

## **Bewertung und Auswertung von Studien bei seltenen Erkrankungen**

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)

### **Hintergrund**

Das Bundesgesundheitsministerium (BMG) hat mit Schreiben vom 10.12.2013 das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) mit der Erstellung eines Rapid Reports zum Thema „Bewertung und Auswertung von Studien bei seltenen Erkrankungen“ beauftragt.

### **Fragestellung**

Die Ziele des vorliegenden Rapid Reports sind

- die Erstellung einer Expertise zu methodischen Aspekten bei der Durchführung und Auswertung sowie zur Bewertung der Ergebnissicherheit von Studien bei seltenen Erkrankungen,
- die Beschreibung der Studiengrundlage für die Zulassung von Orphan Drugs in Europa.

### **Methoden**

#### ***Methodische Expertise***

In der methodischen Expertise wird die besondere Problematik bei der Durchführung und Bewertung von Studien zu seltenen Erkrankungen dargestellt. Dabei wird unterschieden zwischen seltenen ( $\leq 5/10.000$ ) und sehr seltenen Erkrankungen ( $< 2/100.000$  Einwohner). Es werden Kriterien der Ergebnis- und Aussagesicherheit von häufig bei seltenen Erkrankungen eingesetzten bzw. vorgeschlagenen Studiendesigns als Grundlage für Nutzenbewertungen dargestellt und diskutiert.

#### ***Empirische Untersuchung zur Studiengrundlage für die Zulassung von Orphan Drugs in Europa***

Die Recherche für die empirische Untersuchung der Studiengrundlage für die Zulassung von Orphan Drugs in Europa erfolgte im Orphanet. Für die Extraktion der relevanten Daten zu den identifizierten Orphan Drugs wurden European Public Assessment Reports (EPARs) zu den Zulassungsstudien herangezogen. Relevante Charakteristika des Arzneimittels sowie der zugrunde liegenden Zulassungsstudien wurden extrahiert und mithilfe deskriptiver statistischer Verfahren ausgewertet.

### **Ergebnisse**

#### ***Methodische Expertise***

##### *Grundlegende methodische Aspekte bei der Bewertung und Auswertung von Studien*

Nach internationaler Übereinkunft ist die wissenschaftliche Grundlage der evidenzbasierten Medizin (EbM), im Hinblick auf eine klar definierte Fragestellung (gemäß PICOS [Patient, Intervention, Comparator, Outcome, Setting/Study design]-Schema) systematisch die zur Beantwortung dieser Fragestellung geeigneten klinischen Studien zu identifizieren, die Sicherheit der Ergebnisse der identifizierten Studien in

nachvollziehbarer Weise zu bewerten und auf Basis der beobachteten Daten sowie unter Berücksichtigung der Ergebnissicherheit eine zusammenfassende Bewertung abzugeben. Im Hinblick auf die Aussagesicherheit für die Beantwortung der zugrunde liegenden Fragestellung gemäß dem PICOS-Schema spielen 4 Komponenten eine Rolle. Die Sicherheit der Ergebnisse der identifizierten klinischen Studien wird dabei in erster Linie durch 3 Komponenten getragen:

- 1) eine qualitative Komponente, die durch das Verzerrungspotenzial der zu bewertenden Studien gekennzeichnet ist (interne Validität),
- 2) eine quantitative Komponente, die durch die Fallzahl(en), aber auch durch die Varianz der Beobachtungen bestimmt wird (Präzision der Ergebnisse),
- 3) die Größe beobachteter Unterschiede (Effektstärke).

Hinzu kommt noch eine weitere Komponente hinsichtlich der Aussagesicherheit:

- 4) die externe Validität (oder auch Anwendbarkeit), d. h. der Aspekt, inwieweit die Studienbedingungen ebendiese Fragestellung abbilden.

