte führen grundsätzlich nicht zu verzerrten Ergebnissen. Allerdings verringert sich die Präzision der Effektschätzungen [15]. Ein Bespiel hierfür: Ein Teil der Materialproben wird im Labor unsachgemäß behandelt, sodass diese nicht mehr untersucht werden können [16].

Missing not at Random

Als Missing not at Random (MNAR) wird eine Annahme über das Fehlen eines Wertes bezeichnet. Die Wahrscheinlichkeit für das Fehlen eines Wertes hängt vom tatsächlichen, fehlenden Wert der betrachteten Zielgröße selbst ab. Solche fehlenden Werte führen zu verzerrten Ergebnissen. Die Verzerrung kann auch nicht korrigiert werden, da der Ausfallmechanismus von unbeobachteten Daten abhängt. In diesem Fall kann nur mittels Sensitivitätsanalysen auf Basis von nicht prüfbaren Annahmen die Größe der Verzerrung untersucht werden [15]. Ein Beispiel hierfür: Die Wahrscheinlichkeit des Fehlens von Daten zur Schwere der Depression hängt von der Schwere der Depression selbst ab [16].

Multiple Imputation mittels Chained Equations

Die multiple Imputation ist ein statistisches Verfahren zum Umgang mit fehlenden Werten in Datensätzen. Es werden mehrere vollständige Datensätze erzeugt, in denen fehlende Werte unter Modellannahmen (Missing completely at Random [MCAR], Missing at Random [MAR] oder Missing not at Random [MNAR]) ersetzt werden. Hierbei können beliebige Variablen verwendet werden. Die multiple Imputation mittels Chained Equations (MICE) – auch Fully

Conditional Specification (FCS) genannt – hat sich als eine Variante der multiplen Imputation etabliert [17].

Overlap Weights

Overlap Weights ist ein Gewichtungsansatz auf Basis von Propensity Scores. Das Gewicht, das jede Person erhält, entspricht der Wahrscheinlichkeit, die jeweils entgegengesetzte Behandlung bekommen zu haben [5].

Positivität

Unter Positivität ist zu verstehen, dass bei einem Vergleich zweier Patientengruppen für alle Patientinnen und Patienten, die in die Auswertung eingehen, die zu vergleichenden Behandlungsformen (neuer Wirkstoff bzw. Komparator) eine Therapieoption darstellen. Daraus folgt, dass z. B. Patientinnen und Patienten mit einer Kontraindikation gegen eine der untersuchten Therapien nicht in der Auswertung berücksichtigt werden dürfen [18].

Prevalent-New-User-Design

Für die Nachbildung vergleichender Studien mit Randomisierung auf Basis von Beobachtungsdaten kann ein Prevalent-New-User(PNU)-Design gewählt werden. Bei diesem Design können in den Vergleich zweier Behandlungsgruppen 3 verschiedene Patientenpopulationen eingehen: (1) Patientinnen und Patienten, die eine Therapie mit der Intervention beginnen und die zuvor (hinsichtlich der Intervention und des Komparators) therapienaiv waren (New User), (2) Patientinnen und Patienten mit Wechsel vom Komparator auf die Intervention (Prevalent New User) und (3) Patientinnen und Patienten, die ihre Therapie mit dem Komparator fortführen [19]. Ziel ist es, mögliche Verzerrungen durch eine geeignete Wahl des Beobachtungsstarts (t_0) für Patientinnen und Patienten der Kontrollgruppe zu reduzieren. Für Patientinnen und Patienten der Interventionsgruppe (umfasst New User und Prevalent New User) erfolgt die Wahl von t_0 analog zum ACNU-Design (zu ACNU-Design, siehe Glossar oben).

Propensity-Score-Matching

Mit dem Propensity-Score-Matching werden Patientinnen und Patienten der Interventions- und Kontrollgruppe gemäß ihrem Propensity Score einander zugeordnet (d. h. „gematcht“). Das Matching kann mittels verschiedener Algorithmen durchgeführt werden, darunter z. B. Nearest-neighbour-Matching. Die Zuordnung kann 1:1 oder 1:n, mit oder ohne Zurücklegen erfolgen. Je nach Überlappung der Behandlungsgruppen gemessen am Propensity Score kann diese Methode mit einer Reduktion der Fallzahl und dadurch einem Verlust an statistischer Power einhergehen (Personen ohne Matching-Partner werden ausgeschlossen) [10].

Propensity-Score-Verfahren

Propensity-Score-Verfahren sind statistische Methoden, mit denen in nicht randomisierten vergleichenden Studien die Balanciertheit hinsichtlich beobachteter Kovariaten zwischen den

Behandlungsgruppen (Strukturgleichheit) erreicht werden soll. Der Propensity Score ist die durch beobachtete Kovariaten bedingte Wahrscheinlichkeit, die Intervention zu erhalten. Es gibt 4 Ansätze, um den Therapieeffekt mittels Propensity Scores zu schätzen: Gewichtung, Matching, Stratifizierung und Regressionsadjustierung [10].

Selection Bias

Ein Selection Bias entsteht, wenn Patientinnen und Patienten, ein Teil der Beobachtungsdauer (Follow-up) oder Ereignisse von der Analyse ausgeschlossen werden und die Selektion sowohl mit der Exposition als auch dem Endpunkt assoziiert ist. Dies ist unter anderem der Fall, wenn Patientinnen und Patienten nach Beobachtungsstart (t_0) zensiert werden (z. B. bei Wechsel der Behandlung) oder t_0 auf den Zeitpunkt der Erfüllung der Einschlusskriterien, aber nach der Behandlungszuteilung festgesetzt wird. Bei letzterem Beispiel gehen nur die Patientinnen und Patienten in die Analyse ein, die genau zu t_0 die Einschlusskriterien erfüllen [9,20].

Standardised Morbidity (oder Mortality) Ratio Weighting

Standardised Morbidity (oder Mortality) Ratio Weighting (SMRW) ist ein Gewichtungsansatz auf Basis von Propensity Scores. Patientinnen und Patienten aus der Interventionsgruppe erhalten das Gewicht 1 und Patientinnen und Patienten aus der Kontrollgruppe das Gewicht Propensity Score/(1–Propensity Score). Dies bedeutet, dass Patientinnen und Patienten aus der Kontrollgruppe umso stärker gewichtet werden, je ähnlicher sie der Interventionsgruppe hinsichtlich der beobachteten Charakteristika sind [5].

Target-Trial-Emulation

Unter einer Target-Trial-Emulation wird die Nachbildung der Planung vergleichender Studien mit Randomisierung auf Basis von Beobachtungsdaten verstanden [21].

Time-Lag-Bias

Time-Lag-Bias entsteht bei einem Vergleich von Behandlungsgruppen in unterschiedlichen Krankheitsstadien [22]. Dies ist z. B. der Fall, wenn sich die Zeitspanne zwischen der Diagnose einer progredient verlaufenden Erkrankung und dem Beobachtungsstart (t_0) zwischen den Gruppen unterscheidet.

Überdeckungswahrscheinlichkeit

Für die Berechnung von Propensity Scores, die multiple Imputation bei fehlenden Werten und für Effektschätzungen werden Maximum-Likelihood-Verfahren eingesetzt. Um die Güte dieser Verfahren zu vergleichen, werden Simulationsstudien durchgeführt. In diesen Studien werden in einem 1. Schritt regelhaft für verschiedene Szenarien simulierte Datensätze (Replikationen) erzeugt. In einem 2. Schritt erfolgt mittels der Replikationen die Berechnung von empirischen Gütemaße. Die Überdeckungswahrscheinlichkeit als Gütemaß beschreibt, mit welcher Wahr-

scheinlichkeit das (1–Alpha)-Konfidenzintervall einer Effektschätzung den wahren Wert enthält. Hierfür ist die Festlegung des Alpha-Fehlers notwendig, z. B. Alpha = 0,05. In Simulationsstudien wird die empirische Überdeckungswahrscheinlichkeit über den Anteil der Replikationen bestimmt, in denen das (1–Alpha)-Konfidenzintervall den vorgegebenen Effekt enthält. Ein gutes statistisches Verfahren hat eine empirische Überdeckungswahrscheinlichkeit von 1–Alpha. Zu kleine Überdeckungswahrscheinlichkeiten weisen auf eine große Verzerrung (Bias) hin, Überdeckungswahrscheinlichkeiten nahe 1 auf eine fehlende Power aufgrund zu breiter Konfidenzintervalle [23].

Überlappung

Überlappung bezeichnet die Fläche, in der sich die Verteilungen der Propensity Scores der zu vergleichenden Gruppen überschneiden. Eine Überprüfung der Überlappung kann mittels eines visuellen Abgleichs der Propensity-Score-Verteilungen erfolgen (Propensity Distribution Plots) (siehe Abbildung 1) [24]. Wie in Abbildung 1 dargestellt, kann die Fläche der Überlappung gering sein (beträgt hier ca. 10 %), auch wenn der Bereich der Überlappung recht groß ist (reicht hier von ca. 0,2 bis 0,75). Eine gute Überlappung bezieht sich demnach auf eine ausreichend große überlappende Fläche und nicht auf einen breiten Bereich der Überlappung.

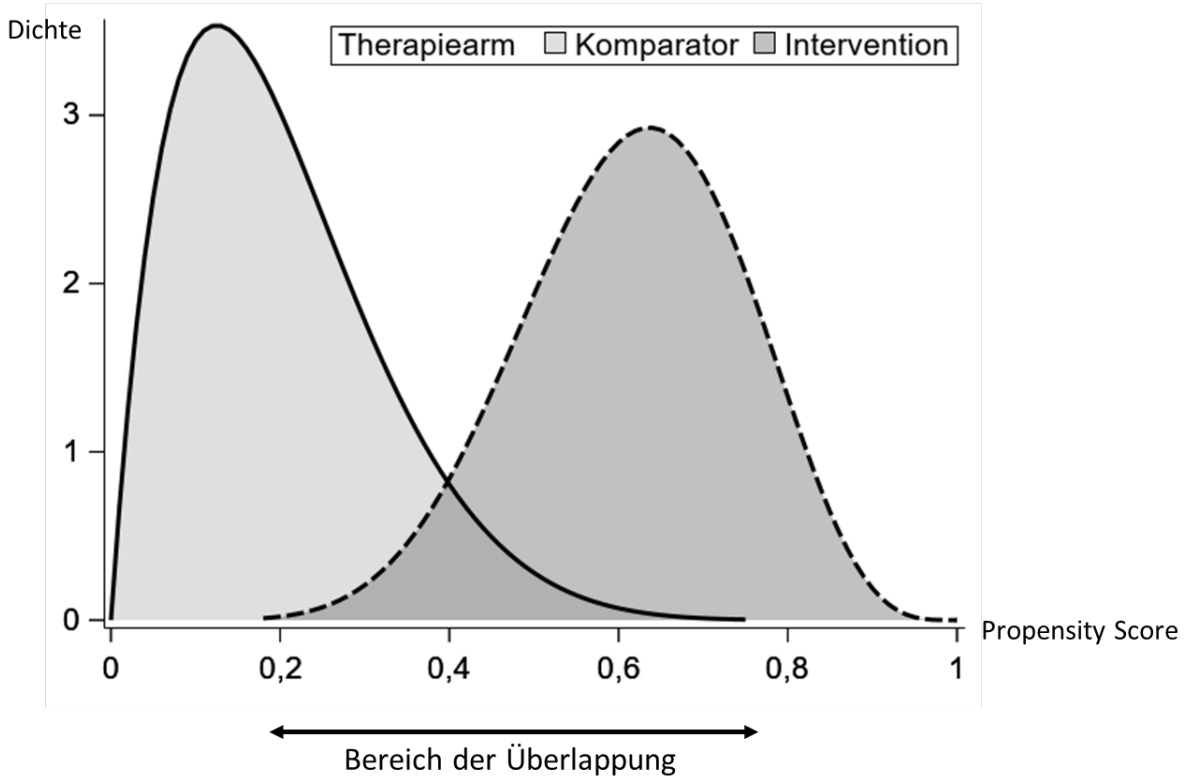


Abbildung 1: Überlappung der Populationen (gemessen am Propensity Score)

Kurzfassung

Der Gemeinsame Bundesausschuss (G-BA) hat am 28.01.2025 das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) mit einer wissenschaftlichen Ausarbeitung zur Methodik im Bereich der anwendungsbegleitenden Datenerhebung (AbD) mit einer Fokussierung auf eine Bewertung von Themenbereichen beauftragt.

Fragestellung

Ziel des vorliegenden Berichts ist eine wissenschaftliche Ausarbeitung zur Methodik (Generierung und Auswertung versorgungsnaher Daten zum Zwecke der Nutzenbewertung) im Bereich der AbD mit einer Fokussierung auf eine Bewertung der folgenden Themenbereiche, jeweils im Kontext nicht randomisierter vergleichender Studien:

- Confounderidentifikation und -auswahl
- Schätzung notwendiger Fallzahlen vor dem Hintergrund unzureichender Vorabinformationen
- Beobachtungsstart
- Behandlungswechsel
- patientenberichtete Endpunkte (PROs)
- fehlende Werte
- Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven

Dabei sollen eine orientierende Recherche nach geeigneter Literatur für die oben genannten Themenbereiche und eine Sichtung der in den bisherigen Prüfungen von Studienprotokoll und statistischem Analyseplan (SAP) identifizierten Mängel (ggf. mit Ergänzung der genannten Themenbereiche) erfolgen.

Für eine AbD eignen sich primär bereits bestehende Register als potenzielle Datenquelle. Die Ausarbeitung fokussiert daher auch auf Patientenregister als Instrumente zur Datenerhebung mit dem Ziel der Generierung versorgungsnaher Daten, da andere Datenquellen (z. B. GKV-Routinedaten) zum aktuellen Zeitpunkt nicht geeignet erscheinen, die Anforderungen an die Daten(qualität) im Rahmen der Nutzenbewertung zu erfüllen.

Methoden

Sichtung der bisherigen Prüfungen von Studienprotokoll und statistischem Analyseplan

Die Änderungsbedarfe, die in bisherigen Prüfungen der von den pUs eingereichten Studienprotokolle und SAPs adressiert wurden, wurden gesichtet. Auf Basis dieser Sichtung sollten

die beauftragten Themenbereiche, sofern erforderlich, um weitere Bereiche ergänzt werden. Da alle identifizierten Änderungsbedarfe einem der 7 Themenbereiche zugeordnet werden konnten, wurden im Rahmen des Auftrages keine weiteren Themenbereiche bearbeitet.

Erarbeiteter Vorschlag für die Anforderung an eine systematische Confounderidentifikation

Für den Themenbereich Confounderidentifikation und -auswahl wurde für den vorliegenden Auftrag auf das IQWiG-Arbeitspapier GA23-02 zurückgegriffen (veröffentlicht am 18.06.2025). Es fand somit keine erneute Erarbeitung der Methodik statt. Die Darstellung im vorliegenden Bericht beschränkt sich im Wesentlichen auf die abgeleiteten Handlungsempfehlungen zur systematischen Confounderidentifikation für nicht randomisierte vergleichende Studien. Für weiterführende Informationen wird auf das Arbeitspapier GA23-02 verwiesen.

Informationsbeschaffung

Für die weiteren 6 Themenbereiche umfasste die Informationsbeschaffung zu methodischen Dokumenten und Übersichtsarbeiten eine orientierende Recherche in PubMed und Website-Suche bei Zulassungsbehörden und anderen HTA-Agenturen. Zusätzlich wurden ausgehend von zuvor identifizierten methodischen Dokumenten und Übersichtsarbeiten Referenzlisten gesichtet sowie die „Similar Articles“-Funktion in PubMed und die „Cited-by“-Funktion in OpenAlex verwendet.

Bewertung ausgewählter Aspekte zur Generierung versorgungsnaher Daten im Kontext nicht randomisierter vergleichender Studien

Confounderidentifikation und -auswahl

In dem Arbeitspapier GA23-02 des IQWiG wurde gemäß dem Ansatz von Pufulete 2022 eine systematische Confounderidentifikation in der Indikation schubförmig remittierende multiple Sklerose (RRMS) an einem geeigneten Vergleich zweier medikamentöser Therapien mit Hinblick auf wesentliche patientenrelevante Endpunkte durchgeführt. Hierfür wurde eine Arbeitsgruppe externer Sachverständiger beauftragt, die ihre Ergebnisse in einem Bericht dargestellt hat. Auf Basis dieses Ergebnisberichts wurden im Rahmen der Bearbeitung von GA23-02 anhand weiterführender Untersuchungen durch das IQWiG die folgenden Handlungsempfehlungen für eine systematische Confounderidentifikation für nicht randomisierte vergleichende Studien (und somit auch für die AbD) abgeleitet:

1) Identifikation

   Potenzielle Confoundervariablen sollen aus einer systematischen Literaturrecherche und aus Interviews mit klinischen Expertinnen und Experten identifiziert werden.

2) Zusammenführung

   Confoundervariablen, die ein inhaltlich gleiches Konstrukt messen oder sich inhaltlich überlappen, sollen zusammengeführt werden. Für jeden dieser potenziellen Confounder ist eine eindeutige Operationalisierung festzulegen.

3) Relevanzbewertung

   Jeder potenzieller Confounder sollte über eine Einbindung von klinischen Expertinnen und Experten in seiner Bedeutung bewertet werden. Das Ergebnis kann für eine inhaltlich gestützte Selektion von Confoundervariablen verwendet werden.

Das Verfahren der systematischen Confounderidentifikation ist trotz der im Arbeitspapier GA23-02 beschriebenen Möglichkeiten zur Aufwandseinsparung ressourcenintensiv. In einem nächsten Schritt soll daher (außerhalb der vorliegenden Ausarbeitung) geprüft werden, ob weitere Anpassungen zur systematischen Identifikation von Confoundern möglich sind, um die Machbarkeit (durch eine Aufwandsreduktion) zu erhöhen. Dabei sind potenzielle Optionen zur Aufwandsreduktion dem damit einhergehenden Informationsverlust und dem möglichen Verlust der Ergebnissicherheit gegenüberzustellen.

Schätzung notwendiger Fallzahlen vor dem Hintergrund unzureichender Vorabinformationen

Der G-BA beauftragt das IQWiG regelhaft mit einer wissenschaftlichen Ausarbeitung von Konzepten für AbDs und deren Auswertung. Die Konzepte sollen auch die Anforderungen an die Dauer und den Umfang der jeweiligen Datenerhebung beschreiben. Dabei geht es um die Beantwortung der Frage, ob die geplante AbD prinzipiell realisierbar und sinnvoll durchführbar ist. Zu diesem Zweck wird keine Fallzahlplanung durchgeführt, da die hierfür notwendige Informationsbasis nicht vorliegt. Die Konzepterstellung verfolgt vielmehr das Ziel, eine Einschätzung zur grundsätzlichen Machbarkeit einer AbD zu geben. Je nach Informationsbasis zum Zeitpunkt der Konzepterstellung wird dabei entweder auf eine orientierende Fallzahlschätzung oder eine orientierende Betrachtung von Fallzahlszenarien als Herangehensweise zurückgegriffen.

Um im Rahmen der orientierenden Fallzahlschätzung die notwendige Fallzahl zu ermitteln, müssen Annahmen für mehrere Parameter getroffen werden. Zu diesen Parametern gehören u. a. die Effektgröße, das Signifikanzniveau (Irrtumswahrscheinlichkeit, auch Fehler 1. Art) und die Power (1–Fehler 2. Art). Die Annahmen für die Effektgröße beruhen bei Fallzahlplanungen für randomisierte kontrollierte Studien (RCTs) in der Regel auf Informationen aus Pilotstudien oder vergleichenden Phase-II-Studien, die eine relativ sichere Abschätzung dieses Parameters erlauben. Für AbDs, die durch den G-BA aufgrund fehlender direkt vergleichender Evidenz gefordert werden, liegt diese Informationsbasis als Ausgangspunkt für eine Fallzahlplanung nicht vor. Aus diesem Grund kann die Fallzahlschätzung im Rahmen einer Konzepterstellung

für eine AbD (häufig auf Basis nicht vergleichender Zulassungsstudien) allenfalls einen orientierenden Charakter haben.

Im Rahmen der Ausarbeitung eines AbD-Konzepts wird eine orientierende Fallzahlschätzung immer dann durchgeführt, wenn sie sinnvoll möglich ist. Zum Zeitpunkt der Konzepterstellung (meist vor der Zulassung) liegen die benötigten Informationen zur Schätzung der Effektgröße für ein neues Arzneimittel jedoch in der Regel nicht vor (z. B. liegen ggf. nur vorläufige, zum Teil widersprüchliche, Ergebnisse zu lediglich selektiv in Kongressabstracts oder Pressemitteilungen dargestellten Endpunkten vor). Damit sind die Voraussetzungen für eine orientierende Fallzahlschätzung nicht erfüllt.

In diesen Fällen kommt eine orientierende Betrachtung von Fallzahlszenarien zur Anwendung, um dennoch Aussagen zum Umfang der Datenerhebung treffen zu können. Diese Herangehensweise folgt im Wesentlichen den Grundsätzen der orientierenden Fallzahlschätzung und hat lediglich einen anderen Ausgangspunkt. In einem 1. Schritt wird die geschätzte Patientenzahl im zu bewertenden Anwendungsgebiet in Deutschland ermittelt. In einem 2. Schritt wird mittels Festlegung der Patientenzahl, des Signifikanzniveaus sowie der Power die Frage beantwortet, welche Effektgrößen auf Basis der im zu bewertenden Anwendungsgebiet grundsätzlich für eine AbD in Deutschland zur Verfügung stehenden Patientenzahl (Präzision der Schätzung) und unter Berücksichtigung der Gefahr systematischer Verzerrung wegen des nicht randomisierten Vergleichs (verschobene Nullhypothese) aufgedeckt werden können.

Aufdeckbar bedeutet, dass das im Rahmen einer AbD geschätzte 95 %-Konfidenzintervall mit 80 % Power bei gegebener Patientenzahl vollständig oberhalb bzw. vollständig unterhalb der verschobenen Hypothesengrenze liegt. Neben der Patientenzahl wird, sofern verfügbar, das Basisrisiko entweder unter Behandlung mit der Intervention oder unter Behandlung mit den Komparatoren für einen patientenrelevanten Endpunkt als Anker herangezogen, um auf diese Weise ein Spektrum der im Anwendungsgebiet möglichen Szenarien aufzuzeigen.

Bei der Entscheidung für oder gegen die orientierende Fallzahlschätzung im Rahmen von AbD-Verfahren werden keine festen Grenzen angelegt. Ob eine orientierende Fallzahlschätzung auf Basis der jeweils vorliegenden Datenkonstellation als sinnvoll durchführbar erachtet wird, ist im Einzelfall zu prüfen.

Unabhängig vom gewählten methodischen Vorgehen für die Konzepterstellung (orientierende Fallzahlschätzung oder orientierende Betrachtung von Fallzahlszenarien) sollte der pU des zu bewertenden Arzneimittels die (finale) Fallzahlplanung der AbD zum Zeitpunkt der Zwischenanalysen auf Basis der dann vorhandenen vorläufigen Ergebnisse vornehmen. Die (finale) Fallzahlplanung unter Berücksichtigung der Ergebnisse von Zwischenanalysen vorzunehmen, wird

als grundsätzlich geeignet beurteilt, da die Schätzung der notwendigen Fallzahl bei versorgungsnahen Datenerhebungen (in Abgrenzung zu klinischen Studien) nicht mit Hinblick auf eine aktive Rekrutierung erfolgt. Es werden stattdessen möglichst alle Patientinnen und Patienten einer Registerpopulation, die die Einschlusskriterien für die AbD erfüllen, nach ihrer Einwilligung in die AbD eingeschlossen. Dies erfolgt auch vor dem Hintergrund, dass, je nach Propensity-Score-Verfahren, ggf. ein relevanter Anteil der eingeschlossenen Patientinnen und Patienten nicht in der Auswertung berücksichtigt werden kann.

Beobachtungsstart und Behandlungswechsel

Beobachtungsstart

Eine Herausforderung bei versorgungsnahen Studien ohne Randomisierung ist eine adäquate Festlegung des Startzeitpunkts der Beobachtung (Index-Datum [t_0]), da es im Gegensatz zu RCTs (Zeitpunkt der Randomisierung) keinen festen Startzeitpunkt für die Beobachtung gibt. Eine Verzerrung der Ergebnisse (z. B. durch Immortal-Time-Bias oder Selection Bias) entsteht, wenn der Beginn der Beobachtung (t_0), die Prüfung der Erfüllung der Einschlusskriterien und die Behandlungszuteilung nicht übereinstimmen.

Beim Active-Comparator-New-User(ACNU)-Design wird für beide Behandlungsgruppen der Tag der Therapieentscheidung (oder eine bestmögliche Näherung daran) als t_0 festgesetzt. Es werden ausschließlich jene Patientinnen und Patienten in die Studie eingeschlossen, die eine Indikation für eine Umstellung ihrer Therapie haben. Zum Umgang mit Situationen, in denen nicht in beiden Behandlungsgruppen zu Beginn der Beobachtung mit einer neuen Therapie begonnen wird, eignet sich das Prevalent-New-User(PNU)-Design. Ziel ist dabei, potenzielle Verzerrungen durch eine falsche Wahl des Beobachtungsstarts zu reduzieren. Die Anwendung dieses Designs ist daran gebunden, dass geeignete Verfahren zur Berücksichtigung erhaltener Vortherapie(n) zum Einsatz kommen. Die Robustheit der Ergebnisse sollte mittels Sensitivitätsanalysen überprüft werden.

Behandlungswechsel

In vergleichenden Studien (mit oder ohne Randomisierung) haben Patientinnen und Patienten die Möglichkeit, bedarfsgesteuert von einer Behandlung zu einer anderen zu wechseln. Dieser Vorgang wird als „Treatment Switching“ bezeichnet. Im Rahmen des Rapid Reports wird der Begriff auf einen Wechsel von der Intervention zum Komparator oder auf einen Wechsel vom Komparator zur Intervention begrenzt.

Ein Wechsel von der Intervention zum Komparator entspricht in einer AbD im Regelfall einem Wechsel zu einer allgemein anerkannten, dem medizinischen Standard entsprechenden Behandlung. Der Wechsel ist somit Teil einer Behandlungsstrategie, die der normalen klinischen Praxis entspricht. In einer solchen Situation ist eine unverzerrte Schätzung des

Therapieeffekts möglich. Ein Wechsel vom Komparator zur Intervention ist dann unproblematisch, wenn die Intervention eine adäquate Folgetherapie im Rahmen der Behandlungsstrategie darstellt (dies trifft z. B. dann zu, wenn für die Intervention bereits zu einem früheren Zeitpunkt eine Zulassung in einer späteren Therapielinie erfolgt ist). Diese Konstellation ist im Rahmen einer AbD jedoch eher nicht zu erwarten. In allen anderen Fällen kann ein Wechsel vom Komparator zur Intervention zu verzerrten Ergebnissen führen und es ist nachzuweisen, dass die Ergebnisse auf die Fragestellung der AbD übertragbar sind.

Unabhängig von einem Behandlungswechsel werden zum Zwecke der Nutzenbewertung von Arzneimitteln primär Auswertungen entsprechend der Treatment-Policy-Strategie (d. h. eine Schätzung des Effekts für die gesamte Behandlungsstrategie) und des Intention-to-treat(ITT)-Prinzips benötigt. Um einen potenziell verzerrenden Einfluss eines Behandlungswechsels auf die ITT-Auswertung unter Anwendung von Verfahren für Überlebenszeiten zu untersuchen, können Sensitivitätsanalysen durchgeführt werden, in denen Patientinnen und Patienten mit Wechsel vom Komparator zur Intervention zum Zeitpunkt ihres Wechsels zensiert werden (Per-Protocol-Analysen) oder in denen die Anwendung komplexer Methoden zur Berücksichtigung von Treatment Switching (wie z. B. Inverse Probability of Censoring Weighting [IPCW]) erfolgt.

Besondere Situation bei der Auswertung versorgungsnah erhobener Daten

Da der G-BA AbDs häufig in Anwendungsgebieten von Arzneimitteln für seltene Erkrankungen (Orphan Drugs) fordert, ergibt sich per se eine niedrige für eine AbD zur Verfügung stehende Patientenzahl. Je nach zu bewertenden Therapien ergibt sich zusätzlich die Schwierigkeit, eine ausreichende Patientenzahl rekrutieren zu können. Um vor diesem Hintergrund dennoch eine möglichst hohe Patientenzahl unter Behandlung mit einer neuen Therapie in der Analyse zu berücksichtigen, besteht im Sinne eines pragmatischen Vorgehens die Option, Patientinnen und Patienten mit Wechsel vom Komparator auf die Intervention den Behandlungsgruppen in Abhängigkeit ihrer Beobachtungszeit unter Therapie mit dem Komparator zuzuordnen (d. h., dass Patientinnen und Patienten, die nach kürzerer Zeit auf die Intervention wechseln und für die noch eine angemessene Beobachtungszeit unter Therapie mit der Intervention zu erwarten ist, in der Interventionsgruppe ausgewertet werden können).

Je nach interessierendem Vergleich (neuer Wirkstoff vs. Komparator) kann es Fälle geben, in denen eine Therapieoption nach der Therapieentscheidung nicht sofort zur Verfügung steht oder nicht sofort wirksam ist. Es kann erforderlich sein, eine Brückentherapie einzusetzen, die als Teil des Therapiekonzepts zu verstehen ist. Entscheidend für einen adäquaten Umgang mit Brückentherapien ist, dass der Tag der Entscheidung für die eigentlich geplante Therapie (z. B. Zeitpunkt der Entscheidung für eine CAR-T-Zelltherapie im Tumorboard) als Startzeitpunkt der Beobachtung (Index-Datum [t_0]) gemäß Target-Trial-Konzept festgesetzt wird.

Patientenberichtete Endpunkte

Um bestehende Evidenzlücken zu neuen Arzneimitteln zu schließen, fordert der G-BA für AbDs regelhaft die Erhebung von Endpunkten zur Symptomatik und gesundheitsbezogenen Lebensqualität. Diese Endpunkte werden üblicherweise mit PROs erhoben. Hierfür sind zum Zwecke der Generierung aussagekräftiger Daten über den Verlauf der Erkrankung unter Therapie für beide Behandlungsgruppen einheitliche Erhebungszeitpunkte zu wählen. Die Erhebung sollte mehrfach im Jahr und in standardisierten Intervallen zu definierten Zeitpunkten erfolgen, mit zu Beginn engeren und später breiteren Intervallen. Dabei sollten für die Erhebungszeitpunkte angemessene Toleranzfenster präspezifiziert werden, die nicht aneinandergrenzen und eine sinnvolle Zuordnung zu einem Erhebungszeitpunkt möglich machen. Im Kontext einer AbD werden Daten in der Versorgungspraxis erhoben, sodass sich die Bedingungen, unter denen PRO-Daten erfasst werden, deutlich von den Bedingungen in einer klinischen Studie unterscheiden. Mittels digitaler Befragungen (z. B. per App und / oder per Patientenportal) ist eine niedrigschwellige Erhebung unabhängig von Arztbesuchen möglich. Digitale Tools werden in Registern und in der Routineversorgung bereits erfolgreich zur Datenerhebung eingesetzt.

Fehlende Werte

Maßnahmen zur Vermeidung fehlender Werte

In Studien lassen sich fehlende Werte in der Regel nicht vollständig vermeiden. Um dennoch hinreichend sichere Aussagen zum Nutzen und Schaden eines Arzneimittels sowohl unter den Bedingungen der Routineversorgung als auch unter Berücksichtigung der Fragestellung der Nutzenbewertung machen zu können, sind Maßnahmen sinnvoll und notwendig, mit denen der Anteil fehlender Werte so gering wie möglich gehalten wird. Um hochwertige Daten aus dem Versorgungsalltag praktikabel und ohne große Hürden vorhalten und bereitstellen zu können, muss eine dauerhaft verfügbare und fortgeführte Dateninfrastruktur und Dokumentation personell und finanziell in den Registern aufgebaut werden. Sowohl der Aufbau als auch der Unterhalt dieser Infrastruktur könnten durch die pUs, die auf die Daten zur Durchführung registerbasierter Studien zurückgreifen wollen, unterstützt werden. Des Weiteren wurden als förderliche Faktoren u. a.

- eine Beschränkung der Erhebung auf jene Daten, die für die Beantwortung der zu untersuchenden Fragestellung relevant sind,
- die Implementierung einer geeigneten Monitoringstrategie,
- ein regelmäßiges Schulungsangebot für das Erhebungspersonal zur Dateneingabe und
- das Zurückspielen der im Register gespeicherten Daten an die Patientinnen und Patienten (inklusive Aufklärung über den Nutzen der Erfassung patientenberichteter Daten) identifiziert.

