

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

A horizontal decorative bar spanning the width of the page, composed of 18 small squares in various shades of blue and grey. The text 'RAPID REPORT' is centered in white on a dark blue background.

RAPID REPORT

Projekt: A25-13

Version: 1.0

Stand: 28.10.2025

IQWiG-Berichte – Nr. 2122

DOI: 10.60584/A25-13

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

DOI-URL

<https://doi.org/10.60584/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

ISSN: 1864-2500

Zitiervorschlag

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Wissenschaftliche Ausarbeitung ausgewählter Aspekte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V; Rapid Report [online]. 2025 [Zugriff: TT.MM.JJJJ]. URL: <https://doi.org/10.60584/A25-13>.

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 einer externen Sachverständigen erstellt.

Für die Inhalte des Berichts ist allein das IQWiG verantwortlich.

Externe Sachverständige, die wissenschaftliche Forschungsaufträge für das Institut bearbeiten, haben gemäß § 139b Abs. 3 Satz 2 Sozialgesetzbuch (SGB) Fünftes Buch (V) – 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 durch das speziell für die Beurteilung der Interessenkonflikte eingerichtete Gremium des Instituts bewertet. Die Selbstangaben der externen Sachverständigen zur Offenlegung von Beziehungen sind in Anhang C dargestellt. Es wurden keine Interessenkonflikte festgestellt, die die fachliche Unabhängigkeit im Hinblick auf eine Bearbeitung des vorliegenden Auftrags gefährden.

Externe Sachverständige

- Anne Regierer, Deutsches Rheuma-Forschungszentrum Berlin – Programmbereich Epidemiologie und Versorgungsforschung, Berlin

Das IQWiG dankt der externen Beteiligten für ihre Mitarbeit am Projekt.

Mitarbeiterinnen und Mitarbeiter des IQWiG

- Caroline Wöhl
- Lars Beckmann
- Katharina Hirsch
- Anne-Kathrin Petri
- Ulrike Seay
- Guido Skipka
- Volker Vervölgyi

Inhaltsverzeichnis

	Seite
Tabellenverzeichnis	vi
Abbildungsverzeichnis	vii
Abkürzungsverzeichnis.....	viii
Glossar	x
Kurzfassung	xvii
1 Hintergrund.....	1
2 Fragestellung.....	2
3 Projektverlauf.....	3
3.1 Zeitlicher Verlauf des Projekts	3
3.2 Spezifizierungen und Änderungen im Projektverlauf	3
4 Methoden	5
4.1 Sichtung der bisherigen Prüfungen von Studienprotokoll und statistischem Analyseplan	5
4.2 Vorschlag für die Anforderung an eine systematische Confounderidentifikation ...	5
4.3 Informationsbeschaffung.....	5
5 Bewertung ausgewählter Aspekte zur Generierung und Auswertung versorgungsnaher Daten im Kontext nicht randomisierter vergleichender Studien	6
5.1 Einleitende Anmerkungen zu den Anforderungen an vergleichende Studien ohne Randomisierung zum Zwecke der Nutzenbewertung	6
5.2 Confounderidentifikation und -auswahl.....	10
5.3 Schätzung notwendiger Fallzahlen vor dem Hintergrund unzureichender Vorabinformationen	13
5.3.1 Grundlagen der Fallzahlschätzung	14
5.3.2 Keine oder unzureichende Vorabinformationen für orientierende Schätzungen zum Umfang der anwendungsbegleitenden Datenerhebung	20
5.3.3 Entscheidungsfindung für oder gegen eine orientierende Fallzahlschätzung	26
5.4 Beobachtungsstart und Behandlungswechsel.....	29
5.4.1 Beobachtungsstart	29
5.4.2 Behandlungswechsel	34
5.5 Patientenberichtete Endpunkte	37
5.6 Fehlende Werte.....	41
5.6.1 Maßnahmen zur Vermeidung fehlender Werte	41

5.6.2	Methodischer Umgang mit fehlenden Werten im Kontext von Propensity-Score-Analysen	45
5.6.2.1	Identifikation und Diskussion von geeigneten statistischen Methoden	45
5.6.2.2	Interpretierbarkeit von Analysen mit fehlenden und / oder imputierten Werten	51
5.6.2.3	Anforderungen an die Darstellung	53
5.7	Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven	54
6	Fazit	60
7	Literatur	63
Anhang A	Beispiele für digitale PRO-Erhebungen in Registern.....	81
Anhang B	Würdigung der Anhörung zum vorläufigen Rapid Report	83
Anhang C	Offenlegung von Beziehungen der externen Sachverständigen	140

Tabellenverzeichnis

	Seite
Tabelle 1: Vergleich eines Studienprotokolls einer RCT und einer Studie auf Basis versorgungsnaher Daten mittels Target-Trial-Emulation.....	8
Tabelle 2: Muster für beobachtete Werte für N Personen im Endpunkt und in den Confoundern L1 bis L7	46
Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report..	84

Abbildungsverzeichnis

	Seite
Abbildung 1: Überlappung der Populationen (gemessen am Propensity Score)	xvi
Abbildung 2: Zeitliche Ausrichtung zentraler Elemente des Studiendesigns am Startzeitpunkt der Beobachtung (t_0)	10
Abbildung 3: Vorgeschlagenes 3-schrittiges Verfahren zur systematischen Confounderidentifikation	11
Abbildung 4: Notwendige Fallzahl (N) bei Festlegung der Parameter erwartete Effektgröße, Signifikanzniveau (Fehler 1. Art) und Power ($1 - \text{Fehler 2. Art}$)	15
Abbildung 5: Szenarien bei der orientierenden Fallzahlschätzung mit verschobener Nullhypothese ($RR_0 = 2$)	19
Abbildung 6: Aufdeckbare Effektgröße bei Festlegung der Parameter Fallzahl (N), Signifikanzniveau (Fehler 1. Art) und Power ($1 - \text{Fehler 2. Art}$)	23
Abbildung 7: Aufdeckbarer Effekt (Effektmaß: Relatives Risiko [RR_1]) mit verschobener Nullhypothese ($RR_0 = 2$)	23
Abbildung 8: Aufdeckbarer Effekt (Effektmaß: Hazard Ratio [HR_1]) in Abhängigkeit von der Stichprobengröße und den Ereignisanteilen in beiden Behandlungsgruppen (Verhältnis von Intervention zu Komparatoren von 1:1)	25
Abbildung 9: Herangehensweisen zur Bewertung der Machbarkeit einer AbD: (A) orientierende Fallzahlschätzung und (B) orientierende Betrachtung von Fallzahlszenarien.....	27

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 Krankenhausgesellschaft e. V.
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
EPAR	European Public Assessment Report
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 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 $1/\text{Propensity Score}$ und Patientinnen und Patienten aus der Kontrollgruppe das Gewicht $1/(1 - \text{Propensity Score})$. 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, für die multiple Imputation bei fehlenden Werten und für Effektschätzungen werden Maximum-Likelihood-Verfahren eingesetzt. Dazu 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. $1 - \text{Propensity Score}$) 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. Die Wahrscheinlichkeit für das Fehlen eines Wertes hängt – unter Berücksichtigung der beobachteten Werte einer Person – nicht vom fehlenden Wert selbst ab. Solche fehlenden Werte führen zu verzerrten Ergebnissen, insofern die Abhängigkeit in der Analyse unberücksichtigt bleibt. Um solche Abhängigkeiten in der Analyse berücksichtigen zu können, muss der Ausfallmechanismus klar ersichtlich mittels der beobachteten Werte erklärbar sein. Die sachgerechte Berücksichtigung in einer Analyse führt zu unverzerrten Schätzungen [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 zwar nicht zu verzerrten Ergebnissen, es verringert sich jedoch die Präzision der Effektschätzungen [15]. Ein Beispiel 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 fehlenden Wert selbst ab und diese Abhängigkeit verschwindet auch nicht durch Berücksichtigung der beobachteten Werte der Person. Solche fehlenden Werte führen zu verzerrten Ergebnissen. Diese Verzerrung kann nicht korrigiert werden, da der Ausfallmechanismus von unbeobachteten Werten abhängt. In diesem Fall kann nur mithilfe von Sensitivitätsanalysen auf Basis von nicht prüfbar Annahmen die Größe der Verzerrung untersucht werden [15]. Ein Beispiel hierfür: Die Wahrscheinlichkeit für das Fehlen von Werten 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 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 $\text{Propensity Score} / (1 - \text{Propensity Score})$. 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 Wahrscheinlichkeit das $(1 - \text{Alpha})$ -Konfidenzintervall einer Effektschätzung den wahren Wert enthält. Hierfür ist die Festlegung des Alpha-Fehlers notwendig, z. B. $\text{Alpha} = 0,05$. In Simulationsstudien wird die empirische Überdeckungswahrscheinlichkeit über den Anteil der Replikationen bestimmt, in denen das $(1 - \text{Alpha})$ -Konfidenzintervall den vorgegebenen Effekt enthält. Ein gutes statistisches Verfahren hat eine empirische Überdeckungswahrscheinlichkeit von $1 - \text{Alpha}$. 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.

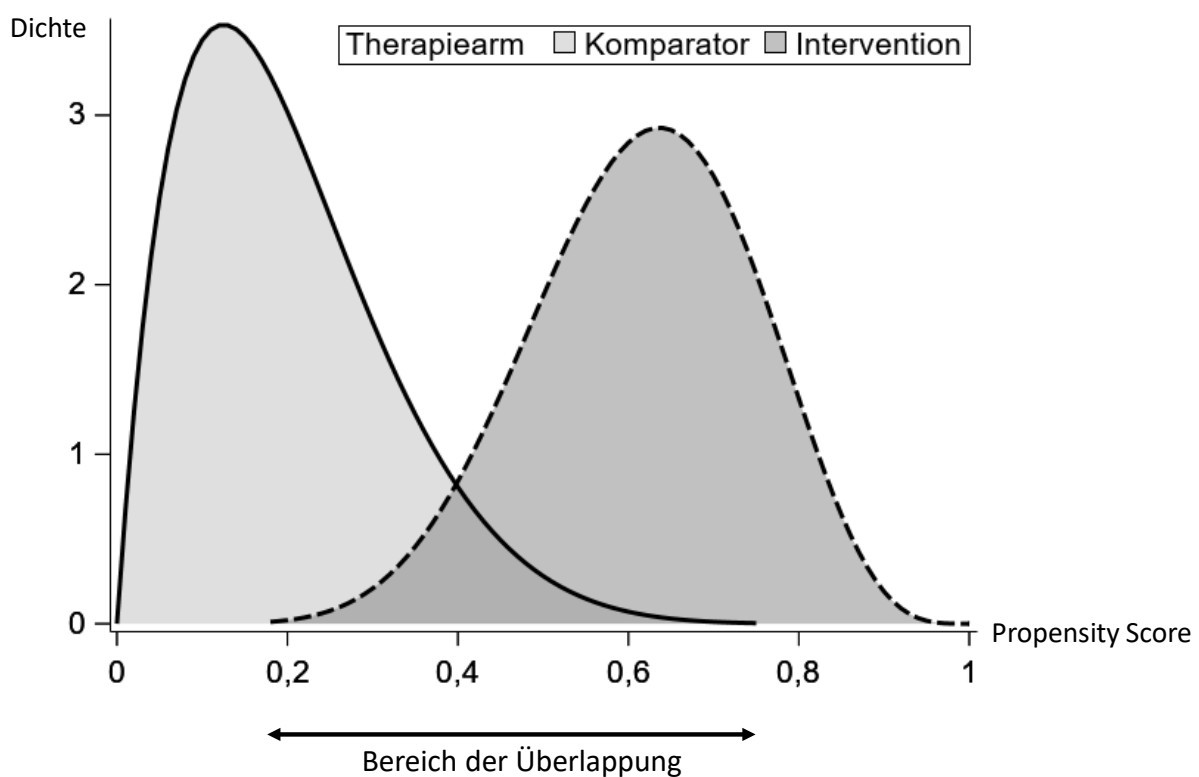


Abbildung 1: Überlappung der Populationen (gemessen am Propensity Score)

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
- Beobachtungswechsel
- 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, eine Website-Suche bei Zulassungsbehörden und anderen HTA-Agenturen sowie eine gezielte Websuche. Zusätzlich wurden ausgehend von den 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 potenzielle 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. Wichtig ist an dieser Stelle jedoch zwischen dem Aufwand im Vorfeld einer AbD (für die Identifikation und Verdichtung potenzieller Confounder) und dem Gesamtaufwand für die AbD (darunter fällt u. a. der Aufwand für die Datenerhebung und für die Dokumentation) zu differenzieren. Es kann zielführend sein, im Vorfeld einen höheren Aufwand zu investieren, um den Aufwand für die AbD in Gänze zu reduzieren. In einem nächsten Schritt soll im Rahmen der Erstellung des Arbeitspapiers GA25-02 geprüft werden, ob weitere Maßnahmen zur Aufwandsreduktion bei der systematischen Identifikation und Erhebung von Confoundern möglich sind. Dabei sind potenzielle Optionen zur Aufwandsreduktion dem damit einhergehenden Informationsverlust und dem potenziellen 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 Fallzahl-schätzung oder eine orientierende Betrachtung von Fallzahl-szenarien als Herangehensweise zurückgegriffen.

Parameter für die Fallzahlschätzung

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 Fallzahl-planungen 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.

Um aus einer nicht randomisierten vergleichenden Studie eine Aussage zum Nutzen oder zum Schaden des zu bewertenden Arzneimittels ableiten zu können, ist ein ausreichend großer beobachteter Effekt erforderlich, von dem angenommen werden kann, dass er nicht allein durch systematische Verzerrung hervorgerufen wird. Dies soll in der AbD statistisch in Form eines Tests auf eine verschobene Nullhypothese (H_0) umgesetzt werden. Dabei muss das Konfidenzintervall des in der AbD beobachteten Effekts vollständig jenseits eines konkreten Schwellenwertes liegen. Einen wissenschaftlichen Standard, wie groß die Schwelle für die verschobene Nullhypothese sein sollte, gibt es derzeit nicht. Als Schwellenwert wurde durch das IQWiG im Rapid Report A19-43 ein Bereich von 2 bis 5 für das relative Risiko (bzw. 0,2 bis 0,5 z. B. bei mortalitätssenkenden Interventionen) bezogen auf die untere (bzw. die obere) Konfidenzintervallgrenze festgelegt. Aufgrund methodischer Weiterentwicklungen erscheint es jedoch angezeigt, sich in einem nächsten Schritt (außerhalb der vorliegenden Ausarbeitung) einer Überprüfung der festgelegten Schwelle anzunehmen.

Keine oder unzureichende Vorabinformationen für eine orientierende Schätzung zum Umfang der anwendungsbegleitenden Datenerhebung

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.

Entscheidungsfindung für oder gegen eine orientierende Fallzahlschätzung

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 für Patientinnen und Patienten in der Kontrollgruppe zu reduzieren. Die Anwendung dieses Designs ist daran gebunden, dass geeignete Verfahren zur Berücksichtigung erhaltener Vorthherapie(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 ist. 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 Gruppenunterschiede 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, bestehende 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. Grundsätzlich kann es zielführend sein, im Vorfeld einer AbD einen höheren Aufwand zur Reduktion der Anzahl zu erhebender Confounder zu investieren, um den Gesamtaufwand für die AbD durch Aufwandseinsparungen bei der Datenerhebung und -auswertung zu reduzieren.
- 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 5.5)
- 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

3.1 Zeitlicher Verlauf des Projekts

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 sollten spezifische Themen, die sich aus den bisherigen Erfahrungen mit dem AbD-Verfahren ergeben haben, bearbeitet werden (siehe hierzu Kapitel 2).

In die Bearbeitung des Projekts wurde eine externe Sachverständige eingebunden.

Der vorläufige Rapid Report in der Version 1.0 vom 28.07.2025 wurde am 30.07.2025 auf der Website des IQWiG veröffentlicht und zur Anhörung gestellt. Bis zum 19.08.2025 konnten schriftliche Stellungnahmen eingereicht werden. Unklare Aspekte aus den schriftlichen Stellungnahmen zum vorläufigen Rapid Report wurden am 16.09.2025 in einer wissenschaftlichen Erörterung mit den Stellungnehmenden diskutiert. Die wesentlichen Argumente aus den Stellungnahmen werden in Anhang B des vorliegenden Rapid Reports gewürdigt.

Der vorliegende Rapid Report beinhaltet die Änderungen, die sich aus der Anhörung ergeben haben.

Im Anschluss an die Anhörung erstellte das IQWiG den vorliegenden Rapid Report, der 4 Wochen nach Übermittlung an den G-BA auf der Website des IQWiG veröffentlicht wird. Die zum vorläufigen Rapid Report eingegangenen Stellungnahmen und das Protokoll der wissenschaftlichen Erörterung werden in einem gesonderten Dokument („Dokumentation der Anhörung zum vorläufigen Rapid Report“) zeitgleich mit dem Rapid Report auf der Website des IQWiG bereitgestellt.

3.2 Spezifizierungen und Änderungen im Projektverlauf

Rapid Report im Vergleich zum vorläufigen Rapid Report

Neben redaktionellen Änderungen ergaben sich folgende Spezifizierungen oder Änderungen im Rapid Report:

- In Abschnitt 5.2 wurde konkretisiert, dass eine Evaluation von Maßnahmen zur Aufwandsreduktion bei der systematischen Identifikation und Erhebung von Confoundern im Rahmen der Erstellung des Arbeitspapiers GA25-02 (Folgeauftrag zu GA23-02) erfolgen wird.

- In Abschnitt 5.3.1 wurde im Textabschnitt zum Parameter „Verschobene Nullhypothese“ ergänzt, dass sich das IQWiG in einem nächsten Schritt einer Überarbeitung der in A19-43 vorgeschlagenen Schwelle für die verschobene Nullhypothese widmen wird.
- In Abschnitt 5.3.2 wurde das Verfahrensbeispiel zu Epcoritamab durch Odronextamab (jeweils in der Indikation diffus großzelliges B-Zell-Lymphom [DLBCL]) ersetzt und ein Textabschnitt „Einordnung der orientierenden Betrachtung von Fallzahlszenarien“ ergänzt.
- In Abschnitt 5.3.2 wurde Abbildung 7 ergänzt, in der eine grafische Darstellung der in einer AbD aufdeckbaren Effektgröße erfolgt.
- Im Glossar und in Abschnitt 5.6.2.1 wurden die Definitionen „Missing at Random“ und „Missing not at Random“ angepasst.
- In Abschnitt 5.6.2.1 wurde im Textabschnitt „Beschreibung geeigneter Sensitivitätsanalysen“ ergänzt, dass die Einbindung von auxiliären Variablen mit der Gefahr der Verzerrung der Studienergebnisse durch Collider verbunden ist.
- In Abschnitt 5.6.2.1 wurde im Textabschnitt „Beschreibung geeigneter Sensitivitätsanalysen“ ergänzt, dass mittels Sensitivitätsanalysen der potenzielle Einfluss von residuellem Confounding abgeschätzt werden kann.
- In Abschnitt 5.6.2.2 wurde der Schwellenwert für fehlende Werte in Endpunkten mit Messwiederholungen präzisiert.
- In Abschnitt 5.7 wurde klargestellt, dass nicht nur Propensity-Score-Verfahren akzeptiert werden, sondern das IQWiG unabhängig von der gewählten Methode der kausalen Inferenz prüft, ob eine adäquate Confounderadjustierung geplant und durchgeführt wurde.
- In Abschnitt 5.7 wurde ein Textabschnitt zur Reihenfolge des Ziehens von Bootstrap-Samples und der Erzeugung multipler Imputationen ergänzt.
- Es wurde Anhang B „Würdigung der Anhörung zum vorläufigen Rapid Report“ ergänzt.
- Es wurde Anhang C „Offenlegung von Beziehungen der externen Sachverständigen“ ergänzt.

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 Rapid Report beschränkt sich auf die konkreten Ergebnisse.

5 Bewertung ausgewählter Aspekte zur Generierung und Auswertung 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 versorgungsnaher Daten zum Zwecke der Nutzenbewertung von Arzneimitteln für den vorliegenden Bericht wie folgt definiert:

- Versorgungsnaher 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 versorgungsnaher 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 versorgungsnaher 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 versorgungsnaher 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 (inklusive 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 versorgungsnahen 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.

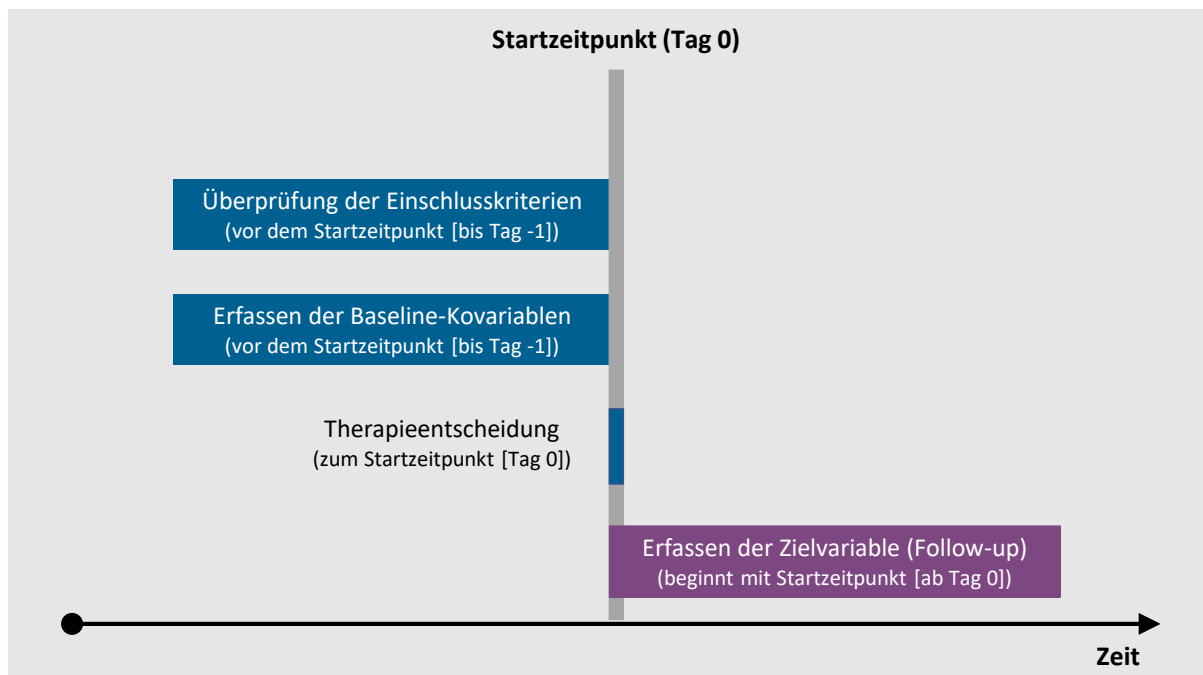


Abbildung 2: Zeitliche Ausrichtung zentraler Elemente des Studiendesigns am Startzeitpunkt der Beobachtung (t_0) (eigene Darstellung, in Anlehnung an Braitmaier 2022 [37])

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.

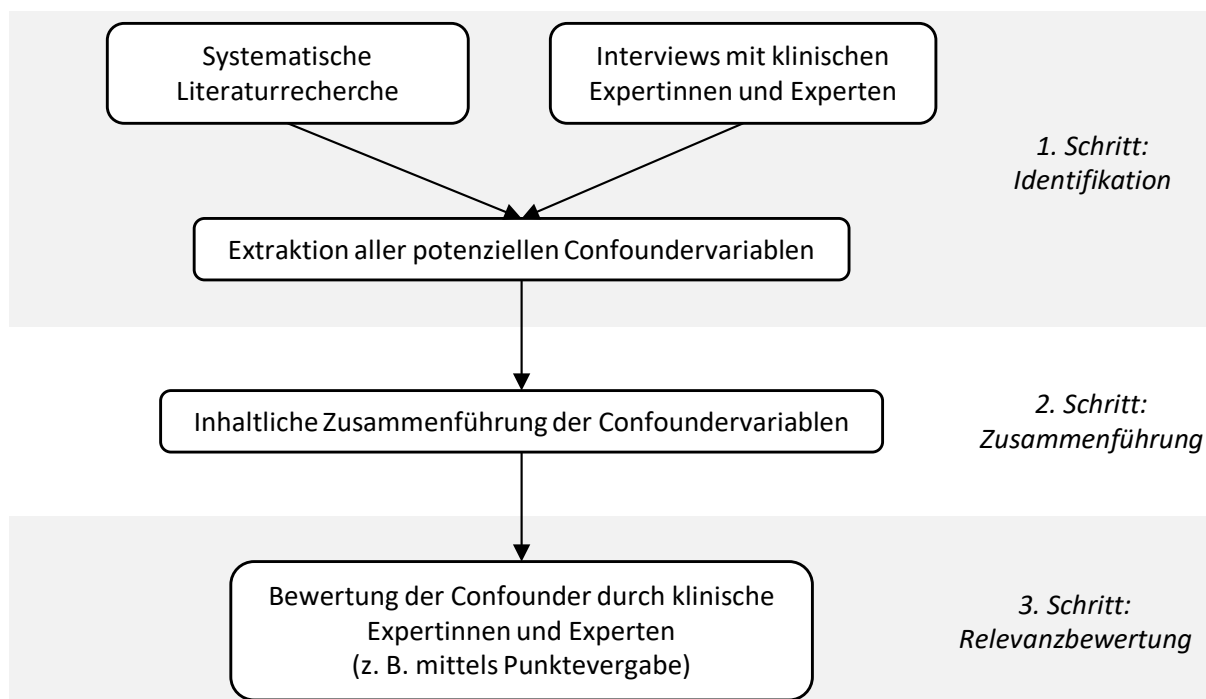


Abbildung 3: Vorgeschlagenes 3-schrittiges Verfahren zur systematischen Confounderidentifikation

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. Gleichmaßen 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 berücksichtigt 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 der systematischen Literaturrecherche sowie 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 Confounder-identifikation als durchführbar angesehen. Die vorliegenden Untersuchungen unterstreichen aber auch, wie aufwendig die Confounderidentifikation für nicht randomisierte vergleichende Studien ist. Wichtig ist an dieser Stelle jedoch zwischen (1) dem Aufwand im Vorfeld einer AbD (für die Identifikation und Verdichtung potenzieller Confounder) und (2) dem Gesamtaufwand für die AbD (darunter fällt u. a. der Aufwand für die Datenerhebung und die Dokumentation) zu differenzieren. Es kann zielführend sein, im Vorfeld einen höheren Aufwand zu investieren, um den Aufwand für die AbD in Gänze zu reduzieren. In einem nächsten Schritt soll im Rahmen der Erstellung des Arbeitspapiers GA25-02 [40] geprüft werden, ob weitere Maßnahmen zur Aufwandsreduktion bei der systematischen Identifikation und Erhebung von Confoundern möglich sind, um die Machbarkeit der AbD 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 Fallzahlsszenarien 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 Fallzahlsszenarien 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

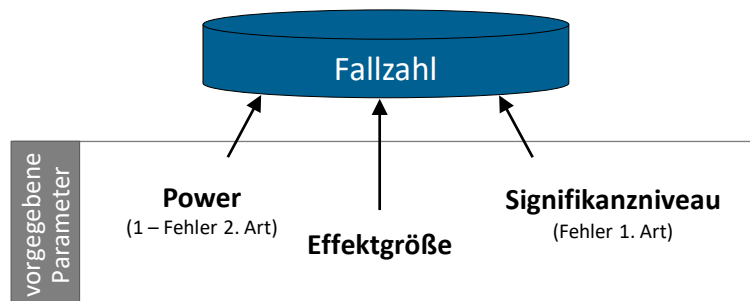
Das Ziel einer jeden Studie ist die Beantwortung einer wissenschaftlichen Fragestellung. Im Rahmen der evidenzbasierten Medizin orientiert sich die wissenschaftliche Fragestellung am sogenannten PICO-Schema (Population, Intervention, Comparison, Outcome). Die Studie soll somit Aussagen für den Vergleich, 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, das Signifikanzniveau (Irrtumswahrscheinlichkeit, auch Fehler 1. Art)

sowie die Power (1 – Fehler 2. Art) [41] (siehe Abbildung 4). Dabei entspricht die in einer AbD erwartete Effektgröße dem wahren Therapieeffekt zuzüglich einer systematischen Verzerrung, die durch das nicht randomisierte Design (z. B. durch unbekannte Confounder) hervorgerufen wird.



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 [42-44]. 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 Verhältnis der Randomisierung (z. B. in Form eines Verhältnisses von 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 wirksamer 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, lediglich 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 [45-47], 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 wegen des nicht randomisierten Studiendesigns. Es ist unstrittig, dass in einer AbD selbst bei sorgfältigster Analyse unter Erfüllung von umfangreichen Qualitätsanforderungen aufgrund der fehlenden Randomisierung relevante Unsicherheiten verbleiben, da u. a. unbekannte Confounder eine Rolle spielen können [48]. Um aus einer nicht randomisierten vergleichenden Studie somit eine Aussage zum Nutzen oder zum Schaden des zu bewertenden Arzneimittels ableiten zu können, ist ein ausreichend großer beobachteter Effekt erforderlich, von dem angenommen werden kann, dass er nicht allein durch systematische Verzerrung hervorgerufen wird. Dies soll in der AbD statistisch in Form eines Tests auf eine verschobene Nullhypothese (H_0) umgesetzt werden. Eine (positive oder negative) Aussage zum Nutzen oder zum Schaden des zu bewertenden Arzneimittels ergibt sich dann, wenn das Konfidenzintervall für den beobachteten Effekt oberhalb bzw. unterhalb einer zu definierenden Schwelle liegt [49]. In der Literatur werden verschiedene alternative Verfahren, z. B. der quantitativen Bias-Analyse, vorgeschlagen [50]. Diese folgen im Wesentlichen einem ähnlichen Ansatz und haben gleichermaßen das Ziel, Aussagen darüber zu erlauben, wie sicher das Ergebnis aus einer nicht randomisierten vergleichenden Studie ist.

Einen wissenschaftlichen Standard, wie groß die Schwelle für die verschobene Nullhypothese sein sollte, gibt es derzeit nicht. Da bei Generierung versorgungsnaher Daten im Rahmen der AbD die Erfüllung umfangreicher Qualitätsanforderungen Voraussetzung für die Betrachtung der Effekte ist, wurde in A19-43 [18] zunächst festgelegt, dass dieser Schwellenwert deutlich unterhalb des Wertes für den „dramatischen Effekt“ (RR von 5 bis 10 [51]) liegen sollte, z. B. in einem Bereich von 2 bis 5 für das RR (bzw. 0,2 bis 0,5 z. B. bei mortalitätssenkenden Interventionen) bezogen auf die untere (bzw. die obere) Konfidenzintervallgrenze.