Für die Bewertung des Verzerrungspotenzials lassen sich 6 Kategorien von Verzerrungen (bias) abgrenzen, wobei es bei den ersten 4 Kategorien um Unterschiede zwischen den zu vergleichenden Interventionsgruppen geht: Selektionsbias (selection bias), Durchführungsbias (performance bias), Entdeckungsbias (detection bias), Abtriebbias (attrition bias), Berichtsbias (reporting bias) und sonstige Verzerrungsmöglichkeiten, z. B. die Verwendung ungeeigneter statistischer Methoden.

Zur Vermeidung der ersten 4 genannten Verzerrungsmöglichkeiten lassen sich im Wesentlichen 3 Strategien einsetzen: randomisierte und verdeckte (concealed) Zuteilung von Patienten in die zu vergleichenden Interventionsgruppen, die Verblindung der Interventionen für Patienten und Studienpersonal (auch als Doppelblindheit bezeichnet) sowie die Auswertung aller in eine Studie eingeschlossenen Patienten gemäß der ihnen zugewiesenen Intervention (Intention-to-treat-Prinzip [ITT-Prinzip]).

#### *Barrieren gegen randomisierte Studien*

Alternativen zur Randomisierung, um insbesondere einen Selektionsbias zu vermeiden, sind derzeit nicht bekannt. Dennoch werden mitunter Argumente gegen randomisierte kontrollierte Studien vorgebracht, die die Durchführbarkeit, häufig aufgrund nicht näher spezifizierter ethischer Bedenken oder aber auch wegen logistischer Probleme, betreffen.

Ethisch zweifelhaft ist eine kontrollierte Studie in der Tat dann, wenn der Nutzen (oder Zusatznutzen im Vergleich zum Standard) einer Intervention in einer bestimmten Indikation mehr oder weniger belegt ist. Dann erübrigen sich aber Studien mit dem Ziel des Erkenntnisgewinnes im Hinblick auf die Frage nach dem Nutzen oder Zusatznutzen ohnehin, und das nicht nur bei seltenen Erkrankungen.

2 Szenarien könnten gegen eine Randomisierung sprechen: Die Durchführung einer adäquaten zentralen Randomisierung mit Wahrung des Concealment ist zu aufwendig oder eine bestimmte Intervention ist aus verschiedenen Gründen (z. B. mengenmäßig oder auch wegen mangelnder Expertise bei der Anwen-

dung) für Studienzwecke nicht ausreichend verfügbar und kann auch nicht kurzfristig verfügbar gemacht werden.

#### *Situation bei seltenen Erkrankungen*

Es ist offensichtlich, dass es immer schwieriger wird, eine für die Beantwortung einer spezifischen Fragestellung ausreichende Fallzahl zu erreichen, je seltener diese Erkrankung ist. Somit wird die Ergebnissicherheit bei seltenen Erkrankungen vor allem im Hinblick auf die Präzision beeinträchtigt. Dies hat u. a. Auswirkungen auf den Fehler 2. Art (einen vorhandenen Unterschied nicht zu finden). Diese Fehlerwahrscheinlichkeit hängt von dem tatsächlichen Unterschied (von der „Wahrheit“) ab und kann somit nur bedingt über die durch Annahmen über diese „Wahrheit“ begründete Fallzahl kontrolliert werden. Demgegenüber kann der Fehler 1. Art (fälschlicherweise einen Unterschied anzunehmen, obwohl er gar nicht existiert) auch bei kleinen Fallzahlen über die Festlegung des Signifikanzniveaus zumindest theoretisch kontrolliert werden.

Die beiden anderen beeinflussbaren Komponenten der Ergebnis- bzw. Aussagesicherheit (interne und externe Validität) sind demgegenüber nicht quantifizierbar und können zu systematischen Fehlern (Abweichungen von der „Wahrheit“) führen. Diesbezügliche Schwächen im Design einer Studie lassen sich insbesondere im Hinblick auf das Verzerrungspotenzial in aller Regel nicht durch noch so ausgefeilte statistische Methoden ausgleichen. Deshalb ist es im Kontext der Bewertung medizinischer Interventionen bei seltenen Erkrankungen naheliegend, zunächst nach Möglichkeiten zu suchen, die Beschränkungen durch die ggf. erniedrigte Präzision durch optimierte statistische Verfahren auszugleichen.