Methodischer Umgang mit fehlenden Werten im Kontext von Propensity-Score-Analysen

Modelle zur Schätzung von Propensity Scores mittels logistischer Regression erfordern für alle potenziell relevanten Confounder vollständige Daten, d. h., dass Patientinnen und Patienten mit (mindestens) 1 fehlendem Wert in der Analyse nicht berücksichtigt werden. Wenn jedoch ausschließlich Patientinnen und Patienten mit vollständigen Daten in die Analysen eingehen (Complete-Case-Analysen), führt dies zu einer deutlichen Verzerrung des Ergebnisses, wenn die Annahme Missing completely at Random (MCAR) nicht erfüllt ist. Davon ist in Studien auf Basis versorgungsnaher Daten in aller Regel auszugehen. Demnach sind bei Propensity-Score-Analysen statistische Verfahren zum Umgang mit fehlenden Werten notwendig. Grundsätzlich werden Verfahren der multiplen Imputation als angemessen angesehen. Es steht eine Reihe von statistischen Verfahren der multiplen Imputation zur Verfügung, von denen die multiple Imputation mittels Chained Equations (MICE) als etablierte Methode zum Umgang mit fehlenden Werten anerkannt. Zur Überprüfung der Robustheit der so berechneten Ergebnisse kommen Sensitivitätsanalysen mit modellbasierten Methoden in Betracht, denen jeweils die MAR- oder Missing-not-at-Random(MNAR)-Annahme zugrunde liegt.

Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven

Bei Studien ohne Randomisierung ist die für einen fairen Vergleich notwendige Strukturgleichheit der zu vergleichenden Gruppen in der Regel nicht gegeben. Demnach müssen Gruppens unterschiede in möglichen Confoundern bei Schätzung von Effekten durch Adjustierung für diese Confounder berücksichtigt werden. Zur Confounderadjustierung im Kontext einer AbD spielt primär die Propensity-Score-Methode eine Rolle. Propensity-Score-Verfahren wurden ursprünglich für große Datensätze (epidemiologische Fragestellungen) entwickelt. Im Rahmen der vorliegenden Ausarbeitung wurden Simulationsstudien zu Propensity-Score-Analysen bei kleinen Fallzahlen identifiziert. Diese Studien zeigen, dass Propensity-Score-Verfahren auch in Anwendungsgebieten mit kleinen Patientenpopulationen unter bestimmten Bedingungen anwendbar sind und zu interpretierbaren Ergebnissen führen können. Ein relevantes Problem besteht bei dieser Methode allerdings in der Gefahr der fehlenden Konvergenz der Modelle für die Effektschätzung. Dem Problem der fehlenden Konvergenz durch eine zu hohe Anzahl von Confoundern kann bis zu einem gewissen Grad dadurch begegnet werden, dass weniger wichtige Confounder schrittweise aus dem Modell entfernt werden. Jenes Vorgehen erfordert eine präspezifizierte Festlegung einer Reihenfolge der Confounder nach ihrer Wichtigkeit und erfolgt unter Inkaufnahme einer erhöhten Unsicherheit. Insgesamt bilden die in den Simulationsstudien untersuchten Szenarien die Situationen, wie sie in den laufenden AbD-Verfahren absehbar auftreten werden, jedoch nur in Teilen ab, da diese vorwiegend andere Effektmaße (wie Odds Ratio oder Risikodifferenz) und anteilig eine geringe Anzahl potenzieller Confounder betrachten. Zum jetzigen Zeitpunkt bleibt aufgrund von fehlender Evidenz daher unklar, in welchen Fällen (und unter welchen Voraussetzungen) Propensity-Score-Analysen in einer AbD zu interpretierbaren Ergebnissen führen werden.

Fazit

Für die Nutzenbewertung von Arzneimitteln bedarf es Daten zum Vergleich mit dem therapeutischen Standard. Da die Zulassung von Orphan Drugs häufig auf Basis nicht vergleichender Daten erfolgt, wurde das Verfahren der AbD eingeführt, mit dem Ziel, bestehenden Evidenzlücken zu schließen und so eine bessere Datenbasis für die Nutzenbewertung zu erhalten. Die Datenerhebungen sind dabei als vergleichende Studien ohne Randomisierung durchzuführen. Unter Einhaltung bestimmter Qualitätsanforderungen können Studien auf Basis von Registerdaten diese Erkenntnislücke schließen.

- Bei der Planung von vergleichenden Studien ohne Randomisierung auf Basis versorgungsnaher Daten ist die Target-Trial-Emulation ein empfohlener Ansatz, um systematische (vermeidbare) Verzerrungen zu minimieren. Voraussetzung für eine optimale Emulation einer hypothetischen RCT mittels Beobachtungsdaten ist, dass die hierfür erforderlichen Daten im Registerdatensatz mit der notwendigen Vollständigkeit und Datentiefe enthalten sind. Eine hohe Datenqualität kann in der Breite nur dann erreicht werden, wenn die Datengenerierung und die Nutzbarmachung von Registerdaten praktikabel und mit möglichst wenig Aufwand möglich sind. Als förderlich wird der Aufbau und der Unterhalt einer dauerhaft verfügbaren (funktionsfähigen) Dateninfrastruktur eingeschätzt. Dies könnte durch pUs, die auf die Register zur Durchführung registerbasierter Studien zurückgreifen wollen, unterstützt werden.
- In vergleichenden Studien ohne Randomisierung zum Zweck des Vergleichs von Therapieeffekten ist für eine adäquate Confounderkontrolle eine systematische Identifikation der relevanten Confounder sowie eine Berücksichtigung dieser Confounder in der Auswertung notwendig. Die Confounderidentifikation nach dem Ansatz von Pufulete 2022 (über eine systematische Literaturrecherche und Einbindung von Klinikerinnen und Klinikern) wird als machbar eingeschätzt und stellt ein grundsätzlich sinnvolles Vorgehen dar. Vor der Relevanzbewertung der Confounder durch klinische Expertinnen und Experten ist zu empfehlen, eine intensive Verdichtung der identifizierten Confounder vorzunehmen.
- In den Konzepten für eine AbD wird abgeschätzt, ob innerhalb eines akzeptablen Zeitraums ausreichend Patientinnen und Patienten eingeschlossen werden können, um mit der AbD grundsätzlich aussagekräftige Ergebnisse für eine Nutzenbewertung generieren zu können. In der Regel liegen nur unsichere Informationen für diese Abschätzung vor. Wenn hinreichende Informationen zur Intervention und dem Komparator vorliegen, wird zu diesem Zwecke eine orientierende Fallzahlschätzung durchgeführt. Fehlen notwendige Informationen für die Annahmen, die einer orientierenden Fallzahlschätzung zugrunde liegen, erfolgt eine orientierende Betrachtung von Fallzahlszenarien, um mögliche aufdeckbare Effekte zu zeigen. Beide Herangehensweisen folgen dabei dem gleichen Grundsatz (sie fußen auf identischen

Parametern) und sie unterscheiden sich lediglich in dem Parameter, der mithilfe der übrigen Parameter geschätzt werden soll.

- Für eine langfristige Datenerhebung (und Begleitung von Patientinnen und Patienten) in der Routineversorgung sind Anreize erforderlich, die den Aufwand der Datengenerierung kompensieren und sowohl die Zentren als auch die Patientinnen und Patienten zur möglichst vollständigen Datenerhebung motivieren.
- Zur Sicherstellung, dass Aufwand und Nutzen bei PRO-Erhebungen in einem angemessenen Verhältnis stehen, werden digitale Befragungen (z. B. per App oder per Patientenportal) empfohlen. Auf diese Weise ist eine niedrigschwellige PRO-Erhebung unabhängig von Arztbesuchen möglich. Digitale Tools werden in Registern und in der Routineversorgung bereits erfolgreich zur Datenerhebung eingesetzt und ermöglichen, dass PRO-Daten für Forschungszwecke barrierearm zur Verfügung gestellt werden können.
- Eine Herausforderung bei der Auswertung versorgungsnaher Daten ohne Randomisierung stellt die Festlegung des Startzeitpunkts der Beobachtung (Index-Datum [t_0]) dar. Wird in beiden Behandlungsgruppen zum Beobachtungsstart (t_0) mit einer neuen Therapie begonnen (Umstellungsindikation), kann in einer Studie das ACNU-Design verwendet werden. Dabei entspricht t_0 idealerweise dem Tag der Therapieentscheidung. In einer Therapiesituation, in der Patientinnen und Patienten in der Kontrollgruppe die Behandlung mit einem etablierten therapeutischen Standard fortführen, stellt das PNU-Design eine geeignete Alternative dar.
- Unabhängig von der Wahl des Designs und der Auswertungsstrategien sollte grundsätzlich immer primär das Ergebnis der ITT-Analyse (entsprechend der Treatment-Policy-Strategie) gezeigt werden. Dies erfordert, dass alle Patientinnen und Patienten (unabhängig von einem Behandlungswechsel im Verlauf der Beobachtung) gemäß ihrer ursprünglichen Gruppenzuteilung ausgewertet werden. Dem verzerrenden Einfluss von Behandlungswechseln im Verlauf der AbD kann mit Sensitivitätsanalysen begegnet werden.
- Eine häufig angewendete Methode zur Berücksichtigung von Confoundern in vergleichenden Studien ohne Randomisierung auf Basis von Registern stellt die Auswertung auf Basis von Propensity Scores dar. Da Modelle zur Schätzung von Propensity Scores mittels logistischer Regression für alle potenziell relevanten Confounder vollständige Daten erfordern, sind statistische Verfahren zum Umgang mit fehlenden Werten notwendig. Als etablierte Methode ist die multiple Imputation mittels Chained Equations (MICE) anerkannt. Propensity-Score-Verfahren sind auch in kleinen Patientenkollektiven unter bestimmten Bedingungen anwendbar und können zu interpretierbaren Ergebnissen führen. Die in den identifizierten Simulationsstudien untersuchten Szenarien bilden die Situationen, wie sie in den laufenden AbD-Verfahren

absehbar auftreten werden, in Teilen jedoch nicht ab. Zum jetzigen Zeitpunkt bleibt daher unklar, in welchen Fällen (und unter welchen Voraussetzungen) Propensity-Score-Analysen in einer AbD zu interpretierbaren Ergebnissen führen werden.

1 Hintergrund

Laut § 35a Absatz 3b Satz 1 Sozialgesetzbuch (SGB) V kann der Gemeinsame Bundesausschuss (G-BA) bei bestimmten Arzneimitteln vom pharmazeutischen Unternehmer (pU) die Vorlage anwendungsbegleitender Datenerhebungen (AbD) und Auswertungen zum Zwecke der Nutzenbewertung fordern [25]. Aufgrund der vergleichenden Fragestellung der Nutzenbewertung handelt es sich dabei regelhaft um eine vergleichende Studie, die unter Berücksichtigung der Vorgaben gemäß § 35a SGB V Absatz 3b ohne Randomisierung durchgeführt wird.

Ein Beschluss zur Forderung einer AbD legt die Anforderungen an eine AbD und deren Auswertung fest. Wenn der G-BA eine AbD fordert, muss der pU ein Studienprotokoll und einen statistischen Analyseplan (SAP) erstellen. In diesen Dokumenten sind Planung, Durchführung und Auswertung der Studie zum Zwecke der AbD zu beschreiben. Der G-BA prüft, ob das in diesen Dokumenten beschriebene Vorgehen geeignet ist, die Anforderungen an die AbD zu erfüllen. Für die Prüfung kann er das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) einbinden [26].

Eine allgemeine Ausarbeitung von Kriterien für die Qualität versorgungsnaher Daten und von methodischen Anforderungen für die Generierung von versorgungsnahen Daten und deren Auswertung wurde durch das IQWiG auf Basis eines Auftrages des G-BA bereits durchgeführt und als Rapid Report A19-43 auf der Website des IQWiG veröffentlicht [18] (im Folgenden als „A19-43“ bezeichnet).

Mittlerweile liegen erste Erfahrungen sowohl mit der Prüfung der von den pUs eingereichten Studienunterlagen hinsichtlich ihrer inhaltlichen und methodischen Eignung zur Durchführung der AbD (einschließlich der zugehörigen Auswertung) als auch mit laufenden AbDs in Registern als Datenquelle vor (registerbasierte Studien). Vor dem Hintergrund ergibt sich der Bedarf für eine weiterführende – über A19-43 hinausgehende – wissenschaftliche Ausarbeitung ausgewählter Aspekte zur Generierung versorgungsnaher Daten und deren Auswertung, jeweils im Kontext nicht randomisierter vergleichender Studien.

2 Fragestellung

Ziel des vorliegenden Berichts ist eine wissenschaftliche Ausarbeitung zur Methodik (Generierung und Auswertung versorgungsnaher Daten zum Zwecke der Nutzenbewertung) im Bereich der AbD mit einer Fokussierung auf eine Bewertung der folgenden Themenbereiche, jeweils im Kontext nicht randomisierter vergleichender Studien:

- Confounderidentifikation und -auswahl (siehe Abschnitt 5.2)
- Schätzung notwendiger Fallzahlen vor dem Hintergrund unzureichender Vorabinformationen (siehe Abschnitt 5.3)
- Beobachtungsstart (siehe Abschnitt 5.4.1)
- Behandlungswechsel (siehe Abschnitt 5.4.2)
- patientenberichtete Endpunkte (PROs) (siehe Abschnitt 5.4.2)
- fehlende Werte (siehe Abschnitt 5.6)
- Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven (siehe Abschnitt 5.7)

Dabei sollen eine orientierende Recherche nach geeigneter Literatur für die oben genannten Themenbereiche und eine Sichtung der in den bisherigen Prüfungen von Studienprotokoll und SAP identifizierten Mängel (ggf. mit Ergänzung der genannten Themenbereiche) erfolgen.

Für eine AbD eignen sich primär bereits bestehende Register als potenzielle Datenquelle. Die Ausarbeitung fokussiert daher auch auf Patientenregister als Instrumente zur Datenerhebung mit dem Ziel der Generierung versorgungsnaher Daten, da andere Datenquellen (z. B. GKV-Routinedaten) zum aktuellen Zeitpunkt nicht geeignet erscheinen, die Anforderungen an die Daten(qualität) im Rahmen der Nutzenbewertung zu erfüllen.

3 Projektverlauf

Der Unterausschuss Arzneimittel hat für den G-BA gemäß 1. Kapitel § 4 Absatz 2 Satz 3 Buchstabe a) Verfahrensordnung in seiner Sitzung am 28.01.2025 beschlossen, das IQWiG mit einer wissenschaftlichen Ausarbeitung zur Methodik der AbD zu beauftragen. Hierbei sollen spezifische Themen, die sich aus den bisherigen Erfahrungen mit der AbD ergeben haben, bearbeitet werden (siehe hierzu Kapitel 2).

In die Bearbeitung des Projekts wurde 1 externe Sachverständige eingebunden.

Auf Basis einer internen Projektskizze fand eine wissenschaftliche Aufarbeitung methodischer Anforderungen an vergleichende Studien ohne Randomisierung statt, die in einem vorläufigen Rapid Report veröffentlicht und auftragsgemäß zur Anhörung gestellt wird. Im Anschluss an die Anhörung erstellt das IQWiG einen finalen Rapid Report. Dieser Bericht wird an den G-BA übermittelt und 4 Wochen später auf der Website des IQWiG veröffentlicht. An dieser Stelle wird auch die Dokumentation der Anhörung zum vorläufigen Rapid Report veröffentlicht.

4 Methoden

4.1 Sichtung der bisherigen Prüfungen von Studienprotokoll und statistischem Analyseplan

Die Änderungsbedarfe, die in bisherigen Prüfungen der von den pUs eingereichten Studienprotokolle und SAPs adressiert wurden, wurden gesichtet. Auf Basis dieser Sichtung sollten die beauftragten Themenbereiche, sofern erforderlich, um weitere Bereiche ergänzt werden. Da alle identifizierten Änderungsbedarfe einem der 7 Themenbereiche zugeordnet werden konnten, wurden im Rahmen des Auftrages keine weiteren Themenbereiche bearbeitet.

4.2 Vorschlag für die Anforderung an eine systematische Confounderidentifikation

Für den Themenbereich Confounderidentifikation und -auswahl wurde für den vorliegenden Auftrag auf das Arbeitspapier GA23-02 [27] zurückgegriffen, das vom IQWiG am 18.06.2025 veröffentlicht wurde. Im Rahmen des vorliegenden Rapid Reports fand somit keine erneute Erarbeitung der Methodik statt. Die Darstellung im vorliegenden Bericht beschränkt sich dabei im Wesentlichen auf die abgeleiteten konkreten Handlungsempfehlungen zur systematischen Confounderidentifikation für nicht randomisierte vergleichende Studien (und somit auch für die AbD). Für weiterführende Informationen (z. B. zur notwendigen Dokumentation) wird auf das Arbeitspapier GA23-02 verwiesen.

4.3 Informationsbeschaffung

Für die weiteren 6 Themenbereiche, auf die sich der vorliegende Rapid Report gemäß Auftrag des G-BA fokussiert, umfasste die orientierende Informationsbeschaffung:

- eine Literaturrecherche nach methodischen Dokumenten und Übersichtsarbeiten in PubMed
- eine Website-Suche nach methodischen Dokumenten der Zulassungsbehörden (European Medicines Agency [EMA] und Food and Drug Administration [FDA]) und anderer HTA-Agenturen (National Institute for Health and Care Excellence [NICE], Canada's Drug Agency [CDA], Agency for Healthcare Research and Quality [AHRQ], Haute Autorité de Santé [HAS] und Austrian Institute for Health Technology Assessment [AIHTA])
- eine gezielte Websuche
- weitere Suchtechniken:
  - Sichten von Referenzlisten der identifizierten methodischen Dokumente und Übersichtsarbeiten
  - Verwendung der „Similar Articles"-Funktion in PubMed sowie der „Cited-by"-Funktion in OpenAlex ausgehend von zuvor identifizierten methodischen Dokumenten und Übersichtsarbeiten

Die Darstellung im Bericht beschränkt sich auf die konkreten Ergebnisse.

5 Bewertung ausgewählter Aspekte zur Generierung versorgungsnaher Daten im Kontext nicht randomisierter vergleichender Studien

5.1 Einleitende Anmerkungen zu den Anforderungen an vergleichende Studien ohne Randomisierung zum Zwecke der Nutzenbewertung

Analog zu A19-43 [18] sind versorgungsnahe Daten zum Zwecke der Nutzenbewertung von Arzneimitteln für den vorliegenden Bericht wie folgt definiert:

- Versorgungsnahe Daten werden bei den Patientenpopulationen erhoben, für die eine Indikation für das zu bewertende Arzneimittel im Rahmen der Zulassung besteht.
- Bei der Erhebung versorgungsnaher Daten erfolgt die Behandlung ohne spezifische Vorgaben.

Da es bei der Bewertung von Arzneimitteln nach SGB V um die Versorgung von Patientinnen und Patienten in Deutschland geht, müssen versorgungsnahe Daten die beiden vorgenannten Kriterien so erfüllen, dass sie Rückschlüsse auf die deutsche Versorgung zulassen.

Gemäß A19-43 sind folgende grundsätzliche Aspekte bei der Erhebung und Auswertung von Daten aus der Versorgung zu beachten:

- Für die Erhebung von versorgungsnahen Daten ist eine adäquate Datenplattform erforderlich. Idealerweise wird hierfür auf bereits bestehende Datenstrukturen zurückgegriffen. Für die versorgungsnahe Datenerhebung zum Zwecke der Nutzenbewertung kommen insbesondere Patientenregister infrage (registerbasierte Studien). Alternativ stellt eine studienindividuelle Datenerhebung eine Option für die Durchführung einer AbD dar.
- Sollen versorgungsnahe vergleichende Studien ohne Randomisierung für die Nutzenbewertung herangezogen werden, muss bereits mit der Studienplanung sichergestellt werden, dass der Ablauf der Studie und die erhobenen Daten die notwendige Qualität haben, um interpretierbare Ergebnisse zu generieren.
  Dabei lassen sich national und international weitgehend übereinstimmende Qualitätskriterien für Patientenregister und registerbasierte Studien ableiten.
- Wesentliche Komponenten der Studienplanung sind Überlegungen zur *Emulation* einer hypothetischen randomisierten kontrollierten Studie (RCT [Target Trial]), die die relevante Fragestellung behandelt, inklusive der Sicherstellung der Erhebung ausreichender Daten für die Confounderkontrolle. Die Aspekte der Studienplanung sind präspezifiziert in einem Studienprotokoll (inklusive SAP) zu dokumentieren.
- Ein zentraler Aspekt der Analyse einer vergleichenden Studie ohne Randomisierung ist zum Ausgleich des Einflusses der Strukturungleichheit der Behandlungsgruppen eine

adäquate Adjustierung für *Confounder*. Die relevanten Confounder müssen systematisch identifiziert und in der Studienplanung präspezifiziert werden.

- Wegen der inhärenten Unsicherheit der Ergebnisse aus vergleichenden Studien ohne Randomisierung durch potenziell unbekannte Confounder kann selbst bei Erfüllung von hohen Qualitätsanforderungen aus den in einer Studie beobachteten Effekten erst ab einer bestimmten Effektstärke eine Aussage zum Nutzen oder Schaden einer Intervention abgeleitet werden. Eine (positive oder negative) Aussage zum Nutzen oder Schaden ergibt sich dann, wenn das Konfidenzintervall für den beobachteten Effekt ober- bzw. unterhalb einer zu definierenden Schwelle liegt (Test auf verschobene Nullhypothese). Die konkrete Schwelle ergibt sich durch die Qualität der Daten im Einzelfall.

Darauf aufbauend werden im Folgenden wesentliche Grundzüge der Planung vergleichender Studien ohne Randomisierung auf Basis versorgungsnaher Daten beschrieben. Dies bildet den Ausgangspunkt für die Aufbereitung der für diesen Bericht beauftragten Themenbereiche, die in den Abschnitten 5.2 bis 5.7 detailliert dargelegt werden.

Für vergleichende Studien ohne Randomisierung auf Basis von versorgungsnahen Daten zum Vergleich der Wirksamkeit von Interventionen wird die explizite Nachbildung der Planung von vergleichenden Studien mit Randomisierung empfohlen. Als Best-Practice-Ansatz hat sich das Konzept der Target-Trial-Emulation bewährt [28,29]. Hierbei wird angestrebt, den Effekt einer hypothetischen RCT (Target Trial), die die zu untersuchende Fragestellung behandelt, mittels Beobachtungsdaten nachzubilden (zu emulieren). In diesem Zuge werden sowohl das Studienprotokoll als auch der SAP so harmonisiert, dass sie möglichst auf den gleichen kausalen Effekt zielen [30]. In Tabelle 1 sind die Komponenten der Studienplanung für die Nachbildung eines Target Trials aus einem nicht randomisierten Datensatz anhand eines Beispiels dargestellt.

Tabelle 1: Vergleich eines Studienprotokolls einer RCT und einer Studie auf Basis versorgungsnaher Daten mittels Target-Trial-Emulation (eigene Darstellung nach Hoffmann 2021 [31])

Komponente	RCT (Target Trial)	Studie mit versorgungsnahen Daten (Target-Trial-Emulation)
Ziel	Untersuchung des Effekts einer Hormonersatztherapie bei postmenopausalen Frauen auf das Brustkrebsrisiko innerhalb von 5 Jahren	identisch zum Target Trial
Einschlusskriterien	▪ Frauen mit Menopause seit mindestens 5 Jahren ▪ keine Tumordiagnose ▪ keine Hormonersatztherapie in den letzten 2 Jahren	identisch zum Target Trial
Interventionen	**A:** Beginn einer Hormonersatztherapie **B:** kein Beginn einer Hormonersatztherapie	identisch zum Target Trial
Zuteilung	Randomisierung zu Baseline, keine Verblindung	**abweichend:** ▪ Zuteilung zu Therapiebeginn (Beginn einer Hormonersatztherapie) bzw. zum Zeitpunkt der Therapieentscheidung (kein Beginn einer Hormonersatztherapie) ▪ Nachbildung der Randomisierung durch umfassende Adjustierung für Baselinecharakteristika (alle relevanten Confounder müssen vorliegen)
Follow-up	**Beginn:** Behandlungszuteilung **Ende:** gesicherte Brustkrebsdiagnose, Tod, Lost-to-Follow-up oder 5 Jahre nach Baseline, je nachdem, was zuerst eintritt	identisch zum Target Trial (besondere Herausforderung bei Beobachtungsstudien ist der Beginn der Beobachtung [des Follow-up] in der Gruppe ohne Therapie)
Endpunkt	Brustkrebsdiagnose, durch Biopsie gesichert, innerhalb von 5 Jahren nach Randomisierung	identisch zum Target Trial
Effekt zwischen Interventionen („causal contrast“) und Analyse	Effekt einer Zuteilung zur Hormonersatztherapie vs. zu keiner Hormonersatztherapie	**abweichend:** Effekt eines Beginns der Hormonersatztherapie vs. kein Beginn einer Hormonersatztherapie (Adjustierung für relevante Confounder ist obligat)
RCT: randomisierte kontrollierte Studie		

Für die Nachbildung vergleichender Studien mit Randomisierung ist es notwendig, dass der Datensatz, der für die vergleichende Studie ohne Randomisierung vorgesehen ist, alle hierfür notwendigen Informationen in hoher Qualität enthält [32].

Die Verwendung des Konzepts der Target-Trial-Emulation zum Zeitpunkt der Studienplanung hilft dabei, Quellen für systematische Verzerrung (Bias) und Emulationsunterschiede (Fehlen einer ausreichenden Übereinstimmung zwischen Target Trial und Emulation) zu identifizieren

und zu vermeiden [9,33]. Häufige Quellen für Verzerrung sind eine strukturelle Verletzung der *Positivität*, eine Strukturungleichheit zwischen den Gruppen (Bias aufgrund von Confounding) sowie eine falsche Wahl des Startzeitpunktes der Beobachtung (führt u. a. zu *Immortal-Time-Bias* und / oder *Selection Bias*).

Sicherstellung der Positivität

In einer RCT wird jeder Teilnehmer durch die Randomisierung zufällig einer Therapiestrategie zugeteilt. Ausgehend vom Target-Trial-Ansatz ist in versorgungsnahen vergleichenden Studien ohne Randomisierung Voraussetzung, dass für alle Patientinnen und Patienten, welche in die Auswertungen eingehen, die zu vergleichenden Behandlungsformen (sowohl neuer Wirkstoff als auch der Komparator) eine potenzielle Therapieoption darstellen (Positivität) [34,35]. Das bedeutet, dass beispielsweise Patientinnen und Patienten mit Kontraindikationen gegen eine der untersuchten Behandlungsformen nicht in der Auswertung berücksichtigt werden dürfen. Dies ist über die Prüfung der Eignung der Patientinnen und Patienten auf Basis der präspezifizierten Einschlusskriterien (von dem Target Trial und der Emulation) sicherzustellen [34].

Adäquate Adjustierung für Confounder

In vergleichenden Studien ohne Randomisierung soll mittels einer adäquaten Adjustierung für Confounder die fehlende Strukturgleichheit zwischen den Behandlungsgruppen ausgeglichen werden (Bias aufgrund von Confounding). Ein Bias aufgrund von Confounding entsteht durch eine Verletzung der Zufallsprinzipien bei der Stichprobenziehung, d. h. während der Zuteilung der Patientinnen und Patienten zu den Behandlungsgruppen. Gerade bei Gruppenvergleichen kann Bias aufgrund von Confounding zu systematischen Unterschieden zwischen den Gruppen führen. Sind infolgedessen wichtige Confounder in den Gruppen ungleich verteilt, so sind die Ergebnisse eines Vergleichs in aller Regel nicht mehr interpretierbar [34]. Eine Variable wird als Confounder (Störgröße) definiert, wenn sie sowohl die Therapieentscheidung als auch den Endpunkt beeinflusst und dadurch einen Behandlungseffekt verfälschen kann (wie Alter oder Schweregrad der Erkrankung) [3]. Beim Vergleich von Gruppen stellt die Randomisierung die beste Methode zur Vermeidung von Bias aufgrund von Confounding dar, da sich die so entstehenden Gruppen weder in bekannten noch in unbekannten Confoundern systematisch unterscheiden [36].

In vergleichenden Studien ohne Randomisierung sind für die Ermittlung von Therapieeffekten eine systematische Identifikation potenziell relevanter Confounder und eine Berücksichtigung dieser Confounder in der Auswertung Voraussetzung für eine adäquate Confounderkontrolle. Ziel dieser Confounderkontrolle ist es, trotz fehlender Randomisierung, eine möglichst niedrig verzerrte Schätzung des interessierenden kausalen Effekts zu erhalten. Dabei ist es notwendig, im Vorfeld alle relevanten Confounder (inklusiver wichtiger Wechselwirkungen) mithilfe eines systematischen Ansatzes zu identifizieren, vollständig zu erheben und bei der Datenanalyse in angemessener Form im Modell zu berücksichtigen [18].

Richtige Wahl des Startzeitpunkts der Beobachtung (Index-Datum [t_0])

Eine Herausforderung bei der Emulation eines Target Trials stellt eine adäquate Festlegung des Startzeitpunkts der Beobachtung (Index-Datum [t_0]) dar. Im Gegensatz zu RCTs (Zeitpunkt der Randomisierung) haben versorgungsnahe Studien ohne Randomisierung keinen festen Startzeitpunkt für die Beobachtung [21]. Um dieses Problem zu adressieren, sollen zu t_0 (in dem Target Trial sowie in der nicht randomisierten Studie) die in Abbildung 2 dargestellten Bedingungen erfüllt sein.

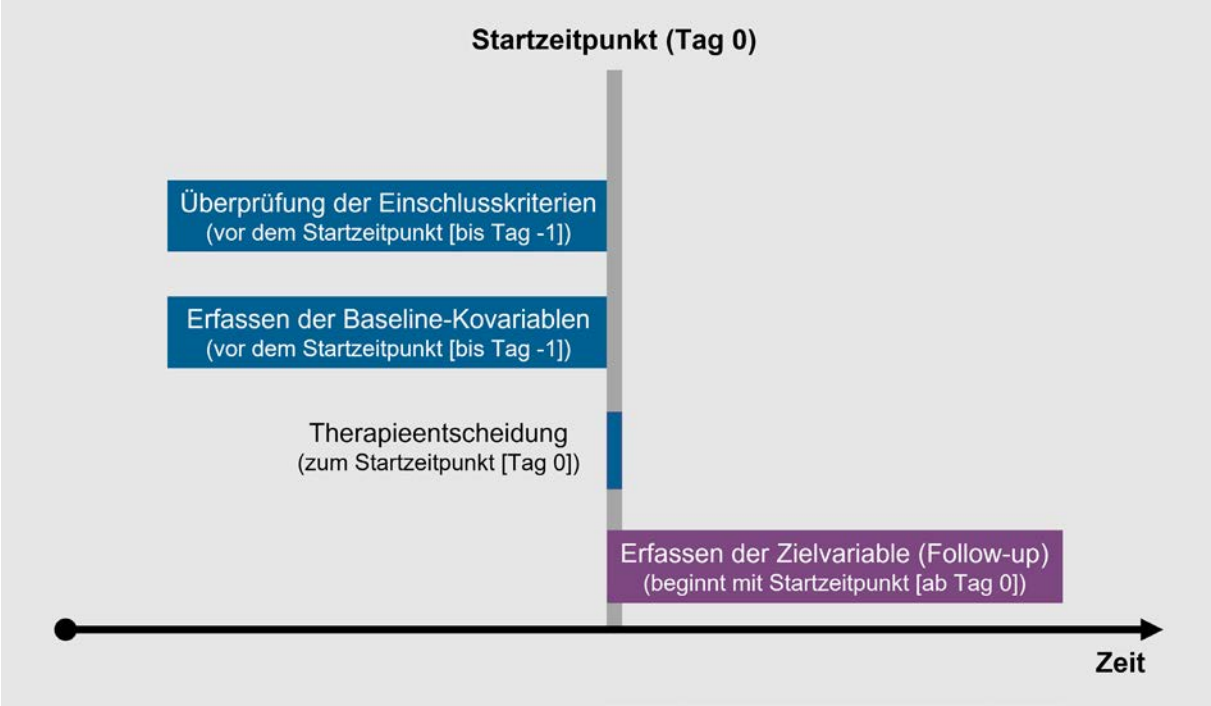


Abbildung 2: Zeitliche Ausrichtung zentraler Elemente des Studiendesigns am Startzeitpunkt der Beobachtung (t_0) (eigene Darstellung, in Anlehnung an Braitmaier 2022 [37])

Die Ein- und Ausschlusskriterien müssen zum Zeitpunkt der Behandlungszuteilung erfüllt sein [9,28]. Idealerweise sollte für t_0 in beiden Gruppen der Tag der Therapieentscheidung gewählt werden. Hierfür sind Informationen über die Behandlungsintention erforderlich (in einer RCT ergibt sich dies aus der Zuteilungssequenz) [33,38]. Mit der Behandlungszuteilung zu t_0 startet die Beobachtung zur Erfassung der Endpunkte (Follow-up).

5.2 Confounderidentifikation und -auswahl

In nicht randomisierten vergleichenden Studien spielen eine adäquate Identifikation von und Adjustierung für potenzielle *Confounder* eine zentrale Rolle. Nach Pufulete 2022 [39] umfasst die Confounderidentifikation 3 methodische Bausteine: (1) systematische Literaturrecherche, (2) Interviews mit Klinikerinnen und Klinikern sowie (3) eine Umfrage unter Klinikerinnen und Klinikern.