Seit der Veröffentlichung von A19-43 (im Mai 2020) hat sich in der Fachwelt zunehmend ein Bewusstsein dafür entwickelt, dass für aussagekräftige Ergebnisse aus Beobachtungsstudien qualitativ hochwertige Daten unverzichtbar sind [52]. Eine methodische Weiterentwicklung stellt die systematische Identifikation von Confoundern gemäß des von Pufulete 2022 [39]

beschriebenen Vorgehens dar. Die bisherigen Erfahrungen mit der Prüfung der von den pUs eingereichten Studienunterlagen hinsichtlich ihrer inhaltlichen und methodischen Eignung zur Durchführung der AbDs zeigen, dass eine gute Planung nicht randomisierter vergleichender Studien auf Basis versorgungsnaher Daten – die die bestehenden Quellen für systematische Verzerrung bestmöglich adressiert – gelingen kann.

Zusammenfassend ist unbestritten, dass in einer vergleichenden Studie ohne Randomisierung relevante Unsicherheiten verbleiben, denen durch geeignete Maßnahmen – wie in der AbD durch Anwendung der verschobenen Nullhypothese (H_0) – Rechnung getragen werden muss. Dennoch scheint es vor dem Hintergrund der methodischen Weiterentwicklungen angezeigt, sich in einem nächsten Schritt (außerhalb der vorliegenden Ausarbeitung) einer Überprüfung der in A19-43 festgelegten Schwelle anzunehmen.

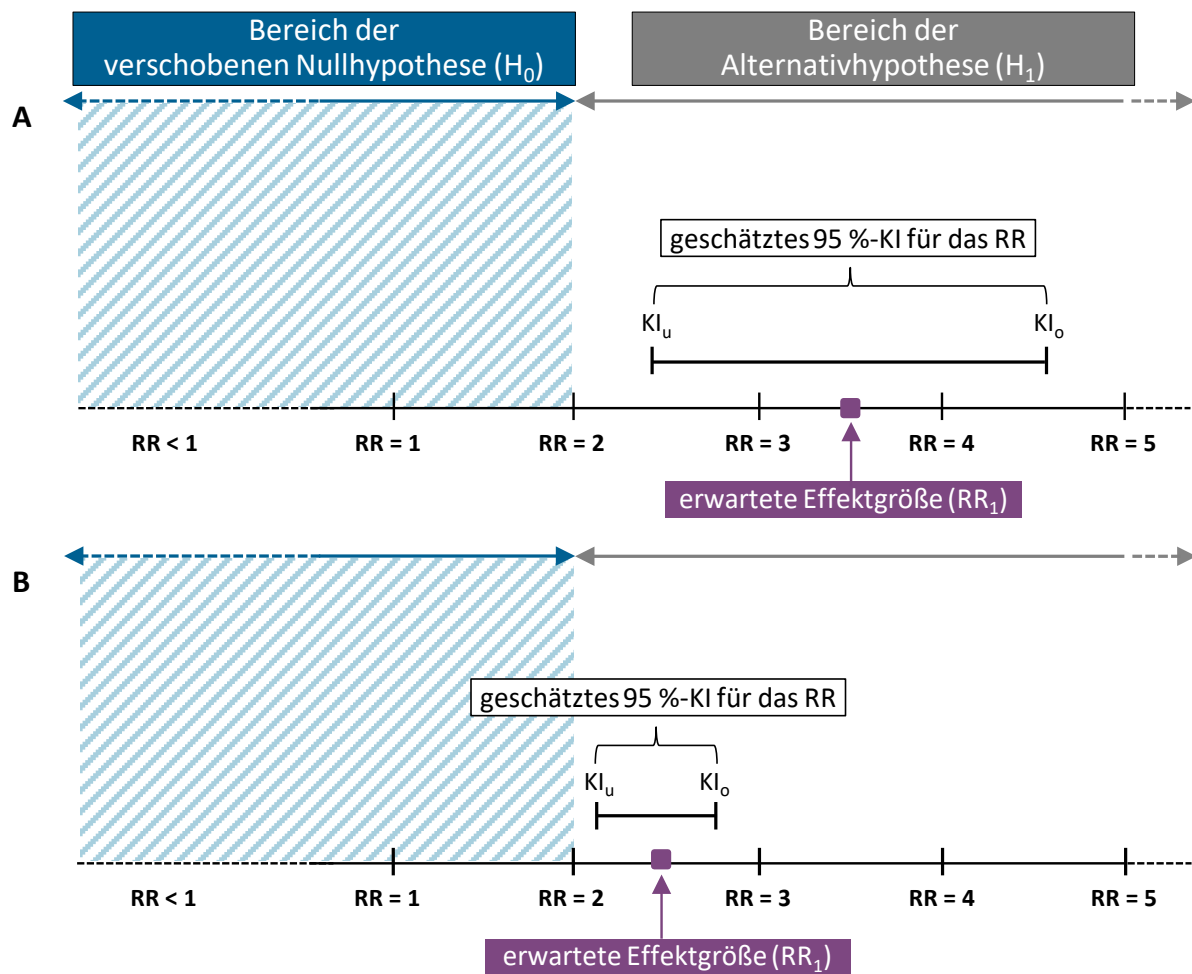
Um die Auswirkungen der verschobenen Nullhypothese auf die Fallzahlschätzung aufzuzeigen, wird im Folgenden der in A19-43 festgelegte Schwellenwert verwendet.

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 ($1 - \text{Fehler 2. Art}$) eine in einer AbD erwartete Effektgröße (RR_1) aufzudecken [53].

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_0 : Schwellenwert (Hypothesengrenze) für die untere Grenze des 2-seitigen 95 %-Konfidenzintervalls für das relative Risiko; RR_1 : erwartetes relatives Risiko

Abbildung 5: Szenarien bei der orientierenden Fallzahlschätzung mit verschobener Nullhypothese ($RR_0 = 2$)

- 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 01.02.2024 mit der Erstellung eines Konzeptes für eine AbD zu Odronextamab zur Behandlung von Erwachsenen mit einem rezidivierenden oder refraktären diffus großzelligen B-Zell-Lymphom (DLBCL) nach mindestens 2 vorherigen systemischen Therapien beauftragt (AbD-Konzept A24-18 [54]).

Im Rahmen der Konzepterstellung wurden gemäß der vom G-BA festgelegten zweckmäßigen Vergleichstherapie 2 Patientenpopulationen betrachtet:

- Erwachsene mit rezidiviertem oder refraktärem DLBCL nach mindestens 2 vorherigen systemischen Therapien, für die eine chimäre Antigenrezeptor(CAR)-T-Zelltherapie oder Stammzelltransplantation infrage kommt (Fragestellung 1)
- Erwachsene mit rezidiviertem oder refraktärem DLBCL nach mindestens 2 vorherigen systemischen Therapien, für die eine CAR-T-Zelltherapie und Stammzelltransplantation nicht infrage kommen (Fragestellung 2)

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 Odronextamab gegenüber dem vom G-BA jeweils festgelegten Komparator vorzunehmen.

Für Odronextamab lagen zum Zeitpunkt der Konzepterstellung (Fertigstellung am 03.06.2024) – vor Zulassung des Wirkstoffes in der Europäischen Union (22.08.2024) – ausschließlich erste vorläufige Ergebnisse der laufenden, pivotalen Phase-2-Studie ELM-2 als Pressemitteilung vor [55]. Für die Komparatoren der oben genannten Fragestellungen zu Odronextamab standen entweder finale Ergebnisse aus in Fachzeitschriften publizierten Studien oder ein Dossier zur Nutzenbewertung gemäß § 35a SGB V zur Verfügung. Daten zum Gesamtüberleben lagen hierbei weder für Odronextamab noch für die durch den G-BA festgelegten Komparatoren differenziert für die beiden Fragestellungen vor. Aufgrund der unterschiedlichen Prognosen der Patientengruppen der beiden Fragestellungen (kurative Intention vs. palliative Intention)

lassen jedoch nur fragestellungsbezogene Fallzahlschätzungen eine sinnvolle Einschätzung zum Umfang der AbD zu. Aufgrund dessen wurde im Rahmen des AbD-Konzepts A24-18 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).

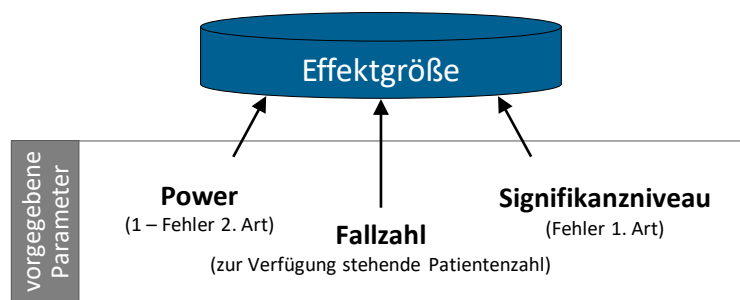
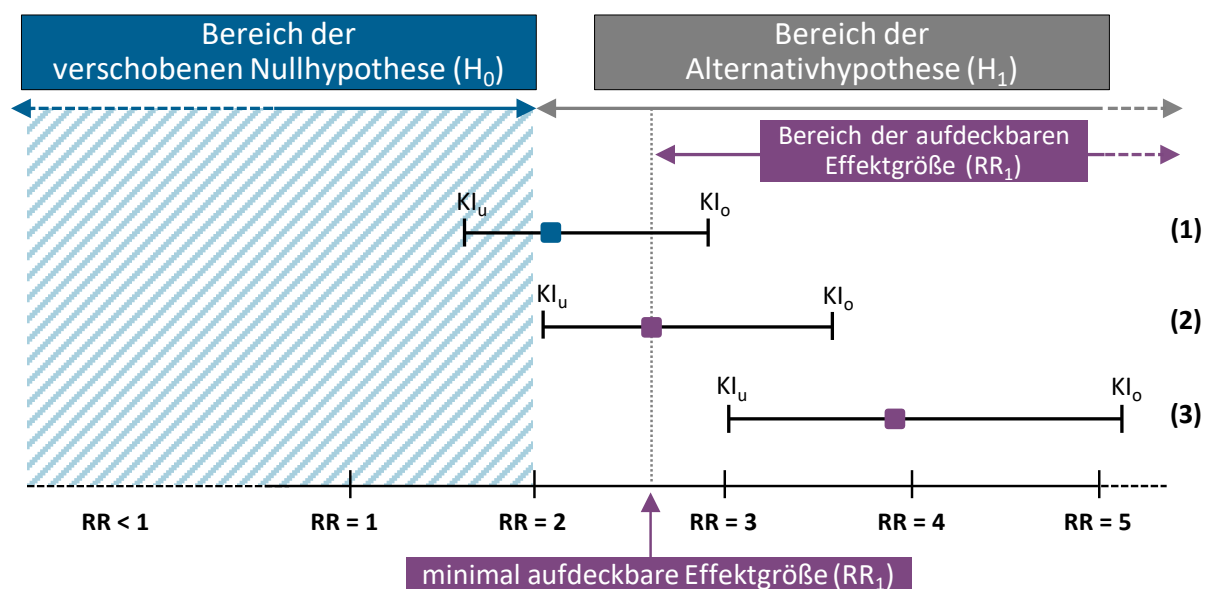


Abbildung 6: Aufdeckbare Effektgröße bei Festlegung der Parameter Fallzahl (N), Signifikanzniveau (Fehler 1. Art) und Power (1 – Fehler 2. Art)

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. Dies wird exemplarisch durch die folgende Abbildung 7 illustriert.



H_0 : Nullhypothese; H_1 : Alternativhypothese; KI_u : untere Grenze des Konfidenzintervalls; KI_o : obere Grenze des Konfidenzintervalls; RR: relatives Risiko; RR_0 : Schwellenwert (Hypothesengrenze) für die untere Grenze des 2-seitigen 95 %-Konfidenzintervalls für das relative Risiko; RR_1 : aufdeckbares relatives Risiko

Abbildung 7: Aufdeckbarer Effekt (Effektmaß: Relatives Risiko [RR_1]) mit verschobener Nullhypothese ($RR_0 = 2$)

Die Punktschätzungen der Effekte (2) und (3) entsprechen der minimal aufdeckbaren Effektgröße bzw. liegen jenseits davon. Das bedeutet, dass in der AbD beobachtete Effekte dieser Größenordnung bei gegebener Patientenzahl mit einer Power von 80 % statistisch signifikant sind, eine Effektschätzung in der Größenordnung von Effekt (1) hingegen nicht.

Neben der geschätzten Patientenzahl wird, insofern entsprechende Angaben verfügbar sind, 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 hierbei auf das Basisrisiko in der Interventionsgruppe oder das in der Kontrollgruppe zurückgegriffen wird, hängt davon ab, für welche der Therapien Informationen für die interessierende Patientengruppe vorliegen (siehe folgender Kasten).

Beispiel für ein Addendum zu einem AbD-Konzept, in dem eine orientierende Betrachtung verschiedener Fallzahlsszenarien erfolgte

Das Addendum A24-120 [56] zum oben erwähnten AbD-Konzept A24-18 zu Odronextamab ist gemäß Folgeauftrag des G-BA auf Patientinnen und Patienten, für die eine CAR-T-Zelltherapie und Stammzelltransplantation nicht infrage kommen, beschränkt (Fragestellung 2). Für die Patientenpopulation von Fragestellung 2 lagen auch zum Zeitpunkt der Bearbeitung des Folgeauftrages keine hinreichenden Informationen für eine orientierende Fallzahlschätzung vor (weder für Odronextamab noch für die vom G-BA festgelegten Komparatoren).

Wie bereits bei der Konzepterstellung war unklar:

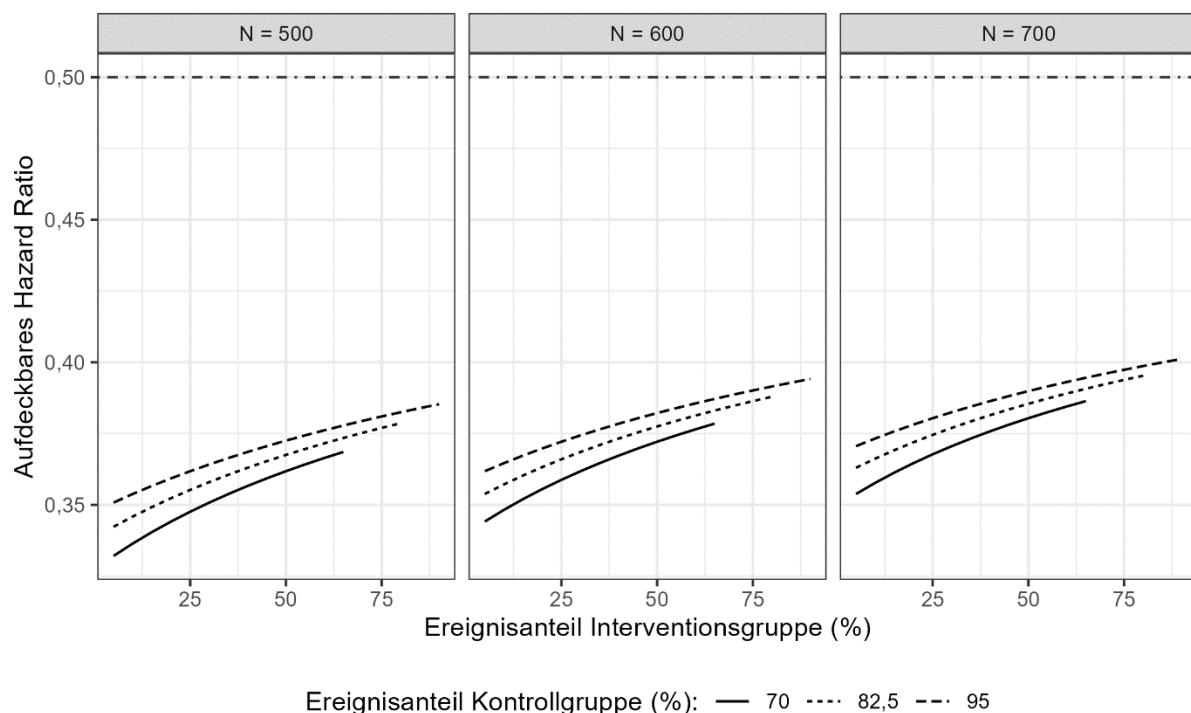
- wie hoch der Anteil der Patientinnen und Patienten in den Studien jeweils ist, für die eine CAR-T-Zelltherapie und Stammzelltransplantation nicht infrage kommen,
- ob die jeweiligen Anteile unter Odronextamab und den Komparatoren vergleichbar sind und
- ob zwischen den Studien zu Odronextamab und den Komparatoren Unterschiede in der Teilpopulation der Patientinnen und Patienten, für die eine CAR-T-Zelltherapie und Stammzelltransplantation nicht infrage kommen, hinsichtlich der Anzahl vorangegangener Therapielinien und somit der Schwere und dem Fortschreiten der Erkrankung bestehen.

Zur Annäherung an die Frage, ob eine AbD zu Odronextamab grundsätzlich sinnvoll durchführbar ist, erfolgte eine orientierende Betrachtung verschiedener Fallzahlsszenarien. Es wurden die Effekte eines mittels Ereigniszeitanalysen ausgewerteten Endpunkts (Effektmaß: Hazard Ratio) – im vorliegenden Fall des Gesamtüberlebens – dargestellt, die bei den im Anwendungsgebiet geschätzten Patientenzahlen ($N = 500$, $N = 600$ und $N = 700$) und einem Verteilungsverhältnis von 1:1 (Intervention zu Komparatoren) mit einer Power von 80 % aufdeckbar sind.

Zur Ermittlung des in einer AbD aufdeckbaren Effektes (HR_1) wurde die für die Kontrollgruppe vorliegende Evidenz im Endpunkt Gesamtüberleben als Anker herangezogen. Basierend auf den gesichteten Daten wurden für die Kontrollgruppe Anteile von verstorbenen Patientinnen

und Patienten (im Folgenden als „Ereignisanteil“ bezeichnet) von 70 %, 82,5 % und 95 % angenommen (siehe „Einordnung der orientierenden Betrachtung der Fallzahlsszenarien“ unten). Für die Interventionsgruppe wurden die sich ergebenden Ereignisanteile von 5 % bis 65 %, 5 % bis 80 % bzw. 5 % bis 90 % dargestellt.

Die folgende Abbildung 8 zeigt den in einer AbD (mit einer Power von 80 %) aufdeckbaren Effekt mit einer Spanne von ca. $HR_1 = 0,35$ bis $HR_1 = 0,40$ zum Vorteil von Odronextamab gegenüber den Komparatoren unter Verwendung sowohl festgelegter Stichprobengrößen als auch Ereignisanteile in beiden Behandlungsgruppen.



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

Abbildung 8: Aufdeckbarer Effekt (Effektmaß: Hazard Ratio [HR_1]) in Abhängigkeit von der Stichprobengröße und den Ereignisanteilen in beiden Behandlungsgruppen (Verhältnis von Intervention zu Komparatoren von 1:1)

Einordnung der orientierenden Betrachtung der Fallzahlsszenarien

Zur Einordnung der orientierenden Betrachtung der Fallzahlsszenarien wurden die Ergebnisse zum Endpunkt Gesamtüberleben aus den jeweils vorhandenen Studien zur Intervention und den Komparatoren (inklusive der seitens des G-BA für den Folgeauftrag zusätzlich benannten Komparatoren) beschrieben.

Daten zu den Komparatoren

Zu den für die Konzepterstellung vom G-BA übermittelten Komparatoren von Fragestellung 2 lagen für das zu bewertende Anwendungsgebiet u. a. publizierte Ergebnisse zu Tafasitamab + Lenalidomid [57] und das Dossier zur Nutzenbewertung von Polatuzumab Vedotin [58] vor.

Für die 3 Komparatoren Glofitamab, Epcoritamab und Loncastuximab tesirin, die laut Auftrag des G-BA ergänzend berücksichtigt werden sollen, lagen für den Endpunkt Gesamtüberleben jeweils Daten im Dossier zur Nutzenbewertung vor (Glofitamab [59], Epcoritamab [60] und Loncastuximab tesirin [61]).

Nach Extrapolation der für alle genannten Komparatoren vorhandenen Angaben auf den für die Datenerhebung relevanten Zeitraum von 36 Monaten [53] ergaben sich Ereignisanteile von 72 % bis 91 %. Hiervon abgeleitet wurden in den orientierenden Fallzahlszenarien für die Kontrollgruppe Ereignisanteile zwischen 70 % und 95 % berücksichtigt.

Daten zu Odronextamab

Für Odronextamab lagen Ergebnisse der beiden Studien ELM-2 und ELM-1 im European Public Assessment Report (EPAR) vor [62]. Nach Extrapolation der Angaben auf den für die Datenerhebung relevanten Zeitraum von 36 Monaten [53] ergaben sich Ereignisanteile von 83 % bis 96 % unter Odronextamab. Diese Spanne entspricht zum Teil den Ereignisanteilen, die für die orientierende Betrachtung möglicher Fallzahlszenarien (siehe Abbildung 8) als Ereignisanteile in der Interventionsgruppe herangezogen wurden.

Unsicherheiten in der Interpretierbarkeit der Studienergebnisse

Die bisherigen Studienergebnisse lassen wegen der oben beschriebenen Unsicherheiten, die die Interpretierbarkeit der Ergebnisse deutlich einschränken, keine hinreichende Einschätzung der Effekte zu, die für eine orientierende Fallzahlschätzung zum Vergleich von Odronextamab mit den Komparatoren notwendig wäre. Die Daten können jedoch als Orientierungspunkte für die Ereignisanteile in der Kontrollgruppe in den dargestellten orientierenden Fallzahlszenarien dienen.

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 9 illustriert.

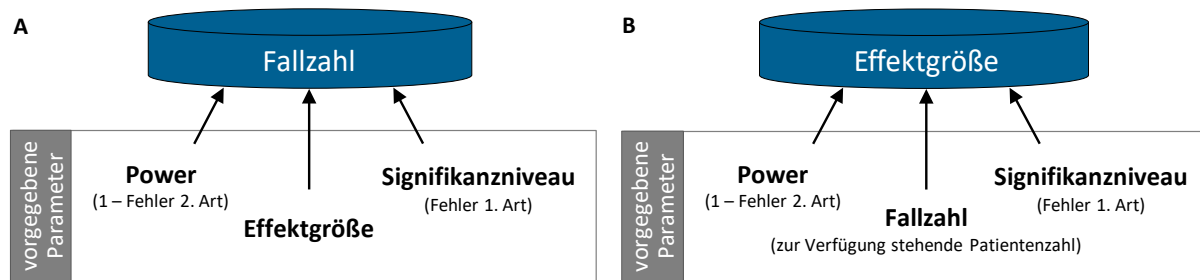


Abbildung 9: Herangehensweisen zur Bewertung der Machbarkeit einer AbD:

(A) orientierende Fallzahlschätzung und (B) orientierende Betrachtung von Fallzahlszenarien

In Abbildung 9 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, versorgungsnaher 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 Fallzahl Szenarien) sollte der pU des zu bewertenden Arzneimittels die (finale) Fallzahlplanung der AbD zum Zeitpunkt der Zwischenanalysen auf Basis der dann vorhandenen vorläufigen Ergebnisse vornehmen [63]. 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 [64]. 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 [64]. 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) therapie-naiv 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 [65]. 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 [66].

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 [64] 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 [64,67]. 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 [64].

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 [64]. 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] [68]). 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 (zum 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 [64]. 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 / der vom Komparator auf die Intervention gewechselt 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 [19,69].
- **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 zeitabhängigen 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 [70]).
- **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 [70].

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 [66]. Dies kann gleichermaßen bei der Definition und Bildung des Expositionssets geschehen [70], 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 Expositionset 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 [65] 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 [64]. 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 [71] (zur Erläuterung von IPCW siehe GA14-04 [72]).

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 [65]) und die Bildung von Klonen [64] 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 [64]. 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 [72].

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 [73,74]. Im Rahmen des vorliegenden Rapid Reports werden ausschließlich ein Wechsel von der Intervention zum Komparator (Szenario A) und 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 [74].
- **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 [72,74].

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 [72]. 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 [75].

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 [76]).

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 Behandlung mit dem Komparator (entspricht der 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 [76]. Gemäß den Angaben in den Studienunterlagen der AbD zu Etranacogen Dezaparvovec in der Indikation Hämophilie B ist Folgendes geplant [77,78]:

- 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 Antitumorthherapie zwischen der Gewinnung von T-Zellen durch Leukapherese zur Überbrückung der Wartezeit bis zur Bereitstellung der CAR-T-Zellen [79]

- **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) [80,81]
- **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 [77]

Eine Brückentherapie ist als Teil des Therapiekonzepts zu sehen [82]. 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 [83]. 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 [84]. Dabei sind, insofern möglich, indikations-spezifische 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 rezidiertes 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 [85,86]. 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) [86]. Hinzu kommt, dass die behandelnden Zentren über die Ergebnisse der jeweiligen PRO-Erhebung nicht informiert werden [85], obwohl der positive Einfluss einer PRO-Datennutzung in der Routineversorgung in Studien gut belegt ist (siehe unten).
- Im Deutschen Hämophileregister (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 [87,88]. 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 [88].

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 [89,90]. 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 [91,92]. Ein weiterer Vorteil einer digitalen Erfassung wird darin gesehen, dass die mit einer PRO-Messung potenziell einhergehende Belastung der Patientinnen und Patienten verringert wird [91] (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 [91,93]. 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 [91]. 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 (Projekt A24-38, AbD zu Etranacogen Dezaparvovec [Hämophilie B]: 3. Addendum zum Auftrag A22-83 [94]).

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 [90,95].
- Erkenntnisse aus vorhandenen PRO-Daten (z. B. durch das Abrufen in tagesaktuellen Dashboards) können in der Patientenberatung und zur gemeinsamen Entscheidungsfindung eingesetzt werden [90,95,96].
- Für einzelne onkologische Entitäten wurde nachgewiesen, dass eine medizinische Betreuung, die patientenberichtete Daten im Verlauf integriert, einer Betreuung ohne solche Informationen überlegen ist [97-103]. 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 [100,103].

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 [90].

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 [86,104].

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 [91,92,105-107]. Ein anderer Erhebungsmodus (wie z. B. telefonisch oder papierbasiert) wird lediglich als Ergänzung für ansonsten nur schwer erreichbare Patientinnen und Patienten diskutiert [92,105,106,108,109].

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 Gesprächen 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) [110,111].
- Um hochwertige Gesundheitsdaten aus dem Versorgungsalltag praktikabel und ohne große Hürden vorhalten und bereitstellen zu können, müssen 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 [112,113]. 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 [114].

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 [90,95,111,115].
- Eine Förderung der Netzerkennung 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 [96,111,115].
- Die Vergütung des Aufwands der teilnehmenden Zentren kann durch Zahlung einer Aufwandsentschädigung, z. B. in Abhängigkeit von der Qualität der erhobenen Daten, erfolgen [111,115,116].
- 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 [117-123]. 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 [90,95]. Die unmittelbare Sichtbarkeit des Nutzens der erhobenen Daten setzt einen Anreiz, diese vollständig und in guter Qualität zur Verfügung zu stellen (siehe auch Abschnitt 5.5).

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) [115,124,125].
- 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 Symptomanagement sind oft genannte Erfolgsfaktoren (im Sinne eines Mehrwerts für die individuelle Patientenbehandlung) [16,90,126,127].
- 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 [110,111,124,126].
- 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 [90,117,119,121,125,126,128-131].
- 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) [119,121,126,128,132].
- Neben immateriellen Anreizen (wie Bereitstellung der Registerdaten) empfinden Patientinnen und Patienten materielle Anreize in Form finanzieller Zuwendungen (Geschenkgutscheine) oder Sachprämien als motivierend [115,128,133].

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, unter Berücksichtigung der beobachteten Werte einer Person, nicht vom fehlenden Wert selbst ab. Solche fehlenden Werte führen zu verzerrten Ergebnissen, insofern die Abhängigkeit in der Analyse unberücksichtigt bleibt [15]. Wenn Angaben zu anderen Werten der Person (z. B. Kovariablen oder Werten zu früheren Zeitpunkten) 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 fehlenden Wert selbst ab und diese Abhängigkeit verschwindet auch nicht durch Berücksichtigung der beobachteten Werte der Person. Solche fehlenden Werte

führen zu verzerrten Ergebnissen. Diese Verzerrung kann auch nicht korrigiert werden, da der Ausfallmechanismus von unbeobachteten Werten 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 Nichterfüllung der MCAR-Annahme zu einer deutlichen Verzerrung führen [134,135]. 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 [134-144]. 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 $N \geq 1000$), 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 [134-140,142,145,146]. 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 [134,135,137-139,141,142,146-149]. 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 [135,138,149]. 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 [134,135]. 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 bayessche 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 [145]. 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 [137,150]. Für den Einbezug des Endpunktes sprechen nicht nur neuere Simulationsergebnisse [135,138], sondern auch theoretische Überlegungen [151].

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 [145] 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 [152], Missing-Indicator-Methode [147], General-Boosting-Modeling sowie

Random-Forest-Based-Missing-Imputation-Methods [139]. 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 modellbasierte Methoden, die auf der Annahme MAR oder MNAR beruhen [136,139,147]. Lee 2021 [146], Ling 2020 [140] sowie Carpenter und Smuk 2021 [153] geben dazu praktische Hinweise. Die relevante Komponente ist dabei generell die MAR-Annahme [152,153]. Daneben kann die Einbindung von sog. auxiliären Variablen erfolgen [140,145], sofern diese im Register erhoben werden. Diese Variablen sind keine Confounder im engeren Sinne, jedoch mit Confoundern assoziiert, und ihre Berücksichtigung im Modell für die multiple Imputation kann letztlich die Imputation der fehlenden Werte bei den Confoundern verbessern [145,153]. Bei der Auswahl von auxiliären Variablen ist jedoch zu beachten, dass das Risiko für einen Collider Bias besteht [154].

In weiteren Sensitivitätsanalysen kann die Operationalisierung der potenziellen Confounder variiert werden, z. B. durch Transformation stetiger Variablen [145] oder durch Kalibrierungsmethoden [143]. 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 [145]. 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 adäquate Sensitivitätsanalysen identifiziert und geplant werden. Zudem besteht die Möglichkeit, mittels Sensitivitätsanalysen den potenziellen Einfluss residuellen Confoundings abzuschätzen, z. B. durch die quantitative Bias-Analyse (umfasst den E-Value, der misst, wie stark ein nicht berücksichtigter Confounder mit Behandlung und Endpunkt assoziiert sein muss, um den Effekt zu erklären) [50,155,156] oder durch Verwendung von Negativ-Kontrollen [157]. Dabei sind die spezifischen Annahmen, die ebendiesen Methoden zugrunde liegen, allerdings hinreichend in den Studienunterlagen zu beschreiben.

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 [137,138,140,146-148].

5.6.2.2 Interpretierbarkeit von Analysen mit fehlenden und / oder imputierten Werten

Im 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 der Unterschied zwischen den Behandlungsgruppen > 15 Prozentpunkte beträgt, 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. Bei Endpunkten mit Messwiederholungen, 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. Zusätzlich zur vorgenannten Bedingung sollten bei Analysen zu einem bestimmten Auswertungszeitpunkt zu ebendiesem Zeitpunkt Werte für mindestens 50 % der Patientinnen und Patienten vorliegen. Diese Schwellenwerte dienen dann als Orientierung, wenn nicht mit ausreichender Sicherheit davon ausgegangen werden kann, dass die Werte zufällig fehlen. Ausnahmen ergeben sich beispielsweise, wenn im Rahmen 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 ist 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 [144,145,148].