#### *Design-Vorschläge für randomisierte Therapiestudien bei seltenen Erkrankungen*

Ausgangspunkt für vorgeschlagene Modifikationen ist das klassische randomisierte Parallelgruppen-Design, bei dem die Patienten auf Basis eines Zufallsmechanismus einer von 2 oder mehr unterschiedlichen Interventionen zugewiesen und über eine vorab definierte Zeit beobachtet werden. Die Verfahren und Designs lassen sich grob in 5 Gruppen unterteilen, wobei die Abgrenzung nicht scharf gezogen werden kann:

- Designs zur Reduktion der Varianz zwischen den Beobachtungen, z. B. Stratifizierung, regressions-analytische Verfahren.
- Verfahren mit Anpassungen des Designs im Studienverlauf, z. B. sequenzielle Designs.
- Verfahren, bei denen Vorinformationen (außerhalb der Studie) in die statistische Auswertung mit einfließen (Bayes'sche Verfahren).
- Designs, bei denen alle Patienten auch die zu prüfende Therapie erhalten.
- Sonstige Designs, z. B. adaptive Randomisierung.

#### *Ergebnissicherheit nicht randomisierter Studien*

Nicht randomisierte Studien bergen inhärent erhebliche Einschränkungen bei der internen Validität (1. Komponente der Ergebnissicherheit). Gegenwärtig ist keine Methode bekannt, die in suffizienter Weise diese Einschränkungen aufheben könnte. Je nach Grund für einen eventuellen Verzicht auf eine Randomisierung ergeben sich unmittelbare Konsequenzen für ersatzweise nicht randomisierte Studien.

Werden ethische Bedenken geltend gemacht, verbieten sich jegliche parallel vergleichenden Studien, und es verbleiben im Prinzip nur noch historische Kontrollen.

Historisch kontrollierte Studien stehen in den gängigen Evidenzhierarchien weit unten. Neben den üblichen Verzerrungsmöglichkeiten (z. B. Selektionsbias) kommt hier als weitere Störgröße der Faktor Zeit oder Chronologie dazu. Ergebnisse aus historisch kontrollierten Studien lassen Aussagen im Sinne eines Interventionseffektes nur dann zu, wenn (nahezu) eine Umkehr eines quasi deterministischen Verlaufs beobachtet wurde („dramatischer Effekt“) vorliegt. In der Literatur wird als Kriterium für das Vorliegen eines dramatischen Effekts ein um den Faktor 10 erhöhtes Risiko in der Interventionsgruppe im Vergleich zur Kontrollintervention in Verbindung mit einem (adäquaten) statistischen Test zum Irrtumsniveau 1 % vorgeschlagen.

Von prospektiv geplanten Interventionsstudien werden Beobachtungsstudien abgegrenzt, deren Sinn und Zweck entweder weniger der Frage nach den Effekten von Interventionen nachgehen oder deren Datengrundlage primär nicht der Gewinnung von medizinisch-klinischen Fragen dient. In den gängigen Evidenzhierarchien werden sie deshalb in ihrer Aussagekraft den randomisierten und nicht randomisierten Interventionsstudien nachgeordnet.

### *Empfehlungen*

Es lässt sich keine wissenschaftliche Begründung für eine unterschiedliche Herangehensweise bei der Bewertung von medizinischen Interventionen für seltene und nicht seltene Erkrankungen ableiten. Studien mit geringer Präzision oder mit nicht ausreichenden Schutzmechanismen vor potenziell verzerrenden Faktoren oder mit Abweichungen der Studienbedingungen von der eigentlich interessierenden Fragestellung (z. B. Verwendung von Surrogatendpunkten anstelle von patientenrelevanten Endpunkten) haben die gleichen Auswirkungen auf die Aussagesicherheit bei seltenen und nicht seltenen Erkrankungen. Umgekehrt existieren keine spezifischen Designs und statistischen Methoden für seltene Erkrankungen, die nicht auch relevant für häufige(re) Erkrankungen sein könnten.