In dem Arbeitspapier GA23-02 des IQWiG [27] wurde gemäß Pufulete 2022 eine systematische Confounderidentifikation in der Indikation schubförmig remittierende multiple Sklerose (RRMS) an einem geeigneten Vergleich zweier medikamentöser Therapien mit Hinblick auf wesentliche patientenrelevante Endpunkte durchgeführt. Das Ziel war es, die grundsätzliche Machbarkeit des von Pufulete 2022 vorgeschlagenen Vorgehens zur Confounderidentifikation einzuschätzen. Hierfür wurde eine Arbeitsgruppe von externen Sachverständigen beauftragt, die ihre Ergebnisse in einem Bericht dargestellt hat. Auf Basis dieses Ergebnisberichts wurden im Rahmen der Bearbeitung von GA23-02 anhand weiterführender Untersuchungen durch das IQWiG konkrete Handlungsempfehlungen zur systematischen Identifikation von Confoundern für nicht randomisierte vergleichende Studien (und somit auch für die AbD) abgeleitet.

Obschon sich der Bericht der externen Sachverständigen (ebenso wie die Arbeit von Pufulete 2022) auf eine Indikation mit umfangreicher Studienlage bezieht (RRMS), wurde bei Ableitung der Handlungsempfehlungen auch die Situation der selten(er)en Erkrankungen berücksichtigt, da dies die regelhafte Situation, z. B. im Rahmen der AbD, darstellt. Insofern hierfür besondere Anforderungen bestehen, wird dies nachfolgend explizit beschrieben.

Übergeordnet gliedert sich der erarbeitete Vorschlag für die konkreten Anforderungen an eine systematischen Confounderidentifikation in 3 Schritte (siehe Abbildung 3):

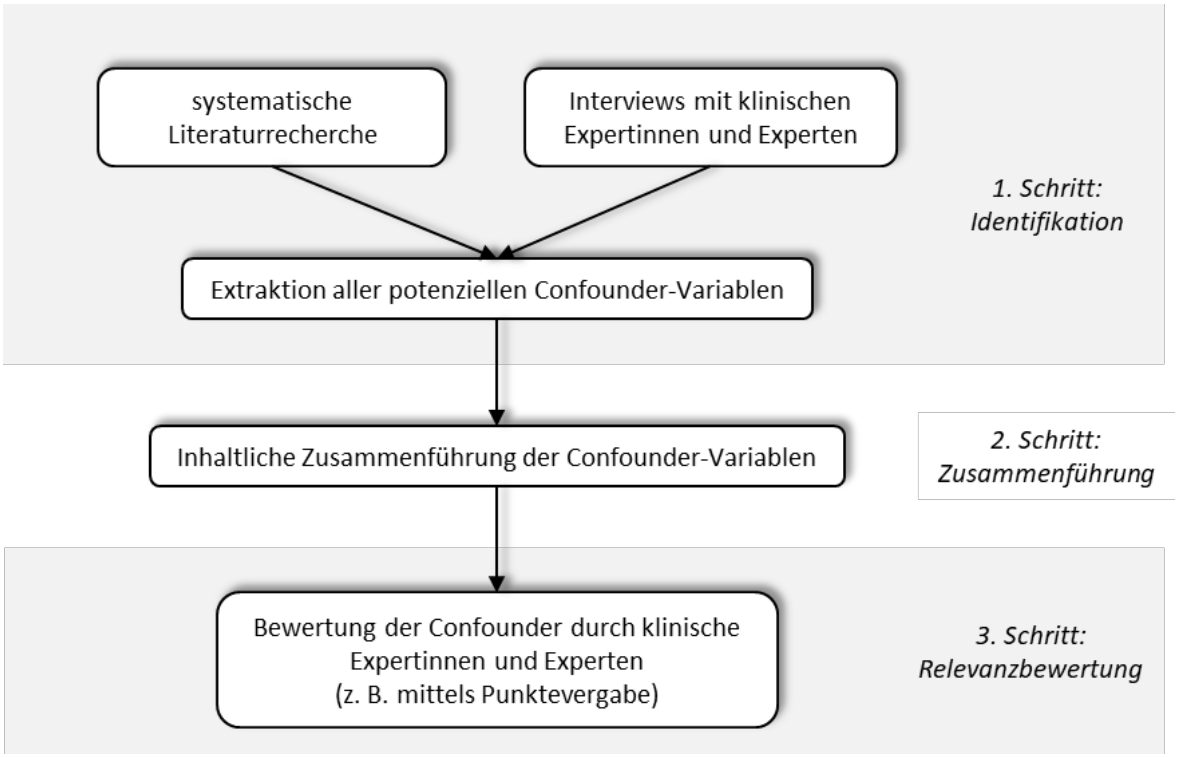


Abbildung 3: Vorgeschlagenes 3-schrittiges Verfahren zur systematischen Confounderidentifikation

Schritt 1: Identifikation mittels systematischer Literaturrecherche und Interviews mit klinischen Expertinnen und Experten

Angelehnt an Pufulete 2022 umfasst der 1. Schritt der Confounderidentifikation 2 Bausteine: (1) eine systematischen Literaturrecherche und (2) Interviews mit klinischen Expertinnen und Experten. Bei der systematischen Informationsbeschaffung sollten Primärpublikationen zu RCTs und Kohortenstudien einbezogen werden. Bei Fragestellungen in Indikationen mit einer absehbar weniger umfassenden Studienlage (beispielsweise bei seltenen Erkrankungen), kann bei Fehlen von geeigneten RCTs und Kohortenstudien auf 1-armige Studien (inklusive Vorher-Nachher-Vergleiche) und die in den jeweiligen Primärpublikationen in tabellarischer Form aufgeführten Patientencharakteristika zurückgegriffen werden. Gleichermaßen besteht in diesen Fällen die Option, z. B. Leitlinien in die Literaturrecherche einzubeziehen und diese nach Prognose- oder Risikofaktoren und nach Faktoren, die die Therapieentscheidung bestimmen, zu durchsuchen. Die Ein- und Ausschlusskriterien, die die Studien erfüllen müssen, um in den Studienpool für die Confounderextraktion eingeschlossen zu werden, sollten sich eng an der zu untersuchenden Fragestellung orientieren. Bei der Extraktion sollten bei vergleichenden Beobachtungsstudien alle Faktoren berücksichtigen werden, die für die Adjustierung in den statistischen Analysen verwendet wurden, losgelöst von der statistischen Signifikanz. Bei RCTs oder (je nach Datenlage) 1-armigen Studien sollten die tabellarisch berichteten Patientencharakteristika („Tabelle 1“) extrahiert werden.

Für die Extraktion aus Primärpublikationen wird die Anwendung eines geeigneten Sättigungskriteriums empfohlen, um insbesondere bei einer hohen Anzahl identifizierter Publikationen Ressourcen einsparen zu können.

Der 2. Baustein zur Confounderidentifikation entspricht der Durchführung von Interviews mit klinischen Expertinnen und Experten. Als Entscheidungskriterien für die Festlegung der Anzahl von durchzuführenden Interviews können z. B. die Indikation, die erwartbare Studienlage oder die Variabilität im Erkrankungsverlauf herangezogen werden. Anzunehmen ist, dass in jenen Indikationen, in denen ggf. eine geringere Menge an Publikationen zur Verfügung steht, den Interviews eine größere Bedeutung zuteilwird. Grundsätzlich sollte transparent dokumentiert werden, wie viele klinische Expertinnen und Experten mit welchem Fachgebiet eingebunden wurden. Die eingebundenen Expertinnen und Experten sind namentlich zu benennen und ihre potenziellen Interessenkonflikte offenzulegen.

Schritt 2: Zusammenführung

Im 2. Schritt wird nach der Extraktion aller identifizierten potenziellen Confoundervariablen aus 1) der systematischen Literaturrecherche und 2) den qualitativen Interviews mit klinischen Expertinnen und Experten empfohlen – wo sinnvoll – eine Zusammenführung von inhaltlich überlappenden Confoundervariablen vorzunehmen. Für jede der identifizierten Confoundervariablen ist dabei zu prüfen, ob sie innerhalb der zu untersuchenden Fragestellung liegt. Bei

Ausschluss einer Variablen aus der Liste der identifizierten potenziellen Confoundervariablen ist es notwendig, eine literaturgestützte Begründung vorzulegen. Grundsätzlich sollte bei der Zusammenführung von potenziellen Confoundervariablen zu potenziellen Confoundern eine transparente und nachvollziehbare Dokumentation gewährleistet sein. Für alle identifizierten potenziellen Confounder ist a priori eine geeignete Operationalisierung festzulegen. Obschon potenzielle Confounder, die nicht messbar und / oder nicht operationalisierbar sind, nicht in die Auswertung einbezogen werden können, sollten ebendiese dennoch benannt und bei der Ergebnisinterpretation berücksichtigt werden.

Schritt 3: Relevanzbewertung der potenziellen Confounder durch klinische Expertinnen und Experten

Im 3. Schritt findet eine quantitative Bewertung der bereits zusammengeführten, potenziellen Confounder hinsichtlich ihrer Relevanz durch die klinischen Expertinnen und Experten statt. Eine Bewertung mittels Punktevergabe wird empfohlen. Das Ergebnis dieses Bewertungsschritts kann in der Datenanalyse für eine inhaltlich gestützte Variablenselektion in solchen Fällen verwendet werden, in denen das statistische Modell nicht konvergiert. Hierbei wird beispielsweise der potenzielle Confounder mit dem niedrigsten Score als erstes nicht mehr im Modell berücksichtigt.

Zusammenfassend wird das im Arbeitspapier GA23-02 erarbeitete Vorgehen zur Confounderidentifikation als durchführbar angesehen. Die vorliegenden Untersuchungen unterstreichen aber auch, wie aufwendig die Confounderidentifikation für nicht randomisierte vergleichende Studien ist. In einem nächsten Schritt soll daher (außerhalb der vorliegenden Ausarbeitung) geprüft werden, ob weitere Anpassungen zur systematischen Identifikation von Confoundern möglich sind, um die Machbarkeit (durch eine Aufwandsreduktion) zu erhöhen. Dabei sind mögliche Optionen zur Aufwandsreduktion dem damit einhergehenden Informationsverlust und potenziellen Verlust der Ergebnissicherheit gegenüberzustellen.

5.3 Schätzung notwendiger Fallzahlen vor dem Hintergrund unzureichender Vorabinformationen

Der G-BA beauftragt das IQWiG regelhaft mit einer wissenschaftlichen Ausarbeitung von Konzepten für AbDs und deren Auswertung. Die Konzepte sollen auch die Anforderungen an die Dauer und den Umfang der jeweiligen Datenerhebung beschreiben. Dabei geht es um die Beantwortung der Frage, ob die geplante AbD prinzipiell realisierbar und sinnvoll durchführbar ist. Auf Basis der vorliegenden Erkenntnisse zum zu bewertenden Arzneimittel und zum Komparator sowie der Anzahl der im jeweiligen Anwendungsgebiet zur Verfügung stehenden Patientinnen und Patienten wird abgeschätzt, ob innerhalb eines akzeptablen Zeitraums ausreichend Patientinnen und Patienten bzw. Ereignisse (d. h. die notwendige Fallzahl bzw. Ereigniszahl) eingeschlossen bzw. beobachtet werden können, um mit der AbD grundsätzlich aussagekräftige Ergebnisse für eine Nutzenbewertung generieren zu können [26,34].

Dabei ist anzumerken, dass zu diesem Zweck im Rahmen der Konzepterstellung keine Fallzahlplanung (wie für eine klinische Studie) durchgeführt werden kann, da die hierfür notwendigen Vorabinformationen nicht vorliegen. Die Fallzahlplanung für eine RCT basiert in der Regel auf Informationen aus Pilotstudien oder vergleichenden Phase-II-Studien, die eine relativ sichere Bestimmung der notwendigen Fallzahl erlauben. Ein Aufgreifkriterium für die Forderung einer AbD ist jedoch gerade, dass keine vergleichenden Daten im zu bewertenden Anwendungsgebiet vorliegen. Daher verfolgt die Konzepterstellung vielmehr das Ziel, eine Einschätzung zur grundsätzlichen Machbarkeit einer AbD zu geben. In Abhängigkeit der Informationsbasis zum Zeitpunkt der Konzepterstellung wird dabei auf eine orientierende Fallzahlschätzung oder auf eine orientierende Betrachtung von Fallzahlszenarien als Herangehensweise zurückgegriffen.

Die orientierende Fallzahlschätzung wird als eine der Herangehensweisen im nachfolgenden Abschnitt 5.3.1 näher erläutert (inklusive einer Darstellung der Grundlagen [Parameter] einer Fallzahlschätzung). Eine Beschreibung der orientierenden Betrachtung von Fallzahlszenarien erfolgt im Abschnitt 5.3.2. Die Einordnung der beiden Herangehensweisen wird abschließend in Abschnitt 5.3.3 diskutiert.

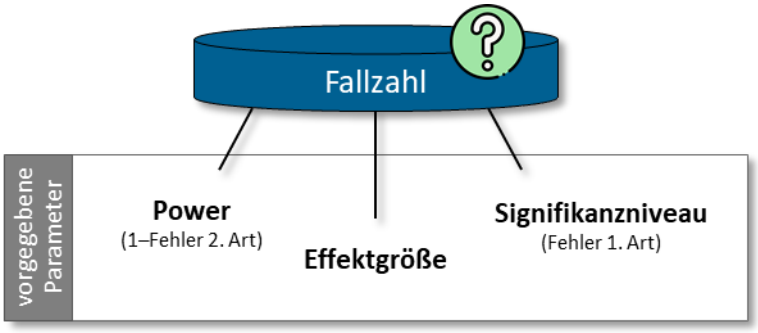
5.3.1 Grundlagen der Fallzahlschätzung

Das Ziel einer jeden Studie ist die Beantwortung einer wissenschaftlichen Fragestellung. Im Rahmen der evidenzbasierten Medizin orientiert sich die wissenschaftliche Fragestellung am sogenannten PICO-Schema (Population, Intervention, Comparison, Outcome). Die Studie soll somit Aussagen für den Vergleich, z. B. des Nutzens, einer Intervention mit einem Komparator für eine bestimmte Patientenpopulation ermöglichen. Im Rahmen einer Studie wird jedoch in der Regel nicht die gesamte interessierende Patientenpopulation (d. h. die Grundgesamtheit) beobachtet, sondern eine möglichst repräsentative Stichprobe eingeschlossen und die Ergebnisse anschließend auf die Grundgesamtheit übertragen.

Eine Fallzahlschätzung wird durchgeführt, um abzuschätzen, wie groß diese Stichprobe sein sollte, um mit einer Studie aussagekräftige Ergebnisse zu erzielen. Sie dient der Ermittlung der notwendigen Fallzahl.

Parameter für die Fallzahlschätzung

Um die notwendige Fallzahl zu ermitteln, müssen Annahmen für verschiedene Parameter getroffen werden. Diese werden nachfolgend für das Effektmaß relatives Risiko (RR) beschrieben. Für weitere Effektmaße (wie z. B. das Hazard Ratio oder die standardisierte Mittelwertdifferenz) gelten jedoch ähnliche Überlegungen. Zu den Parametern gehören die in einer AbD erwartete Effektgröße (die sowohl den wahren Therapieeffekt als auch die systematische Verzerrung aufgrund des nicht randomisierten Designs berücksichtigt), das Signifikanzniveau (Irrtumswahrscheinlichkeit, auch Fehler 1. Art) sowie die Power (1–Fehler 2. Art) [40] (siehe Abbildung 4).



Je nach Effektmaß gehen in die Schätzung der Fallzahl noch weitere Parameter ein, wie z. B. das Basisrisiko beim relativen Risiko.

Abbildung 4: Notwendige Fallzahl (N) bei Festlegung der Parameter erwartete Effektgröße, Signifikanzniveau (Fehler 1. Art) und Power (1−Fehler 2. Art)

Durch Festlegung der Power, des Signifikanzniveaus, des Risikos in der Kontrollgruppe, des in einer AbD erwarteten Effekts (RR_1) sowie des Verteilungsverhältnisses der Patientinnen und Patienten zwischen den Behandlungsgruppen (Intervention zu Komparator) ergibt sich die geschätzte notwendige Fallzahl, um unter Berücksichtigung der verschobenen Nullhypothese einen statistisch signifikanten Unterschied zwischen der Intervention und dem Komparator abzuleiten.

Ein Aufgreifkriterium für eine AbD ist eine unzureichende Evidenzlage für ein neues Arzneimittel im Vergleich zum therapeutischen Standard. Mit der AbD werden demnach erstmals vergleichende Daten gewonnen. Vor diesem Hintergrund kann es für die Situation einer AbD sinnvoll sein, eine vergleichende Studie nicht zwingend mit dem Ziel durchzuführen, einen Vorteil des neuen Arzneimittels zu zeigen und die Fallzahlschätzung auf dieses Ziel auszurichten. Stattdessen ist denkbar, die Schätzung der notwendigen Fallzahl darauf zu beziehen, dass aus einem gewählten *Propensity-Score-Verfahren* unter Sicherstellung einer adäquaten Confounderadjustierung verwertbare Ergebnisse im Sinne einer Effektschätzung mit zugehörigem 95 %-Konfidenzintervall resultieren (siehe hierzu auch Abschnitt 5.7), und zwar losgelöst davon, ob sich unter Berücksichtigung der verschobenen Nullhypothese ein Vorteil des zu bewertenden Arzneimittels ableiten lässt. Unter dieser Voraussetzung stellt jedes Ergebnis einer AbD einen Erkenntnisgewinn dar.

Power und Signifikanzniveau

Die Annahmen für die Power (mindestens 80 %) und das Signifikanzniveau (2,5 % [1-seitige Fragestellung]) beruhen in der Regel auf üblichen Festlegungen.

Erwartete Effektgröße

Um die Frage, in welcher Größenordnung der in einer AbD erwartete Effekt (RR_1) oder das Risiko in der Kontrollgruppe jeweils liegen könnte, hinreichend sicher beantworten zu können, werden Vorabinformationen aus vorhandener Evidenz zu beobachteten Ereignisraten sowohl

unter Therapie mit der Intervention als auch unter Therapie mit den Komparatoren für einen (patienten-)relevanten Endpunkt mit geeigneter Operationalisierung in der interessierenden Patientenpopulation benötigt.

In der Regel liegen für das Arzneimittel, für das ein AbD-Konzept erstellt wird, lediglich (zum Teil vorläufige) Daten aus nicht vergleichenden (1-armigen) Zulassungsstudien vor [41-43]. Für die Komparatoren werden Studiendaten aus verschiedenen Quellen (z. B. aus Publikationen oder aus Herstellerdossiers) herangezogen. Die Informationen für die Intervention und die Komparatoren werden folglich zum einen aus verschiedenen Studien herangezogen, zum anderen sind die verfügbaren Informationen üblicherweise wenig detailliert, im Besonderen hinsichtlich der Beschreibung der interessierenden Population und / oder der Intervention. Insgesamt wird deutlich, dass für die Fallzahlschätzung im Rahmen einer AbD lediglich sehr unsichere Vorabinformationen vorliegen. Aus diesem Grund kann die Fallzahlschätzung im Rahmen eines AbD-Konzepts allenfalls einen orientierenden Charakter haben.

Verteilungsverhältnis zwischen den Behandlungsgruppen

Ein weiterer Parameter, der für die Schätzung der benötigten Fallzahl berücksichtigt werden muss, ist das Verteilungsverhältnis der Patientinnen und Patienten zwischen den zu vergleichenden Behandlungsgruppen, d. h. die Festlegung einer Annahme, wie sich die Patientinnen und Patienten wahrscheinlich auf die Behandlungsgruppen verteilen werden.

In einer RCT ist das Verteilungsverhältnis durch das im Studienprotokoll festgelegte Randomisierungsverhältnis (z. B. Verhältnis Intervention zu Komparator von 1:1 oder 2:1) vorgegeben. Im Rahmen der AbD, in der es aufgrund der fehlenden Randomisierung keine aktive Zuteilung zu einer Behandlungsgruppe gibt, muss abgeschätzt werden, wie sich die Versorgungsanteile in der Zukunft (nach Marktzugang des zu bewertenden Arzneimittels) entwickeln werden. Dies hängt in der Praxis von vielen verschiedenen Faktoren ab. Ist das neue Arzneimittel beispielsweise mit dem Versprechen der Heilung von einer andernfalls nicht heilbaren Erkrankung verbunden oder ist die Verfügbarkeit von wirksamen Therapiealternativen begrenzt, ist die Präferenz der Patientinnen und Patienten für dieses Arzneimittel erwartbar sehr hoch. Im Extremfall kann dies sogar dazu führen, dass sich kaum Patientinnen und Patienten für die Kontrollgruppe der AbD finden lassen. Demgegenüber kann es Situationen geben, in denen die Erkrankung der Patientinnen und Patienten mit den bisher verfügbaren Therapieoptionen gut eingestellt ist, sodass sich das neue Arzneimittel, insbesondere dann, wenn es sich um ein gänzlich neues Therapieprinzip handelt, nur langsam und verzögert durchsetzt.

Wie sich die Versorgungsanteile für ein konkretes Arzneimittel entwickeln werden, ist zum Zeitpunkt der Konzepterstellung für eine AbD, die meist vor der Zulassung erfolgt, nur schwer abschätzbar. Aus diesen Schilderungen wird deutlich, dass die Annahmen für ein realistisches

Verteilungsverhältnis in einer AbD nur sehr unsicher zu treffen sind. Im Rahmen der Konzepterstellung werden deshalb in der Regel verschiedene Verteilungsverhältnisse angenommen, um deren Auswirkungen auf die benötigte Fallzahl zu veranschaulichen. Dabei ist zu beachten, dass sich die benötigte Fallzahl bei Verteilungsverhältnissen jenseits von 5:1 bzw. 1:5 nur noch geringfügig erhöht [44-46], sodass sich die in den AbD-Konzepten dargestellten Verteilungsverhältnisse in der Regel zwischen 5:1 und 1:5 bewegen.

Verschobene Nullhypothese

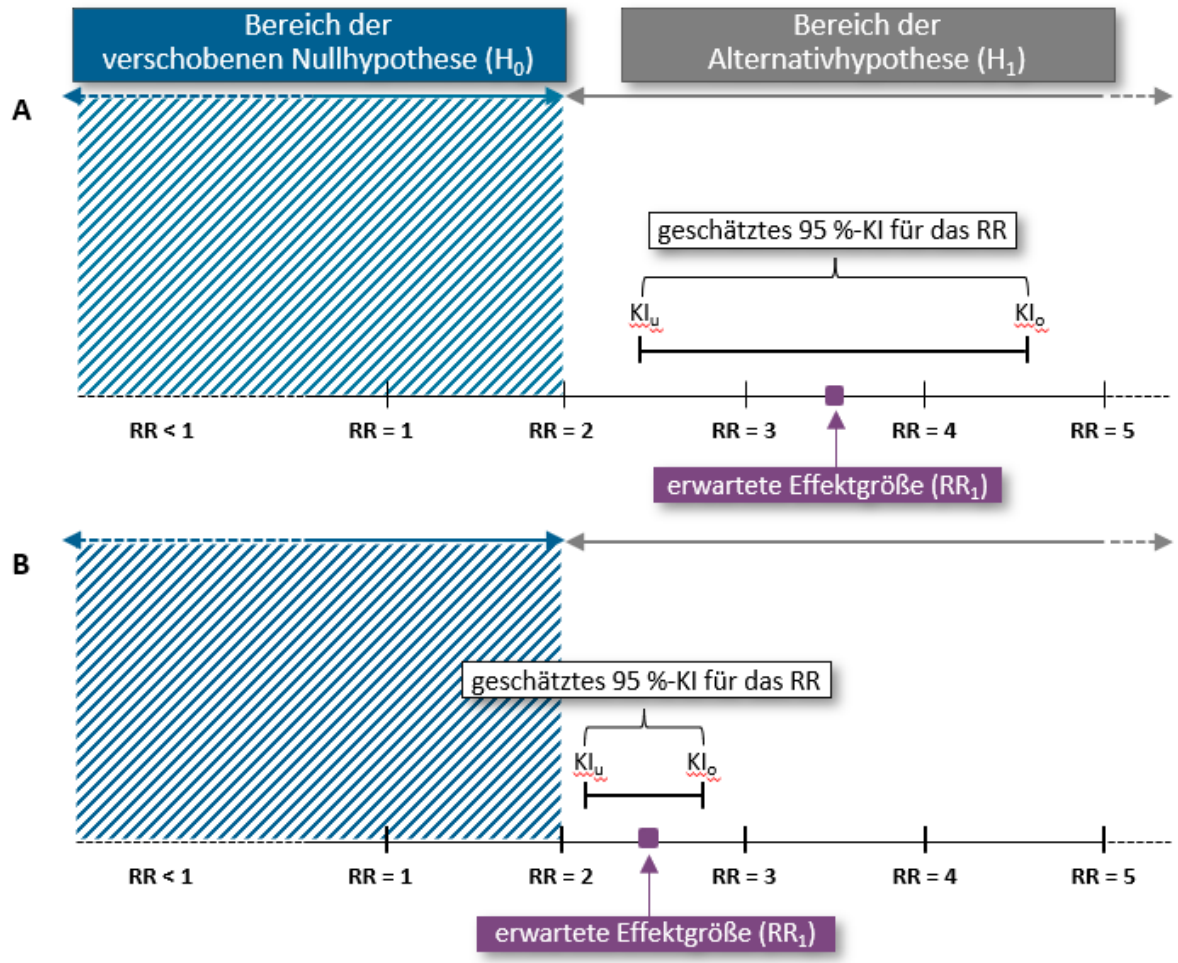
Ein weiterer wesentlicher Aspekt der Fallzahlschätzung in einer AbD betrifft die Ergebnisunsicherheit aufgrund der fehlenden Randomisierung. Aus einer nicht randomisierten Studie kann z. B. aufgrund potenziell unbekannter Confounder aus den in der Studie beobachteten Effekten erst ab einer bestimmten Effektstärke eine Aussage zum Nutzen oder Schaden einer Intervention abgeleitet werden. Eine (positive oder negative) Aussage zum Nutzen oder zum Schaden ergibt sich dann, wenn das Konfidenzintervall für den beobachteten Effekt oberhalb bzw. unterhalb einer zu definierenden Schwelle liegt (Test auf verschobene Nullhypothese H_0). Da bei der Generierung von versorgungsnahen Daten (Studien ohne Randomisierung) die Erfüllung umfangreicher Qualitätsanforderungen Voraussetzung für die Betrachtung der Effekte ist, sollte dieser Schwellenwert deutlich unterhalb des Wertes für den „dramatischen Effekt“ (RR von 5 bis 10 [47]) liegen, z. B. in einem Bereich von 2 bis 5 für das RR bezogen auf die untere bzw. die obere Konfidenzintervallgrenze (bzw. 0,2 bis 0,5 z. B. bei mortalitätssenkenden Interventionen) [34].

Den Ausgangspunkt bildet die Planung einer Studie zur Testung der verschobenen Hypothesen

$$H_0: RR \leq RR_0 \text{ vs. } H_1: RR > RR_0$$

anhand des RR mit dem Schwellenwert $RR_0 = 2$ (bzw. $H_0: RR \geq RR_0$ vs. $H_1: RR < RR_0$ mit dem Schwellenwert $RR_0 = 0{,}5$, z. B. bei mortalitätssenkenden Interventionen). Mit einer Fallzahlschätzung wird sodann ermittelt, wie viele Patientinnen und Patienten bzw. Ereignisse (d. h. notwendige Fallzahl) mindestens eingeschlossen und beobachtet werden müssen, um unter Verwendung eines statistischen Tests mit verschobenen Hypothesengrenzen (oben genannte Schwellenwerte) und unter Vorgabe von Signifikanzniveau (Fehler 1. Art) und Power (1−Fehler 2. Art) eine in einer AbD erwartete Effektgröße (RR_1) aufzudecken [48].

Wie in Abbildung 5 schematisch dargestellt, können sich u. a. folgende Szenarien ergeben:



H_0: Nullhypothese; H_1: Alternativhypothese; KI: Konfidenzintervall; KI_u: untere Grenze des Konfidenzintervalls; KI_o: obere Grenze des Konfidenzintervalls; RR: relatives Risiko; RR_1: erwartetes relatives Risiko

Abbildung 5: Szenarien bei der orientierenden Fallzahlschätzung mit verschobenem Nullhypothesentest ($RR_0 = 2$)

- Wenn die in einer AbD erwartete Effektgröße (RR_1) den Schwellenwert von 2 deutlich überschreitet, ist eine eher kleine Fallzahl bzw. wenige Ereignisse (damit einhergehend geringe Präzision [breites Konfidenzintervall]) ausreichend, damit das 95 %-Konfidenzintervall für das RR vollständig oberhalb der verschobenen Hypothesengrenze liegen kann (Szenario a).
- Wenn die in einer AbD erwartete Effektgröße (RR_1) den Schwellenwert von 2 nur knapp überschreitet, ist eine größere Fallzahl bzw. viele Ereignisse (damit einhergehend hohe Präzision [schmales Konfidenzintervall]) erforderlich, damit das 95 %-Konfidenzintervall für das RR vollständig oberhalb der verschobenen Hypothesengrenze liegen kann (Szenario b).

5.3.2 Keine oder unzureichende Vorabinformationen für orientierende Schätzungen zum Umfang der anwendungsbegleitenden Datenerhebung

Für Arzneimittel, für die das IQWiG mit der wissenschaftlichen Ausarbeitung eines Konzeptes für eine AbD und von Auswertungen beauftragt wird, liegen die benötigten Informationen zur Schätzung der Effektgröße und des Basisrisikos zum Zeitpunkt der Konzepterstellung in der Regel nicht vor, sodass die Voraussetzungen für eine orientierende Fallzahlschätzung nicht erfüllt sind (siehe folgender Kasten).

Beispiel für ein AbD-Konzept, in dem eine orientierende Fallzahlschätzung aufgrund unzureichender Vorabinformationen nicht sinnvoll möglich war

Der G-BA hat das IQWiG am 16.01.2025 mit der Erstellung eines Konzeptes für eine AbD zu Epcoritamab zur Behandlung von Erwachsenen mit einem rezidivierenden oder refraktären diffus großzelligen B-Zell-Lymphom (diffuse large B-cell lymphoma [DLBCL]) nach 2 oder mehr Linien einer systemischen Therapie, für die eine chimäre Antigenrezeptor(CAR)-T-Zelltherapie und Stammzelltransplantation nicht infrage kommen, beauftragt (AbD-Konzept A25-07 [49]).

Da das primäre Therapieziel im Anwendungsgebiet eine Verlängerung des Gesamtüberlebens ist, war bei der Konzepterstellung geplant, eine orientierende Fallzahlschätzung auf Basis des Endpunkts Gesamtüberleben für den Vergleich von Epcoritamab gegenüber den festgelegten Komparatoren des G-BA vorzunehmen.

Für Epcoritamab lagen zum Zeitpunkt der Konzepterstellung im Dossier zur Nutzenbewertung zu Epcoritamab für den Endpunkt Gesamtüberleben Ergebnisse der beiden 1-armigen Studien EPCORE NHL-1 (GCT3013-01) und EPCORE NHL-3 vor (GCT3013-04) [50]. Zu den für die Konzepterstellung durch den G-BA übermittelten Komparatoren standen publizierte finale Ergebnisse zu Tafasitamab (Kombination mit Lenalidomid) [51] und ein Dossier zur Nutzenbewertung von Polatuzumab Vedotin [52] zur Verfügung. Für die 3 Komparatoren, die gemäß Auftrag des G-BA ergänzend berücksichtigt werden sollten, lagen Daten für den Endpunkt Gesamtüberleben im Dossier zur Nutzenbewertung (Glofitamab [53], Loncastuximab tesirin [54]) oder im European Public Assessment Report (Odronextamab [55]) vor.

Auf Basis der vorliegenden Daten (sowohl zu Epcoritimab als auch zu den Komparatoren) war unklar, wie hoch der Anteil der eingeschlossenen Patientinnen und Patienten jeweils war, für die eine CAR-T-Zelltherapie und Stammzelltransplantation nicht infrage gekommen sind, und ob die Anteile unter Behandlung mit Epcoritamab und den Komparatoren vergleichbar waren. Aufgrund dessen war nicht abschätzbar, ob sich die jeweiligen Teilpopulationen der potenziell von der Fragestellung umfassten Patientinnen und Patienten hinsichtlich der Anzahl vorangegangener Therapielinien und somit der Schwere und dem Fortschreiten der Erkrankung sowie weiterer prognostischer Faktoren unterschieden.

In der Gesamtschau ließen die Studienergebnisse keine hinreichende Einschätzung der Effekte zu, die für eine orientierende Fallzahlschätzung für den Vergleich von Epcoritamab gegenüber den Komparatoren notwendig gewesen wäre. Folglich wurde im Rahmen des AbD-Konzeptes A25-07 auf eine orientierende Fallzahlschätzung verzichtet.