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 [137] 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 [138] 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 [140]. 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 bayesschen 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 [145]

- Begründungen für methodische Aspekte, die nicht im Studienprotokoll und / oder SAP präspezifiziert werden konnten [145]
- 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. In den folgenden Abschnitten wird – in Anbetracht ihrer zentralen Bedeutung – ausschließlich auf Propensity-Score-Verfahren zur Confounderadjustierung Bezug genommen. Davon unbenommen wird das IQWiG jedoch in jeder Prüfung von Studienunterlagen und in jeder Bewertung ergebnisoffen prüfen, ob eine adäquate Confounderadjustierung geplant und durchgeführt wurde, unabhängig von der vom pU gewählten Methode der kausalen Inferenz.

Im Rahmen einer AbD spielt zur Confounderadjustierung insbesondere die Propensity-Score-Methode eine Rolle [49], u. a., weil es hier Ansätze gibt, die zentralen Annahmen der kausalen Inferenz (d. h. ausreichende *Positivität*, *Überlappung* und Balanciertheit) zu prüfen und darzustellen [158]. *Propensity-Score-Verfahren* wurden ursprünglich für große Datensätze mit zum Teil mehreren tausend Patientinnen und Patienten entwickelt (für epidemiologische Fragestellungen) [159]. 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 [160] 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 [161] 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 [162]. 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. Bei der Verwendung von Bootstrap-Verfahren stellt sich für den Umgang mit fehlenden Werten die Frage der Reihenfolge des Ziehens von Bootstrap-Samples und der Erzeugung multipler Imputationen. In einer Simulationsstudie von Schomaker 2018 [163] zeigt sich der Ansatz, erst Bootstrap-Samples zu ziehen und anschließend jedes Sample mehrfach zu imputieren, als die zuverlässigste, aber auch rechenintensivste, Methode. Der Ansatz, erst multiple Imputationen zu erzeugen und anschließend Bootstrap-Samples jeder Imputation zu ziehen, scheint den Ergebnissen zufolge ebenfalls valide Ergebnisse zu liefern, vorausgesetzt, es liegt ein nicht zu hoher Anteil fehlender Werte vor.

In einer Simulationsstudie von Wilkinson 2022 [164] 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 [164].

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 [77,79-81,165]) als die, die den genannten Simulationsstudien zugrunde lag (4 bis 10 Confounder [10 Confounder bei Fallzahl von $N \geq 250$]). 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 [162] 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 [160] und Friedrich 2020 [161] 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 Nichtberechenbarkeit 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, bestehende 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 Pufulute 2022 (über eine systematische Literaturrecherche und Einbindung von Klinikern 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. Grundsätzlich kann es zielführend sein, im Vorfeld einer AbD einen höheren Aufwand zur Reduktion der Anzahl zu erhebender Confounder zu investieren, um den Gesamtaufwand für die AbD durch Aufwandseinsparungen bei der Datenerhebung und -auswertung zu reduzieren.
- 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 Fallzahl Szenarien, 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 niedrighschwellige 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_versorgungsnaher-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-related 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 schubförmige 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. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Evaluation von Maßnahmen zur Aufwandsreduktion bei der systematischen Identifikation und Erhebung von Confoundern [online]. 2025 [Zugriff: 04.09.2025]. URL: <https://www.iqwig.de/projekte/ga25-02.html>.
41. 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>.
42. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzept für eine anwendungsbegleitende Datenerhebung – Brexucabtagen autoleucel; Rapid Report [online]. 2022 [Zugriff: 07.03.2025]. URL: https://www.iqwig.de/download/a21-130_anwendungsbegleitende-datenerhebung-brexucabtagen-autoleucel_rapid-report_v1-0.pdf.
43. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzept für eine anwendungsbegleitende Datenerhebung – Valoctocogen Roxaparvovec; Rapid Report [online]. 2022 [Zugriff: 07.03.2025]. URL: https://www.iqwig.de/download/a22-20_anwendungsbegleitende-datenerhebung-valoctocogen-roxaparvovec_rapid-report_v1-0.pdf.
44. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Etranacogen Dezaparvovec (Hämophilie B) – Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2023 [Zugriff: 07.03.2025]. URL: https://www.iqwig.de/download/a22-83_etranacogen-dezaparvovec-haemophilie-b_abd-konzept_v1-1.pdf.

45. 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>.
46. 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>.
47. 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>.
48. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health* 2000; 21: 121-145. <https://doi.org/10.1146/annurev.publhealth.21.1.121>.
49. 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.
50. Brown PJ, Hunnicutt JN, Sanni Ali M et al. Quantifying possible bias in clinical and epidemiological studies with quantitative bias analysis: common approaches and limitations. *BMJ* 2024; 385: e076365. <https://doi.org/10.1136/bmj-2023-076365>.
51. 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>.
52. International Society for Pharmacoeconomics and Outcomes Research. Towards a Vision for HEOR: Opportunities for Enhancement and Evolution; An ISPOR Health Science Policy Council White Paper [online]. 2024 [Zugriff: 26.09.2025]. URL: https://www.ispor.org/docs/default-source/councils/health-science-policy-council/towards-a-vision-for-heor---opportunities-for-enhancement-and-evolution-an-isor-health-science-policy-council-white-papere53f0015-8dd9-4a79-90a5-e8efe89f03e4.pdf?sfvrsn=fdbcb152_1.
53. Chow SC, Wang H, Shao J. Sample Size Calculations in Clinical Research. Boca Raton: Taylor & Francis; 2003.
54. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Odronextamab (DLBCL); Bewertung gemäß § 35a SGB V; AbD-Konzept [online]. 2024 [Zugriff: 05.08.2025]. URL: <https://doi.org/10.60584/A24-18>.
55. Kahl KL. Single-Agent Odronextamab Efficacy, Safety Upheld in Relapsed/Refractory DLBCL [online]. 2023 [Zugriff: 07.03.2025]. URL: <https://www.cancernetwork.com/view/single-agent-odronextamab-efficacy-safety-upheld-in-relapsed-refractory-dlbcl>.

56. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Odronextamab (DLBCL); 1. Addendum zum Projekt A24-18 (AbD-Konzept) [online]. 2025 [Zugriff: 05.08.2025]. URL: <https://doi.org/10.60584/A24-120>.
57. 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>.
58. 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>.
59. 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>.
60. 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>.
61. 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>.
62. European Medicines Agency. Ordspono; Assessment report [online]. 2024 [Zugriff: 26.09.2025]. URL: https://www.ema.europa.eu/en/documents/assessment-report/ordspono-epar-public-assessment-report_en.pdf.
63. 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>.
64. 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>.
65. 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>.
66. 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>.

67. 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>.
68. 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>.
69. 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>.
70. 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>.
71. 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>.
72. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Treatment Switching in onkologischen Studien; Arbeitspapier [online]. 2018 [Zugriff: 07.03.2025]. URL: https://www.iqwig.de/download/ga14-04_treatment-switching-in-onkologischen-studien_arbeitspapier_v1-0.pdf.
73. 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>.
74. 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>.
75. 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.

76. 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-Committee-Meeting.pdf>.

77. 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.

78. 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.

79. 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.

80. 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.

81. Roche Pharma. Evaluation of a real world data collection for the reassessment of the additional benefit of 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.

82. 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>.

83. 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>.

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

85. 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.

86. 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>.

87. Deutsches Hämostasieregister. 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.

88. Deutsches Hämostasieregister. 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.

89. 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>.

90. 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.

91. 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>.

92. 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>.
93. 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>.
94. 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: 07.03.2025]. URL: <https://doi.org/10.60584/A24-38>.
95. 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>.
96. 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.
97. 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>.
98. 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>.
99. 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>.
100. 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>.

101. 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>.
102. 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>.
103. 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>.
104. 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>.
105. 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>.
106. 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.
107. 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>.
108. 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>.
109. 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>.
110. 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>.

111. 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.
112. 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>.
113. 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.
114. 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>.
115. 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>.
116. 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>.
117. 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>.
118. Lissbrant IF, Eriksson 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>.
119. 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>.
120. 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>.

121. 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>.
122. 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>.
123. SCQM Foundation. SCQM-Datenbank [online]. [Zugriff: 23.06.2025]. URL: <https://www.scqm.ch/de/medizinisches-fachpersonal/scqm-datenbank/>.
124. 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>.
125. 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>.
126. 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>.
127. Nelson EC, Dixon-Woods M, Batalden PB et al. Patient focused registries can improve health, care, and science. *BMJ* 2016; 354: i3319. <https://doi.org/10.1136/bmj.i3319>.
128. 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>.
129. 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.
130. 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>.
131. SCQM Foundation. mySCQM Webapplikation [online]. [Zugriff: 23.06.2025]. URL: <https://www.scqm.ch/de/betroffene/myscqm-ihre-webapplikation/>.

132. 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>.

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

134. 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>.

135. 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>.

136. 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>.

137. 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>.

138. 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>.

139. 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>.

140. 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>.

141. 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>.

142. 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>.

143. 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>.
144. 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>.
145. 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>.
146. 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>.
147. 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>.
148. 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>.
149. 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>.
150. 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>.
151. Carpenter JR, Kenward MG, Bartlett JW et al. *Multiple Imputation and its Application*. Chichester: Wiley; 2023.
152. 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>.
153. 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>.
154. Curnow E, Tilling K, Heron JE et al. Multiple imputation of missing data under missing at random: including a collider as an auxiliary variable in the imputation model can induce bias. *Front Epidemiol* 2023; 3: 1237447. <https://doi.org/10.3389/fepid.2023.1237447>.

155. Ioannidis JPA, Tan YJ, Blum MR. Limitations and Misinterpretations of E-Values for Sensitivity Analyses of Observational Studies. *Ann Intern Med* 2019; 170(2): 108-111. <https://doi.org/10.7326/M18-2159>.
156. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med* 2017; 167(4): 268-274. <https://doi.org/10.7326/M16-2607>.
157. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010; 21(3): 383-388. <https://doi.org/10.1097/EDE.0b013e3181d61eeb>.
158. 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>.
159. 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>.
160. 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>.
161. Friedrich S, Friede T. Causal inference methods for small non-randomized studies: Methods and application in COVID-19. *Contemp Clin Trials* 2020; 99: 106213. <https://doi.org/10.1016/j.cct.2020.106213>.
162. 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>.
163. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. *Stat Med* 2018; 37(14): 2252-2266. <https://doi.org/10.1002/sim.7654>.
164. 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>.
165. 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.

166. 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>.
167. 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>.
168. 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>.
169. Karolinska Institutet. Swedish Rheumatology Quality Registry; PER – Patient's Self Registration [online]. [Zugriff: 23.06.2025]. URL: <https://srq.nu/en/about-per-patient/>.
170. Andr n C. SP0141: Ask the patient – distinguishing your symptoms. Ann Rheum Dis 2014; 73: 38. <https://doi.org/10.1136/annrheumdis-2014-eular.6198>.
171. Institut f r Qualit t und Wirtschaftlichkeit im Gesundheitswesen. Exagamglogen Autotemcel (Sichelzellkrankheit): Pr fung des Studienprotokolls (Version 1.0) und des statistischen Analyseplans (Version 1.0); 2. Addendum zum Projekt A23-49 (AbD-Konzept) [online]. 2025 [Zugriff: 18.09.2025]. URL: <https://doi.org/10.60584/A25-82>.
172. Loudon K, Treweek S, Sullivan F et al. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ 2015; 350: h2147. <https://doi.org/10.1136/bmj.h2147>.
173. Food and Drug Administration. Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products; Guidance for Industry [online]. 2024 [Zugriff: 04.09.2025]. URL: <https://www.fda.gov/media/152503/download>.
174. European Medicines Agency. Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence for regulatory purposes [online]. 2025 [Zugriff: 04.09.2025]. URL: https://www.ema.europa.eu/en/documents/other/reflection-paper-use-real-world-data-non-interventional-studies-generate-real-world-evidence-regulatory-purposes_en.pdf.
175. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002; 288(3): 358-362. <https://doi.org/10.1001/jama.288.3.358>.
176. Altman DG. Statistics and ethics in medical research: III How large a sample? Br Med J 1980; 281. <https://doi.org/10.1136/bmj.281.6251.1336>.

177. Gemeinsamer Bundesausschuss. Anwendungsbegleitende Datenerhebung und Beschränkung der Versorgungsbefugnis: Onasemnogen-Abeparvovec – Spinale Muskelatrophie [online]. [Zugriff: 24.09.2025]. URL: <https://www.g-ba.de/anwendungsbegleitende-datenerhebung-verfahren/1/>.

178. Gemeinsamer Bundesausschuss. Anwendungsbegleitende Datenerhebung und Beschränkung der Versorgungsbefugnis: Brexucabtagen-Autoleucel/Autologe Anti-CD19-transduzierte CD3+ Zellen – Mantelzelllymphom [online]. [Zugriff: 24.09.2025]. URL: <https://www.g-ba.de/anwendungsbegleitende-datenerhebung-verfahren/5/>.

179. Lange S, Lauterberg J. Pragmatischere randomisierte Studien mit Fokus auf Registerbasierung. Präz Gesundheitsf 2022. <https://doi.org/10.1007/s11553-022-00974-w>.

180. Austin PC. Informing power and sample size calculations when using inverse probability of treatment weighting using the propensity score. Stat Med 2021; 40: 6150-6163. <https://doi.org/10.1002/sim.9176>.

181. Gemeinsamer Bundesausschuss. Zusammenfassende Dokumentation über die Einstellung eines Beratungsverfahrens nach § 35a Absatz 3b des Fünften Buches Sozialgesetzbuch (SGB V); Marstacimab (Hämophilie A und B); Forderung einer anwendungsbegleitenden Datenerhebung und von Auswertungen; Wortprotokoll des Fachaustausches [online]. 2024 [Zugriff: 04.04.2025]. URL: https://www.g-ba.de/downloads/40-268-11018/2024-12-05_AM-RL-XII_Einstellung-Beratungsverfahren_Marstacimab_2023-AbD-010_ZD.pdf.

182. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching [online]. 2019 [Zugriff: 12.08.2025]. URL: <https://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-for-matching>.

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 enthaltenen 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 [117,166].
- 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 [119].
- 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 [167].
- 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 [130,168].

- 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 [121,169,170]. 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 [121].
- 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 [131]. 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 [123].

Anhang B Würdigung der Anhörung zum vorläufigen Rapid Report

Insgesamt wurden 27 Stellungnahmen zum vorläufigen Rapid Report frist- und formgerecht eingereicht.

Die im Rahmen der Anhörung vorgebrachten Aspekte wurden hinsichtlich valider wissenschaftlicher Argumente überprüft.

Die wesentlichen Argumente der Stellungnahmen werden in folgender Tabelle 3 gewürdigt. Literaturzitate aus dem Wortlaut der Stellungnahmen finden sich im separaten Dokument „Dokumentation der Anhörung zum vorläufigen Rapid Report“.

Zusätzliche Aspekte aus der wissenschaftlichen Erörterung zu Argumenten aus schriftlichen Stellungnahmen werden an den jeweils passenden Stellen gewürdigt. In der Erörterung wurden insbesondere die Darstellung und die Einordnung der orientierenden Betrachtung von Fallzahl-szenarien, die Schwelle für die verschobene Nullhypothese und alternative Methoden zu Propensity-Score-Verfahren in kleinen Patientenkollektiven diskutiert. Dies führte zu einer Überarbeitung in Abschnitt 5.3.1 (Parameter „Verschobene Nullhypothese“), einer Ergänzung in Abschnitt 5.3.2 (Abbildung 7 und „Einordnung der orientierenden Betrachtung von Fallzahl-szenarien“) und zu einer Klarstellung im Abschnitt 5.7.

Eine Zusammenfassung aller Änderungen im Rapid Report gegenüber dem vorläufigen Rapid Report, die sich u. a. durch die Anhörung zum vorläufigen Rapid Report ergeben haben, ist in Abschnitt 3.2 dargestellt.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
Allgemeine Aspekte		
1)	Bristol-Myers Squibb GmbH & Co. KGaA (BMS) [Allgemeine Aspekte] <u>Anmerkung:</u> [...] BMS [sieht] die grundsätzliche Machbarkeit einer AbD in bestimmten Konstellationen kritisch. In diesem Zusammenhang stellt die vom IQWiG erhobene Forderung, der Datenanalyse einen statistischen Test mit verschobener Nullhypothese zugrunde zu legen, aus Sicht von BMS weiterhin eine methodische Herausforderung dar [...]. Aus Sicht von BMS sollte – gemäß Gesetzesgrundlage – die valide Quantifizierung des Zusatznutzens stets das übergeordnete Ziel einer AbD darstellen [2]. In diesem Zusammenhang erscheint die Aussage des IQWiG, dass „[...] jedes Ergebnis einer AbD einen Erkenntnisgewinn dar[stellt] [...] losgelöst davon, ob sich [...] ein Vorteil des zu bewertenden Arzneimittels ableiten lässt“ (S. 14) diskussionswürdig.	<u>Keine Änderung im Rapid Report:</u> <i>Erkenntnisgewinn einer AbD</i> Die AbD und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V verfolgen 2 Ziele: ▪ die Generierung direkt vergleichender Evidenz im Versorgungsalltag und ▪ die Quantifizierung des Zusatznutzens. Aus Sicht des IQWiG stellt jedes Ergebnis einer AbD einen Erkenntnisgewinn dar, um die bestehende Evidenzlücke zu schließen. Der Erkenntnisgewinn beruht darauf, dass erstmals vergleichende Evidenz generiert wird, unabhängig davon, ob sich ein Vorteil des zu bewertenden Arzneimittels gegenüber der vom G-BA festgelegten zweckmäßigen Vergleichstherapie zeigt. Es ist richtig, dass es Ziel der AbD ist, eine valide Quantifizierung des Zusatznutzens zu erreichen. Zur Quantifizierung des Ausmaßes des Zusatznutzens kommt gemäß AM-NutzenV jedoch auch die Kategorie „Zusatznutzen nicht belegt“ in Betracht und nicht ausschließlich jene Zusatznutzen-Kategorien, die eine Überlegenheit des zu bewertenden Arzneimittels abbilden. Die AbD wird in Situationen gefordert, in denen es keine ausreichenden Daten zum Vergleich des zu bewertenden Arzneimittels mit dem Therapiestandard gibt. Es sind somit ausnahmslos Situationen, in denen sich nicht abschätzen lässt, mit welcher Wahrscheinlichkeit sich ein Vorteil des zu bewertenden Arzneimittels zeigen wird. Wenn bereits vor Start einer AbD mit hinreichender Sicherheit feststehen würde, dass sich ein solcher Vorteil zeigen wird, bräuchte es die AbD ggf. nicht.
2)	Bundesverband der Pharmazeutischen Industrie e. V. (BPI) [Allgemeine Aspekte] [...]	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Es ist hervorheben, dass das übergeordnete Ziel einer AbD die Quantifizierung eines Zusatznutzen ist. Die Aussage des IQWiG im vorliegenden vorläufigen Rapid Report, dass „[...] jedes Ergebnis einer AbD einen Erkenntnisgewinn dar[stellt] [...] losgelöst davon, ob sich [...] ein Vorteil des zu bewertenden Arzneimittels ableiten lässt“ (S. 14) steht nicht im Einklang mit der Gesetzesbegründung zum Gesetz für mehr Sicherheit in der Arzneimittelversorgung (GSAV) zu ihrer Einführung [...].	
3)	Deutsche Gesellschaft für Kinder- und Jugendmedizin e. V. (DGKJ) und Gesellschaft für Neuropädiatrie e. V. (GNP) [Zum Verfahrensablauf] <u>Patientenauswahl</u> Die Patientenauswahl für Verum- und Vergleichsgruppe muss auf Basis der Erfassung von in der Versorgungspraxis routinemäßig (nicht-interventionell) erhobener Parameter erfolgen können. Diese in der Versorgungspraxis routinemäßig (nicht-interventionell) erhobenen Parameter müssen eine ausreichende Risikofaktor- bzw. Schweregrad-Adjustierung der Verum- und Vergleichsgruppe für wesentliche Confounder ermöglichen. Wenn nicht beide Anforderungen erfüllt werden können, ist eine AbD unbeachtet Methoden-wissenschaftlicher Betrachtungen nicht durchführbar [...].	<u>Keine Änderung im Rapid Report:</u> Die Patientenpopulation, die vom G-BA im Forderungsbeschluss genannt und konkretisiert wird, bezieht sich auf das zugelassene Anwendungsgebiet. Adäquate Ein- und Ausschlusskriterien zur Abgrenzung der interessierenden Patientenpopulation (einschließlich Operationalisierung) sind durch den pU vor Beginn der AbD im Studienprotokoll festzulegen. Die Erhebung der erforderlichen Daten (z. B. zur Abgrenzung der für die AbD relevanten Patientenpopulation) muss in der gewählten Datenquelle sichergestellt sein. Die Anforderung, Daten versorgungsnah zu erheben, bedeutet nicht, dass sich die Datenerhebung nur auf solche Daten beschränkt, die im Versorgungsalltag per se erfasst werden. Für die Nutzenbewertung sind regelhaft Daten notwendig, die in der klinischen Praxis nicht bei allen Patientinnen und Patienten dokumentiert werden (z. B. Daten zur gesundheitsbezogenen Lebensqualität, zu Nebenwirkungen oder zu relevanten Confoundern) [18]. Zum Zwecke der Nutzenbewertung können somit Anpassungen an der (primären) Datenquelle für die notwendige Datenerhebung erforderlich sein. Dies wird in laufenden AbD-Verfahren auch bereits so umgesetzt.
4)	IGES Institut GmbH [Allgemeine Aspekte] <u>Anmerkung:</u> [...] Das übergeordnete Ziel der anwendungsbegleitenden Datenerhebung (AbD) ist die Quantifizierung eines Zusatznutzens. Im Rapid Report wird jedoch nicht ausreichend betont, dass die methodischen Anforderungen diesem Hauptzweck	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	dienen sollten. Die Aussage des IQWiG, dass „[...] jedes Ergebnis einer AbD einen Erkenntnisgewinn dar[stellt] [...] losgelöst davon, ob sich [...] ein Vorteil des zu bewertenden Arzneimittels ableiten lässt“ (S. 14) sollte dabei im Kontext der bisher angewendeten Effektschwellen zur Anerkennung eines quantifizierbaren Zusatznutzens diskutiert werden (1) [...].	
5)	Lilly Deutschland GmbH [Allgemeine Aspekte] <u>Anmerkung:</u> [...] Aus Sicht von Lilly ist eine praxistaugliche Ausgestaltung und Umsetzung der AbD essenziell für die Realisierung einer adäquaten Evidenzgenerierung. Zum aktuellen Zeitpunkt werfen jedoch die methodischen Vorgaben des G-BA und IQWiG auf Basis der bisher begrenzt verfügbaren Erfahrung mit AbD Fragen zur Umsetzbarkeit auf und wirken teilweise überhöht und praxisfern, so dass daraus eine Überforderung aller Beteiligten resultieren kann. In diesem Zusammenhang sollte der G-BA frühzeitig und verbindlich zum Thema AbD beraten, um die Planungssicherheit für forschende Pharmaunternehmen gewährleisten zu können. Mögliche Forderungen zur Einleitung einer AbD sollten frühzeitig mit dem betroffenen pU kommuniziert werden [...].	<u>Keine Änderung im Rapid Report:</u> <i>Praxistauglichkeit</i> Es liegt grundsätzlich in der Verantwortung des betroffenen pU, vergleichende Daten zu generieren, um eine identifizierte Evidenzlücke zu schließen. Das Verfahren der AbD kommt erst zum Tragen, wenn ein pU sich im Rahmen der Entwicklung eines neuen Arzneimittels dagegen entschieden hat, bis zum Zeitpunkt der frühen Nutzenbewertung eine aussagekräftige vergleichende Studie durchzuführen. In diesem Fall kann der G-BA vom pU eine AbD fordern, damit dieser eine bessere Datenbasis zur Bewertung des Zusatznutzens schafft. Es ist richtig, dass AbDs durch die geforderten Maßnahmen zur Sicherstellung der Datenqualität mit einem hohen Aufwand verbunden sein können. Diese Maßnahmen sind notwendig, um die Ergebnisunsicherheit, die sich ohnehin durch den nicht randomisierten Vergleich ergeben, zu verringern. Bei der Formulierung der Anforderungen an die AbD und von Auswertungen werden Möglichkeiten der Vereinfachung unter Abwägung einer damit einhergehenden größeren Unsicherheit berücksichtigt. Das Ziel muss jedoch stets sein, mit einer AbD aussagekräftige (verwertbare) Ergebnisse für eine Nutzenbewertung zu generieren. Die Anforderungen müssen folglich in dem vom pU erstellten und dem G-BA zur Überprüfung übermittelten Studienprotokoll und statistischen Analyseplan umgesetzt sein. Beispiele zu laufenden AbDs (z. B. zu Risdiplam oder Onasemnogen Apeparvovec) zeigen, dass die (methodischen) Vorgaben für eine adäquate Planung von vergleichenden Studien ohne Randomisierung zum Zwecke des Vergleichs von Therapieeffekten umsetzbar sind.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
		<i>Einbindung des pU</i> Die Erstellung eines AbD-Konzepts verläuft, wie auch bei anderen Produkten des IQWiG (z. B. Dossierbewertung) üblich, ohne Einbindung des pU. Das anschließende Beteiligungsverfahren zu einem AbD-Konzept liegt im Verantwortungsbereich des G-BA und wird durch ihn durchgeführt. Der Ablauf, einschließlich der Stellungnahmefrist, ist in der Verfahrensordnung des G-BA geregelt [26]. Über dieses Verfahren ist eine frühzeitige Einbindung des betroffenen pU möglich. Um Rückfragen bei der Erstellung der Studienunterlagen klären zu können, hat der pU vor Übermittlung der vorzulegenden Unterlagen an den G-BA zudem die Möglichkeit, eine Beratung beim G-BA unter Beteiligung des IQWiG zu beantragen.
6)	Lilly Deutschland GmbH [Weitere methodische Aspekte] [...] Zur Gewährleistung einer hohen Datenqualität verlangt das IQWiG [...] wesentliche Erweiterungen der bestehenden Indikationsregister, beispielsweise bzgl. der Confounder-Erhebung und -Adjustierung, der Erhebungsstruktur und Maßnahmen zur Vollständigkeit der Daten. Die hohen Qualitätsanforderungen werden jedoch in deutschen Patientenregistern derzeit selten erfüllt, so dass derartige Forderungen in der Realität mit einem überdurchschnittlich hohen Aufwand verbunden sein können und u. a. deshalb oftmals schwer umsetzbar sind. Die Erweiterung des Datensatzes um relevante Confounder, die derzeit in den verfügbaren Registern nicht ausreichend erfasst werden, sowie die Sicherstellung einer ausreichenden Fallzahl zur Beobachtung eines signifikanten Effekts stellen häufig erhebliche Herausforderungen für die Umsetzung von AbD dar [...]. <u>Fazit</u> Zusammengefasst ist aus Sicht von Lilly die Berücksichtigung der Praxistauglichkeit, eine Flexibilität aller Beteiligten, das Ausmaß der Anforderungen an Auswertungen sowie eine Bereitschaft zur Digitalisierung essenziell für eine erfolgreiche Umsetzung von AbDs. [...] Die Anforderungen	<u>Keine Änderung im Rapid Report:</u> zu Praxistauglichkeit siehe Würdigung in Zeile 5)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	müssen realisierbar, transparent und nachvollziehbar sein, so dass eine valide und für die Nutzenbewertung verwertbare Evidenzgrundlage zur Beantwortung der konkreten Fragestellung (unter Berücksichtigung der Besonderheiten des Anwendungsgebiets des Arzneimittels) geschaffen werden kann [...].	
7)	MSD Sharp & Dohme GmbH [Allgemeine Aspekte] [Wir] möchten [...] auf wesentliche Herausforderungen und Limitationen des derzeitigen Konzepts der Anwendungsbegleitenden Datenerhebung (AbD) hinweisen, die in der Praxis vielfach zu erheblichen Problemen führen und die Umsetzung in den meisten Fällen verhindern. Die überwiegende Zahl der bisherigen AbD-Verfahren wurde aus methodischen oder logistischen Gründen eingestellt, weil eine praxisorientierte Durchführung nicht möglich ist. MSD plädiert daher für eine pragmatische und flexible Weiterentwicklung des gesamten AbD-Konzepts, die den besonderen Herausforderungen der Real World Evidence Rechnung trägt [...].	<u>Keine Änderung im Rapid Report:</u> Es wurden bisher insgesamt 21 Verfahren zur Forderung einer AbD und deren Auswertung eingeleitet. Darunter wurde in 8 Verfahren (38 %) die Beratung zur Forderung einer AbD eingestellt. Somit wurde – entgegen der Anmerkung in der Stellungnahme – nicht die überwiegende Zahl der bisherigen AbD-Verfahren eingestellt. Gründe für die Einstellung der Beratungsverfahren waren: ▪ eine zu niedrige Patientenzahl (unter 100 Patientinnen und Patienten) mit unverhältnismäßig langer Rekrutierungszeit, ▪ fehlende Kooperation eines geeigneten Patientenregisters und ▪ dass die gesetzlichen Voraussetzungen für die Forderung einer AbD zum Zeitpunkt der Beschlussfassung nicht erfüllt waren. zu Praxistauglichkeit siehe Würdigung in Zeile 5)
8)	Sanofi-Aventis Deutschland GmbH [Allgemeine Aspekte] <u>Analytische und statistische Verfahren</u> <u>Anmerkung:</u> [...] Anstelle der Hervorhebung einzelner Verfahren wäre es zielführender, einen offenen, kontextabhängigen Ansatz für die Auswahl geeigneter statistischer Methoden zu empfehlen [...]. <u>Vorgeschlagene Änderung:</u> [...] Neben etablierten Ansätzen existiert eine Vielzahl neuartiger Methoden zur kausalen Inferenz, die – abhängig von Datenstruktur und Zielsetzung – wertvolle Alternativen zu PS-Verfahren (Matching, Gewichtung, Stratifizierung oder	<u>Änderung im Rapid Report:</u> <i>Alternative Methoden zu Propensity-Score-Verfahren</i> Wenngleich sich die Ausführungen im Rapid Report (gemäß Beauftragung des G-BA) auf Propensity-Score-Verfahren zur Confounderadjustierung in Anwendungsgebieten mit kleinen Patientenkollektiven beziehen, verschließt sich das IQWiG, wie auch im Rahmen der mündlichen Erörterung erläutert, nicht gegenüber weiteren Methoden zur Kontrolle von Confoundern, z. B. gegenüber den von den Stellungnehmenden genannten Verfahren. Entscheidend sind eine gute Planung, Durchführung und Dokumentation dieser Ansätze, um ihre Validität im Einzelfall bewerten zu können. Das IQWiG hat alle in der Stellungnahme genannten Referenzen gesichtet, jedoch keine Informationen oder Belege gefunden, die die