In der Gesamtschau der bisherigen in AbD-Verfahren erstellten Konzepte und der zugehörigen Addenda zeigen sich insbesondere folgende Konstellationen:

- Das zu prüfende, neue Arzneimittel ist zum Zeitpunkt der Konzepterstellung häufig noch nicht zugelassen. Da Publikationen zu den Hauptergebnissen der (pivotalen) Studie in diesen Fällen in der Regel noch nicht veröffentlicht sind, liegen nur vorläufige, teilweise widersprüchliche, Ergebnisse zu lediglich selektiv in Kongressabstracts oder Pressemitteilungen dargestellten Endpunkten vor.
- Die Informationen zu vorhandenen Studien sind nicht ausreichend detailliert, um beurteilen zu können, ob die Behandlung der Patientinnen und Patienten mit den Komparatoren entsprechend der Vorgaben des G-BA zur zweckmäßigen Vergleichstherapie erfolgte (z. B. mit Hinblick auf Vortherapien).
- Es müssen für die Intervention und die Komparatoren im Regelfall Ergebnisse einzelner Arme verschiedener Studien herangezogen werden. Eine Überprüfung der Ähnlichkeit der Studien (z. B. mit Hinblick auf die Krankheitsschwere) ist aufgrund der mangelnden, öffentlichen Verfügbarkeit von für eine Prüfung relevanten spezifischen Daten nicht möglich.
- Unterschiede zwischen den für die Intervention und die Komparatoren verfügbaren Studien in der Operationalisierung eines patientenrelevanten Endpunkts, auf Basis dessen eine orientierende Fallzahlschätzung angestrebt wird, führen zu methodischen Unsicherheiten (betrifft z. B. die Operationalisierung von Blutungsereignissen in der Indikation Hämophilie).
- Der G-BA kann bei Festlegung der zweckmäßigen Vergleichstherapie (der Komparatoren) eine Unterteilung der Patientenpopulation (z. B. nach Eignung für eine Therapieform) vornehmen. Ergebnisse differenziert nach Fragestellung lassen sich für die Intervention und die Komparatoren aus öffentlich zugänglichen Dokumenten in der Regel nicht ableiten.

Auf Basis der beschriebenen Datensituationen zeigt sich, dass im Rahmen der Erarbeitung von AbD-Konzepten in vielen Fällen inhaltlich-methodische Probleme des Ansatzes einer orientierenden Fallzahlschätzung vorliegen. Im Regelfall fehlen notwendige Informationen für die Annahmen, die einer orientierenden Fallzahlschätzung zugrunde liegen. Daher wurde in AbD-

Verfahren zuletzt vermehrt auf eine orientierende Fallzahlschätzung verzichtet, da diese nicht sinnvoll möglich war.

Orientierende Betrachtung von Fallzahlszenarien

Um in jenen Fällen, in denen eine orientierende Fallzahlschätzung nicht sinnvoll möglich ist, dennoch Aussagen zum Umfang der Datenerhebung treffen zu können, erfolgt eine orientierende Betrachtung von Fallzahlszenarien. Diese folgt im Wesentlichen dem Prinzip der orientierenden Fallzahlschätzung. Als Ausgangspunkt für die Einschätzung zur grundsätzlichen Machbarkeit einer AbD dient hier jedoch die für eine AbD potenziell zur Verfügung stehende Patientenzahl im zu bewertenden Anwendungsgebiet. Dies ist darin begründet, dass in Fällen, in denen keine ausreichende Abschätzung des erwarteten Effekts möglich ist, für die Anzahl der Patientinnen und Patienten in der Regel die verlässlichste Schätzung vorliegt.

Für die orientierende Betrachtung von Fallzahlszenarien wird zunächst die geschätzte Anzahl der Patientinnen und Patienten im zu bewertenden Anwendungsgebiet in Deutschland ermittelt. Die Grundlage hierfür bilden – sofern vorhanden – primär frühere G-BA-Beschlüsse zur frühen Nutzenbewertung nach § 35a SGB V im interessierenden Anwendungsgebiet. Alternativ werden Informationen zu Patientenzahlen aus verfügbaren IQWiG-Berichten zur Einschätzung der Patientenzahlen herangezogen, mit denen das Institut vom G-BA Im Rahmen der Erforderlichkeitsprüfung für eine AbD beauftragt werden kann [26]. Liegen weder G-BA-Beschlüsse noch Einschätzungen der Patientenzahlen vor, wird die Anzahl der Patientinnen und Patienten im Rahmen der Konzepterstellung selbst ermittelt. Anders als bei der orientierenden Fallzahlschätzung wird die im Anwendungsgebiet zur Verfügung stehende Anzahl von Patientinnen und Patienten als Grundlage genommen, um die aufdeckbaren Effektgrößen zu ermitteln. Das Vorgehen wird im Folgenden näher erläutert.

Die Informationen zu den Patientenzahlen liegen aufgrund der Unsicherheit der Abschätzung üblicherweise in Form einer Spanne vor. Diese Spanne wird herangezogen, um ein Spektrum unterschiedlicher Patientenzahlen festzulegen, die in die weiteren Berechnungen eingehen. Wie viele Patientinnen und Patienten im zu bewertenden Anwendungsgebiet tatsächlich für eine AbD zur Verfügung stehen, ist dabei in der Regel nicht abschätzbar, da dies von unterschiedlichen Faktoren beeinflusst wird (u. a. von der Teilnahmebereitschaft der Zentren und der Patientinnen und Patienten).

In einem 2. Schritt wird mittels Festlegung der Patientenzahlen, des Signifikanzniveaus (2,5 % [1-seitige Fragestellung]) sowie der Power (mindestens 80 %) sodann die Frage beantwortet, welche erwarteten Effektgrößen (RR_1) auf Basis der im zu bewertenden Anwendungsgebiet grundsätzlich für eine AbD in Deutschland zur Verfügung stehenden Patientenzahl (Präzision der Schätzung) und unter Berücksichtigung der Gefahr systematischer Verzerrung aufgrund

des nicht randomisierten Vergleichs (verschobene Nullhypothese) aufgedeckt werden können (siehe Abbildung 6).

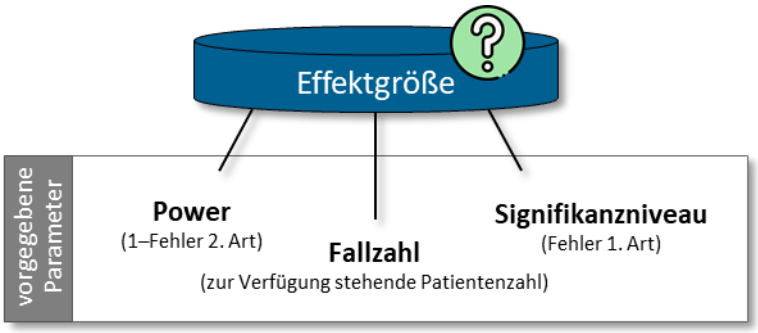
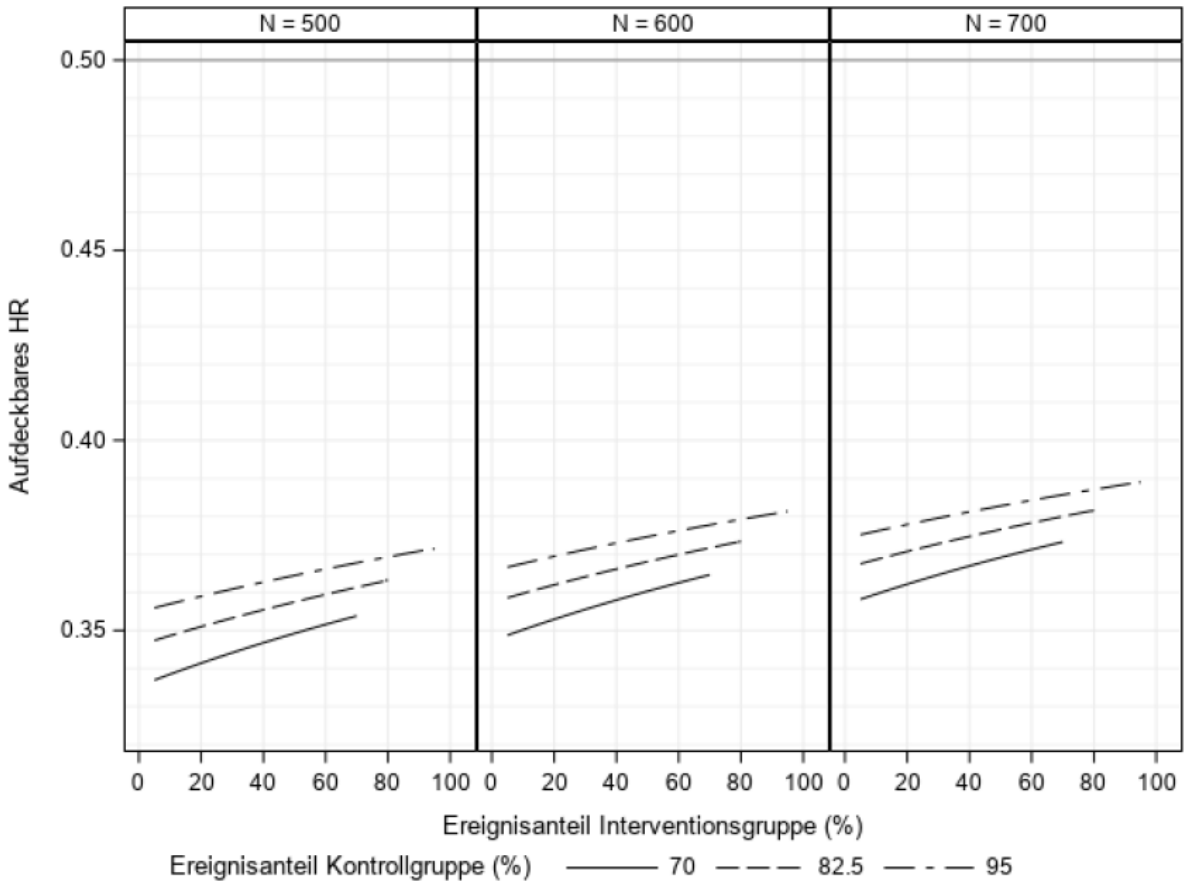


Abbildung 6: Aufdeckbare Effektgröße bei Festlegung der Parameter Fallzahl (N), Signifikanzniveau (Fehler 1. Art) und Power (1−Fehler 2. Art)

Aufdeckbar bedeutet, dass das im Rahmen einer AbD geschätzte 95 %-Konfidenzintervall mit 80 % Power bei gegebener Patientenzahl vollständig oberhalb (RR_0 = 2) bzw. unterhalb (RR_0 = 0,5) der verschobenen Hypothesengrenze liegt. Neben der Patientenzahl wird (wenn Angaben verfügbar) entweder das Basisrisiko unter Behandlung mit der Intervention oder das unter Behandlung mit den Komparatoren für einen patientenrelevanten Endpunkt als Anker herangezogen, um ein Spektrum der im Anwendungsgebiet möglichen Szenarien aufzuzeigen. Ob dabei auf das Basisrisiko der Intervention oder das der Komparatoren zurückgegriffen wird, hängt davon ab, für welche der Therapien Informationen für die interessierende Patientenpopulation vorliegen (siehe folgender Kasten).

Beispiel für ein AbD-Konzept, in dem eine orientierende Betrachtung verschiedener Fallzahlszenarien erfolgte

Im Rahmen der Erstellung des oben bereits erwähnten AbD-Konzeptes A25-07 zu Epcoritamab [49] war eine orientierende Fallzahlschätzung nicht sinnvoll möglich, da die vorhandenen Daten zu Epcoritamab und den Komparatoren hierfür keine hinreichenden Informationen lieferten. Zur Annäherung an die Frage, ob eine AbD zu Epcoritamab grundsätzlich sinnvoll durchführbar ist, erfolgte im AbD-Konzept eine orientierende Betrachtung verschiedener Fallzahlszenarien. Es wurden die Effekte eines mittels Ereigniszeitanalysen ausgewerteten Endpunktes (Effektmaß: Hazard Ratio) – im vorliegenden Fall des Gesamtüberlebens – dargestellt, die bei den im Anwendungsgebiet geschätzten Patientenzahlen (N = 500, N = 600, N = 700) und einem Verteilungsverhältnis von 1:1, 1:3 oder 1:5 (jeweils Intervention zu Komparatoren) mit einer Power von 80 % aufdeckbar sind (siehe Abbildung 7).



36-monatige Beobachtungsdauer; 1-seitiger Test mit verschobener Nullhypothese (H_0: HR ≥ 0,5, gekennzeichnet mit der horizontalen Linie bei HR = 0,5) und Signifikanzniveau 2,5 % (HR: Hazard Ratio; N: Fallzahl)

Abbildung 7: Aufdeckbarer Effekt (Effektmaß: Hazard Ratio) in Abhängigkeit von der Stichprobengröße und den Ereignisanteilen in beiden Behandlungsgruppen (Verhältnis von Intervention zu Komparatoren von 1:3)

Zur Ermittlung des aufdeckbaren Effekts wurde die für die Kontrollgruppe vorliegende Evidenz im Endpunkt Gesamtüberleben als Anker herangezogen. Basierend auf den gesichteten Daten wurden für die Kontrollgruppe Anteile an verstorbenen Patientinnen und Patienten von 70 %, 82,5 % und 95 % zu Monat 36 angenommen (für detailliertere Beschreibung der vorliegenden Evidenz für die Komparatoren, siehe AbD-Konzept A25-07 [49]). Für die Interventionsgruppe wurden die sich ergebenden Ereignisanteile von 5 % bis 70 %, 5 % bis 80 % sowie 5 % bis 90 % dargestellt.

Abbildung 7 zeigt den aufdeckbaren Effekt mit einer Spanne von ca. HR_1 = 0,35 bis HR_1 = 0,40 (mit einer Power von 80 %) zum Vorteil von Epcoritamab gegenüber den Komparatoren unter Vorgabe der Stichprobengrößen und Ereignisanteile in beiden Behandlungsgruppen.

5.3.3 Entscheidungsfindung für oder gegen eine orientierende Fallzahlschätzung

Zusammenfassend wird aus den obigen Schilderungen deutlich, dass die Herangehensweisen der orientierenden Fallzahlschätzung und der orientierenden Betrachtung von Fallzahlszenarien im Wesentlichen den gleichen Grundsätzen folgen und sich lediglich in ihrem Ausgangspunkt unterscheiden. Die 2 Herangehensweisen spiegeln folglich unterschiedliche Spielarten der gleichen Methodik wider, von denen eine – abhängig von den zur Verfügung stehenden Vorabinformationen – angewendet wird. Dies wird durch die folgende Abbildung 8 illustriert.

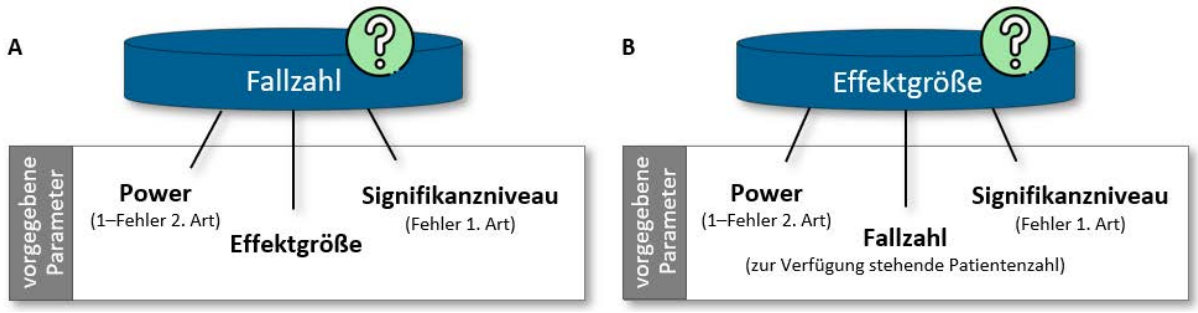


Abbildung 8: Herangehensweisen zur Bewertung der Machbarkeit einer AbD:
(A) orientierende Fallzahlschätzung und (B) orientierende Betrachtung von Fallzahlszenarien

In Abbildung 8 sind die Inhalte von Abbildung 4 und Abbildung 6 einander gegenübergestellt. Durch diese Gegenüberstellung wird deutlich, dass die betrachteten Parameter identisch und lediglich die eingehenden und resultierenden Parameter (Effektgröße und Fallzahl) vertauscht sind.

Im Rahmen der Ausarbeitung eines AbD-Konzepts wird eine orientierende Fallzahlschätzung immer dann durchgeführt, wenn sie sinnvoll möglich ist. Dafür müssen die folgenden Voraussetzungen erfüllt sein:

- Die Patientenpopulationen, auf der die Evidenz für die Intervention und den Komparator beruht, sind hinreichend ähnlich und entsprechen jeweils in ausreichendem Maß der interessierenden Population gemäß Fragestellung.
- Die Intervention und der Komparator werden in den Studien jeweils hinreichend entsprechend der Fragestellung angewendet.
- Die Operationalisierungen des betrachteten Endpunkts sind zwischen der Evidenz für die Intervention und derjenigen für den Komparator hinreichend vergleichbar.
- Die vorliegenden Informationen sind ausreichend, um die vorgenannten Voraussetzungen zu beurteilen.

Bei der Entscheidung für oder gegen die orientierende Fallzahlschätzung im Rahmen von AbD-Verfahren werden durch das IQWiG keine festen Grenzen angelegt. Ob eine orientierende

Fallzahlschätzung bei einer gegebenen Datenkonstellation als sinnvoll durchführbar erachtet wird oder eine orientierende Betrachtung von Fallzahlszenarien erfolgt, ist jeweils im Einzelfall zu prüfen. Grundsätzlich ist jedoch wichtig anzumerken, dass selbst eine orientierende Fallzahlschätzung wegen der beschriebenen Unsicherheiten allenfalls eine grobe Einschätzung der Größenordnung der benötigten Fallzahl liefern kann.

Gewährleistung einer ausreichenden Rekrutierung der Studienpopulation

Es liegt im Verantwortungsbereich des pU, eine ausreichend große Anzahl von Patientinnen und Patienten zu rekrutieren, um im Rahmen einer AbD mit ausreichender Sicherheit einen bestehenden Unterschied zwischen der Intervention und den Komparatoren nachzuweisen.

Insbesondere für seltene Erkrankungen kann es sinnvoll und notwendig sein, versorgungsnahe Daten in internationaler Kooperation zu erheben, um innerhalb eines akzeptablen Zeitraums zur Generierung aussagekräftiger Daten zum Zwecke der Nutzenbewertung nach § 35a SGB V prinzipiell ausreichend Patientinnen und Patienten rekrutieren und beobachten zu können [18]. Neben der Datenerhebung in einem bestehenden (grundsätzlich geeigneten) europäisch oder international angelegten (Indikations-)Register, kann eine Erhebung in einem nationalen Register auch durch eine Einbindung weiterer (internationaler) Register unterstützt werden, jeweils ggf. ergänzt mittels studienindividueller Erhebung der nicht im Register verfügbaren Daten.

Dabei müssen die im jeweiligen Register erhobenen Daten zum Zwecke der Nutzenbewertung von Arzneimitteln geeignet sein. Dies setzt voraus, dass

- die Daten von Umfang und Qualität den Anforderungen an die AbD entsprechen und die Auswertung der Daten gemäß den Anforderungen an die AbD durchgeführt werden und
- die Versorgung in den Ländern, in denen die Daten erhoben werden, der Versorgung in Deutschland ausreichend ähnlich ist bzw. die aus dem jeweiligen Register gewonnenen Erkenntnisse auf die Situation in Deutschland (deutschen Versorgungskontext) übertragbar sind.

Finale Fallzahlplanung im Verlauf der anwendungsbegleitenden Datenerhebung

Unabhängig vom gewählten methodischen Vorgehen für die Konzepterstellung (orientierende Fallzahlschätzung oder orientierende Betrachtung von Fallzahlszenarien) sollte der pU des zu bewertenden Arzneimittels die (finale) Fallzahlplanung der AbD zum Zeitpunkt der Zwischenanalysen auf Basis der dann vorhandenen vorläufigen Ergebnisse vornehmen [56]. Dieses Vorgehen wird als grundsätzlich geeignet beurteilt, da die Schätzung der notwendigen Fallzahl bei versorgungsnahen Datenerhebungen – in Abgrenzung zu klinischen Studien – nicht mit Hinblick auf eine aktive Rekrutierung erfolgt. Stattdessen werden möglichst alle Patientinnen und Patienten einer Registerpopulation, die die Einschlusskriterien für die AbD erfüllen, nach

ihrer Einwilligung in die AbD eingeschlossen. Dies erfolgt auch vor dem Hintergrund, dass, je nach gewähltem *Propensity-Score-Verfahren*, ggf. ein relevanter Anteil der Patientinnen und Patienten nicht in der Auswertung berücksichtigt werden kann. Wie viele Patientinnen und Patienten nicht in die Analysepopulation eingehen, ist dabei im Vorfeld nicht abschätzbar.

5.4 Beobachtungsstart und Behandlungswechsel

5.4.1 Beobachtungsstart

Wie bereits in Abschnitt 5.1 erläutert, ist die richtige Wahl des Startzeitpunkts der Beobachtung (Index-Datum [t_0]) eine der zentralen Herausforderungen bei der *Target-Trial-Emulation*. Wird t_0 für einen Vergleich falsch gewählt, kann dies zu fehlerhaften Ergebnissen führen und darauf beruhende Schlussfolgerungen verfälschen [57]. Grund dafür sind sogenannte zeitbezogene Verzerrungen (Time-Related Bias). Darunter fallen *Immortal-Time-Bias*, *Time-Lag-Bias* sowie *Latency-Time-Bias*. Risiken für zeitbezogene Verzerrungen entstehen, wenn sich die mit t_0 verknüpften Elemente des Studiendesigns zwischen den Behandlungsgruppen unterscheiden [57]. Demnach ist entscheidend, dass der Beginn der Beobachtung (t_0), die Prüfung der Erfüllung der Einschlusskriterien und die Behandlungszuteilung übereinstimmen [9]. Um sich dem zu nähern, liegen verschiedene Ansätze vor, die im Folgenden beschrieben werden.

Active-Comparator-New-User-Design

Das Studiendesign, das als aktueller Standard in der pharmakologischen Forschung angesehen wird, ist das Active-Comparator-New-User(ACNU)-Design [1]. Beim ACNU-Design werden ausschließlich Patientinnen und Patienten in eine Studie eingeschlossen, die eine Indikation für eine Umstellung ihrer Therapie haben (wie z. B. Beginn einer neuen Therapielinie). Dieser Ansatz geht mit einer Reduktion der Verzerrung durch Confounding und *Selection Bias* einher. Ein weiterer Vorteil des Designs wird darin gesehen, dass die Zeitpunkte Prüfung der Erfüllung der Einschlusskriterien, Behandlungszuteilung und Beobachtungsstart in Einklang gebracht werden können und der Tag der Therapieentscheidung für beide Behandlungsgruppen als t_0 festgesetzt werden kann. In Fällen, in denen die Therapieentscheidung nicht dokumentiert ist, sollte eine bestmögliche Näherung daran erfolgen. Was eine bestmögliche Näherung ist, hängt von der konkreten Therapiesituation ab. Wird üblicherweise zeitnah nach der Therapieentscheidung mit der Behandlung begonnen, kann das Datum des Behandlungsbeginns als adäquate Näherung gelten. Sind vor Start der Behandlung zwingende vorbereitende Schritte notwendig (wie z. B. die Zubereitung einer CAR-T-Zelltherapie), sollte ein Zeitpunkt gewählt werden, der möglichst nah am Zeitpunkt der Therapieentscheidung liegt. Im Studienprotokoll sollte festgehalten sein, wie damit umgegangen wird, wenn t_0 vom Zeitpunkt der Therapieentscheidung abweicht. Potenzielle Abweichungen sollten in den Ergebnissen transparent dargestellt werden.

Prevalent-New-User-Design

Ein Design, das für Situationen geeignet ist, in denen die Patientinnen und Patienten entweder auf einem etablierten therapeutischen Standard verbleiben (beispielsweise bei Hämophilie A und B) oder auf ein neues Arzneimittel wechseln können, stellt das Prevalent-New-User(PNU)-Design dar. Hierbei werden neben therapienaiven Patientinnen und Patienten auch bereits mit dem Komparator behandelte Patientinnen und Patienten berücksichtigt. In den Vergleich zweier Behandlungsgruppen können somit 3 verschiedene Patientenpopulationen eingehen:

1) Patientinnen und Patienten, die eine Therapie mit der Intervention beginnen und die zuvor (hinsichtlich der Intervention und des Komparators) therapienaiv waren (New User)
2) Patientinnen und Patienten mit einem Wechsel vom Komparator auf die Intervention (Prevalent New User) sowie
3) Patientinnen und Patienten, die ihre Therapie mit dem Komparator fortführen.

Da davon ausgegangen wird, dass Vortherapien mit dem jeweiligen Endpunkt von Interesse assoziiert sind, kann das PNU-Design zu zeitbezogener Verzerrung führen [58]. Somit müssen im Rahmen des PNU-Designs im Wesentlichen sowohl die Wahl von t_0 als auch die Exposition gegenüber Vortherapie(n) (beispielsweise in Form von Behandlungsdauer und / oder Anzahl von Verschreibungen) als *Confounder* betrachtet werden [59].

Verschiedene Möglichkeiten für die Wahl des Startzeitpunkts der Beobachtung (Index-Datum [t_0])

Als naives Vorgehen kann für beide Behandlungsgruppen der Zeitpunkt des Einschlusses in die Studie als t_0 festgelegt werden. Da dieses Vorgehen jedoch zu einer verzerrten Schätzung des Behandlungseffektes durch Immortal-Time-Bias [57] führt, ist dieser Ansatz regelhaft nicht zu empfehlen.

Eine weitere Möglichkeit ist eine Festlegung des Tages der Therapieentscheidung als t_0 für die Interventionsgruppe, wohingegen für die Kontrollgruppe der Zeitpunkt des Einschlusses in die Studie verwendet wird [57,60]. Dieses Vorgehen kann zu Time-Lag-Bias und Latency-Time-Bias führen, insbesondere dann, wenn die Inzidenz für das interessierende Ereignis über die Zeit sinkt und daher frühe Ereignisse in der Kontrollgruppe ein stärkeres Gewicht im Vergleich erhalten [57].

Als eine mögliche Lösung bietet sich die Durchführung von Sensitivitätsanalysen an, in denen z. B. Patientinnen und Patienten mit Ereignissen zu einem frühen Zeitpunkt im Studienverlauf aus der Analyse ausgeschlossen werden [57]. Eine hierzu ähnliche Empfehlung findet sich im Addendum A23-99 zum AbD-Konzept A22-83 zu Etranacogen Dezaparvovec in der Indikation Hämophilie B (Prüfung von Studienprotokoll und SAP [Version 1.0] [61]). Da eine potenzielle

zeitbezogene Verzerrung nicht ausgeschlossen werden konnte, wurde im Addendum A23-99 (neben der oben vorgeschlagenen Wahl von t_0) die Erforderlichkeit von Sensitivitätsanalysen beschrieben. In diesen soll für die Zeit der Patientinnen und Patienten unter Behandlung mit dem Komparator adjustiert werden (Vorgehen nach Suissa 2017 [19], siehe unten).

Weiterhin ist denkbar, für die Interventionsgruppe den Tag der Therapieentscheidung als t_0 festzulegen, für die Kontrollgruppe hingegen eine zufällige Festlegung anhand von Visitenzeitpunkten (z. B. Krankenhausvisiten) nach Einschluss in die Studie vorzunehmen [57]. Da unklar ist, inwieweit ein solches Vorgehen für zeitbezogene Verzerrung adjustiert, ist dies regelhaft nicht zu empfehlen.

Ein PNU-Ansatz, mit dem die Zeit der Patientinnen und Patienten unter einer Vortherapie mit dem Komparator berücksichtigt wird, stammt von Suissa 2017 [19]. Hierbei wird schrittweise wie folgt vorgegangen:

- **Schritt 1:** Es wird eine Basiskohorte gebildet, die den Ein- und Ausschlusskriterien der Studie entspricht. Für alle Patientinnen und Patienten dieser Basiskohorte gilt die Annahme der *Positivität*.
- **Schritt 2:** Für jede Patientin / jeden Patienten, die vom Komparator auf die Intervention wechselt ist, wird ein Expositionsset gebildet. Hierfür wird pro Patientin / pro Patient zum Zeitpunkt des Wechsels die jeweilige Exposition gegenüber dem Komparator (z. B. Zeit unter Vergleichstherapie oder Anzahl der Verschreibungen seit Beginn der Vergleichstherapie) herangezogen. In das Set gehen anschließend alle Patientinnen und Patienten der Kontrollgruppe mit zu diesem Zeitpunkt ähnlicher Exposition gegenüber dem Komparator ein [19,62].
- **Schritt 3:** Im PNU-Design ist der Propensity Score definiert als die Wahrscheinlichkeit, mit der eine Patientin / ein Patient vom Komparator auf die Intervention wechselt. Dabei wird ein zeitabhängiger Propensity Score berechnet. Die Berechnung kann unter Verwendung eines zeitabhängige Cox-Proportional-Hazards-Modells oder eines äquivalenten bedingten logistischen Regressionsmodells erfolgen, jeweils unter Berücksichtigung aller Patientinnen und Patienten. (In der Literatur wird ein weiterer Ansatz beschrieben, in dem das Modell des zeitabhängigen Propensity Scores separat für therapienaive und vorbehandelte Patientinnen und Patienten angepasst wird [63]).
- **Schritt 4:** Innerhalb eines Expositionssets erfolgt eine Paarbildung (jedem Behandlungswechsler wird die Patientin / der Patient der Kontrollgruppe zugeordnet, die / der dem Wechsler im zeitabhängigen Propensity Score am ähnlichsten ist).
- Die **Schritte 1 bis 4** münden in einer (gematchten) Analysepopulation. Als t_0 wird pro Paar der Tag der Entscheidung, eine Therapie mit der Intervention zu beginnen (z. B. Erstverordnung der Intervention) herangezogen. Um Selection Bias zu vermeiden,

empfehlen Suissa 2017 [19] eine chronologische und systematische Zuteilung ohne Zurücklegen mittels Nearest-Neighbour-Matching im Verhältnis 1:1. Je nach Datensituation (z. B. geringe Anzahl von Patientinnen und Patienten, die die Therapie mit dem Komparator fortführen) können andere Matchingverfahren sowie andere Zuteilungsschemata (1:n-Matching, mit Zurücklegen) gewählt werden [63].

Das von Suissa 2017 vorgeschlagene Vorgehen kann auf unterschiedliche Weise modifiziert werden. Wenn sich Patientinnen und Patienten, die (direkt oder verzögert) vom Komparator auf die Intervention wechseln, hinsichtlich ihrer Vortherapie(n) unterscheiden, sind solche Verfahren zu wählen, in denen durch *Matching*, Stratifizierung oder Confounderadjustierung weitere Merkmale dieser Vortherapie(n) berücksichtigt werden [59]. Dies kann gleichermaßen bei der Definition und Bildung des Expositionssets geschehen [63], führt jedoch potenziell zu kleineren und somit zu restriktiveren Expositionssets.

Eine weitere Modifikation erfolgte durch Yang 2022 [14]. Um die beiden Behandlungsgruppen hinsichtlich aller potenziellen Vortherapien anzugleichen, wird das Vorgehen von Suissa 2017 um 1 Selektionsschritt in Form der Berechnung von *Medication Possession Ratios* (MPRs) mit anschließendem Matchen erweitert. In dem von Yang 2022 verwendeten Datenbeispiel führte dieses Vorgehen jedoch insbesondere in der Kontrollgruppe zu einer starken Verringerung der Fallzahl, die in Anwendungsgebieten mit ohnehin kleinen Patientenkollektiven zu Problemen führen kann.

Das Verfahren nach Suissa 2017 (inklusive genannter Modifikationen) geht mit verschiedenen Anforderungen an die Datenerhebung einher. Bei Bildung von Expositionssets sind folgende Prüfschritte zu beachten [19]:

1) Innerhalb eines Expositionssets muss die Positivität für alle Patientinnen und Patienten geprüft werden und gegeben sein.
2) Alle Patientinnen und Patienten des Expositionssets müssen die Ein- und Ausschlusskriterien der Studie erfüllen. Bei Patientinnen und Patienten, die im Verlauf der AbD vom Komparator auf die Intervention wechseln, erfolgt zum Zeitpunkt des Wechsels eine erneute Überprüfung der Ein- und Ausschlusskriterien. Patientinnen und Patienten, bei denen zu diesem Zeitpunkt die Ein- und Ausschlusskriterien nicht mehr vollständig erfüllt sind, können dennoch potenziell als Matchingpartner der Kontrollgruppe in die Analysepopulation eingehen.
3) Für alle Patientinnen und Patienten, die aus der Kontrollgruppe in ein Expositionsset eingehen, muss eine erneute Baselineerhebung innerhalb eines zuvor festgelegten Erhebungszeitfensters erfolgen.

Die genannten Prüfschritte gehen sowohl mit einer aufwendigen Prüfung der tatsächlichen Eignung der Patientinnen und Patienten als auch mit einem hohen Erhebungsaufwand einher. Dies ist bei der Planung einer AbD zu berücksichtigen.

Daneben werden Verfahren ohne zeitabhängige Propensity Scores diskutiert. Webster-Clark 2022 [58] betrachten in einer Simulationsstudie mit sehr großen Datensätzen unterschiedliche Verfahren. Davon umfasst ist ein Verfahren mit *Standardised Morbidity (oder Mortality) Ratio Weighting* (SMRW) und ein Verfahren mit *Disease Risk Scores* (DRSs). Diese Verfahren gehen, im Vergleich mit dem Verfahren nach Suissa 2017 [19], mit einem geringeren Aufwand der Erhebung einher. Es zeigen sich für beide Verfahren präzise Schätzungen mit einem geringen Rechenaufwand. Da die Simulationen von sehr großen Fallzahlen ausgehen, ist jedoch unklar, ob die genannten Verfahren für die Anwendung im Rahmen der AbD geeignet sein werden, da hier eher kleine Patientenkollektive zu erwarten sind.

Ein weiterer Ansatz stellt die Bildung von Klonen der Patientinnen und Patienten dar [57]. Dabei werden für jede Patientin bzw. jeden Patienten 2 Klone (Kopien) erstellt; davon wird einer der Interventionsgruppe und der andere der Kontrollgruppe zugeordnet. Einer der Klone wird dann zensiert, wenn er mit der Definition der zugeordneten Gruppe nicht mehr vereinbar ist. Da diese künstliche Zensur zu einem Selection Bias (zu informativer Zensierung) führt, kommt bei diesem Ansatz zusätzlich die Inverse-Propability-of-Censoring-Weighting(IPCW)-Methode zum Einsatz [64] (zur Erläuterung von IPCW, siehe GA14-04 [65]).

Einordnung und Empfehlung

Um zeitbezogene Verzerrung bei der Auswertung versorgungsnaher Daten (und somit auch der Daten einer AbD) zu reduzieren, ist eine adäquate Wahl von t_0 von großer Bedeutung. Je nach Anwendungsgebiet stellt die Festlegung von t_0 eine besondere Herausforderung dar. Wenn die Patientinnen und Patienten in beiden Gruppen die Behandlung mit einer neuen Therapie(linie) beginnen, kann in einer Studie das ACNU-Design verwendet werden. Dabei entspricht t_0 idealerweise dem Tag der Therapieentscheidung.

In einer Therapiesituation, in der die Patientinnen und Patienten entweder die Behandlung mit einem etablierten therapeutischen Standard fortführen oder auf ein neues Arzneimittel wechseln können, kann das ACNU-Design an seine Grenzen stoßen. In diesen Fällen stellt das PNU-Design eine geeignete Alternative dar. Bei diesem Design gibt es für die Festlegung von t_0 keinen Goldstandard. Das Vorgehen nach Suissa 2017 [19] mittels Matching (systematisch, ohne Zurücklegen), nach Yang 2022 [14] mittels SMR und DRS (beschrieben bei Webster-Clark 2022 [58]) und die Bildung von Klonen [57] werden als Ansätze zur Reduktion zeitbezogener Verzerrung diskutiert. Die genannten Verfahren stellen eine Möglichkeit dar, um sich der für den interessierenden Vergleich notwendigen Strukturgleichheit der Behandlungsgruppen anzunähern [57]. Bei der Verwendung eines PNU-Designs sollte gewährleistet sein, dass die

Vortherapie(n) durch die Wahl von geeigneten Verfahren hinreichend berücksichtigt werden. Voraussetzung ist, dass die Vollständigkeit und Richtigkeit der Daten zu den Vortherapien im jeweiligen Register sichergestellt sind. Um die Robustheit der jeweiligen Effektschätzung zu überprüfen, sollten Sensitivitätsanalysen durchgeführt werden, in denen weitere der zuvor genannten Verfahren verwendet werden.

5.4.2 Behandlungswechsel

Zuordnung zu den Behandlungsgruppen

In vergleichenden Studien (mit oder ohne Randomisierung) haben Patientinnen und Patienten die Möglichkeit, bedarfsgesteuert von einer Behandlung zu einer anderen zu wechseln. Dieser Vorgang wird als „Treatment Switching“ bezeichnet. Ein solcher Wechsel kann aus mehreren Gründen erfolgen, etwa, weil eine Patientin oder ein Patient nicht auf die Studienbehandlung, die zum Zeitpunkt des Beobachtungsstarts gegeben wird, anspricht oder trotz der Behandlung eine Krankheitsprogression erleidet [65].

Unter dem Begriff „Treatment Switching“ kann ein Behandlungswechsel vom Komparator zur Intervention, ein Behandlungswechsel von der Intervention zum Komparator und ein Wechsel von der Studienbehandlung zu einer Behandlung außerhalb einer protokollgemäßen Studienbehandlung verstanden werden [66,67]. Im Rahmen des vorliegenden Rapid Reports werden ausschließlich (1) ein Wechsel von der Intervention zum Komparator (Szenario a) sowie (2) ein Wechsel vom Komparator zur Intervention (Szenario b) betrachtet. Je nach Szenario ergeben sich üblicherweise unterschiedliche Konsequenzen:

- **Szenario a:** Ein Wechsel von der Intervention zum Komparator entspricht in einer AbD im Regelfall einem Wechsel zu einer allgemein anerkannten, dem medizinischen Standard entsprechenden Behandlung. Der Wechsel ist somit Teil einer Behandlungsstrategie, die der normalen klinischen Praxis entspricht. In einer solchen Situation ist, trotz des Wechsels im Verlauf der Beobachtung, eine unverzerrte Schätzung des Therapieeffekts möglich [67].
- **Szenario b:** Ein Wechsel vom Komparator zur Intervention ist dann unproblematisch, wenn die Intervention eine adäquate Folgetherapie im Rahmen der Behandlungsstrategie darstellt (dies trifft z. B. dann zu, wenn für die Intervention bereits zu einem früheren Zeitpunkt eine Zulassung in einer späteren Therapielinie erfolgt ist). Diese Konstellation ist im Rahmen einer AbD jedoch eher nicht zu erwarten. In allen anderen Fällen kann ein Wechsel vom Komparator zur Intervention zu verzerrten Ergebnissen führen und es ist nachzuweisen, dass die Ergebnisse auf die Fragestellung der AbD übertragbar sind.

Unabhängig von einem Behandlungswechsel werden zum Zwecke der Nutzenbewertung von Arzneimitteln primär Auswertungen entsprechend der Treatment-Policy-Strategie (d. h. eine

Schätzung des Effekts für die gesamte Behandlungsstrategie) und des Intention-to-treat(ITT)-Prinzips benötigt [6,34]. Um einen potenziell verzerrenden Einfluss eines Wechsels auf die ITT-Auswertung bei Anwendung von Verfahren für Überlebenszeiten zu untersuchen, können Sensitivitätsanalysen durchgeführt werden, in denen Patientinnen und Patienten mit Wechsel von der Kontrollgruppe zur Intervention zum Zeitpunkt ihres Wechsels zensiert werden (Per-Protocol-Analysen). Dieser Ansatz führt bei Vorliegen von Störgrößen, die sowohl das Überleben als auch den Mechanismus des Behandlungswechsels beeinflussen, allerdings zu informativen Zensierungen (d. h. die Zensierung ist nicht unabhängig vom Endpunkt und von der Behandlungsgruppe, sondern hängt von Personen- / Behandlungscharakteristika ab, z. B. Beobachtungsende nach Progression). Da Überlebenszeitanalysen (z. B. das Cox-Proportional-Hazards-Modell) nicht-informative Zensierungen voraussetzen, führt der Ansatz, Patientinnen und Patienten, die die Behandlung wechseln, zum Zeitpunkt ihres Wechsels zu zensieren, zu potenziell verzerrten Effektschätzungen [65,67].

Des Weiteren kommen zur Berücksichtigung von „Treatment Switching“ auch komplexere Methoden für die Schätzung des Behandlungseffektes auf Basis parametrischer Modelle des Zensierungsmechanismus zur Anwendung (z. B. IPCW). Diese komplexeren Methoden sind in ihrer Anwendung allerdings an eine Reihe von Voraussetzungen und Annahmen geknüpft, die in der Praxis im Regelfall nicht überprüfbar sind [65]. Aufgrund dessen sollten Ergebnisse aus solchen Analysen ergänzend zur ITT-Analyse dargestellt werden. Unabhängig davon, welche Methoden zur Berücksichtigung von Behandlungswechsel für einen Target Trial vorgesehen sind, kommen entsprechende statistische Verfahren auf Basis von versorgungsnahen Daten nur dann infrage, wenn der Registerdatensatz die dafür notwendigen Daten (insbesondere zu potenziellen *Confoundern*) enthält [68].

Besondere Situationen bei der Auswertung versorgungsnah erhobener Daten

Der G-BA fordert AbDs häufig für Arzneimittel für seltene Erkrankungen (Orphan Drugs). Daraus ergeben sich per se kleine für eine AbD zur Verfügung stehende Patientenzahlen. Dazu kommen weitere Besonderheiten, die die Rekrutierung einer ausreichenden Patientenzahl für AbD-Verfahren zu neuen Arzneimitteln erschweren, wie z. B. eine geringe Inanspruchnahme von Gentherapien in der Indikation Hämophilie (u. a. aufgrund einer verzögerten Entwicklung geeigneter Erstattungsmodelle [69]).

Nachfolgend wird ein pragmatischer Ansatz beschrieben, der ergriffen werden kann, um eine möglichst hohe Patientenzahl unter Behandlung mit der jeweils interessierenden Therapie (im Sinne der Realisierbarkeit einer AbD) in der Analyse zu berücksichtigen. Darüber hinaus wird der Umgang mit Brückentherapien (d. h. Therapien, die nach der Behandlungszuteilung, aber vor der interessierenden Intervention verabreicht werden) erläutert.

Zuordnung zu den Behandlungsgruppen in Abhängigkeit der Beobachtungszeit unter Therapie mit dem Komparator

Wie oben bereits ausgeführt, erfolgt die Auswertung von Studienergebnissen im Rahmen von Nutzenbewertungen in der Regel im Sinne einer ITT-Auswertung. Dies erfordert, dass alle Patientinnen und Patienten, losgelöst von einem Wechsel der Behandlung im Verlauf der Beobachtung, gemäß ihrer ursprünglichen Gruppenzuteilung ausgewertet werden. Um dem Problem teils sehr geringer Patientenzahlen in der Interventionsgruppe (z. B aufgrund von Rekrutierungsproblemen) entgegenzuwirken, ist denkbar, bei Durchführung vergleichender Studien zur Generierung versorgungsnaher Daten die Zuordnung zu den Behandlungsgruppen in Abhängigkeit von der Beobachtungszeit der Patientin bzw. des Patienten unter Therapie mit dem Komparator (Zeit von Beobachtungsstart [t_0] bis Behandlungswechsel) vorzunehmen.

Als Beispiel zur Veranschaulichung dieses Vorgehens kann die Indikation Hämophilie dienen, in der bisher erst wenige Patienten mit einer Gentherapie behandelt werden [69]. Gemäß den Angaben in den Studienunterlagen der AbD zu Etranacogen Dezaparvovec in der Indikation Hämophilie B ist Folgendes geplant [70,71]:

- Patienten, die bereits > 2 Jahre bei einer geplanten Beobachtungszeit von 3 Jahren unter Therapie mit dem Komparator beobachtet wurden, sodass bereits aussagekräftige Daten für die Kontrollgruppe vorliegen, und erst dann auf eine Behandlung mit Etranacogen Dezaparvovec wechseln, verbleiben (im Sinne des ITT-Prinzips) in der Kontrollgruppe, werden bis zum Studienende weiterbeobachtet und in dieser Gruppe ausgewertet. Um den Einfluss von Etranacogen Dezaparvovec in der Kontrollgruppe zu untersuchen, werden Sensitivitätsanalysen durchgeführt, in denen Patienten, die in der Kontrollgruppe auf Etranacogen Dezaparvovec wechseln und in der Primäranalyse nicht der Interventionsgruppe zugeordnet werden, zum Zeitpunkt ihres Wechsels zensiert werden.
- Patienten, die nach einer kürzeren Zeit (≤ 2 Jahre) auf eine Behandlung mit Etranacogen Dezaparvovec wechseln und für die noch eine angemessene Beobachtungszeit unter der Behandlung mit Etranacogen Dezaparvovec zu erwarten ist, sollen hingegen in der Interventionsgruppe ausgewertet werden. Für diese Patienten wird der Zeitpunkt des Behandlungswechsels als Startzeitpunkt der Beobachtung (Index-Datum [t_0]) herangezogen und die Baselinecharakteristika erneut erhoben. Die Beobachtungszeit dieser Patienten unter Therapie mit dem Komparator wird für die AbD nicht berücksichtigt.

Umgang mit Brückentherapien

In Fällen, in denen eine Therapieoption nach der Therapieentscheidung nicht sofort verfügbar oder nicht sofort wirksam ist, kann der Einsatz einer Brückentherapie zur Krankheitskontrolle

angezeigt sein. Im Rahmen bisheriger AbD-Verfahren haben Brückentherapien in folgenden Konstellationen eine Rolle gespielt:

- **B-Zell-Neoplasien:** Einsatz einer Antitumortherapie zwischen der Gewinnung von T-Zellen durch Leukapherese zur Überbrückung der Wartezeit bis zur Bereitstellung der CAR-T-Zellen [72]
- **Spinale Muskelatrophie:** Vorbehandlung mit Nusinersen oder Risdiplam bis zum Einsatz der geplanten Gentherapie (mit einer präspezifizierten klaren Grenze, bis zu welcher Dauer eine Brückentherapie als solche aufgefasst wird) [73,74]
- **Hämophilie:** hämostatische Unterstützung mit exogenem humanen Faktor VIII / Faktor IX nach Infusion der geplanten Gentherapie, um eine ausreichende Faktor-VIII- / Faktor-IX-Versorgung in den ersten Wochen nach der Behandlung zu erreichen [70]

Eine Brückentherapie ist als Teil des Therapiekonzepts zu sehen [75]. Entscheidend für einen adäquaten Umgang mit Brückentherapien ist, dass der Tag der Entscheidung für die eigentlich geplante Therapie (z. B. der Zeitpunkt der Entscheidung für eine CAR-T-Zelltherapie im Tumorboard) als t_0 gemäß Konzept der *Target-Trial-Emulation* festgesetzt wird. Das bedeutet, dass die Beobachtung ab t_0 startet und dass alle interessierenden Ereignisse, die unter Behandlung mit der Brückentherapie auftreten, im Sinne des ITT-Prinzips ebenfalls in die Analysen eingehen. Der Grund für den Einsatz einer Brückentherapie und die Art der Brückentherapie sind dabei zu dokumentieren.

5.5 Patientenberichtete Endpunkte

Der Nutzen eines Arzneimittels ist im Sinne der Arzneimittel-Nutzenbewertungsverordnung der patientenrelevante therapeutische Effekt u. a. hinsichtlich der Verbesserung des Gesundheitszustandes (der Symptomatik) oder einer Verbesserung der Lebensqualität [76]. Um bestehende Evidenzlücken zu einem neuen Arzneimittel zu schließen, fordert der G-BA daher auch für AbDs regelhaft die Erhebung von Endpunkten zur Symptomatik und zur gesundheitsbezogenen Lebensqualität. Diese Endpunkte werden üblicherweise mit PROs erhoben.

Anforderungen an Instrumente zur PRO-Erhebung

Für die Erhebung von PROs sollten Instrumente, die gemäß etablierter Standards entwickelt und validiert wurden, herangezogen werden [77]. Dabei sind, insofern möglich, indikationsspezifische Instrumente zur Erhebung der Symptomatik oder gesundheitsbezogenen Lebensqualität gegenüber generischen Instrumenten zu bevorzugen.

Herausforderungen bei der Erhebung von PROs in der Versorgung

Da PRO-Daten in einer AbD zum Zwecke der Nutzenbewertung erhoben werden, müssen sie, um interpretierbare Ergebnisse zu liefern, bestimmte Qualitätsanforderungen erfüllen. Dabei ist zu beachten, dass Daten im Kontext von AbDs in der Versorgungspraxis erhoben werden.

Damit geht einher, dass sich die Bedingungen für die Datenerhebungen deutlich von jenen in einer klinischen Studie unterscheiden. Beispielsweise erfolgen Visiten unregelmäßig(er) und die Behandlung der Patientinnen und Patienten findet ggf. auf unterschiedlichen Versorgungsebenen (z. B. hausärztliche Versorgung und Versorgung in Zentren) statt. Im Folgenden sollen Wege aufgezeigt werden, wie unter diesen Rahmenbedingungen eine Erhebung von PROs mit möglichst wenig Aufwand gelingen kann. Hinweise auf Ansätze zur Vermeidung von fehlenden Werten insbesondere bei der PRO-Erhebung, mit dem Ziel der Sicherstellung einer hohen Datenqualität, finden sich in Abschnitt 5.6.1.

Bisherige Erfahrungen mit der PRO-Erhebung im Rahmen der anwendungsbegleitenden Datenerhebung

Bei der Konzepterstellung für eine AbD wird regelhaft geprüft, ob in dem infrage kommenden Register die Möglichkeit besteht, PROs zu erheben. Dabei zeigen die bisherigen Erfahrungen ein gemischtes Bild. Neben Registern, in denen keine PROs erfasst werden und dies auch nicht geplant ist, finden sich solche, in denen die PROs vom regulären Datensatz umfasst sind oder über Zusatzmodule erhoben werden. In einigen Fällen wurden eigens für vom G-BA geforderte AbDs Lösungen gefunden, um die Erhebung von PROs zu ermöglichen. Beispiele hierfür sind:

- Für die AbD zu Brexucabtagen Autoleucel in der Indikation rezidiviertes oder refraktäres Mantelzell-Lymphom ist die PRO-Erhebung nicht in das (primäre) Europäische Mantelzell-Lymphom(EMCL)-Register direkt integriert, sondern ein externes Institut mit der Erhebung der Daten beauftragt. Hierfür wurde (angesiedelt am Institut für Medizinische Biometrie Epidemiologie und Information [IMBEI] an der Universitätsmedizin Mainz) eine Vertrauensstelle und zentrale PRO-Unit eingerichtet [78,79]. Auf diese Weise kann der Aufwand im Register geringer gehalten werden und die Erhebung der PRO-Daten ist nicht mehr zwingend an eine Visite im Behandlungszentrum gebunden. Da die Fragebogen in der AbD auf postalischem Weg versendet und eingeholt werden (mit postalischen und ggf. telefonischen Erinnerungen bei fehlender Antwort), können Potenziale, die mit einer digitalen PRO-Datenerfassung einhergehen würden, jedoch nicht ausgeschöpft werden (z. B. müssen die durch die Patientinnen und Patienten beantworteten Fragebogen zunächst durch die PRO-Unit in eine Datenbank überführt werden) [79]. Hinzu kommt, dass die behandelnden Zentren über die Ergebnisse der jeweiligen PRO-Erhebung nicht informiert werden [78], obwohl der positive Einfluss einer PRO-Datennutzung in der Routineversorgung in Studien gut belegt ist (siehe unten).
- Im Deutschen Hämophilieregister (DHR) sind PROs regulär nicht im Datensatz erfasst. Für die AbDs zu Valoctocogen Roxaparvovec (Indikation Hämophilie A) und Etranacogen Dezaparvovec (Indikation Hämophilie B) wurde die Erhebung von PROs – nur für die AbD – in einem externen Modul etabliert [80,81]. Die Funktion für die Beantwortung der Patientenfragebogen wird nach einer Online-Schulung durch die DHR-Geschäftsstelle

freigeschaltet. Den Patienten wird für die Beantwortung ein (patientenspezifisch) generierter Link oder QR-Code am Tag der Einwilligung in die AbD zur Verfügung gestellt. Für jeden Fragebogen (Erfassungszeitpunkte alle 3 Monate) wird über die gesamte Dauer der AbD derselbe Link bzw. QR-Code verwendet. Die Zentren können die ausgefüllten PROs ihrer Patienten in der DHR-Datenbank (mit Status [z. B. „fällig“ oder „Befragung unterbrochen“) einsehen. Wird ein Fragebogen in einer Folgeerhebung ausnahmsweise nicht elektronisch, sondern in Papierform ausgefüllt, kann das Zentrum die Daten vom Papierfragebogen in der Datenbank nachtragen [81].

PRO-Erhebung mittels digitaler Anwendungen

Es zeichnet sich bei der Generierung versorgungsnaher Daten (in Registern und Pilotprojekten in der klinischen Routineversorgung) ein Trend zur digitalen Erfassung von PROs ab [82,83]. In der klinischen Versorgungspraxis ist mittels einer digitalen Erhebung per App und / oder per Patientenportal eine niedrigschwellige, den Arzt-Patienten-Dialog strukturierende, Erhebung unabhängig von Arztbesuchen möglich [84,85]. Ein weiterer Vorteil einer digitalen Erfassung wird darin gesehen, dass die mit einer PRO-Messung potenziell einhergehende Belastung der Patientinnen und Patienten verringert wird [84] (z. B. müssen keine ausgefüllten Papierfragebogen postalisch versendet werden). Anschauliche Beispiele für erfolgreiche digitale Anwendungen in Registern finden sich in Anhang A.

Erhebungszeitpunkte und Toleranzfenster

Zum Zwecke der Generierung aussagekräftiger Daten über den Verlauf der Erkrankung unter Therapie sind für beide Behandlungsgruppen einheitliche Erhebungszeitpunkte notwendig. Dies gilt in besonderem Maße, wenn in ihrer Anwendung sehr unterschiedliche Therapien verglichen werden (z. B. CAR-T-Zelltherapie vs. Chemotherapie). Die Erhebung sollte mehrfach im Jahr und in standardisierten Intervallen zu definierten Zeitpunkten erfolgen [84,86]. Hierbei sollte eine im Studienverlauf angepasste Erhebungsfrequenz gewählt werden (mit zu Beginn engeren und später breiteren Intervallen), um eine Belastung durch häufige Erhebungen zu verringern [84]. Eine niedrigschwellige, jedoch standardisierte Erhebung von PRO-Daten im Rahmen einer AbD kann durch den Einsatz digitaler Tools gelingen.

Für die Erhebungszeitpunkte von PROs sollten in den Studienunterlagen für eine AbD jeweils angemessene Toleranzfenster präspezifiziert werden, die nicht aneinandergrenzen und eine sinnvolle Zuordnung zu einem Erhebungszeitpunkt möglich machen. Es ist davon auszugehen, dass Rückläufe der PRO-Fragebogen in der Regel eher nach einem festgelegten Erhebungszeitpunkt eingehen. Aufgrund dessen kann es sinnvoll sein, asymmetrische Toleranzfenster zu wählen, die ein kürzeres Zeitfenster vor dem geplanten Erhebungszeitpunkt (z. B. 2 Wochen) und einen längeres Nacherhebungszeitfenster (z. B. 4 bis 6 Wochen) umfassen ([A24-38] Etranacogen Dezaparvovec [Hämophilie B]: 3. Addendum zum Auftrag A22-83 [87]).

Nutzbarkeit digital erhobener PRO-Daten im Versorgungsalltag (auch außerhalb einer anwendungsbegleitenden Datenerhebung)

Digitale Tools können (in Abgrenzung zu papierbasierten Befragungen, die postalisch oder per E-Mail übermittelt werden) durch eine digitale Übermittlung von PROs und ihrer Entwicklung im Zeitverlauf im klinischen Setting eine zentrale Rolle einnehmen. Gründe hierfür sind u. a.:

- Das Abrufen von Echtzeitdaten ermöglicht das Symptomtracking von Patientinnen und Patienten. Dabei wird im betreuenden Zentrum beim Überschreiten klinisch relevanter Grenzwerte ein Alarm ausgelöst (Frühwarnsystem). In der Folge können Behandlungsanpassungen frühzeitig eingeleitet oder die Patientinnen und Patienten telefonisch beraten oder einbestellt werden [83,88].
- Erkenntnisse aus vorhandenen PRO-Daten (z. B. durch das Abrufen in tagesaktuellen Dashboards) können in der Patientenberatung und zur gemeinsamen Entscheidungsfindung eingesetzt werden [83,88,89].
- Für einzelne onkologische Entitäten wurde nachgewiesen, dass eine medizinische Betreuung, die patientenberichtete Daten im Verlauf integriert, einer Betreuung ohne solche Informationen überlegen ist [90-96]. Darunter wurde in 2 Studien der Effekt einer Einbindung von digital übermittelten PRO-Daten in die Routineversorgung (in Form eines Symptommonitorings) auf das Gesamtüberleben untersucht. Es zeigt sich für den Endpunkt Gesamtüberleben jeweils ein statistisch signifikanter Unterschied zum Vorteil einer Nutzung von PROs zum Monitoring von Symptomen gegenüber der Standardversorgung [93,96].

Als ein Erfolgsfaktor für eine häufigere Nutzung der Daten im Klinikalltag wird die Integration der digital erfassten PRO-Daten in die elektronischen Fall- oder Patientenakten, inklusive einer adäquaten grafischen Aufbereitung zur schnellen Erfassung der Daten, genannt, da diese mit einer Minimierung des Vorbereitungsaufwands beim zu behandelnden Personal einhergehen kann [83].

Leicht anwendbare, effiziente digitale Lösungen tragen durch eine automatisierte Versendung von Einladungen zum Ausfüllen der Fragebogen (mit Erinnerungen bei fehlender Antwort) und die automatisierte PRO-Datenerhebung, -eingabe und -auswertung zur Minimierung des sehr hohen personellen und logistischen Aufwands in den behandelnden Zentren bei, der z. B. bei einer papierbasierten Erhebung eine große Herausforderung darstellt [79,97].

Auf Patientenebene ist es für eine erfolgreiche Erhebung von PROs notwendig, die Erhebung so aufwandsarm und in die Behandlung integriert wie möglich zu gestalten (siehe Abschnitt 5.6.1). Zur Sicherstellung, dass dabei Aufwand und Nutzen in einem angemessenen Verhältnis stehen, wird, wenn immer möglich, der Einsatz elektronischer PROs empfohlen [84,85,98-100]. Ein anderer Erhebungsmodus (wie z. B. telefonisch oder papierbasiert) wird lediglich als

Ergänzung für ansonsten nur schwer erreichbare Patientinnen und Patienten diskutiert [85,98,99,101,102].

In der Gesamtschau der gesichteten Literatur zeigt sich, dass die Erhebung von PRO-Daten im Behandlungsalltag grundsätzlich möglich und sinnvoll ist. Dabei bietet eine digitale Erhebung von PROs die Möglichkeit, dass diese Daten praktikabel und ohne größere Hürden für Studien auf Basis versorgungsnaher Daten (und somit auch für die AbD) zur Verfügung stehen.

5.6 Fehlende Werte

5.6.1 Maßnahmen zur Vermeidung fehlender Werte

Fehlende Werte bei wissenschaftlichen Untersuchungen gehen mit einem Informationsverlust und, wenn diese fehlenden Werte nicht durch einen Zufallsmechanismus zustande kommen, mit der Gefahr der Verzerrung der Studienergebnisse einher [15]. Bei Bedarf sind statistische Verfahren (Ersetzungsstrategien) zum adäquaten Umgang mit fehlenden Werten erforderlich. Die Validität aller Verfahren sinkt mit zunehmendem Anteil von fehlenden Werten und beruht auf der Plausibilität der zugrunde liegenden Annahmen, warum ein fehlender Wert auftritt. Diese Annahmen sind für die Auswahl von geeigneten Verfahren zum Umgang mit fehlenden Werten jedoch in der Regel nicht überprüfbar. Fehlende Werte lassen sich in Studien in der Regel nicht vollständig vermeiden. Um jedoch hinreichend sichere Aussagen zum Nutzen und zum Schaden der zu bewertenden Arzneimittel sowohl unter den Bedingungen der Routineversorgung als auch unter Berücksichtigung der Fragestellung der Nutzenbewertung machen zu können, sind daher Maßnahmen sinnvoll und notwendig, den Anteil fehlender Werte so gering wie möglich zu halten.

Ursachen fehlender Werte bei einer Datenerhebung in einem Register können thematisch den 3 folgenden Bereichen zugeordnet werden:

- Registerstruktur,
- Motivation der Zentren (des medizinischen Personals) zur Datenerhebung und Datenübermittlung sowie
- Bereitschaft der Patientinnen und Patienten, ihre Daten dem Register zur Verfügung zu stellen.

Rückmeldungen von Registerbetreibenden in A19-43

In A19-43 wurden besonders förderliche und besonders hinderliche Faktoren für den Aufbau und den Betrieb von Registern aufgeführt. In den zur Identifizierung dieser Faktoren durchgeführten Gespräche mit Registerbetreibenden wurden folgende Ansätze genannt, die auf eine Minimierung unvollständiger Daten schließen lassen [18]:

- Etablierung elektronischer Schnittstellen für automatisierte Datenübernahmen aus (externen) Datenquellen, z. B. aus Krankenhausinformationssystemen (KIS), Praxissystemen oder Sterberegistern (Standardisierung der IT-Infrastruktur)
- Sicherstellung semantischer Interoperabilität durch Verwendung einheitlicher Datenstandards (Harmonisierung von Kodiersystemen)
- Etablierung von standardisierten Kommunikationswegen zwischen Registerbetreibenden und teilnehmenden Zentren (Zentrumsbetreuung)
- regelmäßige Schulungen für das Erhebungspersonal, ggf. Einsatz von eigens geschulten Fachkräften (professionelle Unterstützung der Datenerhebung vor Ort)
- Etablierung eines Rückmeldesystems an die teilnehmenden Zentren (Nutzbarkeit einrichtungsbezogener Daten aus dem Register, z. B. zur Darstellung zeitlicher Verläufe oder für Benchmarking)
- Datenbereitstellung für Forschungszwecke und Beteiligung der Zentren (des medizinischen Personals) an wissenschaftlichen Publikationen

Zusätzlich identifizierte Maßnahmen

Aufbauend auf den Ausführungen in A19-43 sollten weitere Ansätze identifiziert werden, um den Anteil fehlender Werte bei der Erhebung von versorgungsnahen Daten möglichst gering zu halten. Zu diesem Zweck wurden zum einen die Ergebnisse der bisherigen Prüfungen von Studienprotokoll und SAP für vom G-BA geforderte AbDs und die im Rahmen der vorliegenden Ausarbeitung recherchierte Literatur gesichtet. Zum anderen wurden Hinweise der externen Sachverständigen auf Maßnahmen, die sich in der Registerpraxis bewährt haben, integriert. Folgende Ansätze bieten Potenzial zur Erhöhung der Vollständigkeit der Datensätze:

Registerstruktur

- Eine hohe Datenqualität ist mit einem hohen Aufwand bei der Dateneingabe verbunden. Aufgrund der ohnehin schon begrenzten Ressourcen für die Versorgung kann eine hohe Datenqualität nur dann in der Breite erreicht werden, wenn die Datenerhebung in einem Register, die durch die Zentren neben dem Behandlungsalltag zu bewältigen ist, mit möglichst geringem Aufwand möglich ist. Dazu ist es erforderlich, die Erhebung auf diejenigen Daten zu beschränken, die den Registerzielen entsprechen und die für die Beantwortung der zu untersuchenden Fragestellung einer Registerstudie relevant sind. Eine einfach zu bedienende, anwenderfreundliche Softwarelösung kann die Dateneingabe und Datenübermittlung unterstützen (Software-Ergonomie) [103,104].
- Um hochwertige Gesundheitsdaten aus dem Versorgungsalltag praktikabel und ohne große Hürden vorhalten und bereitstellen zu können, muss eine dauerhaft verfügbare und fortgeführte Dateninfrastruktur und Dokumentation personell und finanziell in den Registern aufgebaut werden. Der Aufbau und der Unterhalt dieser Infrastruktur könnten

durch die pUs, die auf die Daten zur Durchführung registerbasierter Studien zurückgreifen wollen, unterstützt werden.

- Eine systematische, risikobasierte Monitoringstrategie, in der ein zentrales Monitoring (einschließlich Querymanagement) um Monitoringmaßnahmen in Zentren mit Auffälligkeiten (zielgerichtetes On-Site-Monitoring) ergänzt wird, wird als geeignete Maßnahme zur Erhöhung der Vollständigkeit der Datensätze erachtet [105,106]. Vor dem Hintergrund des mit Monitoringmaßnahmen verbundenen hohen Ressourcenaufwands liegen erste positive Erfahrungen in der Nutzung eines automatisierten Monitoringsystems (IT-gestützte Prüfung [Missing-Analysen] mit Query-System) vor [107].