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Regressions-adjustierung) darstellen können. Dazu zählen optimierungsbasierte Verfahren [Cousineau, 2023], [Zubizarreta, 2012], [Cho, 2018], [Privitera, 2024], G-Computation [Shi, 2024], Targeted Maximum Likelihood Estimation (TMLE) [van der Laan, 2006], Random-Forest-basierte Ansätze [Breiman, 2001], usw [Zubizarreta, 2023]. [...] Die Methoden von [Hainmueller, 2012], [Zubizarreta, 2012], [Cho, 2018] und [Privitera, 2024] ermöglichen eine exakte Kovariatenbalance zwischen den Behandlungsgruppen, ohne dass ein PS-Modell erforderlich ist. Sie eignen sich besonders für kleine Stichproben und reduzieren die Modellabhängigkeit bei der Schätzung von Behandlungseffekten [...].	Aussage, dass diese Verfahren sich „besonders für kleine Stichproben“ eignen und „die Modellabhängigkeit bei der Schätzung von Behandlungseffekten“ reduzieren, stützen. Auch auf Nachfrage in der mündlichen Erörterung wurden von den Stellungnehmenden keine weiteren Referenzen genannt. Im Rapid Report, Abschnitt 5.7, wird klarer herausgestellt, dass das IQWiG neben Propensity-Score-Verfahren auch weitere Verfahren zur Confounderadjustierung ergebnisoffen prüfen wird, wenn sie von den pUs vorgelegt werden.
9)	Sanofi-Aventis Deutschland GmbH [Allgemeine Aspekte] <u>Datenquellen</u> <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> Die Fokussierung auf Patientenregister als primäre Datenquelle erscheint zu einseitig. Statt ein bestimmtes Quelletyp (Register) zu bevorzugen, wäre es zielführender, allgemeine Kriterien für „fit-for-purpose“-Datenquellen zu benennen und deren Eignung systematisch zu bewerten [FDA, 2023]. Routinedaten, andere Quellen als auch Daten-Linkage könnten unter bestimmten Voraussetzungen geeignet sein [...].	<u>Keine Änderung im Rapid Report:</u> Neben einer studienindividuellen Datenerhebung kommen für die versorgungsnaher Datenerhebung zum Zwecke der Nutzenbewertung nach § 35a SGB V derzeit Patientenregister infrage, da diese am ehesten die Möglichkeit der Anpassung an die notwendige Datenerhebung (mit Hinblick auf Spezifizierung der notwendigen Daten und Datenqualität) bieten. Andere Datenquellen, wie GKV-Routinedaten, bieten diese Option bisher nicht. Eine Datenverknüpfung (Datenlinkage) stellt dann eine Option dar, wenn für die Daten, die mit Registerdaten verknüpft werden sollen, dieselben Qualitätsanforderungen wie für das Primärregister erfüllen sind.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
10)	Vertex Pharmaceuticals (Germany) GmbH [Allgemeine Aspekte] Die im vorläufigen Rapid Report des IQWiG dargestellten methodischen Anforderungen können nicht pauschal auf alle Therapiearten angewandt werden. So sind die Vorgaben insbesondere in Studien zu Zell- und Gentherapien (cell and gene therapies, CGT) nicht immer sinnvoll umsetzbar oder mit besonderen Herausforderungen verbunden [...]. CGT [werden] häufig nur für eine eingeschränkte geeignete Patientenpopulation zugelassen. Die Eignung von Patientinnen und Patienten muss durch umfangreiche Tests sichergestellt werden, die in der Routineversorgung nicht standardmäßig durchgeführt werden [...]. Diese Besonderheit erschwert die Gewährleistung der geforderten Positivität, d. h. der Eignung aller Patientinnen und Patienten für beide untersuchten Therapieoptionen (Intervention und Komparator). Im vorläufigen Rapid Report wird der Umgang mit dieser Besonderheit nicht thematisiert [...].	<u>Keine Änderung im Rapid Report:</u> Im Rapid Report ist beschrieben, dass gemäß Konzept der Target-Trial-Emulation dann Positivität gegeben ist, wenn für alle Patientinnen und Patienten, die in die Auswertung eingehen, die zu vergleichenden Behandlungsformen eine grundsätzlich geeignete Therapieoption darstellen. Ein Einschlusskriterium, das lediglich eine der zu vergleichenden Gruppen betrifft, ist problematisch, weil es von vornherein die Positivität infrage stellt. Aufgrund des anwendungsbegleitenden Charakters der AbD ist, wie in der Prüfung von Studienprotokoll und statistischem Analyseplan zu Exagamglogen Autotemcel ausgeführt (Addendum A25-82 [171]), davon auszugehen, dass nicht alle Untersuchungen und Tests auch bei Patientinnen und Patienten ohne geplante Behandlung mit einer Zell- und Gentherapie durchgeführt werden. Die Einschätzung für die Eignung für eine Zell- und Gentherapie kann somit nicht in beiden Gruppen auf Basis identischer Informationen erfolgen. Für die Kontrollgruppe ist in solchen Fällen das Einschlusskriterium zu ergänzen, dass die Zell- und Gentherapie für die Patientin oder den Patienten nach Ermessen der behandelnden Ärztin oder des behandelnden Arztes eine geeignete Therapieoption darstellt. Da die Erfüllung der Positivität (als zentrale Annahme der kausalen Inferenz) eine wesentliche Voraussetzung für eine vergleichende Studie ohne Randomisierung ist und das unabhängig von der Behandlungsform, ergibt sich für den Rapid Report kein Änderungsbedarf.
Einleitende Anmerkungen zu den Anforderungen an vergleichende Studien ohne Randomisierung zum Zwecke der Nutzenbewertung		
11)	Roche Pharma AG [Abschnitt 5.1, S. 8] <u>Anmerkung:</u> Wenn das neue Medikament in der Behandlungsrealität zu einer Einnischung der Patienten führt, ist es nahezu unmöglich Positivität zu erreichen. So können z. B. Patienten mit Ausgangssituation x regelhaft das eine Medikament bekommen,	<u>Keine Änderung im Rapid Report:</u> In einer Studie zur Generierung versorgungsnaher Daten sind unabhängig von der potenziell zur Verfügung stehenden Patientenzahl im zu bewertenden Anwendungsgebiet die Ein- und Ausschlusskriterien so zu wählen, dass die Positivität erfüllt ist (siehe auch Würdigung in Zeile 10).

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	jene mit y regelhaft ein anderes Medikament zugewiesen bekommen. Dies ist nicht beeinflussbar und kann auch nicht durch bloßes zusätzliches Rekrutieren gelöst werden [...]. <u>Vorgeschlagene Änderung:</u> Der Rapid Report sollte auch solche Situationen beschreiben und einen Weg aufzeigen, wie bei nicht herstellbarer Positivität vorgegangen werden soll, und welche Kriterien zu welchen Entscheidungen führen [...].	
12)	SmartStep Consulting GmbH [Abschnitt 5.1, S. 7) <u>Anmerkung:</u> Die Einschlusskriterien für die Studie mit versorgungsnahen Daten (in dem Fall die Anwendungsbegleitende Datenerhebung [AbD]) sollten laut Tabelle 1 identisch zu einer hypothetischen randomisierten Studie (RCT) sein. Die Selektion der Patient*innen, deren Daten in die AbD einfließen können, würden dementsprechend zwei wesentlichen Einschränkungen unterliegen: Erstens müssen die Leistungserbringer*innen (Ärzt*innen oder Versorgungszentren) an der geforderten AbD, gemäß der Beschränkung der Versorgungsbefugnis teilnehmen (§ 35a Absatz 3b Satz 2 SGB V). Zweitens müssen die Patient*innen den Ein- und Ausschlusskriterien unterliegen, die analog zu den Vorgaben einer hypothetischen RCT definiert wären. Im Besonderen im Fall von Orphan-Drug(OD)-Verfahren, stellt es sich als herausfordernd dar, überhaupt ausreichend Patient*innen einzuschließen und die geforderten Fallzahlen zu erfüllen [...]. Zudem soll eine AbD Daten zum Nutzen des Arzneimittels in der Versorgung im deutschen Gesundheitskontext liefern. Eine Restriktion der Patient*innen über das Anwendungsgebiet gemäß Zulassung hinaus, wie es der Fall wäre, wenn Einschlusskriterien einer RCT übernommen bzw. simuliert werden würden, scheint dem zu widersprechen [...]. <u>Vorgeschlagene Änderung:</u> Die Einschlusskriterien sollten dem Anwendungsgebiet gemäß Zulassung entsprechen und maximal weitere verschreibungsrelevante Aspekte der Fachinformation umfassen. Dadurch kann gewährleistet werden, dass alle	<u>Keine Änderung im Rapid Report:</u> Es ist richtig, dass ein Großteil von vergleichenden Studien mit Randomisierung über ihre Ein- und Ausschlusskriterien eng definierte Patientenpopulationen einschließen und dadurch in ihrer externen Validität eingeschränkt sind. Eine Randomisierung geht aber nicht zwingend mit einer derartigen Einschränkung der Patientenpopulation einher. Eine (hypothetische) vergleichende Studie mit Randomisierung kann auch eine Population ohne relevante Einschränkungen einschließen, die möglichst viele Patientinnen und Patienten umfasst, die im Versorgungsalltag mit den zu untersuchenden Therapieoptionen behandelt werden (pragmatische RCTs) [172].

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Patient*innen, die vom Anwendungsgebiet umfasst werden und von dem Arzneimittel profitieren könnten, tatsächlich behandelt werden können [...].	
Confounderidentifikation und -auswahl (Themenbereich 1)		
13)	AMS Advanced Medical Services GmbH [Allgemeine Aspekte] [...] <u>Fazit:</u> Aus Sicht von AMS stellt eine (ggf. zeitlich eingeschränkte) systematische Recherche nach Sekundärliteratur in Kombination mit einer ergänzenden Recherche nach neuesten RCT eine sinnvolle Reduzierung des Aufwands ohne nennenswerten Informationsverlust dar und sollte bei künftigen Überlegungen zur Aufwandsreduzierung der Confounderidentifikation in Betracht gezogen werden.	<u>Keine Änderung im Rapid Report:</u> Die Prüfung einer möglichen Aufwandsreduktion bei der systematischen Informationsbeschaffung wird im Folgeauftrag GA25-02 [40] erfolgen.
14)	Deutsche Krankenhausgesellschaft e. V. (DKG) [Abschnitt 5.2, S. 9-12) <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> [...] Es fällt [...] auf, dass die Qualität der für die Confounderidentifikation herangezogenen Studien in dem vom IQWiG herangezogenen Verfahren nicht bewertet und berücksichtigt wird. Im Arbeitspapier wurde auf eine Indikation zurückgegriffen, für die es viele RCTs und Kohortenstudien gibt. In den Anwendungsgebieten, für die vom G-BA AbDs gefordert werden, ist das meist nicht der Fall und es müssen Studien niedriger Evidenz, herangezogen werden. Aus unserer Sicht sollte die Qualität der Quelle, aus der Confounder extrahiert werden, bei der Bewertung der Relevanz mit einbezogen werden.	<u>Keine Änderung im Rapid Report:</u> Der im Rapid Report beschriebene Ansatz, die Confounder unabhängig von der Qualität der Quelle zu extrahieren, ist ein sehr sensibler Ansatz. Der Vorschlag, die Qualität der Quelle, aus der die Confounder extrahiert werden, bei der Bewertung der Relevanz miteinzubeziehen, würde den Arbeitsaufwand für die Confounderidentifikation erhöhen, könnte jedoch den Gesamtaufwand (durch eine geringere Anzahl zu erhebender Confounder und ggf. besser konvergierende statistische Modelle) reduzieren. Dies wird im Folgeauftrag GA25-02 [40] geprüft.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
15)	Medizinischer Dienst Bund (KöR) [Abschnitt 5.2, S. 9-12] <u>Anmerkung:</u> [...] Der Einfluss von ungemessenem Confounding ist am stärksten, wenn keine Korrelation zwischen gemessenen und nicht gemessenen Confoundern vorliegt [2]. Daher möchten wir anregen, die Korrelation von potenziellen Confoundern bei der Auswahl der für eine AbD zu erhebenden Variablen stärker zu berücksichtigen. [...] Auch potenzielle Confounder-Variablen, die zunächst einmal nicht das gleiche Konstrukt messen sollen, können miteinander korreliert sein. Daher könnte in einem zusätzlichen Schritt durch Berücksichtigung dieser Korrelationen (auch anhand von Auswertungen in Registern) eine weitere Reduktion von Variablen erfolgen. [...] Zudem wird ausgeführt, dass der (anschließende) Ausschluss eines potenziellen Confounders aus inhaltlichen Überlegungen literaturgestützt begründet werden muss. Uns ist unklar, wie solch ein literaturgestützter Ausschluss erfolgen kann bzw. welche Untersuchungen hierzu erforderlich sind. [...] Um die Praktikabilität zu erhöhen, könnte aus unserer Sicht ein aufwandsärmeres Vorgehen, das in weniger zu erhebenden (und in Analysen zu berücksichtigenden) Variablen resultiert, hilfreich sein. Hierzu wäre denkbar, in Schritt 1 auf Faktoren zu fokussieren, die die Therapieentscheidung beeinflussen. Neben Interviews mit klinischen Expertinnen und Experten wäre es aus unserer Sicht sinnvoll, regelhaft Leitlinien (falls vorhanden) nach Faktoren, die die Therapieentscheidung beeinflussen, zu durchsuchen [3, 4]. Um die wechselseitigen Beziehungen der identifizierten Variablen und den Einfluss dieser Variablen auf die Exposition bzw. Therapieentscheidung und das Outcome zu visualisieren und für die Auswahl potenzieller Confounder zu berücksichtigen, kommt die Anwendung eines gerichteten azyklischen Graphen (DAG) in Frage [5]. Bekannte instrumentelle Variablen sollten nicht als potenzielle Confounder angesehen und nicht als Adjustierungsfaktoren berücksichtigt werden [6, 7]. [...] Vor diesem Hintergrund ist ein Anspruch auf Vollständigkeit der	<u>Keine Änderung im Rapid Report:</u> <i>Korrelation</i> Der im Arbeitspapier GA23-02 [27] beschriebene Schritt der Zusammenführung beruht auf der Annahme, dass die Confoundervariablen, die stark miteinander korrelieren, gebündelt betrachtet werden können. Eine weitergehende Berücksichtigung von Korrelationen von Confoundervariablen, die nicht das gleiche Konstrukt messen, ist plausibel und wird im Folgeauftrag GA25-02 [40] geprüft. <i>Weitere genannte Möglichkeiten der Aufwandsreduktion (z. B. DAG / Literaturrecherche)</i> Im Arbeitspapier GA23-02 wurden Möglichkeiten zur Aufwandsreduktion bei der systematischen Confounderidentifikation abgeleitet, die im vorliegenden Rapid Report berücksichtigt sind. Im Folgeauftrag GA25-02 werden weitere Maßnahmen zur Aufwandsreduktion bei der systematischen Identifikation und Erhebung von Confoundern geprüft. Dazu zählt z. B., inwieweit der Aufwand für die Literaturrecherche reduziert werden kann (siehe auch Würdigung in Zeile 13) und die Anwendung eines DAG zur Definition eines Minimalsets an Confoundern zur Aufwandsreduktion beitragen kann. <i>Methoden der Quantitativen Bias-Analyse oder Negativ-Kontrollen</i> In Abschnitt 5.2 des Rapid Reports geht es um den zuvor notwendigen Schritt der Confounderidentifikation im Rahmen der Studienplanung. Eine Abschätzung der Auswirkungen durch residuelles Confounding kann mittels der genannten Methoden, z. B. im Rahmen der Durchführung von Sensitivitätsanalysen, sinnvoll sein (zu Bias-Abschätzung siehe auch Würdigung in Zeile 42).

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Literaturrecherche und -analyse aus unserer Sicht zu überdenken. So könnte beispielsweise die Recherche in lediglich einer bibliographischen Datenbank (Medline) durchgeführt oder eine zeitliche Beschränkung der Recherche vorgenommen werden. Zudem schlagen wir vor, das Sättigungskriterium für RCTs und einarmige Studien zu reduzieren (beispielsweise auf einen Wert von 5). Alternativ (bei weiterer Anwendung eines Sättigungskriteriums von 10) möchten wir anregen, Variablen, die lediglich in einer Quelle als Patientencharakteristikum aufgeführt werden, nur dann als potenzielle Confounder anzusehen, wenn dies inhaltlich plausibel erscheint [...]. Um Auswirkungen durch residuelles / ungemessenes Confounding besser abschätzen zu können, kann es aus unserer Sicht zudem sinnvoll sein, beispielsweise Methoden der Quantitativen Bias-Analyse [8] oder Negativ-Kontrollen [9, 10] anzuwenden. <u>Vorgeschlagene Änderung:</u> ▪ Stärkere Berücksichtigung von Korrelationen zwischen Variablen zur Reduktion der Anzahl potenzieller Confounder ▪ Überdenken von oben genannten Vorschlägen zur Anpassung von Schritt 1 (der Identifikation von Confoundern)	
16)	MSD Sharp & Dohme GmbH [Abschnitt 5.2, S. 9-12] <u>Anmerkung:</u> [...] Vor dem Hintergrund des gesetzlich vorgegebenen engen Zeitrahmens für die Durchführung einer AbD – der nur wenige Monate bis zur Fertigstellung des Studienprotokolls und des statistischen Analyseplans umfasst – ist der im Arbeitspapier geforderte Aufwand zur systematischen Confounderidentifikation in der Praxis kaum zu bewältigen [...].	<u>Änderung im Rapid Report:</u> Wie in der wissenschaftlichen Erörterung erläutert, ist es wichtig, zwischen dem Aufwand im Vorfeld einer AbD (z. B. für die Identifikation und Verdichtung potenzieller Confounder) und dem Gesamtaufwand für die AbD (darunter fällt u. a. der Aufwand für die Datenerhebung) zu differenzieren. Es kann zielführend sein, im Vorfeld einen höheren Aufwand zu investieren, um die Machbarkeit der AbD durch Aufwandseinsparungen ab Beginn der Datenerhebung zu erhöhen. Dies wird im Rapid Report, Abschnitt 5.2, ergänzt. Erfahrungen aus bereits laufenden AbD-Verfahren zeigen, dass der zeitliche Vorlauf ausreichend ist. Im Arbeitspapier GA23-02 [27] wurden bereits Möglichkeiten zur Aufwandsreduktion bei der systematischen Confounderidentifikation abgeleitet, die im

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> [...] Es bedarf alternativer, pragmatischer Ansätze zum Umgang mit Confoundern, um deren Umsetzbarkeit innerhalb von AbD-Verfahren in der Praxis zu verbessern.	vorliegenden Rapid Report berücksichtigt sind. Weitere Optionen zur Aufwandsreduktion werden im Folgeauftrag GA25-02 [40] geprüft.
17)	Pfizer Pharma GmbH [Abschnitt 5.2, S. 9 ff.] <u>Anmerkung:</u> [Es] stellt sich die Frage, inwieweit sich der Ansatz auf Erkrankungen mit wenigen Patienten, die meist Gegenstand einer AbD sind, übertragen lässt. [...] Pfizer bezweifelt, dass die Erhebung zahlreicher potenziell relevanter Confounder, welche bei einer AbD im Rahmen einer Adjustierung zu berücksichtigen wären, in Registern bzw. Studienzentren mit der entsprechenden Datenqualität praktisch umsetzbar ist (z. B. fehlende Werte). Damit einher geht ein erhöhter Dokumentationsaufwand, welcher bereits in vergangenen AbD-Verfahren von Registern grundsätzlich als problematisch erachtet wurde [5, 6]. Weiterhin sollen gemäß den Ausführungen im Arbeitspapier GA23-02 Confoundervariablen, die ein inhaltlich gleiches Konstrukt messen oder sich inhaltlich überlappen, zusammengeführt werden. Pfizer hinterfragt, ob dies inhaltlich in allen Fällen zielführend möglich ist. [...] Des Weiteren möchte Pfizer darauf hinweisen, dass im vorliegenden Rapid Report eine Auseinandersetzung mit dem Konzept der Directed Acyclic Graphs (DAGs) fehlt. <u>Vorgeschlagene Änderung:</u> DAGs sind ein hilfreicher, effektiver und validierter Ansatz, um Confounding zu vermeiden. [...] Pfizer ist daher der Ansicht, dass dieser Ansatz im Rapid Report berücksichtigt und diskutiert werden sollte.	<u>Keine Änderung im Rapid Report:</u> <u>Übertragbarkeit</u> Auch wenn sich im Arbeitspapier GA23-02 der Bericht der externen Sachverständigen (wie auch die Arbeit von Pufulete 2022) auf eine Indikation mit umfangreicher Studienlage bezieht, wurde bei der Ableitung der Handlungsempfehlungen auch die Situation der selten(er)en Erkrankungen berücksichtigt, da dies die regelhafte Situation im Rahmen der AbD darstellt. Gleichzeitig zeigen laufende AbD-Verfahren, dass der Ansatz übertragbar ist [171]. <u>Dokumentationsaufwand</u> Im Arbeitspapier GA23-02 wurden Möglichkeiten zur Aufwandsreduktion bei der systematischen Confounderidentifikation abgeleitet, die im vorliegenden Rapid Report berücksichtigt sind. Im Folgeauftrag GA25-02 [40] werden weitere Optionen geprüft, die Anzahl identifizierter potenzieller Confounder zu verdichten, um den Dokumentationsaufwand zu reduzieren. <u>DAG</u> zu DAG siehe Würdigung in Zeile 15)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
18)	Sanofi-Aventis Deutschland GmbH [Abschnitt 5.2, S. 9 ff.] <u>Anmerkung:</u> [...] [Es] sollte berücksichtigt werden, dass moderne Verfahren existieren, die mit einer großen Anzahl von Kovariaten umgehen können, ohne eine strikte Unterscheidung zwischen Confoundern und anderen Variablentypen (z. B. Prädiktoren für die Behandlung oder für das Outcome) zu erfordern. Beispiele hierfür finden sich bei [Cousineau, 2023], [van der Laan, 2006], [Zubizarreta, 2023]. Der Einsatz solcher Verfahren kann den Aufwand für die Confounderidentifikation reduzieren, ohne die Validität der kausalen Schätzung zu gefährden. <u>Vorgeschlagene Änderung:</u> Ergänzung des Berichts um eine methodenoffene Perspektive, die moderne Verfahren zur Berücksichtigung von Confoundern beinhaltet [...].	<u>Keine Änderung im Rapid Report:</u> Der Rapid Report fokussiert auf den Schritt der Confounderidentifikation. Für die vom pU zitierten Referenzen, die verschiedene Verfahren der statistischen Analyse diskutieren, ist zu berücksichtigen, dass dies zum einen die Erhebung aller Variablen (unabhängig davon, ob Kovariate oder Confounder) in hoher Qualität und zum anderen eine ausreichende, für die Analysen zur Verfügung stehende, Patientenzahl voraussetzt. Es obliegt dem pU, eine angemessene statistische Methode auszuwählen und zu präspezifizieren.
19)	Servier Forschung und Pharma-Entwicklung GmbH [Abschnitt 5.2, S. 9-11] <u>Anmerkung:</u> [...] Die ausschließliche Bezugnahme auf GA23-02 und den Ansatz nach Pufulete 2022 wird aus Sicht aktueller internationaler Methodenguidelines, wie Good Practices for Real-World Data, FDA Real-World Evidence Program Guidance und EMA Guideline, als nicht ausreichend angesehen. Es fehlt eine Einordnung möglicher Limitationen dieser Vorgehensweise sowie eine Diskussion alternativer oder zusätzlicher Verfahren, die zur Akzeptanz der Confounderadjustierung beitragen können. Darüber hinaus wird im Rapid Report zutreffend beschrieben, dass in bestimmten Versorgungssituationen eine vollständige Identifikation und Adjustierung potenzieller Confounder aufgrund begrenzter Datenverfügbarkeit nicht immer realisierbar ist. In solchen Fällen kann eine transparente Darstellung methodischer Restriktionen zur Einordnung der Ergebnisse beitragen, ohne den Erkenntnisgewinn grundsätzlich infrage zu stellen [...].	<u>Keine Änderung im Rapid Report:</u> Das beschriebene Vorgehen deckt sich mit den aktuellen FDA- und EMA-Guidelines [173,174]. Im Rapid Report wird beschrieben, dass in bestimmten Versorgungssituationen ein abweichendes Vorgehen erforderlich ist, z. B. kann beim Fehlen von geeigneten RCTs und Kohortenstudien auf 1-armige Studien oder Leitlinien zurückgegriffen werden. Wir stimmen zu, dass methodische Restriktionen transparent darzustellen sind. In Abhängigkeit vom Ausmaß der methodischen Restriktionen (z. B. mehrere wichtige Confounder sind im Datensatz nicht enthalten), ist es möglich, dass ein Zusatznutzen nur bei Vorliegen von sehr großen Effekten im Sinne eines dramatischen Effekts abgeleitet werden kann.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Ergänzung um eine Einordnung, dass die Identifikation und Adjustierung von Confoundern in versorgungsnahen Studien methodische Herausforderungen mit sich bringen kann, die Diskussion dieser Herausforderungen den Erkenntnisgewinn jedoch nicht infrage stellen. Hinweise zum Umgang mit residualer Verzerrung sowie die Möglichkeit der Nutzung datengetriebener Verfahren könnten zur methodischen Absicherung beitragen.	
20)	SmartStep Consulting GmbH [Abschnitt 5.2, S. 12] <u>Anmerkung:</u> Die vollständige Erhebung aller relevanter Confounder kann zu sehr umfangreichen Fragebögen zu Studienbeginn führen. Dies erhöht die Wahrscheinlichkeit für fehlende Werte [...]. <u>Vorgeschlagene Änderung:</u> Für sehr umfangreiche Sets von Confoundern kann eine Reduktion auf die wichtigsten Elemente die Datenqualität erhöhen und das Risiko fehlender Werte gesenkt werden. Diese Reduktion des Confoundersets kann zum Beispiel durch die Verwendung gerichteter azyklischer Graphen [...] vorgenommen und diskutiert werden.	<u>Keine Änderung im Rapid Report:</u> zu Dokumentationsaufwand siehe Würdigung in Zeile 17)
21)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.2, S. 9 ff.] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> [...] Möchte man auch anhand nicht randomisierter Studien qualitativ hochwertige Evidenz in einem relevanten Zeitraum generieren, ist ein praktikables Verfahren zur Confounderidentifikation notwendig. Nach Möglichkeit sollten Scheinkorrelationen identifiziert und nicht berücksichtigt werden, gesicherte Kausalzusammenhänge wären regelhaft zu berücksichtigen.	<u>Keine Änderung im Rapid Report:</u> zu Übertragbarkeit siehe Würdigung in Zeile 17) <i>Aktualität der Confounderrecherche</i> Die beschriebene Anforderung an die Aktualität der Recherche bezieht sich auf die Suche nach Studien mit dem zu bewertenden Arzneimittel bei der Nutzenbewertung gemäß § 35a SGB V. Dies entspricht jedoch nicht der geforderten Aktualität der Recherche für die Identifikation von Confoundern. Es ist derzeit unklar, welche Auswirkungen die Aktualität der herangezogenen Publikationen bzw. Studien auf die Confounderidentifikation hat. Dies wird im Folgeauftrag GA25-02 [40] geprüft.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Möglichkeiten zur weiteren Reduktion sollten [...] weiterführend diskutiert werden. Ob das dargestellte Verfahren im Bereich seltener Erkrankungen anwendbar ist, wäre ebenfalls noch zu prüfen. [...] Im Gegensatz zu einer AbD steht der pU vor der Dossiereinreichung [...] vor der Herausforderung, dass die systematische Recherche nicht älter als 3 Monate sein darf und somit wiederholt werden muss. Für den Fall, dass dann neue Confounder identifiziert werden, müssten diese auch wieder von Experten evaluiert und in den Analysen im Dossier entsprechend berücksichtigt werden, was in diesem kurzen Zeitraum sehr herausfordernd ist [...].	In den bisherigen AbD-Verfahren lag die Recherche der pUs in der Regel lediglich 3 bis 6 Monate zurück. Dies wird als sachgerecht beurteilt.
22)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.2, S. 11] <u>Anmerkung:</u> Die Auswahl der Expertinnen und Experten erscheint nicht strukturiert vorgenommen worden zu sein. Es sollte klarer werden, wie Expertenmeinungen strukturiert herangezogen werden können [...]. <u>Vorgeschlagene Änderung:</u> Weiterführende Diskussion zur Strukturierung der Expertenidentifikation, z. B. könnte durch eine Versorgungsstrukturanalyse sichergestellt werden, dass auch Ärztinnen und Ärzte Einfluss nehmen können, die nicht an Leitlinien arbeiten, sondern auch außerhalb der großen Zentren die Versorgungsrealität prägen. Mittels Patientenpfades könnte ermittelt werden, welche Facharztgruppen an der Diagnose und Behandlung teilhaben.	<u>Keine Änderung im Rapid Report:</u> Es ist davon auszugehen, dass klinische Expertinnen und Experten bereits aus der Studienplanung und -durchführung bekannt sind. Weitere klinische Expertinnen und Experten können über Fachgesellschaften oder Leitlinien identifiziert werden. Insbesondere in Indikationen mit geringen Patientenzahlen, welche für die AbD den Regelfall darstellen, ist die Zahl klinischer Expertinnen und Experten absehbar gering. Umso mehr ist es in diesen Fällen von Bedeutung, dass mögliche Interessenkonflikte offengelegt werden.
Schätzung notwendiger Fallzahlen vor dem Hintergrund unzureichender Vorabinformationen (Themenbereich 2)		
23)	AbbVie Deutschland GmbH & Co. KG [Abschnitt 5.3, S. 12 ff.] <u>Anmerkung:</u> Es ist anzumerken, dass auch im Falle der Fallzahlplanung für eine RCT in der Praxis nicht immer Daten vorliegen, die exakt den Bedingungen der geplanten RCT entsprechen. In solchen Fällen wird die bestmögliche vorliegende Evidenz genutzt, um eine methodisch möglichst robuste, aber ethisch vertretbare Fallzahl	<u>Keine Änderung im Rapid Report:</u> Im Rapid Report ist beschrieben, dass die Fallzahlplanung für eine RCT in der Regel (nicht immer) auf Informationen aus Pilotstudien oder vergleichenden Phase-II-Studien basiert (zum ethischen Aspekt siehe Würdigung in Zeile 25). Es ist richtig, dass eine orientierende Fallzahlschätzung, insofern sie im Rahmen eines AbD-Verfahrens bei gegebener Datenkonstellation als sinnvoll durchführbar erachtet wird,