Aufrechterhaltung der Motivation der teilnehmenden Zentren sowie Patientinnen und Patienten

Als förderlich für eine möglichst langfristige Aufrechterhaltung der Motivation der am Register teilnehmenden Zentren sowie teilnehmenden Patientinnen und Patienten wurden, ergänzend zu den in A19-43 gemachten Angaben, beispielhaft folgende Faktoren identifiziert:

Zentren

- Eine Anreizmöglichkeit zur nachhaltigen Sicherstellung der Nutzbarkeit der Registerdaten kann die Sichtbarkeit einer hohen Versorgungsqualität für eine breitere Öffentlichkeit sein, die durch Transparenz der erreichten Ergebnisse entsteht. Einrichtungsbezogene Daten aus dem Register können für Auszeichnungen (Zertifikate) oder die Zertifizierung der teilnehmenden Zentren und (auf aggregierter Ebene) für Public Reporting (zur Kommunikation in die Fachcommunity) genutzt werden [83,88,104,108].
- Eine Förderung der Netzwerkbildung innerhalb eines Registers (Register-Community) und des Erkenntnisaustausches zur Weiterentwicklung der Kompetenz (z. B. in Form eines regelmäßigen Treffens zum Austausch zu Best Practices) können zu einer erhöhten Teilnahmebereitschaft des medizinischen Personals beitragen [89,104,108].
- Die Vergütung des Aufwands der teilnehmenden Zentren kann durch Zahlung einer Aufwandsentschädigung, z. B. in Abhängigkeit von der Qualität der erhobenen Daten, erfolgen [104,108,109].
- Ein Ansatz zur Sensibilisierung für die Datenerfassung aufseiten des medizinischen Personals stellt eine Förderung der Einbettung der Daten in die tägliche klinische Praxis und Entscheidungsfindung mittels Abrufen von Echtzeitdaten (z. B. über eine IT-Plattform oder die elektronische Patientenakte [Fallakte]) dar [110-116]. Engagierte Teams, gezielte Schulungsprogramme und eine grafische Aufbereitung der Daten zur schnellen Erfassung und Aufwandsminimierung (durch automatisierte Bereitstellung der

Auswertung) können den Wandel hin zur Nutzung im klinischen Alltag unterstützen [83,88]. Die unmittelbare Sichtbarkeit des Nutzens der erhobenen Daten setzt einen Anreiz, diese vollständig und in guter Qualität zur Verfügung zu stellen (siehe auch Abschnitt 5.4.2).

Patientinnen und Patienten

- Eine Einbeziehung der Interessen von Patientinnen und Patienten in Form einer Interessenvertretung in die Organisation eines Registers stellt sicher, dass die erhobenen Daten für die Patientinnen und Patienten von Relevanz sind und die Teilnahme am Register als sinnvoll erlebt wird (partizipativer Ansatz) [108,117,118].
- Die Patientinnen und Patienten müssen über den Nutzen der Datenerfassung (einschließlich patientenberichteter Daten) für die Qualität ihrer Versorgung und für ihre Gesundheit aufgeklärt werden. Eine adäquate Auswahl von Instrumenten zur Erfassung von PROs (orientiert an der Relevanz für Patientinnen und Patienten), eine Auswahl möglichst kurzer PRO-Instrumente, direktes Feedback auf digitale Dateneingaben und die Nutzung zur unmittelbaren Therapiesteuerung oder das Symptommanagement sind oft genannte Erfolgsfaktoren (im Sinne eines Mehrwerts für die individuelle Patientenbehandlung) [16,83,119,120].
- Eine laienverständliche Aufbereitung und Kommunikation der Registerergebnisse, eine zielgruppengerechte Gestaltung der Patienteninformation und kontinuierliche Presse- und Öffentlichkeitsarbeit (z. B. durch informativen Webauftritt oder Newsletter) wirken sich positiv auf die Teilnahmebereitschaft aus [103,104,117,119].
- Durch die Schaffung eines spezifischen Zugangs (z. B. über eine IT-Plattform [Dashboard] oder per App über Smartphone / Tablet) können die im Register gespeicherten Daten individualisiert in patientenverständlicher und erläuterter Form direkt an die Patientinnen und Patienten zurückgespielt werden (individuelle Verlaufsdaten oder Daten im Vergleich mit einer Peer-Gruppe), um ihr Engagement zu fördern [83,110,112,114,118,119,121-124].
- Die Nutzung einer Plattform mit der Möglichkeit, weiterführende Informationen zum Krankheitsbild oder zum Selbstmanagement zu erhalten und sich untereinander zu vernetzen, kann für Patientinnen und Patienten einen weiteren Motivationsanreiz darstellen (Informations- und Austauschangebot) [112,114,119,121,125].
- Neben immateriellen Anreizen (wie Bereitstellung der Registerdaten) empfinden Patientinnen und Patienten materielle Anreize in Form finanzieller Zuwendungen (Geschenkgutscheine) oder Sachprämien als motivierend [108,121,126].

Empfehlungen zur Vermeidung fehlender Werte im Rahmen der anwendungsbegleitenden Datenerhebung

Nicht alle der oben genannten Maßnahmen zur Vermeidung fehlender Werte lassen sich unter den derzeitigen Rahmenbedingungen zum Zwecke einer AbD durch (einzelne) pUs umsetzen. Darunter fällt beispielhaft die Etablierung von elektronischen Schnittstellen für automatisierte Datenübernahmen aus (externen) Datenquellen. Die Sicherstellung eines Teils der genannten Maßnahmen liegt jedoch im Verantwortungsbereich des jeweiligen pU. Im Rahmen einer AbD wird u. a. die Umsetzung folgender Maßnahmen als machbar eingeschätzt:

- Beschränkung der Erhebung auf diejenigen Daten, die für die Beantwortung der zu untersuchenden Fragestellung der AbD relevant sind
- Etablierung einer geeigneten Monitoringstrategie, ggf. unter Nutzung eines automatisierten Monitoringsystems
- Vergütung des Aufwands der teilnehmenden Zentren durch Zahlung einer Aufwandsentschädigung (z. B. in Abhängigkeit von der Qualität der erhobenen Daten)
- Angebot (regelmäßiger) Schulungen für die Teams in den Behandlungszentren zur Datenerfassung und zur aktiven Anwendung von PRO-Ergebnissen in der klinischen Praxis
- Aufklärung der Patientinnen und Patienten über den Nutzen der Datenerfassung (einschließlich patientenberichteter Daten)
- Auswahl adäquater PRO-Instrumente (u. a. hinsichtlich der Länge des Fragebogens und der Relevanz der Fragen)
- Zurückspielen der im Register gespeicherten Daten direkt an die Patientinnen und Patienten, individualisiert in patientenverständlicher und erläuterter Form
- Einsatz materieller Anreize in Form finanzieller Zuwendungen (Geschenkgutscheine) oder Sachprämien an die Patientinnen und Patienten

5.6.2 Methodischer Umgang mit fehlenden Werten im Kontext von Propensity-Score-Analysen

5.6.2.1 Identifikation und Diskussion von geeigneten statistischen Methoden

Propensity-Score-Verfahren sind komplex. Ausgehend von der Wahl des *Estimands* (*Average Treatment Effect in the Population* (ATE) oder *Average Treatment Effect on the Treated* (ATT)) muss zum einen eine angemessene Confounderidentifikation und -auswahl erfolgen (siehe Abschnitt 5.2) und zum anderen eine Entscheidung sowohl für ein Verfahren zum Umgang mit fehlenden Werten als auch für konkretes Propensity-Score-Verfahren (*Matching*, Gewichtung [z. B. mit *Inverse Probability of Treatment Weighting*, IPTW], Stratifizierung oder Regressions-adjustierung) getroffen werden.

In der klinischen Forschung ergibt sich regelmäßig das Problem fehlender Werte, so etwa bei Endpunkten oder bei weiteren Variablen wie Baselinecharakteristika (siehe Tabelle 2).

Tabelle 2: Muster für beobachtete Werte für N Personen im Endpunkt und in den Confoundern L1 bis L7

Personen	**Behandlung**	**Endpunkt**	**L1**	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**Anteil fehlende Werte pro Person**
1	Intervention	●		●		●	●	●		
2	Intervention	●	●	●	●	●	●	●	●	
3	Komparator		●				●			
4	Intervention	●	●		●	●		●		
5	Komparator	●	●		●	●		●		
…										
N	Komparator	●		●	●	●		●		
Anteil fehlende Werte pro Endpunkt bzw. Confounder										
● beobachteter Wert										

Aus Tabelle 2 wird deutlich, dass bei Betrachtung fehlender Werte verschiedene Perspektiven eingenommen werden können. Neben dem Anteil von Variablen mit fehlenden Werten pro Patientin bzw. Patient kann der Anteil von Patientinnen und Patienten mit fehlenden Werten pro Variable oder auch der fehlende Anteil des gesamten Datenvolumens, das idealerweise hätte erhoben werden sollen, betrachtet werden. Zur Skizzierung des Musters von fehlenden Werten in einem Datensatz werden 3 Mechanismen unterschieden:

- *Missing completely at Random* (MCAR): Die Wahrscheinlichkeit für das Fehlen eines Wertes ist unabhängig von anderen beobachteten Werten einer Person (z. B. Kovariablen oder Werten zu früheren Zeitpunkten) und dem unbeobachteten, fehlenden Wert selbst. Solche fehlenden Werte führen grundsätzlich nicht zu verzerrten Ergebnissen. Allerdings verringert sich die Präzision der Effektschätzungen [15]. Die MCAR-Annahme ist die stärkste der 3 Annahmen, unter der relativ einfache Methoden der Datenanalyse angewendet werden können.
- *Missing at Random* (MAR): Die Wahrscheinlichkeit für das Fehlen eines Wertes hängt von anderen beobachteten Werten einer Person ab, nicht aber vom fehlenden Wert selbst. Solche fehlenden Werte führen zu verzerrten Ergebnissen, insofern die Abhängigkeit in der Analyse unberücksichtigt bleibt [15]. Wenn Angaben zu den anderen Werten (z. B. Kovariablen oder Werten zu früheren Zeitpunkten) einer Person vorliegen, kann diese Information für den Umgang mit fehlenden Werten genutzt werden.

- *Missing not at Random* (MNAR): Die Wahrscheinlichkeit für das Fehlen eines Wertes hängt vom tatsächlichen, fehlenden Wert der betrachteten Zielgröße selbst ab. Solche fehlenden Werte führen zu verzerrten Ergebnissen. Die Verzerrung kann auch nicht korrigiert werden, da der Ausfallmechanismus von unbeobachteten Daten abhängt [15]. Zum Umgang mit fehlenden Werten unter der MNAR-Annahme sind weitreichende Verteilungsannahmen über den Zusammenhang des fehlenden Wertes mit weiteren beobachteten Werten notwendig.

In der Regel ist es nicht möglich, die Annahmen MAR und MNAR mittels statistischer Analysen auszuschließen. Das Ignorieren von MAR und MNAR führt jedoch zu verzerrten Ergebnissen.

Modelle zur Schätzung von Propensity Scores mittels logistischer Regression erfordern für alle potenziell relevanten *Confounder* vollständige Daten, d. h., dass Patientinnen und Patienten mit (mindestens) 1 fehlendem Wert in der Analyse nicht berücksichtigt werden. Wenn jedoch ausschließlich Patientinnen und Patienten mit vollständigen Daten in die Analysen eingehen, sogenannte Complete-Case-Analysen, kann dies bei einer Nicht-Erfüllung der MCAR-Annahme zu einer deutlichen Verzerrung führen [127,128]. Davon ist in Beobachtungsstudien in aller Regel auszugehen. Aufgrund dessen werden nachfolgend statistische Verfahren zum Umgang mit fehlenden Werten in Confoundern im Rahmen von Propensity-Score-Analysen beschrieben und eingeordnet.

Durch eine orientierende Informationsbeschaffung wurde eine Reihe von Simulationsstudien identifiziert [127-137]. In diesen Studien wurden unterschiedliche Methoden zum Umgang mit fehlenden Werten im Rahmen von Propensity-Score-Verfahren untersucht. Komponenten der Simulationen sind die Fallzahl (in der Regel N ≥ 1000), die Ausprägung der Endpunkte (binär oder stetig), die Anzahl (2 bis 10) und die Ausprägung (binär oder stetig) der Confounder sowie der Korrelation untereinander. Zusätzlich variieren die Anzahl der Confounder mit fehlenden Werten und die Anteile von fehlenden Werten sowie der zugrunde liegende Mechanismus für die fehlenden Werte (MCAR, MAR oder MNAR). Untersucht werden Modelle sowohl unter der Nullhypothese als auch mit Effekten der Behandlung. Für einen Vergleich der verschiedenen Verfahren wird in der Regel die Verzerrung und *Überdeckungswahrscheinlichkeit* betrachtet, wohingegen die Balanciertheit der Confounder, die *Überlappung* sowie die *Konvergenz* der zugrunde liegenden Algorithmen deskriptiv dargestellt werden.

Zusammenfassend haben sich nach Sichtung der Literatur Verfahren der multiplen Imputation als Standardverfahren etabliert. Diese werden als angemessen angesehen, um dem Problem fehlender Werte bei Anwendung von Propensity-Score-Verfahren zu begegnen. Hierzu wird insbesondere die Variante *multiple Imputation mittels Chained Equation* (MICE) – auch Fully Conditional Specification (FCS) genannt – vorgeschlagen [127-133,135,138,139]. Neben dem

MICE-Verfahren gibt es noch eine Reihe weiterer Varianten der multiplen Imputation (modellbasiert, regressionsbasiert, Predictive Mean Matching oder bayesianisch), die in einer Studie zur Anwendung kommen können.

Der klassische Ansatz von MICE umfasst folgende Schritte:

1) Für jede in das Modell eingehende Variable wird ein Regressionsmodell spezifiziert, z. B. ein lineares Regressionsmodell für eine stetige Variable.
2) Fehlende Werte werden M-mal imputiert durch Stichprobenziehung der Parameter des Modells für fehlende Daten. Neben der Verwendung der beobachteten Werte werden dabei im Algorithmus imputierte Werte einer Variablen für die Imputation anderer fehlender Werte in anderen Variablen verwendet (Kette [Chain]).
3) Es werden M vollständige Datensätze erzeugt und die Propensity-Score-Berechnung wird für jeden Datensatz getrennt durchgeführt.

Die Anwendung von MICE wurde in allen für die vorliegende Ausarbeitung berücksichtigten Simulationsstudien untersucht und zeigt in einer Vielzahl der dort untersuchten Situationen bessere Eigenschaften mit Hinblick auf Verzerrung und Überdeckungswahrscheinlichkeit als alternative Verfahren. Die Besonderheiten bei der Anwendung von MICE in Propensity-Score-Verfahren werden im Folgenden diskutiert.

Anwendung von MICE

„across“- oder „within“-Ansatz?

Ein Propensity-Score-Verfahren besteht aus 2 Schritten. Zunächst wird für jede Patientin und für jeden Patienten ein Propensity Score berechnet. Anschließend erfolgt in einem 2. Schritt eine Schätzung des Therapieeffekts, in der für den Propensity Score adjustiert wird (z. B. durch Gewichtungsverfahren). Liegen fehlende Werte vor, die eine multiple Imputation erfordern, so stellt sich die Frage, wie die multiple Imputation mit diesen 2 Schritten kombiniert wird. Hierfür werden hauptsächlich 2 Ansätze vorgeschlagen: „within“-Ansatz und „across“-Ansatz [127,128,130-132,134,135,139-142]. Beide Ansätze imputieren fehlende Daten und schätzen individuelle Propensity Scores in jedem imputierten Datensatz. Daran anschließend wird beim „within“-Ansatz der Behandlungseffekt in jedem Datensatz geschätzt und die Schätzungen werden mittels Rubin-Regel gepoolt. Beim „across“-Ansatz wird hingegen der Durchschnitt der individuellen Propensity Scores, die in jedem vollständigen Datensatz geschätzt wurden, gemittelt und die Effektschätzung erfolgt einmalig anhand der zuvor gemittelten Propensity Scores. Beide Ansätze wurden in mehreren Simulationsstudien untersucht. Die Ergebnisse aus früheren Studien ließen keine eindeutige Aussage darüber zu, welcher der beiden Ansätze in der Praxis angewendet werden soll. Neuere Arbeiten kommen hingegen zu dem Ergebnis, dass der „within“-Ansatz zu bevorzugen ist, da dieser mit einer im Vergleich geringeren Verzerrung und höheren Überdeckungswahrscheinlichkeit einhergeht [128,131,142]. Diese Vorteile des

„within"-Ansatz gehen allerdings zu Lasten eines höheren Aufwands. Neben einem höheren Rechenaufwand ist die Beurteilung der Balanciertheit und der *Überlappung* herausfordernd. Während beim „across"-Ansatz die finalen, gemittelten Propensity Scores zur Beurteilung der Balanciertheit und der Überlappung genutzt werden können, sollte dies beim „within"-Ansatz für jeden imputierten Datensatz mit nachfolgender Propensity-Score-Berechnung geschehen [127,128]. Für den Fall, dass in einem Imputationsschritt keine ausreichende Balanciertheit vorliegt oder die Überlappung nicht ausreichend ist, ist in der gesichteten Literatur weder für den „across"-Ansatz noch für den „within"-Ansatz ein Vorgehen beschrieben.

Einfluss der Anzahl der Imputationen

In den untersuchten Simulationsstudien wurden eine Anzahl imputierter Datensätze von 5 bis 20 angesetzt. Insgesamt wird der Einfluss der Anzahl der Imputationen als nicht hoch eingeschätzt. Aufgrund der in AbDs zu erwartenden Datensituationen mit deutlich geringeren Fallzahlen und höherer Anzahl von Confoundern ist im Kontext einer AbD eine deutlich höhere Anzahl von Imputationen anzusetzen (z. B. 50). Wenn die bayesianische Erweiterung von MICE mittels eines iterativen Markov-Chain-Monte-Carlo-Algorithmus (MCMC) angewendet wird, ist dieser mit einer angemessenen Anzahl von Iterationen durchzuführen. In einem Fallbeispiel sind 40 Iterationen für den MCMC vorgeschlagen [138]. Diese Anzahl ist angemessen.

Einbeziehung des Endpunktes in das MICE-Verfahren

In der Literatur wird diskutiert, ob neben den Confoundern zusätzlich der Endpunkt in der Modellierung des MICE-Verfahrens berücksichtigt werden soll [130,143]. Für den Einbezug des Endpunktes sprechen nicht nur neuere Simulationsergebnisse [128,131], sondern auch theoretische Überlegungen [144].

Operationalisierung von Confoundern

Es können Datensituationen vorliegen, in denen der Algorithmus zur Imputation und / oder der Algorithmus zur Berechnung der Propensity Scores nicht angemessen ist. Dies können Situationen sein, in denen keine ausreichende Balanciertheit zwischen beiden Behandlungsgruppen erreicht wird oder aber der Algorithmus nicht konvergiert. In diesen Fällen stellt die Änderung der Operationalisierung der Confounder eine Option dar. Bei stetigen Confoundern, wie z. B. Alter, kann dies über Transformationen (Log-Funktion oder kubische Spline-Funktion) erfolgen. Bei kategorialen Variablen können Ausprägungen zusammengefasst werden. Auch die Verwendung von Interaktionstermen ist möglich. Eiset und Frydenberg 2022 [138] geben diesbezüglich ein anschauliches Beispiel. Ziel ihrer Analyse war es, für alle 5 eingeschlossenen Confounder eine Balanciertheit mit dem Kriterium standardisierte Mittelwertdifferenz < 0,10 zu erreichen. Für die Erreichung dieses Ziels wurden, je nach Confounder, für die zugrunde liegenden Modelle unterschiedliche Operationalisierungen gewählt.

Weitere Ansätze zum Umgang mit fehlenden Werten

Neben dem MICE-Verfahren gibt es nicht nur weitere Varianten der multiplen Imputation, sondern auch methodische Ansätze, die nicht auf multipler Imputation beruhen, z. B. Missing-Pattern-Approach [145], Missing-Indicator-Methode [140], General-Boosting-Modeling sowie Random-Forest-Based-Missing-Imputation-Methods [132]. Obschon sich für diese Methoden ebenfalls gute Ergebnisse in Simulationsstudien zeigen, stehen für diese bisher insgesamt zu wenige Daten zur Verfügung, als dass die Anwendung von MICE zum jetzigen Zeitpunkt infrage gestellt wird.

Beschreibung geeigneter Sensitivitätsanalysen

Wie oben beschrieben, hat sich als Methode der Wahl zum Umgang mit fehlenden Werten für Confounder bei Propensity-Score-Verfahren MICE etabliert. Vorteil dieser Methode ist, dass hierfür ausschließlich die MAR-Annahme erfüllt sein muss und zur Spezifikation nur wenige Parameter, z. B. die zugrunde liegenden Regressionsmodelle und die Anzahl an Imputationen, notwendig sind. Als Sensitivitätsanalysen bieten sich Methoden an, die prinzipiell geeignet sind, die Ergebnisse der primären Analyse infrage zu stellen, insofern nicht ausgeschlossen werden kann, dass die der primären Analyse zugrunde liegenden Annahmen vollständig erfüllt sind. Da MICE lediglich die MAR-Annahme voraussetzt, eignen sich hierfür weitere modell-basierte Methoden, die auf der Annahme MAR oder MNAR beruhen [129,132,140]. Lee 2021 [139], Ling 2020 [133] sowie Carpenter und Smuk 2021 [146] geben dazu praktische Hinweise. Die relevante Komponente ist dabei generell die MAR-Annahme [145,146]. Daneben können sogenannte auxiliare Variablen verwendet werden [133,138]. Diese sind keine Confounder im engeren Sinne, sind aber mit den Confoundern assoziiert und ihre Berücksichtigung im Modell für die multiple Imputation kann die Imputation der fehlenden Werte bei den Confoundern verbessern [138,146]. Bei der Wahl dieses Verfahrens muss die Erhebung der entsprechenden Variablen im Register sichergestellt sein.

In weiteren Sensitivitätsanalysen kann die Operationalisierung der potenziellen Confounder variiert werden, wie z. B. durch eine Transformation von stetigen Variablen [138] oder durch Kalibrierungsmethoden [136]. Ein lehrreiches Beispiel für eine Propensity-Score-Analyse in Kombination mit multipler Imputation zum Umgang mit fehlenden Werten findet sich in der Publikation von Eiset 2022 [138]. In dieser Publikation wird die Komplexität des Verfahrens herausgestellt und die einzelnen Annahmen für das Modell, das den Imputationen zugrunde liegt, diskutiert. Basierend auf den von Eiset 2022 diskutierten Punkten können geeignete Sensitivitätsanalysen geplant werden.

Nicht geeignet als Sensitivitätsanalysen sind Complete-Case-Analysen, die ausschließlich auf Patientinnen und Patienten beruhen, für die vollständige Daten vorliegen. Eine Complete-Case-Analyse beruht auf der MCAR-Annahme, die eine deutlich strengere Annahme ist als die MAR-Annahme. Es zeigt sich in Simulationsstudien wiederholt, dass die Durchführung einer

Complete-Case-Analyse verglichen mit anderen Methoden zu deutlich erhöhten Verzerrungen führt [130,131,133,139-141].

5.6.2.2 Interpretierbarkeit von Analysen mit fehlenden und / oder imputierten Werten

In vorliegenden Abschnitt wird skizziert, bis zu welchen Anteilen von fehlenden und / oder von imputierten Werten Analysen noch interpretierbar sind. Die Ausführungen beziehen sich sowohl auf die Anteile an fehlenden Werten pro Patientin / pro Patient als auch auf die Anteile an fehlenden Werten pro Endpunkt / *Confounder*.

Schwellenwerte für fehlende Werte in Endpunkten

Das Fehlen von Werten in Endpunkten in relevantem Ausmaß führt in der Regel dazu, dass die hiervon betroffenen Patientinnen und Patienten nicht in der Effektschätzung berücksichtigt werden können. Wenn für > 30 % der Studienpopulation keine Werte vorliegen oder beträgt der Unterschied zwischen den Behandlungsgruppen > 15 Prozentpunkte, so werden die Ergebnisse für den entsprechenden Endpunkt im Regelfall nicht berücksichtigt [34]. Ist der Anteil kleiner als 30 %, so wird endpunktspezifisch über die Konsequenzen hinsichtlich der Ergebnissicherheit diskutiert. Im Falle von Endpunkten mit Messwiederholungen, wie z. B. PROs, sollten für mindestens 70 % der Patientinnen und Patienten sowohl der Wert zu Studienbeginn als auch mindestens ein beobachteter Wert zu einem weiteren Zeitpunkt vorliegen. Bei Analysen zu einem bestimmten Zeitpunkt sollten zu diesem Zeitpunkt Werte für mindestens 50 % der Patientinnen und Patienten vorliegen. Diese Schwellenwerte dienen als Orientierung, wenn nicht mit ausreichender Sicherheit davon ausgegangen werden kann, dass die Werte zufällig fehlen. Ausnahmen ergeben sich beispielsweise, wenn in einer Studie für einen Fragebogen Daten von Patientinnen und Patienten aus bestimmten Zentren fehlen und die Zentren nicht als Effektmodifikatoren angesehen werden oder bei Vorliegen großer Behandlungseffekte.

Schwellenwerte für fehlende Werte in Confoundern

Die Ableitung konkreter Schwellenwerte für den Anteil fehlender Werte bei Confoundern, ab dem die Ergebnisse nicht mehr sinnvoll interpretierbar sind, ist anhand der Literatur lediglich eingeschränkt möglich. In keiner der identifizierten Arbeiten wurden konkrete Vorschläge dazu gemacht, ab welchem Anteil von fehlenden Werten in Confoundern sich Konsequenzen für die Interpretation der Effektschätzung ergeben. Die Aussagekraft der Ergebnisse der dargestellten Simulationsstudien sind hinsichtlich der in AbDs zu erwartenden Datensituationen limitiert. In den Studien werden große Fallzahlen mit ≥ 1000 Patientinnen und Patienten bei zugleich begrenzter Anzahl von Confoundern betrachtet (meist im Bereich von 2 bis 5). Im Kontext von AbDs ist jedoch von kleinen Fallzahlen und von deutlich mehr als 5 Confoundern auszugehen. Dabei ist zu beachten, dass die Anzahl der Patientinnen und Patienten sowie die Anzahl relevanter Confounder (und deren Messniveau) die Durchführbarkeit einer multiplen Imputation beeinflussen, d. h. sowohl die *Konvergenz* des zugrunde liegenden Algorithmus als

auch die Güte der multiplen Imputation und damit letztendlich auch die mit dem jeweiligen *Propensity-Score-Verfahren* vorgenommene Effektschätzung. Hinzukommend wurden in den Simulationsstudien weder die Balanciertheit noch die *Überlappung* ausreichend hinsichtlich eines konkreten Vorgehens untersucht [137,138,141].

Die Anteile von fehlenden Werten sind zwischen den Simulationsstudien sehr unterschiedlich. Die Gesamtanteile für fehlende Werte liegen zwischen 10 % und 80 %, die Anteile für einzelne Confounder zwischen 10 % und 60 %. In einer Simulationsstudie von de Vries und Groenwold 2017 [130] zeigen sich für den „across"-Ansatz deutliche Verzerrungen in einem Bereich von 20 % bis 80 % für den Gesamtanteil an fehlenden Werten (über alle Confounder), wohingegen der „within"-Ansatz mit vertretbaren Verzerrungen im Bereich von 20 % bis 40 % einhergeht. Dabei zeigen sich für das *IPTW*-Verfahren, im Vergleich zur Adjustierung mittels Propensity Scores oder *Propensity-Score-Matching*, höhere Verzerrungen. Bei Granger 2019 [131] zeigt sich für alle Propensity-Score-Verfahren hinsichtlich der Verzerrung eine gute Performance über eine große Spanne von fehlenden Werten (10 % bis 75 %). Die Ergebnisse der genannten Simulationsstudien sind allerdings lediglich bedingt vergleichbar, da sich die Studien nicht nur in den verwendeten statistischen Verfahren, sondern auch in den simulierten Szenarien unterscheiden.

In der Gesamtschau ergeben sich aus diesen Simulationsstudien folgende Schwellenwerte als Orientierung:

Eine Complete-Case-Analyse ist angemessen, insofern weniger als 5 % der Patientinnen und Patienten nicht in der Analyse berücksichtigt werden und insgesamt nicht mehr als 10 % fehlende Werte vorliegen. Bei einem höheren Anteil fehlender Werte können die Ergebnisse aufgrund der nichtberücksichtigten Beobachtungen deutlich verzerrt sein [133]. In den Fällen, in denen der Anteil der fehlenden Werten pro Confounder weniger als 10 % beträgt und der Anteil fehlender Werte insgesamt < 20 % ist, kann eine multiple Imputation als angemessen angesehen werden. Höhere Anteile fehlender Werte pro Confounder und / oder insgesamt können bei der Anwendung der multiplen Imputation zu einer niedrigeren Ergebnissicherheit führen. In einem solchen Fall sollen adäquate Sensitivitätsanalysen durchgeführt werden, um die Robustheit der Ergebnisse der primären Analyse zu prüfen. In Situationen, in denen der Gesamtanteil fehlender Werte > 50 % beträgt, werden die Daten als für die Nutzenbewertung nicht geeignet angesehen.

Nicht adäquat ist ein Ausschluss von relevanten Confoundern aufgrund von zu hohen Anteilen fehlender Werte. In diesen Situationen lässt sich ein Zusatznutzen lediglich bei sehr großen Effekten im Sinne eines dramatischen Effekts ableiten.

5.6.2.3 Anforderungen an die Darstellung

Aus der komplexen Fragestellung – dem Umgang mit fehlenden Werten bei *Propensity-Score-Verfahren* – ergeben sich besondere Anforderungen an die Berichterstattung. Imputation von fehlenden Werten kann, selbst wenn der Algorithmus als solcher erfolgreich ist, zu verzerrten Effektschätzungen führen, da die Imputationen unter Annahmen getroffen werden, die in der Regel nicht überprüft werden können. Dies ist bei der Interpretation der Ergebnisse zu den Effektschätzungen zu berücksichtigen.

Anforderungen für die Darstellung der Planung im Studienprotokoll und statistischen Analyseplan

- methodische Details zur Implementierung der multiplen Imputation
  - Begründung für die Wahl des Verfahrens für den Umgang mit fehlenden Werten
  - Beschreibung des Modells für die primären Analysen, inklusive Operationalisierungen der *Confounder*
  - Anzahl der imputierten Datensätze
  - Anzahl der Iterationen für den MCMC-Algorithmus eines bayesianischen Ansatzes und / oder Anzahl der Iterationen für ein Bootstrap-Verfahren (insofern angewendet)
- methodische Details zu den Sensitivitätsanalysen
- Beschreibung des Umgangs mit Balanciertheit und *Überlappung* in den Imputationen
- Beschreibung von Kriterien für die Ergebnisinterpretation hinsichtlich der Ergebnissicherheit

Anforderungen an die Berichterstattung für die Nutzenbewertung

Die Berichterstattung sollte folgende Aspekte umfassen:

- methodische Details zur Implementierung der multiplen Imputation, insbesondere die im Studienprotokoll und / oder SAP präspezifizierten Vorgehensweisen und Parameter
- deskriptive Darstellung der fehlenden Werte vor der multiplen Imputation
- deskriptive Darstellung der Charakteristika nach der multiplen Imputation
- Beschreibung der *Konvergenz* der zugrunde liegenden Algorithmen zur multiplen Imputation und zur Propensity-Score-Berechnung sowie Begründungen für die Operationalisierungen der eingehenden Confounder [138]
- Begründungen für methodische Aspekte, die nicht im Studienprotokoll und / oder SAP präspezifiziert werden konnten [138]

- geeignete Beschreibungen für das Vorgehen bezüglich Balanciertheit und Überlappung sowie deren Darstellung
- vollständige Darstellung der geplanten Sensitivitätsanalysen
- Diskussion der Ergebnissicherheit der Effektschätzungen unter Berücksichtigung möglicher Verzerrungen durch die fehlenden Werte

5.7 Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven

Bei Studien ohne Randomisierung ist die für einen fairen Vergleich notwendige Strukturgleichheit der zu vergleichenden Gruppen in der Regel nicht gegeben. Daher müssen Gruppenunterschiede in potenziellen *Confoundern* bei der Schätzung von Effekten durch Adjustierung für diese Confounder berücksichtigt werden [18]. Für eine Datenanalyse mit einer adäquaten Adjustierung für Confounder gibt es verschiedene kausale Methoden (wie z. B. multifaktorielle Regressionsmodelle, Instrumentenvariablen und Propensity Scores).

Zur Sicherstellung einer adäquaten Confounderadjustierung ist eine ausreichend hohe Anzahl von Patientinnen und Patienten erforderlich. Obgleich es in der Literatur keine Empfehlungen dazu gibt, wie viele Patientinnen und Patienten für die Auswertung benötigt werden, lässt sich grundsätzlich feststellen, dass die notwendige Patientenzahl mit einer steigenden Anzahl von Confoundern größer wird. Die Auswertung nicht randomisierter vergleichender Studien stellt bei kleinen Patientenkollektiven daher eine besondere Herausforderung dar. Dies gilt generell für alle statistische Verfahren zur Confounderkontrolle.

Im Rahmen einer AbD spielt zur Confounderadjustierung insbesondere die Propensity-Score-Methode eine Rolle [147], u. a., weil es hier Ansätze gibt, zentralen Annahmen der kausalen Inferenz (d. h. ausreichende *Positivität*, *Überlappung* und Balanciertheit) zu prüfen und darzustellen [148]. *Propensity-Score-Verfahren* wurden ursprünglich für große Datensätze mit zum Teil mehreren tausend Patientinnen und Patienten entwickelt (für epidemiologische Fragestellungen) [149]. Zunehmend werden sie allerdings auch zur Auswertung von vergleichenden Studien ohne Randomisierung auf Basis kleinerer Fallzahlen eingesetzt. Da die zur Verfügung stehende Patientenzahl im Kontext einer AbD in der Regel gering ist, stellt sich die Frage, inwieweit Propensity-Score-Verfahren auch bei kleinen Fallzahlen im Bereich von 100 bis 500 Patientinnen und Patienten aussagekräftige Ergebnisse liefern können und welche Besonderheiten bei Vorliegen kleiner Fallzahlen zu beachten sind.

In den letzten Jahren wurde eine Reihe von Untersuchungen zu Propensity-Score-Verfahren bei kleinen Fallzahlen publiziert, in denen der Einfluss verschiedener Parameter anhand von zum Teil sehr umfangreichen Simulationsstudien untersucht wurde. Zu diesen Parametern zählen:

- die Gesamtfallzahl und das Verteilungsverhältnis der Patientinnen und Patienten zwischen den Gruppen,
- das Effektmaß (wie Odds Ratio oder Mittelwertdifferenz),
- die Art des Propensity-Score-Verfahrens (*Matching*, Gewichtung, Stratifizierung oder Adjustierung),
- die Anzahl der Kovariaten und deren Ausprägung (binär oder stetig) und
- das Modell für die Abhängigkeiten zwischen Kovariaten, Behandlung und Endpunkten.

Aufgrund der Vielzahl an Möglichkeiten, die sich durch den vieldimensionalen Parameterraum ergeben, können selbst große Simulationsstudien immer nur einen kleinen Ausschnitt (wenige Szenarien) abbilden. Dies betrifft insbesondere die Stärke des Einflusses der Kovariaten auf Behandlung und Endpunkt.

Simulationsstudien zu Propensity-Score-Analysen bei kleinen Fallzahlen

Mittels orientierender Informationsbeschaffung wurden Simulationsstudien zu Propensity-Score-Analysen bei kleinen Fallzahlen identifiziert. Diese werden im Folgenden beschrieben:

In einer umfangreichen Simulationsstudie von Pirracchio 2012 [150] wurden für das Effektmaß Odds Ratio die statistischen Eigenschaften der beiden gängigsten Propensity-Score-Verfahren, Matching und Gewichtung mittels *IPTW*, verglichen. Dabei wurden Fallzahlen von 40 bis 1000 sowie bis zu 4 Kovariaten mit einem unterschiedlichen Einfluss auf Behandlung und Endpunkt betrachtet. Für alle Simulationen hat die Autorengruppe ein Basisrisiko von 50 % angenommen. Aus der Untersuchung geht hervor, dass der Fehler 1. Art auch bei kleinen Fallzahlen gut eingehalten wird, sofern alle Confounder im Modell berücksichtigt sind. Auch der relative Bias (d. h. die prozentuale Abweichung der mittleren Effektschätzung aus den Simulationen vom angenommenen Effekt im Simulationsmodell) liegt selbst bei Fallzahlen von 40 Patientinnen und Patienten bei weniger als 10 %, was eine akzeptable Größe darstellt. Dabei schneidet die Propensity-Score-Gewichtung mittels IPTW im Vergleich zum Propensity-Score-Matching zum Teil etwas besser ab.

Eine Simulationsstudie von Friedrich 2020 [151] untersuchte verschiedene Propensity-Score-Verfahren bei kleinen Fallzahlen (N = 40, N = 100 und N = 1000) für verschiedene Szenarien mit bis zu 9 Kovariaten mit einem wechselnden Einfluss auf Behandlung und Endpunkt. Dabei wurden die Effektmaße Odds Ratio und Risikodifferenz verglichen. Es zeigt sich, dass Modelle unter Verwendung des Odds Ratio, im Gegensatz zur Verwendung der Risikodifferenz, zum Teil stärker verzerrt sind. Zudem geht aus der Untersuchung hervor, dass Matchingverfahren unter Verwendung beider Effektmaße, insbesondere bei sehr kleinen Fallzahlen (N = 40), oft zu Konvergenzproblemen führen. Dabei ist der Anteil der Szenarien, in denen die Modelle für die Effektschätzungen nicht konvergieren, umso höher, je kleiner die gewählte Fallzahl ist. Die

Autorengruppe dieser Simulationsstudie empfiehlt daher bei kleinen Fallzahlen die Wahl der Risikodifferenz als Effektmaß. Von den in ihrer Untersuchung betrachteten Adjustierungsverfahren empfehlen sie Doubly-Robust-Schätzer für Quintile und raten von einer Propensity-Score-Gewichtung mittels IPTW aufgrund einer niedrigen *Überdeckungswahrscheinlichkeit* ab.

Eine niedrige Überdeckungswahrscheinlichkeit bei einer Propensity-Score-Gewichtung mittels IPTW zeigt sich auch in einer Simulationsstudie von Austin 2022 [152]. Anzumerken ist, dass in dieser Studie lediglich Fallzahlen von mindestens 250 betrachtet wurden. Für verschiedene Szenarien mit jeweils 5 stetigen und 5 dichotomen Kovariaten wurde die Überdeckungswahrscheinlichkeit u. a. der 3 Gewichtungsarten IPTW, *Matching Weights* sowie *Overlap Weights* verglichen. Als Effektmaße wurden Mittelwertdifferenz, Risikodifferenz und relatives Risiko gewählt. Ein besonderes Augenmerk lag dabei auf der Frage, inwieweit eine Varianzschätzung mittels Bootstrap der asymptotischen Varianzschätzung überlegen ist. Während sich für die Gewichtungsarten Matching Weights und Overlap Weights jeweils eine gute Überdeckungswahrscheinlichkeit bei Verwendung einer asymptotischen Varianzschätzung zeigt (ohne eine Verbesserung bei Verwendung von Bootstrap), kann die Überdeckungswahrscheinlichkeit bei der Gewichtungsart IPTW mittels Bootstrap deutlich verbessert werden, sodass sie an jene bei Verwendung von Matching Weights und Overlap Weights heranreicht. Dies gilt insbesondere dann, wenn das Verhältnis der Fallzahlen zwischen den Gruppen sehr unterschiedlich ist.

In einer Simulationsstudie von Wilkinson 2022 [153] wurden 5 Szenarien mit variierender Überlappung der Propensity-Scores für 1 dichotome Kovariate und 4 stetige Kovariaten sowie Fallzahlen ab 100 für das Effektmaß Odds Ratio analysiert. Zusätzlich variierte das Verteilungsverhältnis zwischen den Gruppen von 1:20 (sehr unbalanciert) bis 1:1 (ausgeglichen). Neben einer direkten Adjustierung mittels logistischer Regression wurden die 3 Propensity-Score-Verfahren Matching, Gewichtung mittels IPTW und Adjustierung verglichen. Es zeigt sich für alle Verfahren, dass die Überdeckungswahrscheinlichkeit mit zunehmender Unbalanciertheit der Patientinnen und Patienten zwischen den Gruppen abnimmt, wobei die Abnahme beim Matching im Vergleich zum Gewichten und Adjustieren stärker ausfällt. Hinzu kommt, dass eine Schätzung des Effekts mit zunehmender Unbalanciertheit der Patientinnen und Patienten zwischen den Gruppen und kleiner werdender Überlappung der Propensity Scores zum Teil nicht möglich ist, da die Modelle für die Effektschätzungen nicht konvergieren. Während beim Gewichten und Adjustieren die *Konvergenz* sehr hoch ist, kann der Konvergenzanteil bei den Matchingverfahren insbesondere bei einer mäßigen Überlappung, selbst bei ausgeglichenen Gruppengrößen, auf 50 % bis 70 % absinken [153].

Einordnung der Ergebnisse der Simulationsstudien und Empfehlungen

Die Szenarien und Festlegungen der Parameter der zuvor beschriebenen 4 Simulationsstudien unterscheiden sich untereinander teilweise deutlich und keine Untersuchung beschreibt ein umfassendes Bild.

Zudem bilden die in den Arbeiten untersuchten Szenarien die Situationen, die in bisherigen AbD-Verfahren aufgetreten sind, in Teilen nicht ab. Zwar wurden in den 4 Simulationsstudien die in AbDs häufig geplanten Propensity-Score-Verfahren (Matching und Gewichtung) untersucht, die Anzahl relevanter Confounder ist in den bereits laufenden AbD-Verfahren jedoch teilweise höher (5 bis 22 Confounder [70,72-74,154]) als die, die den genannten Simulationsstudien zugrunde lag (4 bis 10 Confounder [10 Confounder bei Fallzahl von N ≥ 250]). Ferner wurden maßgeblich die Effektmaße Odds Ratio und Risikodifferenz untersucht. In der frühen Nutzenbewertung von Arzneimitteln (und damit auch in der AbD) stellen hingegen das RR, das Hazard Ratio oder die (standardisierte) Mittelwertdifferenz die üblichen Effektmaße dar. Nur in der Arbeit von Austin 2022 [152] wurden die beiden Effektmaße Mittelwertdifferenz und RR herangezogen. Für diese Effektmaße zeigen sich in dieser Arbeit im Vergleich zum Odds Ratio oder der Risikodifferenz keine wesentlichen Unterschiede. Ob dies grundsätzlich gilt, lässt sich anhand der identifizierten Simulationsstudien nicht beurteilen. Insgesamt ist somit nicht abschätzbar, inwiefern sich die Ergebnisse der Simulationsstudien auf den Kontext der AbD anwenden lassen.

Zusammenfassend erschweren die genannten Aspekte sowohl den Vergleich der Simulationsergebnisse untereinander als auch eine Ableitung von allgemeinen Handlungsempfehlungen für die AbD.

Im Folgenden werden verschiedene Einflussfaktoren diskutiert, die bei der Verwendung eines Propensity-Score-Verfahrens berücksichtigt werden müssen.

Einfluss der Confounderanzahl

Um Verzerrungen durch Confounding zu minimieren, ist es bei Propensity-Score-Verfahren essenziell, alle potenziell relevanten Confounder zu identifizieren und diese in der Analyse zu berücksichtigen (siehe Abschnitt 5.2). Dies wird durch die Ergebnisse der Simulationsstudien von Pirracchio 2012 [150] und Friedrich 2020 [151] gestützt, in denen die Modelle, die nicht alle Confounder enthielten, deutlich schlechtere Ergebnisse zur Folge hatten, insbesondere hinsichtlich Verzerrung und Überdeckungswahrscheinlichkeit. Aus den Untersuchungen geht jedoch auch hervor, dass unter Anwendung von Matchingverfahren bei kleinen Fallzahlen die Anzahl der Confounder, die in die Modelle aufgenommen werden können, begrenzt ist, da es sonst zu einer fehlenden Konvergenz der Modelle für die Effektschätzungen und infolgedessen zu einer Nicht-Berechenbarkeit von Effektschätzungen kommen kann. Daneben sollten stetige Confounder für die Schätzung der Propensity Scores auch als stetige Variablen in das Modell eingehen und nicht dichotomisiert werden. Aus der gesichteten Literatur lässt sich dabei nicht ableiten, wie viele Patientinnen und Patienten pro einzubeziehenden Confounder in eine Studie eingeschlossen werden müssen, um dem Konvergenzproblem zu begegnen.

Dem Problem der fehlenden Konvergenz aufgrund einer zu hohen Anzahl von Confoundern kann bis zu einem gewissen Grad durch eine präspezifizierte Festlegung einer Reihenfolge der Confounder nach ihrer Wichtigkeit begegnet werden (zu einem Beispiel, siehe Abschnitt 5.2, Schritt 3). Wenn die Modelle bei Aufnahme aller ursprünglich identifizierten Confounder nicht konvergieren, müssen weniger wichtige Confounder schrittweise aus dem Modell entfernt werden. Dabei ist zu beachten, dass die Ergebnisunsicherheit mit zunehmendem Weglassen von Confoundern ansteigt. Diese erhöhte Ergebnisunsicherheit muss bei der Ergebnisinterpretation, z. B. durch weitere Verschiebung der Nullhypothese, berücksichtigt werden. Dieses Vorgehen ist aber nur begrenzt sinnvoll einsetzbar. Wenn auf diese Weise zu viele Confounder nicht berücksichtigt werden, ist nicht mehr von einer ausreichenden Confounderadjustierung auszugehen und die Ergebnisse sind nur bei sehr großen Effekten im Sinne eines dramatischen Effekts interpretierbar. Dafür muss die Datenlage im Einzelfall betrachtet werden.

Auswahl des Propensity-Score-Verfahrens bei kleinen Fallzahlen

Hinsichtlich der Auswahl eines adäquaten Propensity-Score-Verfahrens bei kleinen Fallzahlen zeigt sich in den Untersuchungen, dass die Propensity-Score-Verfahren, die auf Gewichtung oder auf Adjustierung beruhen, insbesondere mit Hinblick auf Konvergenzprobleme Vorteile gegenüber Matchingverfahren haben. Bei Verwendung der gängigen Gewichtungsmethode IPTW ist die Überdeckungswahrscheinlichkeit bei kleinen Fallzahlen und kleinen Basisrisiken zu klein, dies kann jedoch über eine Varianzschätzung mittels Bootstrap ausgeglichen werden. Für die Risikodifferenz zeigt das Verfahren des Doubly-Robust-Schätzers für Quintile Vorteile gegenüber IPTW.

Die identifizierten Simulationsstudien umfassen nur einen kleinen Anteil möglicher Szenarien. Aus diesem Grund kann anhand der vorhandenen Literatur keine für alle Situationen klare Empfehlung für das Vorgehen unter Verwendung von Propensity Scores bei kleinen Fallzahlen abgeleitet werden. Die Studien zeigen im Grundsatz, dass die Anwendung von Propensity-Score-Verfahren in kleinen Patientenkollektiven möglich ist. Zum jetzigen Zeitpunkt liegen jedoch keine Simulationsstudien vor, die die Bedingungen, unter denen eine AbD durchgeführt wird, hinreichend abbilden, da diese vorwiegend andere Effektmaße und anteilig eine geringe Anzahl potenzieller Confounder betrachten. Folglich gibt es bisher keine Evidenz, um abschätzen zu können, wann Propensity-Score-Analysen in einer AbD zu interpretierbaren Ergebnissen führen.

Zusammenfassung

Propensity-Score-Verfahren wurden ursprünglich für sehr große Datensätze entwickelt. Die identifizierten Simulationsstudien zeigen, dass diese Verfahren auch in Anwendungsgebieten mit kleinen Patientenpopulationen unter bestimmten Bedingungen anwendbar sind und zu interpretierbaren Ergebnissen führen können. Ein relevantes Problem besteht allerdings in der Gefahr der fehlenden Konvergenz der Modelle für die Effektschätzung. Diesem Problem

kann durch verschiedene Maßnahmen begegnet werden, teilweise aber unter Inkaufnahme einer erhöhten Unsicherheit. Die in den Simulationsstudien untersuchten Szenarien bilden die Situationen, wie sie in den laufenden AbD-Verfahren absehbar auftreten werden, jedoch in Teilen nicht ab. Zum jetzigen Zeitpunkt bleibt folglich unklar, in welchen Fällen (und unter welchen Voraussetzungen) Propensity-Score-Analysen in einer AbD zu interpretierbaren Ergebnissen führen werden.

6 Fazit

Für die Nutzenbewertung von Arzneimitteln bedarf es Daten zum Vergleich mit dem therapeutischen Standard. Da die Zulassung von Orphan Drugs häufig auf Basis nicht vergleichender Daten erfolgt, wurde das Verfahren der AbD eingeführt, mit dem Ziel, bestehenden Evidenzlücken zu schließen und so eine bessere Datenbasis für die Nutzenbewertung zu erhalten. Die Datenerhebungen sind dabei als vergleichende Studien ohne Randomisierung durchzuführen. Unter Einhaltung bestimmter Qualitätsanforderungen können Studien auf Basis von Registerdaten diese Erkenntnislücke schließen.

- Bei der Planung von vergleichenden Studien ohne Randomisierung auf Basis versorgungsnaher Daten ist die Target-Trial-Emulation ein empfohlener Ansatz, um systematische (vermeidbare) Verzerrungen zu minimieren. Voraussetzung für eine optimale Emulation einer hypothetischen RCT mittels Beobachtungsdaten ist, dass die hierfür erforderlichen Daten im Registerdatensatz mit der notwendigen Vollständigkeit und Datentiefe enthalten sind. Eine hohe Datenqualität kann in der Breite nur dann erreicht werden, wenn die Datengenerierung und die Nutzbarmachung von Registerdaten praktikabel und mit möglichst wenig Aufwand möglich sind. Als förderlich wird der Aufbau und der Unterhalt einer dauerhaft verfügbaren (funktionsfähigen) Dateninfrastruktur eingeschätzt. Dies könnte durch pUs, die auf die Register zur Durchführung registerbasierter Studien zurückgreifen wollen, unterstützt werden.
- In vergleichenden Studien ohne Randomisierung zum Zweck des Vergleichs von Therapieeffekten ist für eine adäquate Confounderkontrolle eine systematische Identifikation der relevanten Confounder sowie eine Berücksichtigung dieser Confounder in der Auswertung notwendig. Die Confounderidentifikation nach dem Ansatz von Pufulete 2022 (über eine systematische Literaturrecherche und Einbindung von Klinikerinnen und Klinikern) wird als machbar eingeschätzt und stellt ein grundsätzlich sinnvolles Vorgehen dar. Vor der Relevanzbewertung der Confounder durch klinische Expertinnen und Experten ist zu empfehlen, eine intensive Verdichtung der identifizierten Confounder vorzunehmen.
- In den Konzepten für eine AbD wird abgeschätzt, ob innerhalb eines akzeptablen Zeitraums ausreichend Patientinnen und Patienten eingeschlossen werden können, um mit der AbD grundsätzlich aussagekräftige Ergebnisse für eine Nutzenbewertung generieren zu können. In der Regel liegen nur unsichere Informationen für diese Abschätzung vor. Wenn hinreichende Informationen zur Intervention und dem Komparator vorliegen, wird zu diesem Zwecke eine orientierende Fallzahlschätzung durchgeführt. Fehlen notwendige Informationen für die Annahmen, die einer orientierenden Fallzahlschätzung zugrunde liegen, erfolgt eine orientierende Betrachtung von Fallzahlszenarien, um mögliche aufdeckbare Effekte zu zeigen. Beide Herangehensweisen folgen dabei dem gleichen Grundsatz (sie fußen auf identischen

Parametern) und sie unterscheiden sich lediglich in dem Parameter, der mithilfe der übrigen Parameter geschätzt werden soll.

- Für eine langfristige Datenerhebung (und Begleitung von Patientinnen und Patienten) in der Routineversorgung sind Anreize erforderlich, die den Aufwand der Datengenerierung kompensieren und sowohl die Zentren als auch die Patientinnen und Patienten zur möglichst vollständigen Datenerhebung motivieren.
- Zur Sicherstellung, dass Aufwand und Nutzen bei PRO-Erhebungen in einem angemessenen Verhältnis stehen, werden digitale Befragungen (z. B. per App oder per Patientenportal) empfohlen. Auf diese Weise ist eine niedrigschwellige PRO-Erhebung unabhängig von Arztbesuchen möglich. Digitale Tools werden in Registern und in der Routineversorgung bereits erfolgreich zur Datenerhebung eingesetzt und ermöglichen, dass PRO-Daten für Forschungszwecke barrierearm zur Verfügung gestellt werden können.
- Eine Herausforderung bei der Auswertung versorgungsnaher Daten ohne Randomisierung stellt die Festlegung des Startzeitpunkts der Beobachtung (Index-Datum [t_0]) dar. Wird in beiden Behandlungsgruppen zum Beobachtungsstart (t_0) mit einer neuen Therapie begonnen (Umstellungsindikation), kann in einer Studie das ACNU-Design verwendet werden. Dabei entspricht t_0 idealerweise dem Tag der Therapieentscheidung. In einer Therapiesituation, in der Patientinnen und Patienten in der Kontrollgruppe die Behandlung mit einem etablierten therapeutischen Standard fortführen, stellt das PNU-Design eine geeignete Alternative dar.
- Unabhängig von der Wahl des Designs und der Auswertungsstrategien sollte grundsätzlich immer primär das Ergebnis der ITT-Analyse (entsprechend der Treatment-Policy-Strategie) gezeigt werden. Dies erfordert, dass alle Patientinnen und Patienten (unabhängig von einem Behandlungswechsel im Verlauf der Beobachtung) gemäß ihrer ursprünglichen Gruppenzuteilung ausgewertet werden. Dem verzerrenden Einfluss von Behandlungswechseln im Verlauf der AbD kann mit Sensitivitätsanalysen begegnet werden.
- Eine häufig angewendete Methode zur Berücksichtigung von Confoundern in vergleichenden Studien ohne Randomisierung auf Basis von Registern stellt die Auswertung auf Basis von Propensity Scores dar. Da Modelle zur Schätzung von Propensity Scores mittels logistischer Regression für alle potenziell relevanten Confounder vollständige Daten erfordern, sind statistische Verfahren zum Umgang mit fehlenden Werten notwendig. Als etablierte Methode ist die multiple Imputation mittels Chained Equations (MICE) anerkannt. Propensity-Score-Verfahren sind auch in kleinen Patientenkollektiven unter bestimmten Bedingungen anwendbar und können zu interpretierbaren Ergebnissen führen. Die in den identifizierten Simulationsstudien untersuchten Szenarien bilden die Situationen, wie sie in den laufenden AbD-Verfahren

absehbar auftreten werden, in Teilen jedoch nicht ab. Zum jetzigen Zeitpunkt bleibt daher unklar, in welchen Fällen (und unter welchen Voraussetzungen) Propensity-Score-Analysen in einer AbD zu interpretierbaren Ergebnissen führen werden.

7 Literatur

1. Stürmer T, Wang T, Golightly YM et al. Methodological considerations when analysing and interpreting real-world data. Rheumatology 2020; 59(1): 14-25. https://doi.org/10.1093/rheumatology/kez320.

2. Greifer N, Stuart EA. Choosing the Causal Estimand for Propensity Score Analysis of Observational Studies [online]. 2021 [Zugriff: 18.06.2025]. URL: https://arxiv.org/pdf/2106.10577.

3. Rhodes AE, Lin E, Streiner DL. Confronting the Confounders: The Meaning, Detection, and Treatment of Confounders in Research. Can J Psychiatry 1999; 44(2): 175-179. https://doi.org/10.1177/070674379904400209.

4. Arbogast PG, Ray WA. Performance of disease risk scores, propensity scores, and traditional multivariable outcome regression in the presence of multiple confounders. Am J Epidemiol 2011; 174(5): 613-620. https://doi.org/10.1093/aje/kwr143.

5. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. BMJ 2019; 367: l5657. https://doi.org/10.1136/bmj.l5657.

6. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials [online]. 2020 [Zugriff: 18.03.2025]. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf.

7. Lévesque LE, Hanley JA, Kezouh A et al. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 2010; 340: b5087. https://doi.org/10.1136/bmj.b5087.

8. Suissa S. Immortal Time Bias in Pharmacoepidemiology. Am J Epidemiol 2008; 167(4): 492-499. https://doi.org/10.1093/aje/kwm324.

9. Hernán MA, Sauer BC, Hernández-Díaz S et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016; 79: 70-75. https://doi.org/10.1016/j.jclinepi.2016.04.014.

10. Kuss O, Blettner M, Börgermann J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. Dtsch Arztebl Int 2016; 113: 597-603. https://doi.org/10.3238/arztebl.2016.0597.

11. Puhr R, Heinze G, Nold M et al. Firth's logistic regression with rare events: accurate effect estimates and predictions? Stat Med 2017; 36(14): 2302-2317. https://doi.org/10.1002/sim.7273.

12. Pottegård A, Friis S, Stürmer T et al. Considerations for pharmacoepidemiological studies of drug-cancer associations. Basic Clin Pharmacol Toxicol 2018; 122(5): 451-459. https://doi.org/10.1111/bcpt.12946.

13. Suissa S, Dell'Aniello S. Time-related biases in pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2020; 29(9): 1101-1110. https://doi.org/10.1002/pds.5083.

14. Yang CY, Kuo S, Lai ECC et al. Three-step matching algorithm to enhance between-group comparability and minimize confounding in comparative effectiveness studies. Nature 2022; 12: 214. https://doi.org/10.1038/s41598-021-04014-z.

15. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Hoboken: Wiley; 2020.

16. Mack C, Su Z, Westreich D. Managing Missing Data in Patient Registries; White Paper; Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [online]. 2018 [Zugriff: 27.06.2025]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/wysiwyg/missing-data-registries-guide-3rd-ed-addendum-white-paper.pdf.

17. Azur MJ, Stuart EA, Frangakis C et al. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011; 20(1): 40-49. https://doi.org/10.1002/mpr.329.

18. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzepte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V; Rapid Report [online]. 2020 [Zugriff: 07.03.2025]. URL: https://www.iqwig.de/download/a19-43_versorgungsnahe-daten-zum-zwecke-der-nutzenbewertung_rapid-report_v1-1.pdf.

19. Suissa S, Moodie EE, Dell'Aniello S. Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. Pharmacoepidemiol Drug Saf 2017; 26(4): 459-468. https://doi.org/10.1002/pds.4107.

20. Sterne JAC, Hernán MA, McAleenan A et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.5; Chapter 25: Assessing risk of bias in a non-randomized study [last updated October 2019] [online]. 2024 [Zugriff: 13.03.2025]. URL: https://training.cochrane.org/handbook/current/chapter-25.

21. Hernán MA, Wang W, Leaf DE. Target Trial Emulation: A Framework for Causal Inference From Observational Data. JAMA 2022; 328(24): 2446-2447. https://doi.org/10.1001/jama.2022.21383.

22. Suissa S, Azoulay L. Metformin and the risk of cancer; Time-relared biases in observational studies. Diabetes Care 2012; 35(12): 266-273. https://doi.org/10.2337/dc12-0788.

23. Hartung J, Elpert B, Klösener KH. Statistik; Lehr- und Handbuch der angewandten Statistik. München: Oldenbourg Wissenschaftsverlag; 2009.

24. Hashimoto Y, Yasunaga H. Theory and practice of propensity score analysis. Ann Clin Epidemiol 2022; 4(4): 101-109. https://doi.org/10.37737/ace.22013.

25. SGB V Handbuch: Sozialgesetzbuch V; Krankenversicherung. Altötting: KKF-Verlag; 2020.

26. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses [online]. URL: https://www.g-ba.de/richtlinien/42/.

27. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Systematische Confounderidentifikation in der Indikation schubförmiger remittierende multiple Sklerose (RRMS); Arbeitspapier [online]. 2025 [Zugriff: 27.06.2025]. URL: https://doi.org/10.60584/GA23-02.

28. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016; 183(8): 758-764. https://doi.org/10.1093/aje/kwv254.

29. Matthews AA, Danaei G, Islam N et al. Target trial emulation: applying principles of randomised trials to observational studies. BMJ 2022; 378: e071108. https://doi.org/10.1136/bmj-2022-071108.

30. Lodi S, Phillips A, Lundgren J et al. Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples With Apples. Am J Epidemiol 2019; 188(8): 1569-1577. https://doi.org/10.1093/aje/kwz100.

31. Hoffmann F, Kaiser T, Apfelbacher C et al. Versorgungsnahe Daten zur Evaluation von Interventionseffekten: Teil 2 des Manuals. Gesundheitswesen 2021; 83: 470-480. https://doi.org/10.1055/a-1484-7235.

32. Desai RJ, Wang SV, Kattinakere Sreedhara S et al. Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center. BMJ 2023; 384: e076460. https://doi.org/10.1136/bmj-2023-076460.

33. Mathes T. Anforderungen an die Daten für die Target-Trial-Emulation: Eine Diskussion unter Betrachtung von Patientenregistern. GMS Med Inform Biom Epidemiol 2024; 20: Doc03. https://doi.org/10.3205/mibe000259.

34. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden; Version 7.0 [online]. 2023 [Zugriff: 07.03.2025]. URL: https://www.iqwig.de/methoden/allgemeine-methoden_version-7-0.pdf.

35. Zhu Y, Hubbard RA, Chubak J et al. Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches. Pharmacoepidemiol Drug Saf 2021; 30(11): 1471-1485. https://doi.org/10.1002/pds.5338.

36. Higgins JPT, Savović J, Page MJ et al. Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 205-228.

37. Braitmaier M, Didelez V. Emulierung von "target trials" mit Real-world-Daten; Ein allgemeines Prinzip, um den Herausforderungen von Beobachtungsdaten zu begegnen. Prävention und Gesundheitsförderung 2022. https://doi.org/10.1007/s11553-022-00967-9.

38. Brookhart MA. Counterpoint: The Treatment Decision Design. Am J Epidemiol 2015; 182(10): 840-845. https://doi.org/10.1093/aje/kwv214.

39. Pufulete M, Mahadevan K, Johnson TW et al. Confounders and co-interventions identified in non-randomized studies of interventions. J Clin Epidemiol 2022; 148: 115-123. https://doi.org/10.1016/j.jclinepi.2022.03.018.

40. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990; 9(12): 1447-1454. https://doi.org/10.1002/sim.4780091208.

41. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzept für eine anwendungsbegleitende Datenerhebung – Brexucabtagen autoleucel; Rapid Report [online]. 2022 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/download/a21-130_anwendungsbegleitende-datenerhebung-brexucabtagen-autoleucel_rapid-report_v1-0.pdf.

42. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzept für eine anwendungsbegleitende Datenerhebung – Valoctocogen Roxaparvovec; Rapid Report [online]. 2022 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/download/a22-20_anwendungsbegleitende-datenerhebung-valoctocogen-roxaparvovec_rapid-report_v1-0.pdf.

43. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Etranacogen Dezaparvovec (Hämophilie B) – Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2023 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/download/a22-83_etranacogen-dezaparvovec-haemophilie-b_abd-konzept_v1-1.pdf.

44. Walter SD. Matched case-control studies with a variable number of controls per case. Appl Statist 1980; 29(2): 172-179. https://doi.org/10.2307/2986303.

45. Breslow NE, Lubin JH, Marek P et al. Multiplicative models and cohort analysis. J Am Stat Assoc 1983; 78(381): 1-12. https://doi.org/10.1080/01621459.1983.10477915.

46. Taylor JM. Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. Stat Med 1986; 5(1): 29-36. https://doi.org/10.1002/sim.4780050106.

47. Glasziou P, Chalmers I, Rawlins M et al. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334(7589): 349-351. https://doi.org/10.1136/bmj.39070.527986.68.

48. Chow SC, Wang H, Shao J. Sample Size Calculations in Clinical Research. Boca Raton: Taylor & Francis; 2003.

49. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Epcoritamab (DLBCL) – Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2025 [Zugriff: 21.07.2025]. URL: https://doi.org/10.60584/A25-07.

50. AbbVie Deutschland. Epcoritamab (Tepkinly); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Zugriff: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/1004/#dossier.

51. Duell J, Abrisqueta P, Andre M et al. Tafasitamab for patients with relapsed or refractory diffuse large B-cell lymphoma: final 5-year efficacy and safety findings in the phase II L-MIND study. Haematologica 2024; 109(2): 553-566. https://doi.org/10.3324/haematol.2023.283480.

52. Roche Pharma. Polatuzumab Vedotin (POLIVY); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Zugriff: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/1041/#dossier.

53. Roche Pharma. Glofitamab (Columvi); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Zugriff: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/978/#dossier.

54. Swedish Orphan Biovitrum. Loncastuximab tesirin (Zynlonta); Dossier zur Nutzenbewertung gemäß § 35a SGB V [online]. 2023 [Zugriff: 18.07.2025]. URL: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/955/#dossier.

55. European Medicines Agency. Ordspono; Assessment report [online]. 2024 [Zugriff: 18.07.2025]. URL: https://www.ema.europa.eu/en/documents/assessment-report/ordspono-epar-public-assessment-report_en.pdf.

56. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Risdiplam: Prüfung des Studienprotokolls (Version 3.0) und des statistischen Analyseplans (Version 3.0); 3. Addendum zum Projekt A21-131 [online]. 2024 [Zugriff: 07.03.2025]. URL: https://doi.org/10.60584/A24-57.

57. Wakabayashi R, Hirano T, Laurent T et al. Impact of "time zero" of Follow-Up Settings in a Comparative Effectiveness Study Using Real-World Data with a Non-user Comparator: Comparison of Six Different Settings. Drugs Real World Outcomes 2023; 10(1): 107-117. https://doi.org/10.1007/s40801-022-00343-1.