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	zu berechnen. Dabei wird in Kauf genommen und entsprechend als Unsicherheit berücksichtigt, dass es zwischen den zugrunde gelegten Studiendaten (z. B. Phase II Studien) und der zu planenden Studie Abweichungen (z. B. bezüglich der Populations-Charakteristika) geben kann. Dies kann sowohl die im klinischen Entwicklungsprogramm bislang generierten Daten zur Intervention als auch den geplanten Komparator betreffen, zu dem in der Regel nur veröffentlichte Studiendaten bekannt sind. [...] Daraus ergibt sich eine Fallzahlabeschätzung auf möglichst umfassenden Informationen, die im Zuge einer Interimsanalyse angepasst werden kann. Auch die AbD könnte von diesem pragmatischeren Ansatz profitieren. Zwar sind die Fallzahlen mit erhöhter Unsicherheit behaftet, können jedoch einen ersten groben Aufschluss über die grundsätzliche Machbarkeit vor Beginn der AbD liefern [...]. <u>Vorgeschlagene Änderung:</u> Die grundsätzliche Machbarkeit einer AbD sollte <u>immer</u> vorab mittels einer Fallzahlplanung geprüft werden. Das Konzept der orientierenden Betrachtung von Fallzahl Szenarien ist dazu nicht geeignet, da dieses Konzept keinerlei Aufschluss über die eigentlich benötigte Fallzahl zur Detektion eines für die Nutzenbewertung relevanten Effektes gibt, und sollte daher gestrichen werden. Zur Fallzahlplanung sollte ein pragmatischer Ansatz verfolgt und analog zu klinischen Studien die beste verfügbare Evidenz genutzt werden [...].	aufgrund der im Rapid Report beschriebenen Unsicherheiten allenfalls eine grobe Einschätzung der Größenordnung der benötigten Fallzahl liefern kann. 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 im Regelfall jedoch nicht vor, sodass die Voraussetzungen für eine orientierende Fallzahlschätzung nicht erfüllt sind (siehe Rapid Report, Abschnitt 5.3.2).
24)	AbbVie Deutschland GmbH & Co. KG [Abschnitt 5.3, S. 12 ff.] <u>Anmerkung:</u> Es erscheint fraglich, ob bei der Fallzahlplanung für eine AbD der Fokus auf die Gewinnung von Effektschätzungen ohne den Nachweis eines Vorteils des zu bewertenden Arzneimittels tatsächlich einen Erkenntnisgewinn darstellt.	<u>Keine Änderung im Rapid Report:</u> Weder 1-armige Studien noch Studien zum (direkten) Vergleich des zu bewertenden Arzneimittels gegenüber einem Komparator, der keine Therapieoption der zweckmäßigen Vergleichstherapie darstellt, sind allein geeignet, eine bestehende Evidenzlücke zu schließen und so die Datenbasis für eine erneute Nutzenbewertung zu verbessern (zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1). Für eine (erneute) Nutzenbewertung sind zunächst alle verwertbaren Ergebnisse zu patientenrelevanten Endpunkten aus einer aussagekräftigen Studie zum Vergleich des zu bewertenden

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Wenn die Fallzahlplanung einer AbD nicht mehr darauf ausgerichtet wäre, einen Vorteil des Medikaments zu zeigen, wäre innerhalb der AbD ein nicht signifikanter Überlegenheitstest zu erwarten [...]. Dies wirft die Frage auf, welchen Erkenntnisgewinn Effektschätzungen bieten, wenn diese statistisch nicht gerechtfertigt sind und keine Aussagekraft haben [...]. Das Ziel einer AbD ist es, Erkenntnisse zu sammeln, die im Rahmen von Nutzenbewertungen nach § 35a SGB V genutzt werden können. Angesichts der bisher stark evidenzbasierten Methodik des IQWiG ist es überraschend, dass solche Effektschätzer, die keine Aussage bzgl. Signifikanz (also Überlegenheit) liefern, überhaupt Teil einer Nutzenbewertung werden könnten. Die bisherigen Methoden des IQWiG bieten hierfür keinen Ansatz. Zudem liegen in diesen Anwendungsbereichen [...] oft Daten aus einarmigen Studien oder vergleichenden Studien, deren Komparator nicht Teil der ZVT ist, vor. Diese Studien, mit einem höheren Evidenzgrad als eine AbD, könnten theoretisch und sollten vorrangig für solche Bewertungen verwendet werden. Dazu sollten methodische Grundlagen (außerhalb der kaum akzeptierten indirekten Vergleiche und dramatischer Effekte) geschaffen werden, bevor Daten mit geringerem Evidenzgrad in die Nutzenbewertung eingehen. <u>Vorgeschlagene Änderung:</u> Insgesamt muss die Fallzahlplanung einer AbD auf den Nachweis eines Vorteils einer Therapieoption ausgerichtet sein. Der Abschnitt zur Fallzahlplanung auf Basis eines (nicht vorhandenen) Evidenzgewinns unabhängig von einem Therapieoptionenvorteil sollte gestrichen werden, und stattdessen die Bedeutung der Ausrichtung der Fallzahlplanung auf einen statistischen Test betont werden.	Arzneimittels mit dem Therapiestandard relevant. Dazu gehören nicht nur statistisch signifikante Effekte (zum Vor- oder Nachteil eines zu bewertenden Arzneimittels). Ein statistisch nicht signifikantes Ergebnis ist dabei jedoch nicht als Evidenz für das Nichtvorhandensein eines Effekts zu werten (im Sinne einer Äquivalenz). Für die Ableitung des Zusatznutzens werden hingegen die positiven und negativen statistisch signifikanten Effekte gegenübergestellt.
25)	AbbVie Deutschland GmbH & Co. KG [Abschnitt 5.3, S. 12 ff.] [...] Ebenso fraglich ist die ethische Komponente für die Patienten, die bei der Fallzahlplanung klinischer Studien eine Rolle spielt und auch in der AbD berücksichtigt werden muss. Es ist keineswegs so, dass innerhalb einer AbD nur Routinedaten erhoben werden. Im Gegenteil, die höheren Anforderungen an eine AbD (z. B. im Hinblick auf Patient Reported Outcomes) zur Generierung nutzbarer	<u>Keine Änderung im Rapid Report:</u> In einer klinischen Studie (RCT) unterliegt die Fallzahlplanung ethischen Anforderungen: Weder eine zu hohe Fallzahl (zu viele Patientinnen und Patienten werden unnötigerweise einer nicht zugelassenen Therapie [mit unbekannter Nutzen-Schaden-Bilanz] ausgesetzt) noch eine zu kleine Fallzahl (liefert keine aussagekräftigen Ergebnisse) sind im Kontext klinischer Studien akzeptabel [175,176].

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Daten für die Nutzenbewertung erfordern ein größeres Maß an Mitarbeit der Patienten, die im Allgemeinen schwer krank sind und sich in einer Situation mit einhergehender Unsicherheit und psychischem Stress befinden. Von diesen Patienten in dieser schwierigen Therapiesituation wird ein erhöhter Aufwand verlangt, ohne dass jedoch, wie in klinischen Studien, vorab zumindest ein quantifizierbarer Vorteil für die Allgemeinheit (ausgedrückt durch ein Fallzahlplanung und der Abschätzung zur Realisierbarkeit der Fallzahl) absehbar ist.	Im Kontext einer AbD werden die Daten in der Versorgungspraxis nach Entscheidungsfindung für eine Behandlung mit dem zu bewertenden Arzneimittel oder eine Behandlung mit dem therapeutischen Standard erhoben. Sowohl das Arzneimittel, für das die AbD gefordert wurde, als auch die Therapien, die als Komparatoren eingesetzt werden, sind in der Regel zugelassen und erstattungsfähig. Vor diesem Hintergrund sind ethische Bedenken in einer AbD mit Hinblick auf die Schätzung der notwendigen Fallzahl auszuschließen.
26)	AbbVie Deutschland GmbH & Co. KG [Abschnitt 5.3, S. 12 ff.] <u>Anmerkung:</u> Ausgehend von der Prämisse, dass die Diskussion zur Machbarkeit einer AbD wichtig, notwendig und in der Verfahrensordnung des G-BA [...] vorgesehen ist, und diese Machbarkeit von der benötigten Fallzahl abhängt, ist das Konzept der „orientierenden Betrachtung von Fallzahlsszenarien“ fragwürdig. Die durchgeführten Berechnungen sind keineswegs eine orientierende Betrachtung der Fallzahlsszenarien, sondern vielmehr generische Berechnungen aufdeckbarer Effekte, die vom Kontext der konkreten AbD und der Indikation losgelöst sind. Bei dieser Darstellung des IQWiG fehlt die tatsächliche Einordnung der Ergebnisse in den Kontext der spezifischen Indikation. Solange die „aufdeckbaren Effekte“ nicht in Beziehung zu den erwartbaren Effekten aus der besten verfügbaren Evidenz gesetzt werden, sind sie nicht interpretierbar [...]. Die orientierende Betrachtung von Fallzahlsszenarien liefert schlussendlich keine Aussage über die Durchführbarkeit einer AbD, da die betrachteten Szenarien theoretischer Natur sind und durch die getätigten Annahmen noch mehr durch Unsicherheiten belastet sind als eine Fallzahlplanung basierend auf Studiendaten. Es werden Annahmen getroffen; diese können im Rahmen der AbD aber nur schwerlich bis gar nicht diskutiert werden, dazu gehört das Verhältnis zwischen Interventions- und Kontrollarm und insbesondere, ob ausreichend Patienten in Registern rekrutierbar und somit für eine AbD verfügbar sind. Die Frage, ob die verfügbare Zahl an Patienten geeignet ist, Daten für die Nutzenbewertung zu	<u>Keine Änderung im Rapid Report:</u> <i>Abschätzung der Patientenzahl</i> Zunächst ist klarzustellen, dass es in der Verantwortung des betroffenen pU liegt, eine ausreichende Patientenzahl zu rekrutieren, um sicherzustellen, dass grundsätzlich aussagekräftige Ergebnisse für die Nutzenbewertung generiert werden. Wie viele Patientinnen und Patienten von der potenziell für eine AbD zur Verfügung stehenden Patientenzahl tatsächlich rekrutiert (eingeschlossen) werden können, hängt von unterschiedlichen Faktoren ab und ist im Rahmen der Konzepterstellung in der Regel nicht sinnvoll beurteilbar. Dies ist unter anderem darauf zurückzuführen, dass das neue Arzneimittel zum Zeitpunkt der Konzepterstellung in der Regel noch nicht zugelassen und auf dem Markt verfügbar ist. <u>Änderung im Rapid Report:</u> <i>Einordnung der orientierenden Betrachtung von Fallzahlsszenarien</i> Die orientierende Betrachtung von Fallzahlsszenarien erfolgt nicht losgelöst vom zu bewertenden Anwendungsgebiet: Zum einen fließt in die orientierende Betrachtung von Fallzahlsszenarien die für das jeweilige Anwendungsgebiet geschätzte Patientenzahl in Deutschland als Parameter ein, zum anderen wird (je nach Verfügbarkeit) das Basisrisiko unter Behandlung mit der Intervention oder unter Behandlung mit dem Komparator für einen patientenrelevanten Endpunkt als Anker herangezogen, um das Spektrum möglicher

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	generieren, wird unter diesen zusätzlichen methodischen Unsicherheiten gerade nicht beantwortet [...]. <u>Vorgeschlagene Änderung:</u> Eine orientierende Betrachtung von Fallzahlszenarien, die ohne die Einordnung der Ergebnisse (aufdeckbarer Effekt versus realistische, in einer AbD zeigbare Effekte) und ohne Diskussion der zur Abschätzung genutzten Parameter (z. B. Population im Anwendungsgebiet versus realistisch in die AbD rekrutierbare Patienten) erfolgt, kann keinerlei Aussagen zum notwendigen Umfang der Datenerhebung treffen und ist daher nicht sinnvoll [...].	Fallzahlszenarien im jeweiligen Anwendungsgebiet aufzuzeigen. Eine Aussage über die Wahrscheinlichkeit, dass in einer AbD ein Effekt gezeigt werden kann, ist unabhängig vom gewählten methodischen Vorgehen für die Konzepterstellung (orientierende Fallzahlschätzung oder orientierende Betrachtung von Fallzahlszenarien) nicht möglich, da der wahre Effekt und die Größe der Verzerrung unbekannt sind. Ergänzend sei darauf hingewiesen, dass nicht der wahre Therapieeffekt in der Größenordnung des aufdeckbaren Effekts liegen muss, sondern der in einer AbD beobachtete Effekt (unter Berücksichtigung der Verzerrung aufgrund des nicht randomisierten Vergleichs). Zur Einordnung der orientierenden Betrachtung von Fallzahlszenarien werden im Konzept die Ergebnisse für einen patientenrelevanten Endpunkt aus vorhandenen Studien zur Intervention und den Komparatoren dargestellt. Dies wird im Rapid Report, Abschnitt 5.3.2, bei Darstellung des Beispielfahrens, in dem eine orientierende Betrachtung verschiedener Fallzahlszenarien erfolgte, ergänzt.
27)	AbbVie Deutschland GmbH & Co. KG [Abschnitt 5.3, S. 12 ff.] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> Auch wenn sich Ungewissheiten nicht vollständig auflösen lassen, muss zur Abwägung der Machbarkeit einer AbD realistisch abgeschätzt werden, wie viele Patienten aus dem Anwendungsgebiet in angemessener Zeit im Sinne der Nutzenbewertung in die AbD eingeschlossen werden können. Jegliche orientierende Abschätzung in Form von Fallzahlszenarien auf Basis der theoretisch verfügbaren Patientenzahl im Anwendungsgebiet bietet keinen Erkenntnisgewinn und sollte daher gestrichen werden.	<u>Keine Änderung im Rapid Report:</u> zur Abschätzung der Patientenzahl siehe auch Würdigung in Zeile 26) Wie im Rapid Report, Abschnitt 5.3.2, erläutert, ist in jenen Fällen, in denen aufgrund unzureichender Vorabinformationen keine sichere Abschätzung der erwarteten Effektgröße möglich ist, davon auszugehen, dass für die Patientenzahl als Parameter noch die verlässlichste Schätzung erfolgen kann. Um einer potenziellen Unsicherheit der Parameterschätzung Rechnung zu tragen, wird für die orientierende Betrachtung von Fallzahlszenarien üblicherweise eine Spanne von Patientenzahlen herangezogen.
28)	AbbVie Deutschland GmbH & Co. KG [Abschnitt 5.3, S. 12 ff.]	<u>Keine Änderung im Rapid Report:</u> Unabhängig vom gewählten methodischen Vorgehen für die Konzepterstellung (orientierende Fallzahlschätzung oder orientierende

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Anmerkung:</u> Der Vorschlag, Interimsanalysen zur Fallzahlplanung zu verwenden, ist begrüßenswert und bei der Durchführung einer AbD auch bereits regelhaft vorgesehen. Allerdings ist es auch im Falle geplanter Interimsanalysen sinnvoll, vor Beginn der AbD mit Hilfe der verfügbaren Evidenz zu prüfen, ob eine mögliche Fallzahl zumindest grob innerhalb einer AbD realisierbar ist. Erst wenn dies der Fall ist, sollte die AbD durchgeführt und diese Fallzahl mit Hilfe einer Interimsanalyse besser eingegrenzt bzw. angepasst werden. Dazu ist jedoch zumindest eine Fallzahlplanung vor Beginn der AbD notwendig; eine orientierende Betrachtung von Fallzahlszenarien kann dies nicht leisten. Selbst eine konkrete Fallzahlabeschätzung vor Start einer AbD gewährleistet nicht, dass im Rahmen einer Interimsanalyse eine Fallzahlplanung durchgeführt werden kann: Weder in der laufenden AbD zu Onasemnogen-Abeparvovec (Zolgensma®, spinale Muskelatrophie) noch zu Brexucabtagen autoleucel (Tecartus®, Mantelzelllymphom) waren die bisher vorgelegten Zwischenanalysen „reif“ genug dafür, obschon das IQWiG im Vorfeld der Forderung einer AbD im AbD-Konzept eine Fallzahlabeschätzung durchgeführt hatte [...]. <u>Vorgeschlagene Änderung:</u> Interimsanalysen zur Aktualisierung der Fallzahlplanung sind wichtig. Sie ersetzen jedoch nicht eine Fallzahlplanung vor Beginn der AbD, die abschätzen muss, ob eine AbD überhaupt durchführbar ist. Das Konzept der orientierenden Betrachtung von Fallzahlszenarien ist dazu nicht geeignet und sollte gestrichen werden.	Betrachtung von Fallzahlszenarien), sollte der betroffene pU die (finale) Fallzahlplanung zum Zeitpunkt von Zwischenanalysen vornehmen, d. h. zu einem Zeitpunkt, zu dem die bis zu diesem Zeitpunkt prospektiv erhobenen Daten zum Vergleich des neuen Arzneimittels mit dem therapeutischen Standard bezüglich patientenrelevanter Endpunkte in die Schätzung eingehen können. Insofern eine finale Fallzahlplanung zum Zeitpunkt der 1. Zwischenanalyse nicht möglich ist, ist durch den pU in den Studienunterlagen eine transparente und nachvollziehbare Begründung vorzulegen. Bei Bedarf kann durch den G-BA im Anschluss eine Anpassung des Forderungsbeschlusses erfolgen. In den laufenden AbDs zu Onasemnogen Abeparvovec [177] und Brexucabtagen Autoleucel [178] ist geplant, dass die finale Fallzahlplanung jeweils zum Zeitpunkt der 2. Zwischenanalyse vorgelegt wird.
29)	Bayer Vital GmbH [Abschnitt 5.3.1, S. 14] Zitat: „Ein Aufgreifkriterium für eine AbD ist eine unzureichende Evidenzlage für ein neues Arzneimittel im Vergleich zum therapeutischen Standard. [...] Unter dieser Voraussetzung stellt jedes Ergebnis einer AbD einen Erkenntnisgewinn dar“.	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Anmerkung:</u> Dies mag wissenschaftlich zwar stimmen, verkennt aber die gesetzlich intendierte Veranlassung einer AbD mit dem expliziten Ziel einer Quantifizierung des Zusatznutzens von bspw. Orphan Drugs im Rahmen einer anwendungsbegleitenden Datenerhebung [...]. <u>Vorgeschlagene Änderung:</u> Ab „vor diesem Hintergrund ...“ Passus löschen.	
30)	Bayer Vital GmbH [Abschnitt 5.3.1, S. 16 ff.] <u>Anmerkung:</u> Da anders als in einer RCT über Confounderkontrolle nur „obvious bias“ kontrolliert werden kann, ein Teil potenziellen Bias aber als „hidden“ und damit als nicht kontrollierbar auftreten kann, ist damit zu rechnen, dass keine idealtypisierte Strukturgleichheit der Vergleichsarme über eine AbD erzielt werden kann, sondern letztere über entsprechende statistische Verfahren approximiert wird. Dies induziert eine inhärente Heterogenität der Vergleichsarme, die somit zu einer geringeren Präzision der daraus abgeleiteten Effekte führt. [...] Aufgrund der [...] inhärent bedingten größeren Heterogenität zwischen den Studienarmen im Rahmen der AbD ist die verschobene Nullhypothese hier nicht zielführend und es sollte in Anerkennung dieser Konstellation primär auf die Effektschätzer als solche, insofern statistisch signifikant, zur Ableitung eines endpunktbezogenen Zusatznutzens abgestellt werden. <u>Vorgeschlagene Änderung:</u> Die verschobene Nullhypothese nicht im Rahmen der AbD zur Ableitung eines endpunktbezogenen Zusatznutzens anwenden und direkt aus den Effektschätzern den endpunktspezifischen Zusatznutzen ableiten.	<u>Änderung im Rapid Report:</u> Das Vorgehen, dass ein 95 %-Konfidenzintervall einen zu definierenden Schwellenwert (Test auf verschobene Nullhypothese) überschreiten (oder unterschreiten) soll, hat gegenüber der Betrachtung von Punktschätzern den wesentlichen Vorteil, dass die Präzision der Schätzung in die Bewertung einfließt und auf diese Weise auch die Irrtumswahrscheinlichkeit (Wahrscheinlichkeit für einen Fehler 1. Art) auf 5 % beschränkt wird. Vor dem Hintergrund methodischer Weiterentwicklungen im Kontext nicht randomisierter vergleichender Studien aus Beobachtungsdaten erscheint es angebracht, sich einer Prüfung der im Rapid Report A19-43 [18] vorgeschlagene Schwelle für die verschobene Nullhypothese anzunehmen. Dies wird in Abschnitt 5.3.1, Parameter „Verschobene Nullhypothese“, des Rapid Reports aufgenommen.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
31)	Bristol-Myers Squibb GmbH & Co. KGaA (BMS) [Abschnitt 5.3.1, S. 12 ff.] <u>Anmerkung:</u> <u>Verschobene Nullhypothese</u> Das IQWiG hält im vorliegenden vorläufigen Rapid Report weiterhin an statistischen Tests mit verschobener Nullhypothese fest. Aus Sicht von BMS ist dieser methodische Ansatz kritisch zu bewerten. Der vorgeschlagene Schwellenwert für das relative Risiko im Bereich von 2 bis 5, bezogen auf die untere Grenze des Konfidenzintervalls, ist wissenschaftlich nicht etabliert und berücksichtigt weder die Art der Erkrankung noch den zugrunde liegenden Endpunkt. Vor dem Hintergrund der für die AbD vorgesehenen parallelen, prospektiven Datenerhebung, einer einheitlichen Messung von Endpunkten und der Berücksichtigung von Confoundern erscheint dieser Ansatz methodisch sehr anspruchsvoll und nicht verhältnismäßig. BMS ist der Auffassung, dass verbleibende Unsicherheiten nach erfolgter Adjustierung nicht pauschal durch die Forderung großer Effektstärken kompensiert werden sollten [...]. Stattdessen sollten die getroffenen Annahmen zur Strukturgleichheit kritisch überprüft und durch geeignete Sensitivitätsanalysen ergänzt werden. <u>Verteilungsverhältnis zwischen den Behandlungsgruppen</u> [...] BMS möchte [...] darauf hinweisen, dass stark asymmetrische Verteilungsverhältnisse wie 1:5 (oder extremer) praktische Herausforderung und Risiken für die Validität kausaler Schlussfolgerungen mit sich bringen. So kann es vorkommen, dass bestimmte Kombinationen von Patientenmerkmalen – insbesondere seltene Konstellationen – im unterrepräsentierten Arm kaum oder gar nicht vertreten sind. Dies führt zu extremen Propensity Scores nahe 0 oder 1, was instabile Gewichtungen oder mangelnde Überlappung der zugrundeliegenden Verteilungen zur Folge haben kann [...]. <u>Vorgeschlagene Änderung:</u> [...] Das vorgeschlagene Vorgehen des IQWiG, in dem Fall, dass Studienergebnisse keine hinreichende Einschätzung der Effekte zulassen, Fallzahlszenarien zu	<u>Änderung im Rapid Report:</u> <i>Verschobene Nullhypothese</i> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30) Auch bei sorgfältigster Analyse unter Erfüllung umfangreicher Qualitätsanforderungen verbleiben in einer AbD aufgrund der fehlenden Randomisierung (u. a. wegen potenziell unbekannter Confounder) relevante Unsicherheiten. Daher stellt der gänzliche Verzicht auf eine verschobene Nullhypothese keine Option dar. <u>Keine Änderung im Rapid Report:</u> <i>Verteilungsverhältnis zwischen den Behandlungsgruppen</i> Die in der Stellungnahme beschriebenen statistischen Herausforderungen, die mit der Anwendung von Propensity-Score-Verfahren in Anwendungsgebieten mit kleinen Patientenkollektiven aufgrund einer nicht ausreichenden Balanciertheit der Gruppen verbunden sein können, sind nachvollziehbar. Dies ist jedoch kein Argument, um per se auf die Durchführung einer AbD zu verzichten. Inwiefern sich methodische Herausforderungen ergeben, zeigt sich erst anhand der konkreten Datensituation. Da die Arzneimittel, für die ein AbD-Konzept beauftragt wird, entweder noch nicht zugelassen oder erst seit kurzer Zeit auf dem Markt verfügbar sind, ist, wie im Rapid Report, Abschnitt 5.3.1, beschrieben, das Verteilungsverhältnis zwischen den Gruppen in einer AbD schwer abschätzbar. Um dieser Unsicherheit Rechnung zu tragen, wird in einem Konzept ein Spektrum möglicher Szenarien dargestellt, darunter auch Szenarien mit Verteilungsverhältnissen von 1:5 oder 5:1 (Intervention zu Komparator). <u>Keine Änderung im Rapid Report:</u> zur Einordnung der orientierenden Betrachtung von Fallzahlszenarien siehe Würdigung in Zeile 26)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	beschreiben und eine „aufdeckbare Effektgröße“ anzugeben, ohne deren Erreichbarkeit einzuordnen greift aus Sicht von BMS zu kurz [...].	
32)	Bristol-Myers Squibb GmbH & Co. KGaA (BMS) [Abschnitt 5.1, S. 6 ff.; Abschnitt 5.3.2, S. 18 ff.] <u>Anmerkung:</u> [...] [Es] ist zu beachten, dass das Design der emulierten hypothetischen randomisierten kontrollierten Studie (Target Trial), die die relevante Fragestellung behandelt und an dem sich die AbD orientiert, notwendigerweise einem pragmatischen Ansatz folgt [5,6]. Im Gegensatz zu klassischen hochkontrollierten randomisierten klinischen Studien sind pragmatische Studien durch größere Heterogenität (etwa in Patientencharakteristika oder -adhärenz) gekennzeichnet, was typischerweise zu einem höheren Maß an „Hintergrundrauschen“ (noise) und somit auch tendenziell zu kleineren Behandlungseffekten führen kann. Dies gilt umso mehr für die nicht-randomisierten Emulationen solcher Studien. Vor diesem Hintergrund erscheint die Annahme großer Effektstärken in der Fallzahlplanung methodisch fragwürdig: Sie kann zu einer systematischen Unterschätzung der erforderlichen Fallzahl führen und damit Fehleinschätzungen begünstigen – etwa hinsichtlich der Dauer oder der generellen Machbarkeit der AbD. <u>Vorgeschlagene Änderung:</u> Eine umfassende Berücksichtigung dieser Aspekte erscheint notwendig, um die Vorteile der Target Trial Emulation adäquat und praktikabel in die Nutzenbewertung zu integrieren. Insbesondere sollte die Validität der für die Fallzahlplanung angenommenen Effektgrößen kritisch hinterfragt werden [...].	<u>Keine Änderung im Rapid Report:</u> Pragmatische Studien zielen darauf ab, die Wirksamkeit und Sicherheit von Arzneimitteln unter (möglichst) realen Versorgungsbedingungen zu untersuchen. Es ist richtig, dass in pragmatischen RCTs im Vergleich mit RCTs, wie sie typischerweise für Zulassungsverfahren von Arzneimitteln durchgeführt werden, aufgrund des breiten Einschlusses von Patientinnen und Patienten (möglichst unselektierte Population) eine Heterogenität erzeugt wird („Praxisrauschen“). Wegen dieser erwünschten Heterogenität ergibt sich häufiger die Notwendigkeit des Einschlusses einer höheren Patientenzahl, um auch kleinere, aber relevante Effekte in einem Vergleich zweier medikamentöser Therapien ausreichend sicher zeigen zu können [179]. Dies trifft folgerichtig im Regelfall auch auf eine Nachbildung einer hypothetischen RCT mit pragmatischen Designelementen zu. Bezüglich der Annahme über die Effektgröße zur Schätzung der notwendigen Fallzahl sei an dieser Stelle anzumerken, dass nicht der wahre Therapieeffekt in dieser Größenordnung liegen muss, sondern der in einer AbD beobachtete Effekt (unter Berücksichtigung der Verzerrung aufgrund des nicht randomisierten Vergleichs).
33)	Bundesverband der Pharmazeutischen Industrie e. V. (BPI) [Abschnitt 5.3] <u>Anmerkung:</u> [...] Der vorläufige Rapid Report weist [...] darauf hin, dass im Rahmen einer AbD auch ein Erkenntnisgewinn als Studienziel ausreichend sein könne. Diese Perspektive steht nicht mit der gesetzlichen Vorgabe nach § 35a SGB V im Einklang die explizit auf die Quantifizierung des Zusatznutzens abzielt.	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Es sollte ergänzend darauf eingegangen werden, dass die Auswahl und Priorisierung der Methodik stets dem Ziel dienen, eine Quantifizierung des Zusatznutzen zu ermöglichen [...].	
34)	CSL Behring GmbH [Abschnitt 5.3.1, S. 16] <u>Anmerkung:</u> [...] Die verschobene Nullhypothese wird laut Rapid Report eingeführt, um mögliche Verzerrungen durch eine fehlende Randomisierung zu kompensieren. Es wird dabei vorgegeben, dass ein Konfidenzintervall vollständig außerhalb der verschobenen Hypothesengrenze liegen muss, um einen Effekt mit ausreichender Aussagesicherheit ableiten zu können. Dieses Vorgehen ist sehr konservativ mit Blick auf die Aussagesicherheit und führt zu der Vervielfachung der benötigten Anzahl an Patienten, die rekrutiert und ausgewertet werden müssen [...]. <u>Vorgeschlagene Änderung:</u> Aus Sicht von CSL Behring sollte versucht werden einen „Mittelweg“ zu finden, der eine ausreichend hohe Aussagesicherheit auf der einen Seite bei gleichzeitig regelhaft guter Realisierbarkeit auf der anderen Seite gewährleistet [...]. Einzelheiten zu diesem Vorschlag, einschließlich entsprechender Fallzahlkalkulationen, sind im eingereichten Studienprotokoll zur AbD von Etranacogen dezaparvovec beschrieben (CSL Behring, 2024).	<u>Änderung im Rapid Report:</u> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30) Auch bei sorgfältigster Analyse unter Erfüllung umfangreicher Qualitätsanforderungen verbleiben in einer AbD aufgrund der fehlenden Randomisierung (u. a. wegen potenziell unbekannter Confounder) relevante Unsicherheiten. Daher stellt der gänzliche Verzicht auf eine verschobene Nullhypothese bei gleichzeitigem Absenken der Irrtumswahrscheinlichkeit – wie im Studienprotokoll zur AbD zu Etranacogen Dezaparvovec [77] vorgeschlagen – keine Option dar.
35)	CSL Behring GmbH [Abschnitt 5.3.3, S. 25] <u>Anmerkung:</u> [...] Es ist nachvollziehbar, dass der anteilige „Verlust“ an Patienten bei der Ergebnisauswertung durch eine entsprechende Confounderadjustierung im Vorfeld der AbD nicht zuverlässig abschätzbar ist und dieser eine relevante Größenordnung haben kann. Es ist jedoch kritisch, dass dieser Umstand bislang bei den orientierenden Fallzahlplanungen im Rahmen der Machbarkeitsanalyse vor Beginn der AbD komplett unberücksichtigt bleibt [...].	<u>Keine Änderung im Rapid Report:</u> Wie im Rapid Report, Abschnitt 5.3.3, erläutert, hängt der Anteil der Patientinnen und Patienten, die in der Auswertung durch Adjustierung für Confounder nicht berücksichtigt werden können (z. B. aufgrund von Matching oder Trimming), vom gewählten Propensity-Score-Verfahren ab. Im Rahmen der mündlichen Erörterung wurde auf eine Simulationsstudie von Austin 2021 zu Propensity-Score-Verfahren mit Gewichtung hingewiesen, in der die geschätzte Fallzahl mit einem Varianzinflationsfaktor (VIF) multipliziert wird [180]. Hierfür muss eine Annahme über den Wert der c-Statistik getroffen werden. Es ist vorab