58. Webster-Clark M, Mavros P, Garry EM et al. Alternative analytic and matching approaches for the prevalent new-user design: A simulation study. Pharmacoepidemiol Drug Saf 2022; 31(7): 796-803. https://doi.org/10.1002/pds.5446.

59. Webster-Clark M, Ross RK, Lund JL. Initiator Types and the Causal Question of the Prevalent New-User Design: A Simulation Study. Am J Epidemiol 2021; 190(7): 1341-1348. https://doi.org/10.1093/aje/kwaa283.

60. Her QL, Rouette J, Young JC et al. Core Concepts in Pharmacoepidemiology: New-User Designs. Pharmacoepidemiol Drug Saf 2024; 33: e70048. https://doi.org/10.1002/pds.70048.

61. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Etranacogen Dezaparvovec: Prüfung des Studienprotokolls (Version 1.0) und des statistischen Analyseplans (Version 1.0); 2. Addendum zum Projekt A22-83 [online]. 2023 [Zugriff: 10.06.2025]. URL: https://doi.org/10.60584/A23-99.

62. Wintzell V, Svanström H, Pasternak B. Selection of Comparator Group in Observational Drug Safety Studies; Alternatives to the Active Comparator New User Design. Epidemiology 2022; 33(5): 707-714. https://doi.org/10.1097/EDE.0000000000001521.

63. Tazare J, Gibbons DC, Bokern M et al. Prevalent new user designs: A literature review of current implementation practice. Pharmacoepidemiol Drug Saf 2023; 32(11): 1252-1260. https://doi.org/10.1002/pds.5656.

64. Willems S, Schat A, van Noorden MS et al. Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator. Stat Methods Med Res 2018; 27(2): 323-335. https://doi.org/10.1177/0962280216628900.

65. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Treatment Switching in onkologischen Studien; Arbeitspapier [online]. 2018 [Zugriff: 11.07.2023]. URL: https://www.iqwig.de/download/ga14-04_treatment-switching-in-onkologischen-studien_arbeitspapier_v1-0.pdf.

66. Latimer NR, Abrams KR, Lambert PC et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials – an economic evaluation context: methods, limitations, and recommendations. Med Decis Making 2014; 34(3): 387-402. https://doi.org/10.1177/0272989X13520192.

67. Gorrod HB, Latimer NR, Abrams KR. NICE DSU Technical Support Document 24: Adjusting survival time estimates in the presence of treatment switching: an update to TSD 16; Report by the Decision Support Unit [online]. 2024 [Zugriff: 18.03.2025]. URL: https://www.sheffield.ac.uk/media/65536/download?attachment.

68. Aslop J, Latimer N, Olson M et al. Estimating comparative effectiveness when patients are switching treatments: a real-world challenge. Value & Outcomes Spotlight 2020: 31-33.

69. World Federation of Hemophilia, International Society on Thrombosis and Haemostasis, European Association for Haemophilia and Allied Disorders. Critical Juncture in the Development of Hemophilia Therapies; A Statement and Call-to-Action from the International Society on Thrombosis and Haemostasis (ISTH), European Association for Haemophilia and Allied Disorders (EAHAD), World Federation of Hemophilia (WFH) for the importance of continued research into and development of effective and safe hemophilia treatments [online]. 2025 [Zugriff: 11.06.2025]. URL: https://wfh.org/wp-content/uploads/2025/05/WFH-Statement-at-WHO-25th-EML-Expert-Committe-Meeting.pdf.

70. CSL Behring. Routine Practice Data Collection and Evaluation of etranacogene dezaparvovec (Hemgenix) and prophylactic factor IX (FIX) replacement in severe and moderately severe haemophilia B without a history of FIX inhibitors; a prospective, non-interventional study mandated by GBA; Study Protocol; Version 3.0 [online]. 2024 [Zugriff: 02.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10667/2024-07-18_AM-RL-XII_Etranacogen-Dezaparvovec_2022-AbD-005_Studienunterlagen.pdf.

71. CSL Behring. Routine Practice Data Collection and Evaluation of etranacogene dezaparvovec (Hemgenix) and prophylactic factor IX (FIX) replacement in severe and moderately severe haemophilia B without a history of FIX inhibitors; a prospective, non-interventional study mandated by GBA; Statistical Analysis Plan (SAP); Version 3.0 [online]. 2024 [Zugriff: 02.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10667/2024-07-18_AM-RL-XII_Etranacogen-Dezaparvovec_2022-AbD-005_Studienunterlagen.pdf.

72. Gilead Sciences. Real world effectiveness and safety of brexucabtagene autoleucel versus patient-individual therapy in relapsed/refractory mantle cell lymphoma: A European Mantle Cell Lymphoma Network (EMCL) registry study mandated by the G-BA; Study Project Plan; Project Plan Number: RW-X19-2206; Version 4.0 [online]. 2025 [Zugriff: 20.06.2025]. URL: https://www.g-ba.de/downloads/40-268-11585/2025-06-18_AM-RL-XII_BrexCel_2021-AbD-008_Feststellung_Studienunterlagen.pdf.

73. Novartis. Routine data collection and evaluations of onasemnogene abeparvovec in Germany; Study Protocol; Protocol Number: COAV101A1DE01; Version 4.01 [online]. 2024 [Zugriff: 23.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10961/2024-06-06_AM-RL-XII_Onasemnogen-Abeparvovec_2020-AbD-001_Ueberpruefung-SP-SAP_Studienunterlagen.pdf.

74. Roche Pharma. Evaluation of a real world data collection for the reassessment of the additional benefit or evrysdi (risdiplam); Protocol; Protocol Number: ML44661; Version 3.0 [online]. 2024 [Zugriff: 23.06.2025]. URL: https://www.g-ba.de/downloads/40-268-10838/2024-09-19_AM-RL-XII_Risdiplam_AbD-004_Feststellung_Studienunterlagen.pdf.

75. Di Staso R, Casadei B, Locke FL et al. Is CAR T a drug or a therapeutic pathway? Intention to treat versus per protocol analysis of real world studies of CAR-T cell therapy in relapsed refractory diffuse large B cell lymphoma. Blood Cancer J 2024; 14: 197. https://doi.org/10.1038/s41408-024-01183-8.

76. Bundesministerium für Gesundheit. Verordnung über die Nutzenbewertung von Arzneimitteln nach § 35a Absatz 1 SGB V für Erstattungsvereinbarungen nach § 130b SGB V (Arzneimittel-Nutzenbewertungsverordnung - AM-NutzenV) [online]. 2023 [Zugriff: 09.10.2023]. URL: https://www.gesetze-im-internet.de/am-nutzenv/AM-NutzenV.pdf.

77. de Vet H, Terwee CB, Mokkink LB et al. Measurement in Medicine; A Practical Guide. Cambridge: Cambridge University Press; 2011.

78. Gilead Sciences. Real world effectiveness and safety of brexucabtagene autoleucel versus patient-individual therapy in relapsed/refractory mantle cell lymphoma: A European Mantle Cell Lymphoma Network (EMCL) registry study mandated by the G-BA; Study Project Plan; Project Plan Number: RW-X19-2206; Version 3.0 [online]. 2023 [Zugriff: 10.06.2025]. URL: https://www.g-ba.de/downloads/40-268-9958/2023-11-16_AM-RL-XII_BrexCel_2021-AbD-008_finale-Ueberpruefung_Studienunterlagen.pdf.

79. Singer S, Bayer O, Schranz M et al. "Patient-reported outcomes" in medizinischen Registern; Erfahrungen aus einer Anwendungsbegleitenden Datenerhebung. Onkologie 2024; 30: 304-311. https://doi.org/10.1007/s00761-024-01494-2.

80. Deutsches Hämophilieregister. Gesamtdatensatz DHR 2.0 [online]. 2024 [Zugriff: 23.06.2025]. URL: https://www.pei.de/SharedDocs/Downloads/DE/regulation/meldung/dhr-deutsches-haemophilieregister/dhr-20-datensatz.pdf?__blob=publicationFile&v=20.

81. Deutsches Hämophilieregister. Handbuch; Version 2.7 [online]. 2025 [Zugriff: 23.06.2025]. URL: https://www.pei.de/SharedDocs/Downloads/DE/regulation/meldung/dhr-deutsches-haemophilieregister/dhr-20-handbuch.pdf?__blob=publicationFile&v=15.

82. Wang K, Eftang CN, Jakobsen RB et al. Review of response rates over time in registry-based studies using patient-reported outcome measures. BMJ Open 2020; 10(8): e030808. https://doi.org/10.1136/bmjopen-2019-030808.

83. Amelung V, Arnold M, Altendorf M et al. Patient-Reported Outcomes (PROs) in der Routineversorgung bei Krebserkrankungen; Anwendungsbeispiele aus ausgewählten Ländern [online]. 2024 [Zugriff: 30.05.2025]. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Studie_ST-G_QS_PROM__Krebserkrankung.pdf.

84. Food and Drug Administration. Core Patient-Reported Outcomes in Cancer Clinical Trials; Guidance for Industry [online]. 2024 [Zugriff: 10.06.2025]. URL: https://www.fda.gov/media/149994/download.

85. Di Maio M, Basch E, Denis F et al. The role of patient-reported outcome measures in the continuum of cancer clinical care: ESMO Clinical Practice Guideline. Ann Oncol 2022; 33(9): 878-892. https://doi.org/10.1016/j.annonc.2022.04.007.

86. Nielsen LK, King M, Möller S et al. Strategies to improve patient-reported outcome completion rates in longitudinal studies. Qual Life Res 2020; 29(2): 335-346. https://doi.org/10.1007/s11136-019-02304-8.

87. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anwendungsbegleitende Datenerhebung zu Etranacogen Dezaparvovec: Prüfung des Studienprotokolls (Version 3.0) und des statistischen Analyseplans (Version 3.0); 3. Addendum zum Projekt A22-83 [online]. 2024 [Zugriff: 22.07.2024]. URL: https://doi.org/10.60584/A24-38.

88. Fürchtenicht A, Wehling H, Grote Westrick M et al. Patient-Reported Outcomes; Wie die Patientenperspektive die Versorgung transformieren wird [online]. 2023 [Zugriff: 30.05.2025]. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/Patient-Reported_Outcomes_Bertelsmann_Stiftung_2023.pdf.

89. Steinbeck V, Ernst SC, Pross C. Patient-Reported Outcome Measures (PROMs): ein internationaler Vergleich; Herausforderungen und Erfolgsstrategien für die Umsetzung von PROMs in Deutschland [online]. 2021 [Zugriff: 30.05.2025]. URL: https://www.bertelsmann-stiftung.de/fileadmin/files/user_upload/BSt_PROMs-Implementierung_final.pdf.

90. Lindberg-Scharf P, Steinger B, Koller M et al. Long-term improvement of quality of life in patients with breast cancer: supporting patient-physician communication by an electronic tool for inpatient and outpatient care. Support Care Cancer 2021; 29(12): 7865-7875. https://doi.org/10.1007/s00520-021-06270-1.

91. Klinkhammer-Schalke M, Steinger B, Koller M et al. Diagnosing deficits in quality of life and providing tailored therapeutic options: Results of a randomised trial in 220 patients with colorectal cancer. Eur J Cancer 2020; 130: 102-113. https://doi.org/10.1016/j.ejca.2020.01.025.

92. Absolom K, Warrington L, Hudson E et al. Phase III Randomized Controlled Trial of eRAPID eHealth Intervention Drug Chemotherapy. J Clin Oncol 2021; 39: 734-747. https://doi.org/10.1200/jco.20.02015.

93. Basch E, Deal AM, Dueck AC et al. Overall Survival Results of a Trial Assessing Patient-Reported Outcomes for Symptom Monitoring During Routine Cancer Treatment. JAMA 2017; 318(2): 197-198. https://doi.org/10.1001/jama.2017.7156.

94. Basch E, Schrag D, Henson S et al. Effect of Electronic Symptom Monitoring on Patient Reported Outcomes Among Patients With Metastatic Cancer; A Randomized Clinical Trial. JAMA 2022; 327(24): 2413-2422. https://doi.org/10.1001/jama.2022.9265.

95. Basch E, Schrag D, Jansen J et al. Symptom monitoring with electronic patient-reported outcomes during cancer treatment: final results of the PRO-TECT cluster-randomized trial. Nat Med 2025; 31(4): 1225-1232. https://doi.org/10.1038/s41591-025-03507-y.

96. Denis F, Basch E, Septans AL et al. Two-Year Survival Comparing Web-Based Symptom Monitoring vs Routine Surveillance Following Treatment for Lung Cancer. JAMA 2019; 321(3): 306-307. https://doi.org/10.1001/jama.2018.18085.

97. Kowalski C, Wesselmann S, van Oorschot B et al. „Patient-reported outcomes“ in der onkologischen Versorgung – aktuelle Anwendungsfelder und Initiativen der Deutschen Krebsgesellschaft. Onkologie 2024; 30: 411-420. https://doi.org/10.1007/s00761-024-01508-z.

98. Aiyegbusi OL, Cruz Rivera S, Roydhouse J et al. Recommendations to address respondent burden associated with patient-reported outcome assessment. Nat Med 2024; 30(3): 650-659. https://doi.org/10.1038/s41591-024-02827-9.

99. Winter LM, Sztankay M, Giesinger JM et al. Manual for the use of EORTC measures in daily clinical practice [online]. 2016 [Zugriff: 10.06.2025]. URL: https://www.eortc.org/app/uploads/sites/2/2018/02/EORTC_QLQ_Clinical_Practice_User_Manual-1.0.pdf.

100. Calvert MJ, O'Connor DJ, Basch EM. Harnessing the patient voice in real-world evidence: the essential role of patient-reported outcomes. Nat Rev Drug Discov 2019; 18(10): 731-732. https://doi.org/10.1038/ d41573-019-00088-7.

101. Al-Antary N, Tam S, Alzouhayli S et al. Interventions influencing patient-reported outcomes (PROs) response rates in cancer: a scoping review. J Cancer Surviv 2025. https://doi.org/10.1007/s11764-025-01801-9.

102. Meirte J, Hellemans N, Anthonissen M et al. Benefits and Disadvantages of Electronic Patient-reported Outcome Measures: Systematic Review. JMIR Perioper Med 2020. https://doi.org/10.2196/15588.

103. Korngut L, MacKean G, Casselman L et al. Perspectives on neurological patient registries: a literature review and focus group study. BMC Med Res Methodol 2013; 13: 135. https://doi.org/10.1186/1471-2288-13-135.

104. Niemeyer A, Semler SC, Veit C et al. Gutachten zur Weiterentwicklung medizinischer Register zur Verbesserung der Dateneinspeisung und -anschlussfähigkeit [online]. 2021 [Zugriff: 30.05.2025]. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/5_Publikationen/Gesundheit/Berichte/REG-GUT-2021_Registergutachten_BQS-TMF-Gutachtenteam_2021-10-29.pdf.

105. Food and Drug Administration. A Risk-Based Approach to Monitoring of Clinical Investigations; Questions and Answers; Guidance for Industry [online]. 2023 [Zugriff: 30.05.2025]. URL: https://www.fda.gov/media/121479/download.

106. European Medicines Agency. Guideline on registry-based studies [online]. 2021 [Zugriff: 30.05.2025]. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-registry-based-studies_en.pdf.

107. Lindner L, Weiß A, Reich A et al. Implementing an automated monitoring process in a digital, longitudinal observational cohort study. Arthritis Res Ther 2021; 23(1): 181. https://doi.org/10.1186/s13075-021-02563-2.

108. Gliklich RE, Leavy MB, Dreyer NA. Registries for Evaluating Patient Outcomes: A User's Guide; Fourth Edition [online]. 2020 [Zugriff: 27.06.2025]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/registries-evaluating-patient-outcomes-4th-edition.pdf.

109. Hageman IC, van Rooij I, de Blaauw I et al. A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance. Orphanet J Rare Dis 2023; 18: 106. https://doi.org/10.1186/s13023-023-02719-0.

110. Nährlich L, Burkhart M, Registry Working Group of the German CF Registry. Success of the German Cystic Fibrosis Registry. Pharmacoepidemiol Drug Saf 2025; 34(1): e70076. https://doi.org/10.1002/pds.70076.

111. Lissbrant IF, Erikkson MH, Lambe M et al. Set-up and preliminary results from the Patient-overview Prostate Cancer. Longitudinal registration of treatment of advanced prostate cancer in the National Prostate Cancer Register of Sweden. Scand J Urol 2020; 54(3): 227-234. https://doi.org/10.1080/21681805.2020.1756402.

112. Universitätsklinikum Heidelberg. INTEGRATE ATMP; Innovative Versorgungsstrukturen für neue Therapien; Informationsbroschüre [online]. 2023 [Zugriff: 05.06.2025]. URL: https://integrate-atmp.de/integrate-broschuere-2023-final.pdf.

113. Hansen S, Hildberg O, Suppli Ulrik C et al. The Danish severe asthma register: an electronic platform for severe asthma management and research. Eur Clin Respir J 2020; 8(1): 1842117. https://doi.org/10.1080/20018525.2020.1842117.

114. Oliver BJ, Nelson EC, Kerrigan CL. Turning Feed-forward and Feedback Processes on Patient-reported Data into Intelligent Action and Informed Decision-making: Case Studies and Principles. Med Care 2019: S31-S37. https://doi.org/10.1097/MLR.0000000000001088.

115. Lehmann J, Wintner LM, Sztankay M et al. Patient-reported outcomes and psycho-oncological screening in hematology: a practical example of routine electronic monitoring. Magazine of European Medical Oncology 2020; 13: 285-293. https://doi.org/10.1007/s12254-020-00628-7.

116. SCQM Foundation. SCQM-Datenbank [online]. [Zugriff: 23.06.2025]. URL: https://www.scqm.ch/de/medizinisches-fachpersonal/scqm-datenbank/.

117. Klein TL, Bender J, Bolton S et al. A rare partnership: patient community and industry collaboration to shape the impact of real-world evidence on the rare disease ecosystem. Orphanet J Rare Dis 2024. https://doi.org/10.1186/s13023-024-03262-2.

118. Stubbs E, Exley J, Wittenberg R et al. How to establish and sustain a disease registry: insights from a qualitative study of six disease registries in the UK. BMC Med Inform Decis Mak 2024; 24(1): 361. https://doi.org/10.1186/s12911-024-02775-x.

119. Gliklich RE, Dreyer NA, Leavy MB et al. Registries for Evaluating Patient Outcomes: A User's Guide; Addendum – 21st Century Patient Registries [online]. 2018 [Zugriff: 27.06.2025]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/wysiwyg/registries-guide-3rd-ed-addendum-research-2018-revised.pdf.

120. Nelson EC, Dixon-Woods M, Batalden PB et al. Patient focused registies can improve health, care, and science. BMJ 2016; 354: i3319. https://doi.org/10.1136/bmj.i3319.

121. Lee SB, Zak A, Iversen MD et al. Participation in Clinical Research Registries: A Focus Group Study Examining Views From Patients With Arthritis and Other Chronic Illnesses. Arthritis Care Res 2016; 68(7): 974-980. https://doi.org/10.1002/acr.22767.

122. European Medicines Agency. Patient Registries Workshop, 28 October 2016; Observations and recommendations arising from the workshop [online]. 2017 [Zugriff: 13.06.2025]. URL: https://www.ema.europa.eu/en/documents/report/report-patient-registries-workshop_en.pdf.

123. World Federation of Hemophilia. WFH Gene Therapy Registry; User guide for people with hemophilia [online]. 2022 [Zugriff: 23.06.2025]. URL: https://www1.wfh.org/publications/files/pdf-2214.pdf.

124. SCQM Foundation. mySCQM Webapplikation [online]. [Zugriff: 23.06.2025]. URL: https://www.scqm.ch/de/betroffene/myscqm-ihre-webapplikation/.

125. Osara Y, Coakley K, Devarajan A et al. Development of newborn screening connect (NBS connect): a self-reported patient registry and its role in improvement of care for patients with inherited metabolic disorders. Orphanet J Rare Dis 2017; 12(1): 132. https://doi.org/10.1186/s13023-017-0684-3.

126. Deutsches Rheuma-Forschungszentrum. ActiMON; Aktivitätsmonitoring bei Jugendlichen und jungen Erwachsenen mit Rheuma; Flyer [online]. [Zugriff: 05.06.2025]. URL: https://www.drfz.de/wp-content/uploads/2025/02/Flyer-Actimon_compressed3.pdf.

127. Leite WL, Aydın B, Cetin-Berber DD. Imputation of Missing Covariate Data Prior to Propensity Score Analysis: A Tutorial and Evaluation of the Robustness of Practical Approaches. Eval Rev 2021. https://doi.org/10.1177/0193841x211020245.

128. Leyrat C, Seaman SR, White IR et al. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res 2019; 28(1): 3-19. https://doi.org/10.1177/0962280217713032.

129. Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. Eur J Epidemiol 2019; 34(1): 23-36. https://doi.org/10.1007/s10654-018-0447-z.

130. de Vries B, Groenwold R. A comparison of two approaches to implementing propensity score methods following multiple imputation. Epidemiology, Biostatistics, and Public Health 2017; 14(4). https://doi.org/10.2427/12630.

131. Granger E, Sergeant JC, Lunt M. Avoiding pitfalls when combining multiple imputation and propensity scores. Stat Med 2019; 38(26): 5120-5132. https://doi.org/10.1002/sim.8355.

132. Lee Y, Leite W. A Comparison of Random Forest-Based Missing Imputation Methods for Covariates in Propensity Score Analysis. PsyArXiv Preprints 2023. https://doi.org/10.31234/osf.io/a47w6.

133. Ling A, Montez-Rath M, Mathur M et al. How to Apply Multiple Imputation in Propensity Score Matching with Partially Observed Confounders: A Simulation Study and Practical Recommendations. J Mod Appl Stat Methods 2020; 19(1). https://doi.org/10.22237/jmasm/1608552120.

134. Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res 2016; 25(1): 188-204. https://doi.org/10.1177/0962280212445945.

135. Ségalas C, Leyrat C, Carpenter JR et al. Propensity score matching after multiple imputation when a confounder has missing data. Stat Med 2023; 42(7): 1082-1095. https://doi.org/10.1002/sim.9658.

136. Vader DT, Mamtani R, Li Y et al. Inverse Probability of Treatment Weighting and Confounder Missingness in Electronic Health Record-based Analyses: A Comparison of Approaches Using Plasmode Simulation. Epidemiology 2023; 34(4): 520-530. https://doi.org/10.1097/ede.0000000000001618.

137. Yücel S, Ünal I. Balance diagnostics in propensity score analysis following multiple imputation: A new method. Pharm Stat 2024; 23(5): 763-777. https://doi.org/10.1002/pst.2389.

138. Eiset AH, Frydenberg M. Considerations for Using Multiple Imputation in Propensity Score-Weighted Analysis - A Tutorial with Applied Example. Clin Epidemiol 2022; 14: 835-847. https://doi.org/10.2147/CLEP.S354733.

139. Lee KJ, Tilling KM, Cornish RP et al. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol 2021; 134: 79-88. https://doi.org/10.1016/j.jclinepi.2021.01.008.

140. Bottigliengo D, Lorenzoni G, Ocagli H et al. Propensity Score Analysis with Partially Observed Baseline Covariates: A Practical Comparison of Methods for Handling Missing Data. Int J Environ Res Public Health 2021; 18(13): 6694. https://doi.org/10.3390/ijerph18136694.

141. Leyrat C, Seaman SS, White IR et al. Propensity score analysis with partially observed confounders: How should multiple imputation be used? arXiv (Cornell University) 2016. https://doi.org/10.48550/arxiv.1608.05606.

142. Nguyen T, Stuart E. Multiple imputation for propensity score analysis with covariates missing at random: some clarity on “within” and “across” methods. Am J Epidemiol 2024; 193(10): 1470-1476. https://doi.org/10.1093/aje/kwae105.

143. Moons KGM, Donders RART, Stijnen T et al. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol 2006; 59(10): 1092-1101. https://doi.org/10.1016/j.jclinepi.2006.01.009.

144. Carpenter JR, Kenward MG, Bartlett JW et al. Multiple Imputation and its Application. Chichester: Wiley; 2023.

145. Blake H, Leyrat C, Mansfield KE et al. Propensity scores using missingness pattern information: a practical guide. Stat Med 2020; 39(11): 1641-1657. https://doi.org/10.1002/sim.8503.

146. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. Biom J 2021; 63(5): 915-947. https://doi.org/10.1002/bimj.202000196.

147. Member State Coordination Group on Health Technology Assessment. Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons [online]. 2024 [Zugriff: 04.07.2025]. URL: https://health.ec.europa.eu/publications/practical-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons_en.

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

149. Stürmer T, Joshi M, Glynn RJ et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006; 59(5): 437-447. https://doi.org/10.1016/j.jclinepi.2005.07.004.

150. Pirracchio R, Resche-Rigon M, Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. BMC Med Res Methodol 2012; 12: 70. https://doi.org/10.1186/1471-2288-12-70.

151. Friedrich S, Friede T. Causal inference methods for small non-randomized studies: Methods and and application in COVID-19. Contemp Clin Trials 2020; 99: 106213. https://doi.org/10.1016/j.cct.2020.106213.

152. Austin PC. Bootstrap vs asymptotic variance estimation when using propensity score weighting with continuous and binary outcomes. Stat Med 2022; 41(22): 4426-4443. https://doi.org/10.1002/sim.9519.

153. Wilkinson JD, Mamas MA, Kontopantelis E. Logistic regression frequently outperformed propensity score methods, especially for large datasets: a simulation study. J Clin Epidemiol 2022; 152: 176-184. https://doi.org/10.1016/j.jclinepi.2022.09.009.

154. BioMarin. Comparative Effectiveness of Roctavian to Standard of Care Hemostatic Therapies in Germany Among People with Severe Hemophilia A: A Prospective Non-Interventional Study Utilizing Data Collected in the German Hemophilia Register; Study Protocol; Protocol Number: 270-603; Version 3.0 [online]. 2024 [Zugriff: 09.07.2025]. URL: https://www.g-ba.de/downloads/40-268-10669/2024-07-18_AM-RL-XII_Valoctocogen-Roxaparvovec_2020-AbD-002_Studienunterlagen.pdf.

155. Mukoviszidose – Bundesverband Cystische Fibrose. MUKOme – das Patientenportal des Deutschen Mukoviszidose-Registers [online]. [Zugriff: 10.06.2025]. URL: https://www.muko.info/was-wir-tun/register/mukome.

156. Iorio A, Grenier S, Page D et al. Reflections on the Canadian Bleeding Disorders Registry: Lessons Learned and Future Perspectives. Canadian Journal of Health Technologies 2022; 2(8): 1-15. https://doi.org/10.51731/cjht.2022.425.

157. World Federation of Hemophilia. Gene Therapy Registry; myGTR [online]. [Zugriff: 23.06.2025]. URL: https://wfh.org/research-and-data-collection/gene-therapy-registry/#myGTR.

158. Karolinska Institutet. Swedish Rheumatology Quality Registry; PER – Patient's Self Registration [online]. [Zugriff: 23.06.2025]. URL: https://srq.nu/en/about-per-patient/.

159. Andrén C. SP0141: Ask the patient – distinguising your symptoms. Ann Rheum Dis 2014; 73: 38. https://doi.org/10.1136/annrheumdis-2014-eular.6198.

Anhang A Beispiele für digitale PRO-Erhebungen in Registern

- Im Deutschen Mukoviszidose-Register bietet das Patientenportal „MUKOme“ Patientinnen und Patienten mit Mukoviszidose die Möglichkeit, die im Register enthaltenden medizinischen Gesundheitsdaten um PROs zu ergänzen und ihre im Register gespeicherten Daten auf digitalen Endgeräten (Computer, Tablet, Smartphone) in Echtzeit einzusehen und anteilig zu bearbeiten. Die Weboberfläche „MUKOme“ ist direkt mit dem Deutschen Mukoviszidose-Register verknüpft [110,155].
- Das INTEGRATE-ATMP-Projekt, das mit Mitteln des Innovationsfonds gefördert wird, verfolgt unter Federführung des Universitätsklinikums Heidelberg das Ziel, die Versorgung von Patientinnen und Patienten, die Arzneimittel für neuartige Therapien (ATMP) erhalten, zu optimieren und eine bestmögliche Behandlungsqualität sicherzustellen. Eine telemedizinische Plattform, die als App über das Smartphone und / oder Tablet zur Verfügung steht, ermöglicht es den Patientinnen und Patienten, PROs zu erfassen, Informationen zur Therapie zu erhalten, Termine zu vereinbaren und mit dem behandelnden Personal in Kontakt zu treten. Push-Nachrichten erinnern an Folgeuntersuchungen und auszufüllende Fragebogen. Die PRO-Daten werden im System erfasst, visuell aufbereitet und automatisch in das INTEGRATE-ATMP-Register überführt [112].
- In das kanadische Register für Blutgerinnungsstörungen (Canadian Bleeding Disorders Registry [CBDR]) ist das Patientenportal „myCBDR“, das aus einer App- und Webanwendung besteht, integriert. Zur regelmäßigen Erfassung von PRO-Daten besteht eine Verknüpfung mit der App „myPROBE“. Patientinnen und Patienten können über das Patientenportal in Echtzeit ihre Behandlung im häuslichen Umfeld (Anwendung von Blutgerinnungsfaktor-VIII/IX-Präparaten [Uhrzeit und Dosis]) und Blutungsereignisse erfassen. Zudem sind die individuellen Gesundheitsdaten (Behandlungsplan, Häufigkeit, Schwere und Lokalisation der Blutungen, Bedarfe für Behandlungsanpassungen für eine verbesserte Kommunikation im Rahmen der gemeinsamen Entscheidungsfindung) über das Portal einsehbar [156].
- Im Gentherapie-Register (Gene Therapy Registry [GTR]) der World Federation of Hemophilia (WFH) werden von Patienten mit Hämophilie, die eine Gentherapie erhalten haben, regelmäßig per Smartphone über die webbasierte App „myGTR“ PRO-Daten, Blutungsereignisse und Therapien (insofern erhalten) erfasst. Die Patienten werden über ihren bevorzugten Weg der Kontaktaufnahme (E-Mail oder Textnachricht) daran erinnert, die Fragebogen auszufüllen. Die Daten werden automatisiert an die WFH-GTR-Datenbank übermittelt und sind anschließend vom behandelnden Zentrum einsehbar. Zudem stehen den Patienten die Ausgabedaten in ihrer App zur Verfügung, sodass sie ihre Fortschritte im Zeitverlauf verfolgen können [123,157].

- Im schwedischen Rheumatologie-Qualitätsregister (Swedish Rheumatology Quality Register [SQR]) werden PROs im Zeitverlauf (jeweils vor einem Arztkontakt) von zu Hause aus oder vor Ort über das Portal „Patient's Self Registration" per Computer erfasst, automatisch in das SQR überführt und in einem benutzerfreundlichen Dashboard aufbereitet (Echtzeitauswertung und -präsentation). Diese Informationen fließen durch einen modularen Aufbau in die klinische Versorgung, die gemeinsame Entscheidungsfindung und das Selbstmanagement der Patientinnen und Patienten ein [114,158,159]. Das Patientenmodul trägt zudem zur Verbesserung von Versorgungsprozessen bei, indem es den Patientinnen und Patienten eine Echtzeitüberwachung ihrer Krankheitsaktivität ermöglicht. Dies unterstützt das frühzeitige Erkennen von ungünstigen Krankheitsverläufen und gibt Anstoß, bei Bedarf einen Behandlungstermin zu vereinbaren [114].
- Im Schweizer Register für entzündlich-rheumatische Erkrankungen (Swiss Clinical Quality Management in Rheumatic Diseases [SCQM]) können Patientinnen und Patienten über die Webapplikation „mySCQM" (nutzbar über Smartphone und / oder Computer), zusätzlich zu ihrer Medikamenteneinnahme, PRO-Daten in regelmäßigen Intervallen erfassen. Diese werden direkt in die SCQM-Datenbank eingespeist. Neben einer Beantwortung des Fragebogens vor jedem Arzttermin (Dauer: etwa 20 Minuten) ist 1-mal pro Tag eine Teilnahme an Kurzbefragungen (Dauer: etwa 3 Minuten) möglich. Durch diese lückenlose Erfassung (zwischen den Konsultationen) wird eine bessere Beschreibung des eigenen Krankheitsverlaufs und die Therapieoptimierung unterstützt. In „mySCQM" können die Patientinnen und Patienten ihre Krankheitsaktivität im Zeitverlauf visualisiert in Form eines Diagramms anhand verschiedener Werte (z. B. Schmerzen, Funktionseinschränkungen oder Schmerzmittelbedarf) einsehen [124]. Dem medizinischen Personal werden die Daten über das SCQM-Scoreboard zugänglich gemacht. Durch eine einfache Datennutzung mithilfe von Automatisierung und übersichtlicher Visualisierungen (u. a. automatisch berechnete Scores inklusive Ampelsystem) können die PRO-Daten für die Entscheidungsfindung und zur Arzt-Patienten-Kommunikation herangezogen werden [116].