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Der durch Confounderadjustierung anteilige „Verlust“ an auswertbaren Patienten sollte auf Basis von konservativen Annahmen oder mithilfe von Szenarien bereits bei der Machbarkeitsanalyse einer AbD (vor deren Beginn) bei der Fallzahlplanung berücksichtigt werden, um das Risiko einer Unterschätzung der Fallzahl und fehlerhaft Rückschlüsse zur Realisierbarkeit zu minimieren [...].	unklar, welchen Wert die c-Statistik annehmen wird und wie hoch in der Folge die Überlappung der Patientinnen und Patienten im Propensity-Score-Verfahren mit Gewichtung sein wird. Da die Ergebnisse der Simulationsstudie somit keine sichere Abschätzung erlauben, wie viele Patientinnen und Patienten im Falle von Gewichtungsverfahren zusätzlich rekrutiert werden müssen, um mit einer AbD grundsätzlich aussagekräftige Ergebnisse generieren zu können, bleibt dies für die orientierende Fallzahlschätzung im Rahmen der Konzepterstellung ohne Konsequenz.
36)	Deutsche Gesellschaft für Kinder- und Jugendmedizin e.V. (DGKJ) und Gesellschaft für Neuropädiatrie e.V. (GNP) [Abschnitt 5.3] Bei Einbezug der verfügbaren Fallzahlen plädieren wir dafür, bereits im Vorfeld folgende Faktoren miteinzubeziehen: ▪ [...] ▪ Realistische Betrachtung des Anteils der an „AbD-fähigen“-Gesundheitseinrichtungen behandelten Patientenanteile an der identifizierten Gesamtkohorte mit Indikation für die Verum-Therapie ▪ [...]	<u>Keine Änderung im Rapid Report:</u> zur Abschätzung der Patientenzahl siehe Würdigung in Zeile 26)
37)	Deutsche Krankenhausgesellschaft e. V. (DKG) [Abschnitt 5.3, S. 12-25] <u>Anmerkung:</u> [...] Das Ergebnis der Fallzahlsszenarien ist die aufdeckbare Effektgröße, diese wird vom IQWiG nicht weiter in den gegebenen klinischen Kontext eingeordnet. Da sie durch die verschobene Nullhypothese im Bereich von unter 0,5 bzw. über 2 liegt, wäre es auch wichtig zu wissen, wie wahrscheinlich im gegebenen Kontext das Auftreten der durch dieses Vorgehen aufdeckbaren Effektgröße ist. Auch dies bleibt im methodischen Vorgehen des IQWiG aus [...]. Dass eine Fallzahlschätzung bei AbDs herausfordernd ist [...] bleibt unbestritten. Dass eine Fallzahlplanung für eine randomisierte kontrollierte Studie dagegen	<u>Keine Änderung im Rapid Report:</u> zum ethischen Aspekt siehe Würdigung in Zeile 25) <u>Änderung im Rapid Report</u> zur Einordnung der orientierenden Betrachtung von Fallzahlsszenarien siehe Würdigung in Zeile 26)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	immer auf validere oder vergleichende Daten, zurückgreifen kann [...], halten wir nicht in allen Fällen für zutreffend. Auch in der Planung von Phase-III-Studien kann es Unsicherheiten in der Fallzahlplanung geben, die entsprechend in die Berechnung einfließen müssen. Dennoch müssen im Rahmen der Studienplanung schon aus ethischen Gründen Fallzahlplanungen vorliegen, um nicht unnötig viele oder auch zu wenige Probanden einzuschließen. Andernfalls ist der Einschluss von Patientinnen und Patienten in eine Studie bzw. auch in eine AbD nicht zu rechtfertigen. Ziel jeder Studie, egal ob AbD oder nicht, muss die Generierung statistisch belastbarer Ergebnisse sein, was ohne Fallzahlplanung ad absurdum geführt wird [...]. <u>Vorgeschlagene Änderung:</u> Streichung der Ausführungen zu orientierender Betrachtung der Fallzahlszenarien	
38)	Ecker + Ecker GmbH [Abschnitt 5.3.2, S. 20] <u>Anmerkung:</u> Das IQWiG schlägt als Alternative zu einer Fallzahlplanung eine „orientierende Betrachtung von Fallzahlszenarien“ vor [...]. Dieser Ansatz ist aus Sicht des IQWiG verständlich, birgt aber erhebliche Gefahren: Nur weil eine aufdeckbare Effektgröße berechnet wurde, bedeutet dies nicht, dass dieser Effekt in der Realität auch beobachtet werden kann. In dem vom IQWiG genannten Beispiel auf S. 21 fehlt jegliche Einordnung der ermittelten Kenngröße [...]. <u>Vorgeschlagene Änderung:</u> Bei denen durch das IQWiG im Rahmen der AbD-Konzepterstellung ermittelten Fallzahlszenarien sollte immer auch eine Einordnung zur Größenordnung der angenommenen Patientenzahlen und der ermittelten aufdeckbaren Effekte vorgenommen werden. Dabei sollte der Fokus darauf liegen, ob die beschriebenen Effekte im Rahmen einer AbD realistisch zu beobachten sind.	<u>Keine Änderung im Rapid Report:</u> zur Abschätzung der Patientenzahl siehe Würdigung in Zeile 26) <u>Änderung im Rapid Report:</u> zur Einordnung der orientierenden Betrachtung von Fallzahlszenarien siehe Würdigung in Zeile 26)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
39)	Ecker + Ecker GmbH [Abschnitt 5.3.3, S. 24] <u>Anmerkung:</u> [...] Die Machbarkeit einer ausreichenden Rekrutierung von Patienten für die AbD und damit eine solide Fallzahlplanung muss ein zentraler Bestandteil des IQWiG-Konzepts im Rahmen der Forderung einer AbD durch den G-BA sein. Die AbD wird naturgemäß oft in Indikationen gefordert, die durch sehr kleine Patientenpopulationen definiert werden. In solchen Fällen eine AbD zu starten ohne belastbar abschätzen zu können, ob die Rekrutierung einer ausreichenden Patientenzahl überhaupt gelingen kann, ist fahrlässig. <u>Vorgeschlagene Änderung:</u> Eine Fallzahlplanung darf nicht erst im Rahmen einer Zwischenanalyse einer laufenden AbD erfolgen. Vielmehr muss eine belastbare Fallzahlplanung Teil der Entscheidungsfindung über die Forderung einer AbD sein, also zu einem viel früheren Zeitpunkt vorliegen als vom IQWiG vorgeschlagen [...].	<u>Keine Änderung im Rapid Report:</u> zur (finalen) Fallzahlplanung zum Zeitpunkt von Zwischenanalysen siehe Würdigung in Zeile 28)
40)	IGES Institut GmbH [Abschnitt 5.3, S. 12 ff.] <u>Anmerkung:</u> [...] Die AbD verkörpert einen Mittelweg zwischen nicht adjustierten indirekten Vergleichen – also der „naiven“ Verwendung einzelner Studienarme unterschiedlicher Studien für einen Vergleich – und dem Goldstandard der RCT. Die aktuell geforderten Effektschwellen für AbD werden diesem Charakter einer AbD als Mittelweg zwischen nicht adjustierten indirekten Vergleichen und einer RCT bisher nicht gerecht. [...] Zur Adressierung des Ziels einer möglichst guten Evidenzgenerierung für die Nutzenbewertung in vertretbaren Zeiträumen erscheint es daher zielführend, mögliche Alternativen zu den bisher Anwendung findenden Effektschwellen wissenschaftlich zu untersuchen [...].	<u>Änderung im Rapid Report:</u> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
41)	Janssen-Cilag GmbH [Abschnitt 5.3, S. 12-25] <u>Anmerkung:</u> [...] die Fallzahl [ist] eine wesentliche Größe [...], um die Durchführbarkeit einer AbD beurteilen zu können. In einem Konzept für eine AbD ist daher die Fallzahl als eine für die Frage der Durchführbarkeit einer AbD wesentliche abhängige Größe zu diskutieren. Folglich ist eine orientierende Fallzahlschätzung prinzipiell geeignet, um den Auftrag des G-BA zur Erstellung eines AbD-Konzepts für eine der Nutzenbewertung dienende Datenerhebung zu erfüllen. Es ist jedoch fraglich, ob auch die Herangehensweise der orientierenden Fallzahlszenarien ebenso geeignet ist, den Auftrag zur Beurteilung der Durchführbarkeit einer AbD zu erfüllen. [...] eine theoretisch nachweisbare Effektstärke [ist] eine generische Größe ohne Bezug zu einer konkreten Fragestellung in einem Indikationsgebiet mit Intervention und zugehörigen Komparatoren [...]. Wenn bereits vor Durchführung einer AbD feststeht, dass ein Vorteil des zu bewertenden Arzneimittels und damit ein Zusatznutzen nicht nachweisbar ist, kann keine ergebnisoffene Nutzenbewertung durchgeführt werden. Eine solche AbD ist damit nicht geeignet, den Auftrag des Gesetzgebers zu erfüllen, dass eine AbD der Nutzenbewertung dienen soll. [...] Aus Sicht des IQWiG ist in Abwesenheit vergleichender Daten jede aus vergleichender Evidenz gewonnene Erkenntnis ein Gewinn. Es wird nicht in Abrede gestellt, dass jeder Erkenntnisgewinn zu begrüßen ist. Jedoch kann ein AbD-Konzept, welches nicht geeignet ist, die Evidenz für eine Nutzenbewertung zu liefern, nicht als Grundlage für die Forderung einer AbD dienen. <u>Vorgeschlagene Änderung:</u> Die Grundlage eines AbD-Konzepts muss mindestens eine orientierende Fallzahlschätzung sein. Wenn das IQWiG zu dem Schluss kommt, dass eine orientierende Fallzahlschätzung nicht möglich ist, kann in einem Konzept nicht gleichzeitig auch auf die Durchführbarkeit einer AbD geschlussfolgert werden [...].	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1) <u>Änderung im Rapid Report:</u> zur Einordnung der orientierenden Betrachtung von Fallzahlszenarien siehe Würdigung in Zeile 26)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
42)	Janssen-Cilag GmbH [Abschnitt 5.3.1, S. 16] <u>Anmerkung:</u> Zum Umgang mit der Unsicherheit bei nicht-randomisierten Vergleichen, z. B. durch residuelles Confounding, greift das IQWiG auf die Testung der geschätzten Effekte gegen eine verschobene Nullhypothese zurück. Die Verwendung dieses Kriteriums mit einem praktisch fixen, willkürlich gewählten Schwellenwert, stellt [...] ein nicht geeignetes Kriterium dar, um die Validität von Analyseergebnissen aus nicht-randomisierten Vergleichen zu beurteilen [...]. [...] Aus Sicht von Johnson & Johnson ist der E-Value (2) ein geeignetes Instrument, um Aussagen über die Validität von Analyseergebnissen aus nicht-randomisierten Studien zu beurteilen und eine Entscheidung für das Heranziehen solcher Ergebnisse in einer Nutzenbewertung zu treffen. [...] Ein Kritikpunkt am E-Value ist, dass durch die Berechnung ein annähernd linearer Zusammenhang zwischen dem beobachteten Effektschätzer und dem berechneten E-Value besteht und dass ein gegebener Effektschätzer immer zu dem gleichen E-Value führt (3). Kern der Kritik ist damit die fehlende Abhängigkeit vom spezifischen Kontext der Fragestellung bzw. Indikation, in dem der E-Value berechnet wird. Um den Kontext für die Interpretierbarkeit des E-Value herzustellen, schlägt Johnson & Johnson vor, Informationen aus der im Rahmen einer AbD ohnehin geforderten Confounder-Analyse zu verwenden. [...] Die von Johnson & Johnson durchgeführten Confounder-Analysen beinhalten dabei u. a. zu jedem relevanten Confounder Informationen zur Stärke des Zusammenhangs der Confounder mit dem Endpunkt, die aus den identifizierten Publikationen extrahiert wurden (siehe z. B. in den Verfahren 2022-01-15-D-788: Amivantamab und 2023-02-15-D-919: Ciltacabtagene autoleucel). [...] Das bedeutet, dass ein für die Fragestellung spezifischer Schwellenwert für den E-Value hergeleitet werden kann, ab dem es sehr unwahrscheinlich ist, dass residuelles Confounding die Analyseergebnisse infrage stellen würde [...].	<u>Änderung im Rapid Report:</u> Wie von den Stellungnehmenden richtig angemerkt, ist ein Kritikpunkt am E-Value, dass dieser eine annähernd lineare Funktion von der beobachteten Effektschätzung ist. Somit bietet der E-Value keine über die Effektschätzung hinausgehenden Information. Es lassen sich vielmehr Schwellenwerte für einen E-Value direkt auf die Effektebene zurückrechnen. Unabhängig davon wird das Institut im Rahmen von AbD-Verfahren Sensitivitätsanalysen, die den Einfluss residuellen Confoundings untersuchen (wie z. B. quantitative Bias-Analyse [darunter auch E-Values] oder Verwendung von Negativ-Kontrollen), bei der Bewertung berücksichtigen. Im Rapid Report, Abschnitt 5.6.2.1, erfolgt eine Ergänzung im Textabschnitt „Beschreibung geeigneter Sensitivitätsanalysen“. zur verschobenen Nullhypothese siehe Würdigung in Zeile 30)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Die Methode der verschobenen Nullhypothese ist aufgrund der nicht gegebenen allgemeinen Anwendbarkeit als alleiniges Kriterium für den Umgang mit Unsicherheit in einer AbD kritisch zu hinterfragen. Mit der oben skizzierten E-Value-basierten Herangehensweise liegt ein alternatives Konzept für ein Kriterium vor, das allgemein anwendbar ist und fragestellungs- und Indikationsspezifische Information aus der Confounder-Analyse nutzt [...].	
43)	Lilly Deutschland GmbH [Schätzung von Fallzahlen – Methodik, Herausforderungen und Kritik] [...] <u>Kommentar:</u> [...] Die Darstellung der beiden Varianten der Fallzahlschätzung im Bericht bleibt unpräzise und klare Kriterien für die Auswahl der jeweiligen Methode werden nicht genannt [...].	<u>Keine Änderung im Rapid Report:</u> Bei der Entscheidung für oder gegen eine orientierende Fallzahlschätzung werden keine festen Grenzen angelegt. Die Voraussetzungen, die für eine orientierende Fallzahlschätzung erfüllt sein müssen, sind im Rapid Report, Abschnitt 5.3.3, beschrieben. Inwiefern eine (Nicht-)Erfüllung dieser Voraussetzungen im Rahmen eines AbD-Konzepts dazu führt, dass eine orientierende Fallzahlschätzung als sinnvoll möglich eingeschätzt wird oder nicht, lässt sich nicht algorithmisch festlegen. Die Entscheidung über die Sinnhaftigkeit einer orientierenden Fallzahlschätzung wird im Einzelfall auf Basis der für das zu bewertende Anwendungsgebiet vorliegenden Daten getroffen.
44)	Medizinischer Dienst Bund (KÖR) [Abschnitt 5.3.1, S. 14] <u>Anmerkung:</u> [...] Wir teilen die Ansicht, dass Daten aus qualitativ hochwertigen nicht randomisierten Studien vor allem beim Fehlen von Daten aus aussagekräftigen randomisierten Studien einen Erkenntnisgewinn darstellen können. Es stellt sich allerdings die Frage, ob es dann nicht konsequenter und informativer wäre, bei der orientierenden Betrachtung von Fallzahlsszenarien (Seite 20 ff.) nicht „aufdeckbare Effekte“, sondern bei gegebenen Fallzahlen für bestimmte Konstellationen (Basisrisiko und Effektschätzungen) die erzielbare Präzision (z. B. Breite von 95%-Konfidenzintervallen) darzustellen.	<u>Änderung im Rapid Report:</u> Im Rapid Report, Abschnitt 5.3.2, wurde Abbildung 7 ergänzt, in der beispielhaft jeweils die Breite des 95 %-Konfidenzintervalls von 2 in einer AbD aufdeckbaren Effekten illustriert wird. Die erzielbare Präzision (Breite des Konfidenzintervalls) lässt sich jeweils auch indirekt aus den bisher erstellten Abbildungen zur orientierenden Betrachtung von Fallzahlsszenarien durch den Abstand der minimal aufdeckbaren Effektgröße zum Schwellenwert für die verschobene Nullhypothese ($RR_0 = 2$ oder $RR_0 = 0,5$) ablesen (siehe hierzu z. B. Abbildung 8 im Rapid Report).

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Bei der orientierenden Betrachtung von Fallzahlszenarien bei gegebenen Fallzahlen für bestimmte Konstellationen (Basisrisiko und Effektschätzungen) die erzielbare Präzision (z. B. Breite von 95%-Konfidenzintervallen) darstellen.	
45)	Medizinischer Dienst Bund (KöR) [Abschnitt 5.3.1, S. 16] Zitat: „[...] 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].“ (eigene Hervorhebung) <u>Anmerkung:</u> Grundsätzlich unterstützen wir die Idee, statistische Hypothesentests mit einer verschobenen Nullhypothese durchzuführen, um Mindesteffektstärken statistisch abzusichern [11, 12]. Ebenso sinnvoll ist es, bei nicht randomisierten Studien ein „Handicap“ in die Formulierung der zu prüfenden Hypothese(n) einzubauen, um der den nicht randomisierten Studien inhärenten Unsicherheit zu begegnen. Gleichzeitig hat das IQWiG unseres Wissens nach bei seinen bisherigen Konzepten für eine AbD noch nie von dem „z. B.“ dergestalt Gebrauch gemacht, dass eine weniger ambitionierte Schwelle als 2 (bzw. 0,5) vorgeschlagen wurde (beispielsweise 1,5 bzw. 0,67). Wir regen an, über dermaßen abgesenkte Schwellen vor dem Hintergrund einer „Erfüllung umfangreicher Qualitätsanforderungen als Voraussetzung“ nachzudenken. Alternativ wäre denkbar, auf Basis geeigneter Methoden (beispielsweise Nutzung von Negativ-Kontrollen [hier: Outcomes, für die keine kausale Beeinflussung durch die zu prüfende Intervention anzunehmen ist]) eine quantitative Bias-Abschätzung vorzunehmen, um daraus eine Schwelle abzuleiten [9, 10]. Dies wäre in einem Studienprotokoll adäquat a priori zu planen.	<u>Änderung im Rapid Report:</u> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30) zu Bias-Abschätzung siehe Würdigung in Zeile 42)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Erweiterung der Methodik um Kriterien für die „Erfüllung umfangreicher Qualitätsanforderungen“, die gegebenenfalls geeignet sein könnten, in statistische Hypothesenprüfungen abgesenkte Schwellen jenseits von 2 (bzw. 0,5) aufzunehmen. Alternativ könnte die Möglichkeit eröffnet werden, auf Basis von geeigneten Bias-Abschätzungen Schwellen in Studien festzulegen.	
46)	Merck Healthcare Germany GmbH [Abschnitt 5.3.1, S. 13 f.] <u>Anmerkung:</u> [...] Daten sollten generell nicht nur zur Erfüllung formaler Anforderungen gesammelt werden, sondern um einen tatsächlichen Erkenntnisgewinn zu erzielen. Die Generierung von lediglich statistisch signifikanten Ergebnissen, ohne ein Erreichen der Schwellenwerte für eine Aussage anzustreben, ist zudem unter ethischen Gesichtspunkten problematisch. Immerhin wird den Patienten ein zusätzlicher Aufwand durch eine Datenerhebung zugemutet. Dieser sollte in jedem Fall mit einem Erkenntnisgewinn für die medizinische Gesellschaft einhergehen und daher auf die Demonstration eines Vorteils ausgerichtet werden. <u>Vorgeschlagene Änderung:</u> [Streichung der folgenden Passage]: „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 [...]. Unter dieser Voraussetzung stellt jedes Ergebnis einer AbD einen Erkenntnisgewinn dar.“	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1)
47)	MSD Sharp & Dohme GmbH [Abschnitt 5.3, S. 12-25] <u>Anmerkung:</u> Die Forderung, dass der pharmazeutische Unternehmer die finale Fallzahlplanung zum Zeitpunkt der Zwischenanalysen auf Basis vorläufiger Ergebnisse anpassen soll, ist kritisch zu bewerten. Ein derartiges adaptives Vorgehen verschiebt	<u>Keine Änderung im Rapid Report:</u> Vor Beginn einer AbD ist, wie im Rapid Report, Abschnitt 5.3, aufbereitet, aufgrund unsicherer Vorabinformationen allenfalls eine grobe Einschätzung der Größenordnung der benötigten Fallzahl möglich. zur (finalen) Fallzahlplanung zum Zeitpunkt von Zwischenanalysen siehe Würdigung in Zeile 28)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	lediglich den Zeitpunkt der Entscheidung über die Realisierbarkeit der AbD nach hinten, ohne das grundsätzliche Problem zu lösen [...]. <u>Vorgeschlagene Änderung:</u> MSD schlägt daher vor, dass eine realistische mit klinischen Expert:innen und allen beteiligten Institutionen abgestimmte Fallzahlschätzung bereits vor Beginn der AbD vorliegen muss, um eine fundierte und praxisnahe Planung zu gewährleisten [...].	
48)	Pfizer Pharma GmbH [Abschnitt 5.3, S. 12 ff.] <u>Anmerkung:</u> [Bei der orientierenden Betrachtung von Fallzahlszenarien] [...] wird, anders als bei einer orientierenden Fallzahlschätzung, vorab keine plausible bzw. zu erwartende Effektgröße festgelegt, sondern basierend auf der in der Erkrankung zur Verfügung stehenden Patientenzahl berechnet, welche Effektgrößen aufdeckbar sind. Ob diese Effektgrößen auch nur ansatzweise realistisch sind, bleibt unberücksichtigt. [...] Pfizer ist der Ansicht, dass die Fallzahlszenarien auf realistischen Annahmen bezüglich der Größe an der AbD vermutlich teilnehmenden Patientenzahl beruhen sollten. Gleiches gilt für die orientierende Fallzahlschätzung, die auf realistischen Annahmen bezüglich der Effektstärke beruhen sollte. Dementsprechend ist es nicht zielführend, von einer Vollerhebung der gesamten Population auszugehen [...]. Weiterhin bleibt offen, inwiefern es adäquat ist, die gemäß IQWiG für eine Fallzahlschätzung unzureichende Evidenz der Abschätzung von Ereignisanteilen als Bestandteil einer orientierenden Betrachtung von Fallzahlszenarien für Intervention und Vergleich zugrunde zu legen. Die Evidenz zur Abschätzung der Ereignisanteile könnte dann auch im Rahmen einer orientierenden Fallzahlschätzung mit verschiedenen Szenarien hinsichtlich der Ereignisanteile genutzt werden, um die Machbarkeit der AbD einschätzen zu können [...]. <u>Vorgeschlagene Änderung:</u> Für die Beurteilung der Realisierbarkeit und Angemessenheit einer AbD gemäß § 54 Abs. 1 Nr. 3 VerfO ist im Rahmen der Konzepterstellung regelhaft eine	<u>Keine Änderung im Rapid Report:</u> Im Rahmen der orientierenden Betrachtung von Fallzahlszenarien wird bei gegebener Patientenzahl unter Berücksichtigung der verschobenen Nullhypothese (und ggf. unter Verwendung eines Ankers) ein Spektrum der in einer AbD im zu bewertenden Anwendungsgebiet aufdeckbaren Effektgrößen aufgezeigt. Die Ergebnisse aus vorhandenen Studien zur Intervention und zum Komparator (oder den Komparatoren) gehen – in Abgrenzung zur orientierenden Fallzahlschätzung – nicht als Parameter in die Berechnung ein. Diese werden in einem AbD-Konzept lediglich zur Einordnung der orientierenden Betrachtung von Fallzahlszenarien dargestellt. zur Abschätzung der Patientenzahl siehe Würdigung in Zeile 26)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	orientierende Fallzahlschätzung durchzuführen. Bei Unsicherheiten hinsichtlich der zu erwartenden Effekte können für die Fallzahlschätzung verschiedene Szenarien für den Effektschätzer verwendet werden. Diese sollten einen realistischen Bereich abdecken.	
49)	Pfizer Pharma GmbH [Abschnitt 5.3.1, S. 16 ff.] <u>Anmerkung:</u> Das IQWiG geht bei der Fallzahlschätzung im Rahmen einer AbD von einer verschobenen Nullhypothese von 2 bis 5 für das Relative Risiko (bzw. 0,2 bis 0,5 bei mortalitätsenkenden Interventionen) aus. Aus Sicht von Pfizer ist dieses strikte Kriterium überflüssig, da die Anpassung für mögliche Verzerrungen zwischen dem Interventionsarm und dem Kontrollarm bereits durch die Berücksichtigung aller relevanten Confounder erfolgt. Es ist zu bezweifeln, dass der Einfluss möglicher weiterer unbekannter Confounder so groß ist, dass die Effektschätzung maßgeblich verzerrt ist [...]. <u>Vorgeschlagene Änderung:</u> Pfizer schlägt daher vor, im Fall einer geplanten AbD mit einer hohen Kontrolle von Verzerrungen von einer verschobenen Nullhypothese abzusehen.	<u>Änderung im Rapid Report:</u> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30) Auch bei sorgfältigster Analyse unter Erfüllung umfangreicher Qualitätsanforderungen verbleiben in einer AbD aufgrund der fehlenden Randomisierung (u. a. wegen potenziell unbekannter Confounder) relevante Unsicherheiten. Daher stellt der gänzliche Verzicht auf eine verschobene Nullhypothese keine Option dar.
50)	Roche Pharma AG [Abschnitt 5.3.2, S. 20] <u>Anmerkung:</u> Die „Orientierende Betrachtung von Fallzahlserien“ [...] setzt voraus, dass es eine Abschätzung der Patientenpopulation gibt, die durch verschiedene Herangehensweisen angenähert werden. Wird durch das Medikament oder eine sonstige Entscheidung im Gesundheitssystem (z. B. für/gegen ein Screening-Verfahren) die Population so verändert, dass einzelne Teilpopulationen gar nicht mehr entstehen können, hilft eine solche (retrospektive) Patientenpopulationsbestimmung und ihre Fortschreibung nicht. Eine Internationalisierung der Patientenrekrutierung [...] ist hier keine Hilfe: Sollte der Behandlungsalgorithmus dort wesentlich anders sein (z. B. Screening-	<u>Keine Änderung im Rapid Report:</u> Die Patientenzahl in Deutschland wird für das jeweils zu bewertende Anwendungsgebiet geschätzt. Dabei kann der G-BA die für die Forderung einer AbD relevante Patientenpopulation (z. B. unter Berücksichtigung von Leitlinienempfehlungen und des aktuellen Versorgungskontextes) einschränken. Informationen zu möglichen Veränderungen über die Zeit, z. B. mit Hinblick auf Versorgungsanteile, werden bei der Konzepterstellung berücksichtigt, sofern sie belastbar sind. In der Regel liegen sie zum Zeitpunkt der Konzepterstellung jedoch nicht vor und können demnach nicht berücksichtigt werden. <u>Einbindung weiterer (internationaler) Register</u> Die AbD (Rekrutierung einer ausreichenden Patientenzahl) kann durch die Einbindung weiterer (internationaler) Register unterstützt werden.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Verfahren nicht verfügbar), können die Daten aus inhaltlichen Gründen nicht übertragen werden. Ist der Behandlungsalgorithmus gleich (z. B. Screening-Verfahren ebenfalls eingeführt), gibt es keine auch dort keine Daten, die in die AbD integriert werden könnten. <u>Vorgeschlagene Änderung:</u> Es sollte im Rapid Report klar beschrieben werden, wie in solchen Situationen vorgegangen werden soll und welche Kriterien zu welchen Entscheidungen führen [...].	Voraussetzung hierfür ist, dass die Versorgung in dem Land, in dem die Daten erhoben werden, der Versorgung in Deutschland ausreichend ähnlich ist bzw. die aus diesem Register gewonnenen Erkenntnisse auf die Situation in Deutschland (deutschen Versorgungskontext) übertragbar sind (siehe Rapid Report, Abschnitt 5.3.3). Auf diese Weise soll sichergestellt werden, dass die Einbindung von Daten aus weiteren Ländern die Datenqualität nicht beeinträchtigt. Sofern relevante Unterschiede im Versorgungsstandard in einem anderen Land bestehen, sollten Registerdaten dieses Landes nicht für die AbD herangezogen werden. In diesem Zusammenhang ist u. a. auf ggf. unterschiedliche Versorgungspfade und Begleittherapien zu achten. Die Einbindung von Ländern (Zentren) außerhalb Deutschlands muss vor Beginn der Datenerhebung geklärt und im Studienprotokoll beschrieben werden.
51)	Roche Pharma AG [Abschnitt 5.3.2, S. 20] <u>Anmerkung:</u> Die „Orientierende Betrachtung von Fallzahlserien“ [...] kann zu unerreichbaren Effektannahmen führen, so dass eine Fallzahlerhöhung für die Fragestellung keine Lösung ist. Auch hier wäre eine Beschreibung des dann zu erfolgenden Vorgehens hilfreich. <u>Vorgeschlagene Änderung:</u> Für die Fragestellung wäre eine Art gesamthafter Entscheidungsbaum im Rapid Report [...] hilfreich.	<u>Keine Änderung im Rapid Report:</u> Die in einer AbD aufdeckbare Effektgröße berücksichtigt den wahren Therapieeffekt auf einen patientenrelevanten Endpunkt und eine mögliche systematische Verzerrung aufgrund des nicht randomisierten Designs. Eine Einschätzung dazu, welcher Effekt sich in einer AbD schlussendlich zeigt, ist nicht möglich, da der wahre Effekt und die Größe der Verzerrung unbekannt sind. Vor diesem Hintergrund wird im Konzept nicht nur ein Szenario, sondern ein Spektrum der im Anwendungsgebiet möglichen Szenarien aufgezeigt.
52)	Sanofi-Aventis Deutschland GmbH [Abschnitt 5.3, S. 12 ff.] <u>Anmerkung:</u> [...] Aus wissenschaftlicher Sicht ist eine Fallzahlschätzung, die sich primär an der Möglichkeit einer Effektschätzung orientiert, nicht ausreichend zielgerichtet. Sie	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	birgt das Risiko, dass Studien durchgeführt werden, deren Ergebnisse zwar methodisch verwertbar sind, aber keine klare Aussagekraft im Kontext der Nutzenbewertung entfalten [...]. <u>Vorgeschlagene Änderung:</u> Die Fallzahlschätzung sollte sich am Erkenntnisinteresse der AbD orientieren und insbesondere darauf abzielen, eine belastbare Aussage zur Nutzenbewertung des neuen Arzneimittels zu ermöglichen. Eine Effektschätzung ohne Bezug zu einem konkreten Bewertungsziel sollte nicht als alleinige Rechtfertigung für die Durchführung einer AbD dienen.	
53)	Sanofi-Aventis Deutschland GmbH [Abschnitt 5.3, S. 16] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> Die im Bericht formulierte Schwellenanforderung eines starken Behandlungseffekts (z. B. HR = 2) erscheint [...] im Kontext der modernen Arzneimittelentwicklung nicht realistisch [Rothwell, 2018]. Sie ignoriert die typischen Effektgrößen, wie sie in aktuellen klinischen Studien beobachtet werden, und kann dazu führen, dass klinisch relevante, aber moderate Effekte nicht angemessen berücksichtigt werden [Moon, 2018].	<u>Änderung im Rapid Report:</u> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30)
54)	Servier Forschung und Pharma-Entwicklung GmbH [Abschnitt 5.3, S. 12-24] <u>Anmerkung:</u> [...] Der Rapid Report weist [...] darauf hin, dass im Rahmen einer AbD auch ein Erkenntnisgewinn als Studienziel ausreichend sein kann. Diese Perspektive unterscheidet sich jedoch von der gesetzlichen Zielvorgabe nach §35a SGB V, die explizit auf die Ermöglichung einer Quantifizierung des Zusatznutzens abzielt [...].	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Ergänzung eines expliziten Hinweises, dass die Auswahl und Priorisierung der Methodik stets dem Ziel dienen sollten, eine Quantifizierung des Zusatznutzen zu ermöglichen [...].	
55)	SmartStep Consulting GmbH [Abschnitt 5.3.1, S. 16] <u>Anmerkung:</u> [...] Es wird nicht diskutiert, wie mit zu erwartenden Effekten von 1 – 2 oder einer zu kleinen Fallzahl in einer der Behandlungsgruppen umgegangen werden soll. Da AbDs häufig bei OD-Verfahren angefordert werden, ist die Überschreitung der vorgeschlagenen Grenzen von 2 – 5 in der Effektstärke in Zusammenspiel mit nur wenigen verfügbaren Patient*innen und freier Behandlungswahl (im Gegensatz zum RCT-Setting) nur in seltenen Fällen zu erwarten [...]. <u>Vorgeschlagene Änderung:</u> Es sollte auf alle zu erwartende Szenarien eingegangen werden, und das Vorgehen für entweder zu geringe zu erwartende Effekte unterhalb eines RR von 2, oder zu große Konfidenzintervalle erläutert werden [...].	<u>Änderung im Rapid Report:</u> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30)
56)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.3.1, S. 14] <u>Anmerkung:</u> Es wird ausgeführt, es könne bei „[...] 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“ und weiter: es stelle „[...] jedes Ergebnis einer AbD einen Erkenntnisgewinn dar“. Die Durchführung einer AbD, die nicht das Ziel hat, einen Vorteil des neuen Arzneimittels zu zeigen, widerspricht jedoch ihrer Intention [...]. Das Ziel der AbD ist [...] die Quantifizierung des Zusatznutzens, d. h. die Beantwortung der Frage nach der Überlegenheit der neuen Therapie gegenüber der Vergleichstherapie. Nicht jedes Ergebnis einer AbD stellt somit einen gesetzlich intendierten Erkenntnisgewinn dar, sondern nur jenes, das auf eine Quantifizierung des	<u>Keine Änderung im Rapid Report:</u> zu Erkenntnisgewinn einer AbD siehe Würdigung in Zeile 1) zum ethischen Aspekt siehe Würdigung in Zeile 25)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Zusatznutzens abzielt. Zudem tangiert eine solche nicht zielführende Datenerhebung im Versorgungskontext auch ethische Fragen, nämlich inwiefern sie für Leistungserbringer:innen und Patient:innen eine zumutbare Belastung darstellt. Insgesamt ist die Durchführung solcher AbD strikt abzulehnen. <u>Vorgeschlagene Änderung:</u> Streichung des Absatzes	
57)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.3.1, S. 16 f.] <u>Anmerkung:</u> Die Ausführungen zur verschobenen Nullhypothese sind kritisch einzuordnen. Nicht sachgerecht ist die Anforderung eines dramatischen Effekts, da sich die Vorgaben auf Konfidenzintervallgrenzen und nicht auf die Punktschätzer beziehen. [...] Die AbD-Vorgaben für die Effekte berücksichtigen somit weder die zusätzlich zu erfolgende Confounder-Adjustierung noch die Besonderheiten der Therapiesituation (insb. die kleinen Patientenzahlen), die gemäß Auftrag des G-BA zur AbD jeweils vorliegt. Unter der zusätzlichen Berücksichtigung der Vorgaben an die Confounder-Adjustierung sind die Anforderungen an die Effekte in der AbD damit sogar konservativer als in der Nutzenbewertung nach §35 SGB V [...]. <u>Vorgeschlagene Änderung:</u> [...] [Es] muss von den der Anforderung eines dramatischen Effekts (im Bereich von 2 bis 5 bezogen auf die untere bzw. obere Konfidenzintervallgrenze) Abstand genommen werden, da es sich hier um eine praxisferne, überhöhte Anforderung handelt. Alternativ könnte beispielsweise als Kriterium der Überschreitung des Schwellenwertes eines Relativen Risikos von 2 der Punktschätzer anstelle der Grenze des 95%-Konfidenzintervalls verwendet wird [...].	<u>Änderung im Rapid Report:</u> zur verschobenen Nullhypothese siehe Würdigung in Zeile 30)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
58)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.3.2, S. 18 ff.] <u>Anmerkung:</u> Eine orientierende Fallzahlschätzung sollte stets Startpunkt der Überlegungen zur Beauftragung einer AbD sein. Der zunehmende Verzicht auf orientierende Fallzahlschätzungen in Ermangelung gesicherter Vorabinformationen ist daher kritisch zu sehen. Eine verlässliche Fallzahlplanung stellt das Fundament für eine erfolgreiche, ergo umsetzbare und zielführende, AbD dar. Es bleibt unklar, wie aus der „orientierenden Betrachtung von Fallzahlszenarien“ eine konkrete Fallzahl abgeleitet werden kann; die Szenarien sind hypothetisch und können nicht direkt zur Studienplanung verwendet werden [...]. <u>Vorgeschlagene Änderung:</u> Ergänzung eines Hinweises, dass auch unter Unsicherheiten sinnvolle orientierende Fallzahlschätzungen möglich sind und zur Diskussion beitragen [...].	<u>Keine Änderung im Rapid Report:</u> Einer orientierenden Betrachtung von Fallzahlszenarien liegt die Frage zugrunde, welche Effekte in einer AbD im zu bewertenden Anwendungsgebiet bei gegebener Patientenzahl unter Berücksichtigung der verschobenen Nullhypothese aufgedeckt werden können. Bei dieser Herangehensweise ist, wie im Rapid Report erläutert, die in einer AbD aufdeckbare Effektgröße der Parameter, der mithilfe der übrigen Parameter geschätzt werden soll, nicht die Fallzahl.
Beobachtungsstart und Behandlungswechsel (Themenbereiche 3 und 4)		
Beobachtungsstart		
59)	Vertex Pharmaceuticals (Germany) GmbH [Abschnitt 5.4.1, S. 25-27] <u>Anmerkung:</u> [...] Im vorläufigen Rapid Report wird vorgeschlagen, dass t0 als Zeitpunkt der Therapieentscheidung gewählt werden sollte. Vertex merkt an, dass dieser Ansatz nicht für alle Therapien gleichermaßen geeignet ist. Im Beispiel von CGT, die auf autologen Zellen als Ausgangsmaterial basieren, liegt regelmäßig eine teils monatelange Wartezeit zwischen Therapieentscheidung und tatsächlicher Verabreichung der Intervention. Diese Wartezeit führt zu methodischen Herausforderungen, da die Patientinnen und Patienten in der Interventionsgruppe während dieser Zeit weiterhin eine der als Komparatoren festgelegten Therapien erhalten, genau wie die Patientinnen und Patienten in der Kontrollgruppe. Zudem kann der erwartete Zeitpunkt des	<u>Keine Änderung im Rapid Report:</u> Bei Einsatz einer (potenziell kurativen) Zell- und Gentherapie ist eine Unterteilung des Beobachtungszeitraums in 2 Phasen dann inhaltlich nachvollziehbar, wenn die Vermeidung eines Ereignisses durch die interessierende Therapie (z. B. Vermeidung von vasoookklusiven Krisen durch Behandlung mit Exagamglogen Autotemcel [Addendum A25-82 [171]]) erst ab einem bestimmten Zeitpunkt nach Beginn der Behandlung mit der Intervention möglich ist. In solchen Situationen ist bei Unterteilung des Beobachtungszeitraums Folgendes zu beachten:

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Wirkeintritts von CGT deutlich nach dem Zeitpunkt der Verabreichung liegen, z. B. weil zunächst ein Engraftment der übertragenen Zellen stattfinden muss. Wird dennoch der Zeitpunkt der Therapieentscheidung als t0 definiert, entsteht eine systematische Verzerrung, da die Zeit zwischen Therapieentscheidung und Gabe der Therapie bzw. erwartetem Wirkeintritt als Teil der „exponierten“ Beobachtungszeit in der Interventionsgruppe gewertet wird, ohne dass eine tatsächliche Exposition gegenüber der Therapie vorliegt. <u>Vorgeschlagene Änderung:</u> Vertex schlägt daher vor, bei CGT (und vergleichbaren Interventionen mit Wartezeit zwischen Therapieentscheidung und -verabreichung, bzw. erwartetem Wirkeintritt) von einer einheitlichen Definition des t0 als Zeitpunkt der Therapieentscheidung abzusehen. Stattdessen wird empfohlen, die Analyse in zwei Phasen zu unterteilen und die Vorbereitungsperiode bei CGT separat zu analysieren und darzustellen [...].	■ Für einen Endpunkt, der operationalisiert ist durch die Vermeidung eines Ereignisses, dessen Eintritt erst im Zeitraum von Phase 2 möglich ist, sind Auswertungen separat für Phase 1 und Phase 2 denkbar. ■ Für alle weiteren Endpunkte sind primär Auswertungen über den gesamten Beobachtungszeitraum (Phase 1 und Phase 2) durchzuführen. ■ Für alle Endpunkte erfolgen die Auswertungen nach dem Intention-to-treat(ITT)-Prinzip. Mit Blick auf eine adäquate Festlegung des Startzeitpunkts der Beobachtung (Index-Datum [t_0]) für versorgungsnahen Studien ohne Randomisierung stellt die vorgenannte Konstellation einen Sonderfall dar.
60)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.4.1, S. 25 ff.] <u>Anmerkung:</u> Den Tag der Therapieentscheidung als Beobachtungsbeginn (t_0) festzulegen, ist in der Versorgungspraxis häufig nicht realisierbar. Die uneingeschränkte Darstellung und hohe Gewichtung der ITT-Analyse erscheint im Kontext anwendungsbegleitender Datenerhebungen daher kritisch, insbesondere bei begrenzter Interpretierbarkeit durch Therapie-Switches. In der Versorgungsrealität kommen Behandlungswechsel in vielen Indikationen wesentlich häufiger vor und sind früher zu erwarten als im Kontext von RCT. Für den Wechsel vom Komparator auf die Intervention bedeutet dies, dass die erwarteten Effekte deutlich kleiner werden. Analysen, die den Behandlungswechseln Rechnung tragen, werden vom IQWiG lediglich als Sensitivitätsanalysen zugelassen [...]. Im Fall von Behandlungswechseln in wesentlichem Umfang kann es jedoch dazu kommen, dass die Hauptanalyse keinen signifikanten Effekt nachweisen kann, hingegen Analysen, die einen anderen Estimand bedienen, dies sehr wohl erreichen. Es	<u>Keine Änderung im Rapid Report:</u> Alle Methoden zur Berücksichtigung von Treatment Switching sind anfällig für Verzerrungen, deren Richtung und Ausmaß in der Regel nicht ohne zusätzliche Analysen einzuschätzen sind. Die Ergebnisse der ITT-Analyse (Treatment-Policy-Estimand) sollten daher immer dargestellt werden. Weitere Ergebnisse aus Analysen, die den Behandlungswechsel explizit berücksichtigen, können ergänzend dargestellt werden. Allerdings sind für die Bewertung dieser Analysen zusätzliche Informationen notwendig, die ausreichend beschrieben sein müssen (siehe hierzu GA14-04 [72]).

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	wird aus dem Rapid Report jedoch nicht ersichtlich, wie mit solchen Fällen umgegangen werden soll [...]. <u>Vorgeschlagene Änderung:</u> [...] Eine differenziertere Betrachtung wäre geboten, da ITT-Analysen durch die Behandlungswechsel im Versorgungskontext nur begrenzt interpretierbar sind. Ausführungen sind zu ergänzen, wie mit Analysen umzugehen ist, die – anders als die Hauptanalyse – einen signifikanten Effekt nachweisen können, weil sie einen anderen Estimand bedienen.	
Behandlungswechsel		
61)	Deutsche Gesellschaft für Kinder- und Jugendmedizin e.V. (DGKJ) und Gesellschaft für Neuropädiatrie e.V. (GNP) [Abschnitt 5.4.2] Das Beispiel spinale Muskelatrophie zeigt, dass eine überbrückende Therapie (bridging Therapie), wenn auch nur wenige Tage oder Wochen mit anschließendem Therapiewechsel, z. B. auf Zolgensma, die Krankheitstrajektorien erheblich verändern kann. Die Gruppen mit oder ohne bridging Therapie können somit nicht 1:1 und ohne Confounderadjustierung miteinander verglichen werden. Die Festlegung der Interventions- und Comparatorgruppen sollten daher im Vorfeld der AbD nicht nur mit dem pU abgestimmt werden, sondern mit von der jeweiligen med. Fachgesellschaft bestimmten Expert/-innen. Hier muss v. a. die „Praxistauglichkeit“ der gewählten Interventions- und Comparatorstichproben geprüft werden.	<u>Keine Änderung im Rapid Report:</u> Eine Brückentherapie ist als Teil des Therapiekonzepts zu sehen. Der Tag der Entscheidung für die eigentlich geplante Therapie (z. B. für eine Gentherapie) wird in einer AbD als Beobachtungsstart (Indexdatum [t_0]) festgesetzt. Da eine Brückentherapie nicht auf Weiterbehandlung ausgelegt ist, kann sie den nachfolgenden Therapien zugeordnet werden. Die Entscheidung darüber, ob eine Therapie als Brückentherapie eingesetzt wird, muss dokumentiert werden. Ferner sollte der pU im Studienprotokoll einen angemessenen Zeitraum definieren, in dem eine Therapie noch als Brückentherapie gewertet wird. Ausschlaggebend sind in vorliegender Situation Auswertungen nach der Treatment-Policy-Strategie (d. h. eine Schätzung des Effekts für die gesamte Therapiestrategie). Im Rahmen einer AbD spielt aufgrund des nicht randomisierten Designs der Studie eine adäquate Adjustierung für Confounder eine zentrale Rolle, um eine Vergleichbarkeit der Gruppen bezüglich potenzieller prognostischer Faktoren zu erreichen (siehe Rapid Report, Abschnitt 5.1). Für die AbD ist sicherzustellen, dass alle für die Fragestellung relevanten Confounder vorab identifiziert und spätestens mit Start der AbD erhoben werden.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
62)	Vertex Pharmaceuticals (Germany) GmbH [Abschnitt 5.4.2, S. 30-32] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> Aus Sicht von Vertex ist es notwendig, in Szenarien mit hoher Wahrscheinlichkeit für den Behandlungswechsel zur Intervention eine Analyse nach der Hypothetical Estimand Strategy als Hauptanalyse durchzuführen. Dabei werden Patientinnen und Patienten in der Kontrollgruppe zum Zeitpunkt des Behandlungswechsels zur Intervention zensiert, sodass der Effekt unter der Annahme geschätzt wird, dass sie die ursprüngliche Behandlung beibehalten hätten [...].	<u>Keine Änderung im Rapid Report:</u> Für die Auswertung der Ergebnisse aus der AbD ist eine ITT-Auswertung zu bevorzugen. Der von den Stellungnehmenden vorgeschlagene Estimand nach dem Hypothetical-Strategy-Ansatz führt, wie im Rapid Report, Abschnitt 5.4.2, beschrieben, zu Verzerrung durch informative Zensierung. Davon unbenommen können Therapiewechsel durch den pU im Rahmen von Sensitivitätsanalysen adressiert werden.
Patientenberichtete Endpunkte (Themenbereich 5)		
63)	Bundesverband der Pharmazeutischen Industrie e. V. (BPI) [Abschnitt 5.5] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> [...] Es wäre hilfreich, wenn das IQWiG sich auch dazu äußert, wie damit umzugehen ist, wenn keine spezifischen Fragebögen vorliegen bzw. diese nicht in relevanten Sprachen vorliegen und / oder nicht validiert sind. Zudem wäre es hilfreich, wenn das IQWiG Stellung bezieht, wie Schwierigkeiten bei der Erhebung über einheitliche Zeitpunkte insbesondere im SoC-Arm aufgelöst werden können, da SoC-Patienten mit hoher Wahrscheinlichkeit nicht in gleichem Umfang ein Interesse an der Teilnahme haben.	<u>Keine Änderung im Rapid Report:</u> <i>Auswahl von PRO-Fragebogen</i> Wie im Rapid Report beschrieben, sind, insofern möglich, indikationsspezifische Instrumente gegenüber generischen Instrumenten zu bevorzugen. Falls kein indikationsspezifisches Messinstrument zur Erhebung der gesundheitsbezogenen Lebensqualität z. B. der Symptomatik identifiziert werden kann, können somit grundsätzlich auch generische Instrumente eingesetzt werden. Die Auswahl geeigneter Instrumente zur Erhebung patientenberichteter Endpunkte zur Symptomatik und zur gesundheitsbezogenen Lebensqualität in einer AbD sollte im Zuge der Erstellung des Studienprotokolls dargelegt werden.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
		<i>Motivation zur Datenerhebung</i> Die Patientinnen und Patienten zur Datenerhebung und Datenübermittlung, insbesondere in der Kontrollgruppe, zu motivieren, stellt nicht nur in einer AbD, sondern auch in offenen (z. B. nicht verblindeten) RCTs eine Herausforderung dar. Entscheidend ist, die Patientinnen und Patienten zum Zeitpunkt ihrer Einwilligung in die Studienteilnahme über den direkten Nutzen der Datenerfassung (einschließlich patientenberichteter Daten) aufzuklären (siehe Rapid Report, Abschnitt 5.6.1).
64)	Deutsche Krankenhausgesellschaft e. V. (DKG) [Abschnitt 5.5, S. 33-37] <u>Anmerkung:</u> [...] Die Befragungen könnten durch die postalische Versendung der Fragebögen oder digital durchgeführt werden. Wir sehen es trotzdem insbesondere bei Patienten und Patientinnen des Vergleichsarmes, die weniger motiviert sein könnten, oder bei Patientinnen und Patienten, die durch Symptomatik oder unerwünschte Nebenwirkungen in ihrem Allgemeinbefinden stark beeinträchtigt sind, als schwieriger an, solche Daten auf diesem Wege möglichst lückenlos zu erfassen, als wenn die Befragung im zeitlichen Zusammenhang mit dem gewohnten Arztkontakt steht. Die Verwertbarkeit von PRO-Daten ist daran gebunden, dass die Rücklaufquoten über 70% liegen. Dieses Qualitätskriterium ist unter den erschwerten Bedingungen einer AbD nur schwer zu erfüllen. Das IQWiG beschreibt darüber hinaus, dass für die Erhebungszeitpunkte angemessene Toleranzfenster, die nicht aneinandergrenzen, präspezifiziert werden sollen [...]. An dieser Stelle adressiert das IQWiG ein Problem, das auch bei der Bewertung von PRO-Daten, die in RCTs erhoben wurden, und im Rahmen des AMNOG-Verfahrens für die Nutzenbewertung eingereicht wurden, wiederholt aufgetreten ist. Problemen traten auf, wenn zumeist im Interventionsarm komplexe Therapieregime wie bei der Behandlung mit CAR-T-Zellen oder mit Gentherapien durchgeführt wurden und mit Standardtherapien, die wie zum	<u>Keine Änderung im Rapid Report:</u> In einer AbD ist die Datenerhebung, insbesondere von PROs, bei einem Vergleich von unterschiedlichen Therapieansätzen (z. B. [1] CAR-T-Zelltherapie vs. Chemotherapie oder [2] Einmalgabe einer Gentherapie vs. kontinuierliche Therapie) mit Herausforderungen verbunden, da sich die Erhebungshäufigkeit und -frequenz bei alleiniger Erhebung zu Visiten unterscheiden kann. Da zur Generierung aussagekräftiger Ergebnisse eine in beiden Gruppen möglichst gleiche (standardisierte) Erhebungsfrequenz erforderlich ist, ist eine alleinige Erhebung in den Zentren nicht praktikabel. In solchen Therapiesituationen stellen digitale Lösungen einen aussichtsreichen Ansatz zur Erhebung aussagekräftiger PRO-Daten dar (zur Motivation zur Datenerhebung siehe Würdigung in Zeile 63). Zur Generierung aussagekräftiger Daten zum Zwecke der Nutzenbewertung liegt es in der Verantwortung des betroffenen pU, geeignete Toleranzfenster zu wählen, die nicht aneinandergrenzen und eine sinnvolle Zuordnung zu einem Erhebungszeitpunkt möglich machen.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Beispiel Chemo-Immuntherapien völlig anderen zeitlich strukturierten Therapieschemata folgen, verglichen werden mussten [...]. Zudem gab es im Rahmen von onkologischen AMNOG-Verfahren, in denen deutlich weniger komplexe Therapieregime miteinander verglichen wurden, dann Probleme, wenn in den Studienarmen Chemo(immun)Therapien mit unterschiedlicher Zykluslänge eingesetzt wurden. Auch in solchen Fällen war aus Sicht des IQWiG eine genaue Abstimmung der Erhebungszeitpunkte mit den Zeitpunkten der Therapiephasen erforderlich, damit die Belastung durch die Behandlung im Zyklusverlauf in den Studienarmen nicht ungleich abgebildet wird [...]. Unter den Bedingungen einer AbD, in der mit Toleranzfenstern gearbeitet werden muss, halten wir solche Qualitätsanforderungen für nicht umsetzbar.	
65)	MSD Sharp & Dohme GmbH [Abschnitt 5.5, S. 33-37] <u>Anmerkung:</u> Die bestehenden bürokratischen Hürden stellen derzeit eine wesentliche Herausforderung bei der Erfassung von patientenberichteten Endpunkten (Patient-Reported Outcomes, PRO) dar und sind im gesetzlich vorgegebenen engen Zeitrahmen für Planung und Durchführung einer AbD nur schwer zu bewältigen. Die zeitlichen Restriktionen erschweren es, eine umfassende und qualitativ hochwertige PRO-Erfassung sicherzustellen. Die Publikation von Singer et al. (Singer et al. 2024) zur Erfassung von PRO-Daten im Rahmen von AbD-Verfahren bei CAR-T-Therapien zeigt, dass die PRO-Erhebung grundsätzlich im Kontext von Registern möglich ist. Gleichzeitig wird jedoch deutlich, dass die Umsetzung mit einem erheblichen personellen, logistischen und vor allem bürokratischen Aufwand verbunden ist. <u>Vorgeschlagene Änderung:</u> Aus den oben genannten Gründen ist es aus Sicht von MSD notwendig, die Rahmenbedingungen für AbD insbesondere im Hinblick auf die Schwierigkeiten bei der PRO-Erhebung zu überdenken und zu vereinfachen. Nur so kann eine praktikable Integration patientenberichteter Endpunkte gewährleistet werden, die den Anforderungen an eine aussagekräftige Nutzenbewertung gerecht wird.	<u>Keine Änderung im Rapid Report:</u> Die Festlegung des zeitlichen Verfahrensablaufs zur Forderung einer AbD gemäß § 35a Abs. 3b SGB V obliegt nicht dem IQWiG, sondern erfolgt durch den G-BA. Davon losgelöst sei erwähnt, dass eine stärkere Nutzung von PRO-Instrumenten und -Ergebnissen im Versorgungsalltag grundsätzlich wünschenswert wäre, um Hürden und Hemmnisse der Erhebung im Rahmen einer AbD, sowohl aufseiten der Patientinnen und Patienten als auch aufseiten der Register, abzubauen. Damit einhergehend wäre auch das angeführte Problem des relativ kurzen Zeitrahmens der Umsetzung gelöst. Im Rapid Report werden mit der digitalen Erhebung von PRO-Daten (in Abgrenzung zu papierbasierten Befragungen, die postalisch oder per E-Mail übermittelt werden), Wege aufgezeigt, wie diese Daten in der Routineversorgung mit möglichst wenig Aufwand (und möglichst lückenlos unabhängig von Arztbesuchen) im Zeitverlauf erhoben werden können.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
66)	Roche Pharma AG [Abschnitt 5.5, S. 33] <u>Anmerkung:</u> Gemäß Rapid Report sollen für die Patientenberichteten Endpunkte Fragebögen verwandt werden, die gemäß etablierter Standards validiert wurden [...]. Wenn in der AbD auf Instrumente zurückgegriffen muss, die in dieser Indikation nicht gängige Praxis sind, wird die AbD interventionell. Es stellt sich die Frage, wie mit Situationen umgegangen werden soll, in denen die klinische Routine keine Fragebögen verwendet, die gemäß der zitierten Standards validiert wurden. Es geht bei dieser Frage insbesondere um die Implikationen für die AbD, die sich durch Änderung der regulatorischen Einordnung ergeben. <u>Vorgeschlagene Änderung:</u> Hier ist eine Klärung der Szenarien und der verwandten Kriterien durch das IQWiG notwendig und der Ablauf einer interventionellen AbD deutlich zu beschreiben.	<u>Keine Änderung im Rapid Report:</u> zur Verwendung von validierten Erhebungsinstrumenten siehe Würdigung in Zeile 63) Gemäß Einschätzung der Bundesbehörde (Paul-Ehrlich-Institut) ist für den regulatorischen Charakter einer AbD nicht der Endpunkt an sich, sondern die zur Erhebung erforderliche Prozedur entscheidend. Patientenberichtete Endpunkte begründen eine klinische Prüfung eher nicht [181] (zu Beispielen, in denen PRO-Daten im Versorgungsalltag [in Registern] erhoben werden siehe Anhang A im Rapid Report).
Fehlende Werte (Themenbereich 6)		
Maßnahmen zur Vermeidung fehlender Werte		
67)	Bundesverband der Pharmazeutischen Industrie e. V. (BPI) [Abschnitt 5.6] <u>Anmerkung:</u> Das IQWiG schlägt eine Vergütung des Aufwands der teilnehmenden Zentren durch Zahlung einer Aufwandsentschädigung vor, um fehlende Werte zu vermeiden. Pharmazeutischen Unternehmen (pU) sind durch Vergütung auf Fair Market Value-Basis Grenzen gesetzt. Es ist zweifelhaft, inwiefern Zentren hierdurch incentiviert werden. Der Vorschlag des IQWiG für finanzielle Zuwendungen oder Sachprämien für Patienten durch den pU ist ebenfalls rechtlich zweifelhaft. Insgesamt sind pU gesetzlich enge Grenzen gesetzt. <u>Vorgeschlagene Änderung:</u> Das IQWiG wird gebeten, den Passus zu überarbeiten und dabei auf die gesetzlichen Vorgaben bezüglich Zuwendungen durch den pU stärker einzugehen.	<u>Keine Änderung im Rapid Report:</u> Der Anspruch sollte in einer AbD, analog zum Aufsetzen klinischer Studien, darauf ausgerichtet sein, qualitativ hochwertige prospektive Daten zu erhalten. Materielle Anreize werden in Registern und in nicht interventionellen Studien bereits zielführend zur Steigerung der Motivation zur Datenerhebung eingesetzt und können einen förderlichen Faktor zur Erreichung einer hohen Datenqualität darstellen. Weder in der Stellungnahme noch im Rahmen der mündlichen Erörterung wurden genauere Angaben gemacht, welche rechtlichen Hürden es bei Nutzung materieller Anreize als Instrument zur Sicherstellung der Vollständigkeit der Daten geben könnte.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Es sollten keine Optionen in den Raum gestellt werden, die rechtlich nicht möglich sind.	
68)	SmartStep Consulting GmbH [Abschnitt 5.6.1, S. 41] <u>Anmerkung:</u> Die Empfehlung des Einsatzes materieller Anreize in Form finanzieller Zuwendungen oder Sachprämien zur Vermeidung fehlender Werte im Rahmen der AbD ist kritisch zu betrachten. Direkte finanzielle Zuwendungen oder Sachprämien können zwar möglicherweise einen wichtigen Anreiz für Patient*innen darstellen, Fragebögen vollständig auszufüllen und schließlich zu einer höheren Qualität der AbD führen. Trotz dessen könnten diese Anreize eine unzulässige Beeinflussung der Patient*innen darstellen und die Neutralität der Datenerhebung gefährden, insofern sie nicht eingeschränkt werden. <u>Vorgeschlagene Änderung:</u> Für direkte finanzielle Zuwendungen oder Sachprämien an die Patient*innen sollten folgende Rahmenbedingungen gesetzt werden: ▪ Alle Patient*innen sollten unabhängig von der erhaltenen Behandlung gleichermaßen entschädigt werden. ▪ Die Höhe der finanziellen Zuwendungen oder Sachprämien sollte auf Basis des geschätzten Zeitaufwands für die Erhebung der Endpunkte, beispielsweise durch Fragebögen, begrenzt werden. Denkbar wäre hier eine Obergrenze in Euro pro Stunde, die zudem die Vollständigkeit des Fragebogens oder der erhobenen Daten mit einbezieht. ▪ Es muss sichergestellt werden, dass die Antworten und Aussagen der Patient*innen nicht durch die Zuwendungen beeinflusst werden. Die Zuwendungen sollten vollständig unabhängig von den gegebenen Antworten erfolgen und ausschließlich auf der Vollständigkeit der erhobenen Daten basieren.	<u>Keine Änderung im Rapid Report:</u> Im Rapid Report ist erwähnt, dass Patientinnen und Patienten materielle Anreize in Form finanzieller Zuwendungen (Geschenkgutscheine) oder Sachprämien als motivierend empfinden. Es ist richtig, dass materielle Anreize in beiden Gruppen in gleicher Höhe einzusetzen sind, daher erfolgt im Rapid Report auch keine Eingrenzung auf Patientinnen und Patienten unter Behandlung mit dem zu bewertenden Arzneimittel. Die Festlegung, finanzielle Zuwendungen in beiden Gruppen an die (Nicht-)Erfüllung vorab definierter Kriterien zu knüpfen, hat durch den pU im Studienprotokoll zu erfolgen.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
69)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.6.1, S. 37 ff.] <u>Anmerkung:</u> Der Einsatz materieller Anreize in Form finanzieller Zuwendungen (Geschenkgutscheine) oder Sachprämien an die Patientinnen und Patienten (S. 41) muss kritisch hinterfragt werden [...]. <u>Vorgeschlagene Änderung:</u> Die Motivation der Patientinnen und Patienten zum Ausfüllen der PRO-Fragebögen kann durchaus zielführend sein. Hier sind auch andere Lösungen zu diskutieren, der Rahmen des Rapid Reports erscheint dafür jedoch ungeeignet.	<u>Keine Änderung im Rapid Report:</u> siehe Würdigung in Zeile 67)
70)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Kapitel 6, S. 56] <u>Anmerkung:</u> [...] Der vfa erachtet eine nachhaltige Finanzierung für die langfristige Entwicklung und Pflege der Registerlandschaft als erforderlich. Nutzungsentgelte können dabei als eine zusätzliche Finanzierungssäule dienen. Grundlage ist jedoch eine Basisfinanzierung, die beispielsweise durch die öffentliche Hand geleistet werden muss. Dieser Aspekt fehlt hier. <u>Vorgeschlagene Änderung:</u> Ergänzung um die Notwendigkeit einer Basisfinanzierung, die industrieunabhängig auszugestalten ist.	<u>Keine Änderung im Rapid Report:</u> Die Nutzbarkeit von Registerdaten durch pUs setzt bereits einen erfolgreichen Ausbau und Betrieb eines Patientenregisters (und somit eine Tätigkeit von Investitionsmaßnahmen) voraus. Die Erhebung von Nutzungsentgelten ist eine Möglichkeit, um pUs punktuell an der Finanzierung eines laufenden Registers zu beteiligen. Für eine (anteilige) Deckung der Kosten für den initialen Registerrückbau und die langfristige Pflege einer funktionsfähigen Dateninfrastruktur wird eine finanzielle Beteiligung der pUs über die alleinige Zahlung von Nutzungsgebühren jedoch als nicht zielführend angesehen.
Methodischer Umgang mit fehlenden Werten im Kontext von Propensity-Score-Analysen		
71)	AMS Advanced Medical Services GmbH [Abschnitt 5.6.2.1., S. 43] <u>Anmerkung:</u> [...] [Es] fehlen konkrete Entscheidungskriterien, ab wann eine Überlappung oder eine Balanciertheit als ausreichend anzusehen ist. In bisherigen AbD-Verfahren wurden hierzu unterschiedliche Maßstäbe angelegt [...].	<u>Keine Änderung im Rapid Report:</u> Für die Bewertung der Überlappung gibt es keine etablierten Entscheidungskriterien. Bei geringer Überlappung wird die interessierende Fragestellung nicht zufriedenstellend beantwortbar sein. Hierbei ist auch zu bedenken, dass eine unzureichende Überlappung dazu führt, dass die Auswertungspopulation (z. B. durch Matching oder Trimming) nur einen Teilbereich der zugrunde liegenden Fragestellung abdeckt. Um zu beurteilen, auf welche

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> AMS schlägt vor, im Rapid Report klare Entscheidungshilfen für die Bewertung von Überlappung und Balanciertheit zu geben. Dabei könnte dargestellt werden, unter welchen methodischen und inhaltlichen Voraussetzungen bestimmte Schwellenwerte geeignet sind und in welchen Fällen Abweichungen begründet werden sollten [...].	Population die Ergebnisse der Studie übertragbar sind, muss dieser Teilbereich hinreichend charakterisierbar sein. Dies stellt insbesondere bei Anwendung von Gewichtungungsverfahren eine Herausforderung dar. Der Ausschluss von Patientinnen und Patienten aus der Auswertungspopulation hat insbesondere bei seltenen Erkrankungen zusätzlich zur Folge, dass die überlappende Population ggf. nur noch sehr wenige Patientinnen und Patienten umfasst, sodass keine aussagekräftigen Ergebnisse mehr erzielt werden können. Es ist daher jeweils im Einzelfall zu prüfen, ob die Überlappung ausreichend groß ist, um verlässliche Aussagen treffen zu können. Bei Bewertung der Balanciertheit wird im Rahmen der Prüfung von Studienprotokoll und statistischem Analyseplan eine SMD < 0,2 grundsätzlich akzeptiert. Im Falle einer SMD < 0,25 wird angemerkt, dass die vorgeschlagene Grenze zwar sachgerecht, aber ggf. zu liberal ist. Dies ist bei der Interpretation der Ergebnisse entsprechend zu berücksichtigen.
72)	CSL Behring GmbH [Abschnitt 5.6] <u>Anmerkung:</u> Die Forderung nach vollständigen Daten für alle Confounder ist realitätsfern, insbesondere bei seltenen Erkrankungen oder komplexen Registerstrukturen. MICE, als vorgeschlagenes Standardverfahren, funktioniert gut bei großen Datensätzen aber bei kleinen Fallzahlen, wie in AbDs, kann es zu Konvergenzproblemen kommen. Die Qualität der Imputation hängt stark von der Verfügbarkeit geeigneter Hilfsvariablen ab. Diese sind in Registern oft nicht vorhanden. Es wird nicht erklärt, wie die Schwellenwerte (z. B. 30 % für Endpunkte, 50 % insgesamt) zustande kommen. Es fehlen Verweise auf Studien oder methodische Standards, die diese Grenzwerte stützen und dadurch wirken sie willkürlich und nicht wissenschaftlich fundiert [...].	<u>Keine Änderung im Rapid Report:</u> Es gibt keine Standards für die Anteile von fehlenden Werten in Confoundern, sondern lediglich einzelne Untersuchungen, die nur einen Ausschnitt möglicher Datenkonstellationen abbilden. Die vorgeschlagenen Schwellenwerte im Rapid Report, Abschnitt 5.6.2.2, sind als orientierend (in Anlehnung an das bekannte methodische Vorgehen des IQWiG) anzusehen. Die Planung und Durchführung geeigneter Sensitivitätsanalysen obliegt dem pU.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Die im Bericht dargestellten Anforderungen an den Umgang mit fehlenden Werten sind methodisch nachvollziehbar, aber in der Praxis schwer umsetzbar. Insbesondere bei seltenen Erkrankungen und komplexen Registerstrukturen sind vollständige Daten für alle relevanten Confounder kaum realisierbar. Die Empfehlung zur Anwendung von MICE ist sinnvoll, sollte aber durch eine realistische Einschätzung der Grenzen bei kleinen Fallzahlen ergänzt werden. Die vorgeschlagenen Schwellenwerte für die Verwertbarkeit von Daten erscheinen nicht ausreichend empirisch fundiert. Wir empfehlen, die Anforderungen an Datenvollständigkeit und Imputation stärker kontextabhängig zu gestalten und durch praxisnahe Guidance zur Durchführung von Sensitivitätsanalysen zu ergänzen [...].	
73)	iOMEDICO AG [Abschnitt 5.6.2.1, S. 46] <u>Anmerkung:</u> Im Absatz unter „Beschreibung geeigneter Sensitivitätsanalysen“ wird erwähnt, dass auxiliäre Variablen, welche mit Confoundern assoziiert sind, die Qualität von Imputationen erhöhen können. Jedoch besteht hierbei das Risiko von Collider Bias. [1]. <u>Vorgeschlagene Änderung:</u> Ich schlage vor, in diesem Zusammenhang auf das Risiko von Collider Bias bei multipler Imputation hinzuweisen und die Studie [1] zu zitieren.	<u>Änderung im Rapid Report:</u> Der in der Stellungnahme aufgeworfene Hinweis ist berechtigt. Im Rapid Report, Abschnitt 5.6.2.1, wird im Textabschnitt „Beschreibung geeigneter Sensitivitätsanalysen“ ein Hinweis auf das Risiko von Collider Bias bei Verwendung von auxilären Variablen im Rahmen der multiplen Imputation ergänzt und die Studie Curnow 2023 [154] zitiert.
74)	Jansen-Cilag GmbH [Abschnitt 5.6.2.2, S. 48] Zitat: „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.“	<u>Änderung im Rapid Report:</u> siehe Würdigung in Zeile 42)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Anmerkung:</u> Es gibt Situationen, in denen für einen Confounder ein hoher Anteil an fehlenden Werten nicht zu vermeiden ist [...]. Aus Sicht von Johnson & Johnson kann in einem solchen Fall eine E-Value-basierte Herangehensweise verwendet werden, um zu begründen, dass Analyse-Ergebnisse unter Ausschluss eines relevanten Confounders, z. B. aufgrund eines hohen Anteils fehlender Werte, immer noch als interpretierbar anzusehen sind [...]. <u>Vorgeschlagene Änderung:</u> Eine E-Value-basierte Herangehensweise bietet aus Sicht von Johnson & Johnson die Möglichkeit für eine Begründung, Analyseergebnisse auch unter Ausschluss eines relevanten Confounders für die Nutzenbewertung heranziehen zu können, ohne dass das Vorliegen eines dramatischen Effektes vorausgesetzt werden muss. Das Konzept des E-Values sollte im Rapid Report diskutiert werden.	
75)	Medizinischer Dienst Bund (KöR) [Glossar, S. xiv; Abschnitt 5.6.2.1, S. 42-43] <u>Anmerkung:</u> Definition „Missing at Random“ (MAR): Im Glossar steht auf S. xiv: „Unter dieser Annahme hängt die Wahrscheinlichkeit für das Fehlen eines Wertes von anderen beobachteten Werten einer Person ab, nicht aber vom fehlenden Wert selbst.“ Ein gleichbedeutender Satz steht in Kap. 5.6.2.1 auf S. 42. Diese Definition ist nicht adäquat, stattdessen sollte es im Glossar heißen: „Unter dieser Annahme hängt die Wahrscheinlichkeit für das Fehlen eines Wertes bedingt auf die beobachteten Werte der betreffenden Person nicht vom fehlenden Wert selbst ab.“ Entsprechend sollte auch der Satz in Kap. 5.6.2.1 auf S. 42 angepasst werden. Bei der weiteren Beschreibung von MAR wird zwar jeweils gesagt, dass die Berücksichtigung der beobachteten Daten zu einer unverzerrten Schätzung führen kann, aber dies gehört bereits in die Definition. S. zum Beispiel in Carpenter et al 2023 [13], S. 13.	<u>Änderung im Rapid Report:</u> Der Stellungnahme wird zugestimmt. Sowohl im Glossar als auch in Abschnitt 5.6.2.1 des Rapid Reports wird die Definition präzisiert.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	Definition „Missing not at Random“ (MNAR): Im Glossar steht auf S. xiv: „Die Wahrscheinlichkeit für das Fehlen eines Wertes hängt vom tatsächlichen, fehlenden Wert der betrachteten Zielgröße selbst ab.“ Derselbe Satz steht in Kap. 5.6.2.1 auf S. 43. Auch diese Definition ist nicht ganz ausreichend, stattdessen sollte es jeweils heißen: „Die Wahrscheinlichkeit für das Fehlen eines Wertes hängt vom tatsächlichen, fehlenden Wert der betrachteten Zielgröße selbst ab, und diese Abhängigkeit verschwindet auch nicht durch Bedingen auf die beobachteten Werte der betreffenden Person.“ Bei der weiteren Beschreibung von MNAR wird zwar gesagt, dass die Verzerrung nicht korrigiert werden kann, aber dies gehört bereits in die Definition. S. zum Beispiel in Carpenter et al 2023 [13], S. 17. <u>Vorgeschlagene Änderung:</u> Anpassung der Definitionen im Glossar und im Fließtext entsprechend obiger Anmerkungen	
76)	Medizinischer Dienst Bund (KÖR) [Abschnitt 5.6.2.2, S. 47] Zitat: „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. <u>Bei Analysen zu einem bestimmten Zeitpunkt sollten zu diesem Zeitpunkt Werte für mindestens 50 % der Patientinnen und Patienten vorliegen.</u> Diese Schwellenwerte dienen als Orientierung, wenn nicht mit ausreichender Sicherheit davon ausgegangen werden kann, dass die Werte zufällig fehlen.“ [eigene Unterstreichung]	<u>Änderung im Rapid Report:</u> Es wird zugestimmt, dass die bisherige Formulierung zu Missverständnissen führen kann. Der genannte Satz im Rapid Report, Abschnitt 5.6.2.2, wird präzisiert.

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Anmerkung:</u> Wir vermuten, dass der unterstrichene Satz als zusätzliche notwendige Bedingung für Auswertungen mit Messwiederholungen gemeint ist. Insbesondere vermuten wir, dass auch bei Erfüllen dieser Bedingung weiterhin die 70 %-Regel der ausgewerteten Patientinnen und Patienten gelten soll. Dies ist jedoch nicht ganz eindeutig formuliert [...]. <u>Vorgeschlagene Änderung:</u> Sprachliche Präzisierung, um möglichen Missverständnissen vorzubeugen	
77)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.6.2, S. 41 ff.] <u>Anmerkung:</u> Der vorgeschlagene methodische Umgang mit fehlenden Werten erfordert Mindestanforderungen an Datenqualität, um valide Modelle zu ermöglichen. Alternativen zur multiplen Imputation (z. B. MMRM, IPW, Zensierung) bleiben unberücksichtigt [...]. <u>Vorgeschlagene Änderung:</u> Mindeststandards für die Datenqualität müssen gewährleistet sein, bevor komplexe Imputationsverfahren eingesetzt werden. Alternativen sollten diskutiert werden.	<u>Keine Änderung im Rapid Report:</u> Es gibt keine Standards für die Anteile von fehlenden Werten in Imputationsmodellen, sondern lediglich einzelne Untersuchungen, die nur einen Ausschnitt möglicher Datenkonstellationen abbilden. Die vorgeschlagenen Schwellenwerte im Rapid Report, Abschnitt 5.6.2.2, sind als orientierend (in Anlehnung an das bekannte methodische Vorgehen des IQWiG) anzusehen. Das MICE-Verfahren ist anerkannt und vergleichsweise gut untersucht. Es obliegt dem pU, ggf. Alternativen zu MICE zu benennen. Die genannten Alternativen sind nur in speziellen Fällen anwendbar (MMRM) und / oder anfällig für Verzerrung (Zensierung).
78)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.6, S. 48] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> Die skizzierten Schwellenwerte sollten einer praktischen Prüfung zur Übertragbarkeit auf die Datensituation der AbD unterzogen werden. Vorher sind sie lediglich unter Vorbehalt anzuwenden.	<u>Keine Änderung im Rapid Report:</u> siehe Würdigung in Zeile 72)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven (Themenbereich 7)		
79)	AMS Advanced Medical Services GmbH [Abschnitt 5.7, S. 50] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> In diesem Kontext begrüßt AMS den Vorschlag, bei fehlender Konvergenz des Modells die Anzahl an Confoundern gezielt zu reduzieren, in dem sukzessive die als weniger relevant eingestuften Confounder aus dem Modell entfernt werden, immer unter Berücksichtigung des spezifischen Einzelfalls. Kritisch sehen wir jedoch den Vorschlag, in solchen Fällen eine „weitere Verschiebung der Nullhypothese“ vorzunehmen. [...] AMS schlägt vor, diesen Zusatz zu entfernen.	<u>Keine Änderung im Rapid Report:</u> Im Rapid Report, Abschnitt 5.7, wird erläutert, dass die erhöhte Ergebnisunsicherheit bei Nichtberücksichtigung von Confoundern bei der Ergebnisinterpretation zu berücksichtigen ist und dies <u>z. B.</u> durch eine weitere Verschiebung der Nullhypothese erfolgen kann. Da in diesem Textabschnitt ferner darauf hingewiesen wird, dass dieses Vorgehen nur begrenzt sinnvoll einsetzbar und die Datenlage im Einzelfall zu betrachten ist, wird eine Streichung als nicht erforderlich angesehen.
80)	Bristol-Myers Squibb GmbH & Co. KGaA (BMS) [Abschnitt 5.2, S. 9 ff.; Abschnitt 5.7, S. 50 ff.] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> Mit dem Ziel eine valide Quantifizierung des Zusatznutzen zu ermöglichen, stellen Propensity-Score-Analysen in Anwendungsgebieten mit kleinen Patientenkollektiven auch aus Sicht von BMS eine besondere Herausforderung dar – insbesondere im Hinblick auf die Datenerhebung und -verfügbarkeit sowie die Modellstabilität. BMS wünscht sich hier eine weiterführende Diskussion zu alternativen Methoden.	<u>Änderung im Rapid Report:</u> zu alternativen Methoden zu Propensity-Score-Verfahren siehe Würdigung in Zeile 8)
81)	iOMEDICO [Abschnitt 5.7, S. 51] <u>Anmerkung:</u> Im Absatz unter „Simulationsstudien zu Propensity-Score-Analysen bei kleinen Fallzahlen“ wird auf unterschiedliche Methoden zur Confounderadjustierung mittels Propensity Scores eingegangen. In [3] werden theoretische Argumente	<u>Keine Änderung im Rapid Report:</u> Die in der Stellungnahme zitierte Arbeit von King und Nielsen 2019 [182] zeigt, dass mit Propensity-Score-Matching unter Umständen keine adäquate Confounderadjustierung und ausreichende Balanciertheit erreicht werden kann. Dennoch nimmt das Propensity-Score-Matching in der klinischen Forschung einen großen Stellenwert ein und viele Beispiele zeigen die Funktionalität und Berechtigung

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	präsentiert, welche gegen die Verwendung von Propensity Scores zum Matching sprechen. <u>Vorgeschlagene Änderung:</u> Ich schlage vor, an dieser Stelle auf die in [3] vorgebrachten Argumente hinzuweisen und diese gegebenenfalls einzuordnen.	dieses Propensity-Score-Verfahrens. Insgesamt lässt sich aus dieser Arbeit nicht ableiten, dass Propensity-Score-Matching grundsätzlich ungeeignet ist.
82)	iOMEDICO [Abschnitt 5.7, S. 52] <u>Anmerkung:</u> [Es] wird im zweiten Absatz auf die Varianzschätzung mittels Bootstrapping bei der Verwendung von Inverse Probability of Treatment Weighting (IPTW) zur Confounderadjustierung verwiesen. Falls zusätzlich fehlende Daten imputiert werden müssen, stellt sich die Frage der Reihenfolge von a) Ziehen der Bootstrap-Samples und b) Erzeugung multipler Imputationen. In [2] werden diesbezüglich unterschiedliche Varianten im Rahmen einer Simulationsstudie miteinander verglichen. Insbesondere wird die Variante, bei der zunächst Bootstrap-Samples gezogen werden und anschließend jedes Sample mehrfach imputiert wird (Boot MI), mit der Variante verglichen, bei der zunächst multiple Imputationen erzeugt werden und anschließend Bootstrap-Samples jeder Imputation gezogen werden (MI Boot). In dieser Studie erwies sich Boot MI als die zuverlässigste, wenn auch rechenintensivste Variante. Allerdings lieferte MI Boot ebenfalls valide Ergebnisse, sofern der Anteil fehlender Daten nicht zu hoch war. <u>Vorgeschlagene Änderung:</u> Ich rege an, an dieser Stelle auf die Fragestellung der Reihenfolge von Bootstrapping und Generierung multipler Imputationen einzugehen und die Arbeit [2] zu zitieren.	<u>Änderung im Rapid Report:</u> Die in der Stellungnahme aufgeworfene Frage zur Reihenfolge von multipler Imputation und Bootstrapping ist für die Auswertung der Ergebnisse aus der AbD von Bedeutung. Der Textabschnitt im Rapid Report, Abschnitt 5.7, wurde entsprechend ergänzt und die Studie Schomaker 2018 [163] zitiert.
83)	MSD Sharp & Dohme GmbH [Abschnitt 5.7, S. 50-55] <u>Anmerkung:</u> [...] Während im Rapid Report an mehreren Stellen das Propensity-Score-Verfahren als geeignete Methode hervorgehoben wird, wird gleichzeitig die	<u>Keine Änderung im Rapid Report:</u> Im Rapid Report wurde der aktuelle wissenschaftliche Erkenntnisstand zur Anwendung von Propensity-Score-Verfahren in kleinen Patientenkollektiven aufbereitet und die damit einhergehenden methodischen Herausforderungen (z. B. Problem der fehlenden Konvergenz) diskutiert. Weder in den Stellungnahmen noch im

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	praktische Umsetzbarkeit in solchen Settings infrage gestellt. Diese Diskrepanz führt zu Unsicherheiten hinsichtlich der methodischen Anforderungen [...]. <u>Vorgeschlagene Änderung:</u> Vor diesem Hintergrund wäre aus Sicht von MSD eine klarere methodische Orientierung wünschenswert, die sowohl die praktischen Herausforderungen kleiner Patientenkollektive als auch die Anforderungen an eine valide Confounderadjustierung berücksichtigt.	Rahmen der mündlichen Erörterung wurde Evidenz zur Anwendung alternativer Methoden zur kausalen Inferenz in kleinen Patientenkollektiven präsentiert.
84)	Pfizer Pharma GmbH [Abschnitt 5.7, S. 50 ff.] <u>Anmerkung:</u> [...] <u>Vorgeschlagene Änderung:</u> Im Rahmen der sich ständig weiterentwickelnden Methodendiskussion sollte auch der mögliche Einsatz anderer adäquater Analysemethoden eruiert bzw. die Möglichkeit eines Einsatzes offengelassen werden.	<u>Änderung im Rapid Report:</u> zu alternativen Methoden zu Propensity-Score-Verfahren siehe Würdigung in Zeile 8)
85)	Sanofi-Aventis Deutschland GmbH [Abschnitt 5.7, S. 50 ff.] <u>Anmerkung:</u> Die im Bericht vorgeschlagene Strategie [...], weniger wichtige Confounder schrittweise aus dem Modell zu entfernen, stellt [...] einen potenziell riskanten Ansatz dar, der vor allem aus der Praxis der PS-Modellierung stammt. Diese Vorgehensweise wird häufig gewählt, um Konvergenzprobleme bei einer großen Anzahl von Kovariaten zu vermeiden. Wie es in [VanderWeele, 2019] betont ist, kann eine rein statistisch motivierte Variablenselektion jedoch zu Verzerrungen führen, wenn relevante Confounder versehentlich ausgeschlossen werden. Alternativ stehen andere Verfahren zur Verfügung, die auch mit einer großen Anzahl von Kovariaten umgehen können, ohne die Modellstabilität zu gefährden. Dazu zählen baumbasierte Verfahren wie Random Forests [Breiman, 2001] sowie Regularisierungsansätze wie Lasso-Regression [Franklin, 2015], sehe auch [Zubizarreta, 2023] [...].	<u>Änderung im Rapid Report:</u> zu alternativen Methoden zu Propensity-Score-Verfahren siehe Würdigung in Zeile 8)

Tabelle 3: Argumentliste / Würdigung der Stellungnahmen zum vorläufigen Rapid Report (Sortierung nach Themenbereich; mehrseitige Tabelle)

Index	Wortlaut der Stellungnahme	Würdigung der Stellungnahme und Konsequenz
	<u>Vorgeschlagene Änderung:</u> Überarbeitung der Empfehlung zur schrittweisen Entfernung von Confoundern bei Konvergenzproblemen zugunsten einer Darstellung alternativer Verfahren, die auch bei hoher Kovariatenanzahl stabile Modellschätzungen ermöglichen. Eine solche Ergänzung würde die methodische Vielfalt abbilden und die Validität der Effektschätzungen stärken.	
86)	Verband Forschender Arzneimittelhersteller e. V. (vfa) [Abschnitt 5.7, S. 50–56] <u>Anmerkung:</u> Propensity-Score-Verfahren sind bei kleinen Fallzahlen methodisch anfällig. Es wird zudem darauf verwiesen, dass bisher keine Evidenz vorläge, um abschätzen zu können, wann Propensity-Score-Analysen in einer AbD zu interpretierbaren Ergebnissen führen. <u>Vorgeschlagene Änderung:</u> Auch hier ist weitergehend zu prüfen, ob eine Anwendung im Kontext der AbD, bei erwartbar kleinen Fallzahlen, zu interpretierbaren Ergebnissen führt.	<u>Keine Änderung im Rapid Report:</u> Eine Prüfung, ob Propensity-Score-Analysen im Kontext einer AbD zu interpretierbaren Ergebnissen führen, kann frühestens dann erfolgen, sobald (erste) Auswertungen der mit AbDs erhobenen Daten für die Durchführung einer erneuten Nutzenbewertung vorgelegt werden.
AbD: anwendungsbegleitende Datenerhebung; AMNOG: Arzneimittelmarktneuordnungsgesetz; AM-NutzenV: Arzneimittel-Nutzenbewertungsverordnung; BMS: Bristol-Myers Squibb GmbH & Co. KGaA; BPI: Bundesverband der Pharmazeutischen Industrie e. V.; CAR: chimärer Antigenrezeptor; CGT: Zell- und Gentherapie; DAG: gerichteter azyklischer Graph; DGKJ: Deutsche Gesellschaft für Kinder- und Jugendmedizin e. V.; DKG: Deutsche Krankenhausgesellschaft e. V.; EMA: Europäische Arzneimittel-Agentur; FDA: Food and Drug Administration; G-BA: Gemeinsamer Bundesausschuss; GKV: gesetzliche Krankenversicherung; GNP: Gesellschaft für Neuropsychologie e. V.; GSAV: Gesetz für mehr Sicherheit in der Arzneimittelversorgung; HR: Hazard Ratio; IQWiG: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen; ITT: Intention to treat; IPTW: Inverse Probability of Treatment Weighting; IPW: Inverse Probability Weighting; KöR: Körperschaft des öffentlichen Rechts; MAR: Missing at Random; MICE: multiple Imputation mittels Chained Equations; MMRM: gemischtes Modell mit Messwiederholungen; MNAR: Missing not at Random; OD: Orphan Drug; PRO: patientenberichteter Endpunkt; PS: Propensity Score; pU: pharmazeutischer Unternehmer; RCT: randomisierte kontrollierte Studie; RR: relatives Risiko; SGB: Sozialgesetzbuch; SMD: standardisierte Mittelwertdifferenz; SoC: Standard of Care; VerFO: Verfahrensordnung; vfa: Verband Forschender Arzneimittelhersteller e. V.; VIF: Varianzinflationsfaktor; zVT: zweckmäßige Vergleichstherapie		

Anhang C Offenlegung von Beziehungen der externen Sachverständigen

Im Folgenden sind die Beziehungen der externen Sachverständigen zusammenfassend dargestellt. Alle Informationen beruhen auf Selbstangaben der einzelnen Personen anhand des „Formblatts zur Offenlegung von Beziehungen“. Die in diesem Formblatt aufgeführten Fragen finden sich im Anschluss an diese Zusammenfassung.

Externe Sachverständige

Name	Frage 1	Frage 2	Frage 3	Frage 4	Frage 5	Frage 6	Frage 7
Regierer, Anne	ja	ja	ja	ja	nein	nein	nein

Im „Formblatt zur Offenlegung von Beziehungen“ wurden folgende 7 Fragen gestellt:

Frage 1: Sind oder waren Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor bei einer Einrichtung des Gesundheitswesens (z. B. einer Klinik, einer Einrichtung der Selbstverwaltung, einer Fachgesellschaft, einem Auftragsforschungsinstitut), einem pharmazeutischen Unternehmen, einem Medizinproduktehersteller oder einem industriellen Interessenverband angestellt oder für diese / dieses / diesen selbstständig oder ehrenamtlich tätig bzw. sind oder waren Sie freiberuflich in eigener Praxis tätig?

Frage 2: Beraten Sie oder haben Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor eine Einrichtung des Gesundheitswesens (z. B. eine Klinik, eine Einrichtung der Selbstverwaltung, eine Fachgesellschaft, ein Auftragsforschungsinstitut), ein pharmazeutisches Unternehmen, einen Medizinproduktehersteller oder einen industriellen Interessenverband beraten (z. B. als Gutachter/-in, Sachverständige/r, in Zusammenhang mit klinischen Studien als Mitglied eines sogenannten Advisory Boards / eines Data Safety Monitoring Boards [DSMB] oder Steering Committees)?

Frage 3: Haben Sie innerhalb des laufenden Jahres und der 3 Kalenderjahre davor direkt oder indirekt von einer Einrichtung des Gesundheitswesens (z. B. einer Klinik, einer Einrichtung der Selbstverwaltung, einer Fachgesellschaft, einem Auftragsforschungsinstitut), einem pharmazeutischen Unternehmen, einem Medizinproduktehersteller oder einem industriellen Interessenverband Honorare erhalten (z. B. für Vorträge, Schulungstätigkeiten, Stellungnahmen oder Artikel)?

Frage 4: Haben Sie oder hat Ihr Arbeitgeber bzw. Ihre Praxis oder die Institution, für die Sie ehrenamtlich tätig sind, innerhalb des laufenden Jahres und der 3 Kalenderjahre davor von einer Einrichtung des Gesundheitswesens (z. B. einer Klinik, einer Einrichtung der Selbstverwaltung, einer Fachgesellschaft, einem Auftragsforschungsinstitut), einem pharmazeutischen Unternehmen, einem Medizinproduktehersteller oder einem industriellen Interessenverband sogenannte Drittmittel erhalten (d. h. finanzielle Unterstützung z. B. für Forschungsaktivitäten, die Durchführung klinischer Studien, andere wissenschaftliche Leistungen oder Patentanmeldungen)? Sofern Sie in einer größeren Institution tätig sind, genügen Angaben zu Ihrer Arbeitseinheit, z. B. Klinikabteilung, Forschungsgruppe.

Frage 5: Haben Sie oder hat Ihr Arbeitgeber bzw. Ihre Praxis oder die Institution, für die Sie ehrenamtlich tätig sind, innerhalb des laufenden Jahres und der 3 Kalenderjahre davor sonstige finanzielle oder geldwerte Zuwendungen, z. B. Ausrüstung, Personal, Unterstützung bei der Ausrichtung einer Veranstaltung, Übernahme von Reisekosten oder Teilnahmegebühren für Fortbildungen / Kongresse erhalten von einer Einrichtung des Gesundheitswesens (z. B. einer Klinik, einer Einrichtung der Selbstverwaltung, einer Fachgesellschaft,

einem Auftragsforschungsinstitut), einem pharmazeutischen Unternehmen, einem Medizinproduktehersteller oder einem industriellen Interessenverband? Sofern Sie in einer größeren Institution tätig sind, genügen Angaben zu Ihrer Arbeitseinheit, z. B. Klinikabteilung, Forschungsgruppe.

Frage 6: Besitzen Sie Aktien, Optionsscheine oder sonstige Geschäftsanteile einer Einrichtung des Gesundheitswesens (z. B. einer Klinik, einem Auftragsforschungsinstitut), eines pharmazeutischen Unternehmens, eines Medizinprodukteherstellers oder eines industriellen Interessenverbands? Besitzen Sie Anteile eines sogenannten Branchenfonds, der auf pharmazeutische Unternehmen oder Medizinproduktehersteller ausgerichtet ist? Besitzen Sie Patente für ein pharmazeutisches Erzeugnis, ein Medizinprodukt, eine medizinische Methode oder Gebrauchsmuster für ein pharmazeutisches Erzeugnis oder ein Medizinprodukt?

Frage 7: Sind oder waren Sie jemals an der Erstellung einer medizinischen Leitlinie oder klinischen Studie beteiligt, die eine mit diesem Projekt vergleichbare Thematik behandelt/e? Gibt es sonstige Umstände, die aus Sicht von unvoreingenommenen Betrachtenden als Interessenkonflikt bewertet werden können, z. B. Aktivitäten in gesundheitsbezogenen Interessengruppierungen bzw. Selbsthilfegruppen, politische, akademische, wissenschaftliche oder persönliche Interessen?