Im Kontext der Bewertung medizinischer Interventionen bei seltenen Erkrankungen wird es somit (besonders) notwendig sein, ein möglichst effizientes methodisches Vorgehen zu wählen und dabei auch effiziente Strukturen zu schaffen. Darüber hinaus kann es, insbesondere bei sehr seltenen Erkrankungen, erforderlich sein, Kompromisse bei der Aussagesicherheit einzugehen, die auch auf externen (politischen) Vorgaben beruhen können.

Im Rahmen des Arzneimittelmarktneuordnungsgesetzes (AMNOG) hat der Gesetzgeber in Deutschland politische Vorgaben gemacht, indem der Zusatznutzen von Arzneimitteln für die Behandlung seltener Erkrankungen (Orphan Drugs) qua Zulassung als belegt gilt, allerdings nur so lange, wie der Umsatz eines solchen Arzneimittels mit der gesetzlichen Krankenversicherung innerhalb eines Zeitraums von 12 Kalendermonaten einen Betrag von 50 Millionen € nicht übersteigt. Einerseits findet sich in der Begründung das inhaltliche Argument, dass regelmäßig davon auszugehen ist, dass es keine therapeutisch gleichwertige Behandlungsalternative gibt. Andererseits macht die Begründung auch die politische Vorgabe deutlich, indem der Zusatznutzen am Umsatz des Arzneimittels festgemacht wird. Wissenschaftlich ist es nicht zu begründen, warum inhaltliche Kriterien, die – wenn sie denn zutreffen – eine Bewertung (des [Zusatz-]Nutzens) obsolet machen, ab einem gewissen Umsatzvolumen nicht mehr zutreffen (sollen).

Bei seltenen Erkrankungen wird es mit abnehmender Häufigkeit der Erkrankung zunehmend schwieriger, der quantitativen Komponente der Ergebnissicherheit ausreichend Rechnung zu tragen. Eine diesbezüglich im Prinzip triviale Feststellung lautet daher, dass es für die klinische, patientenorientierte Erforschung

seltener Erkrankungen besonders notwendig ist, in vernetzten, überregionalen und supranationalen Strukturen zu arbeiten. Daraus lässt sich ein starkes Plädoyer für Krankheitsregister mit klaren Qualitätskriterien bezüglich Vollständigkeit und Vollständigkeit als Basis für hochwertige klinische, vor allem nicht randomisierte Studien ableiten.

Wenn es schon inhärent Beschränkungen bezüglich einer Komponente der Ergebnissicherheit gibt, scheint es wenig zielführend auch die anderen Komponenten zu kompromittieren. Von den zuvor beschriebenen Strategien wären somit diejenigen zu bevorzugen, die bei weitgehendem Erhalt von interner und externer Validität Fallzahleinsparungen ermöglichen, also sequenzielle Designs.

Bei sehr seltenen Erkrankungen könnte alternativ oder zusätzlich zum gängigen methodischen Vorgehen erwogen werden, ein größeres statistisches Irrtumsniveau für regulative Entscheidungen zuzulassen, z. B. das übliche zweiseitige Irrtumsniveau von den üblichen 5 % auf 10 % anzuheben. Der Vorteil einer solchen Vorgehensweise wäre, die Irrtumsmöglichkeit zumindest quantifizieren zu können. Ein derartiges Vorgehen kann auch als Annäherung an Bayes'sche Verfahren verstanden werden: Ob nämlich die Daten aus einer randomisierten Studie mit geringer Fallzahl mit einem erhöhten Irrtumsniveau zur Entscheidungsfindung herangezogen werden oder ob bei Verknüpfung mit z. B. optimistischem Vorwissen aus einer anderen Indikation (Prior) das übliche Irrtumsniveau beibehalten wird, dürfte auf vergleichbare Ergebnisse im Hinblick auf resultierende Entscheidungen hinauslaufen.

In absteigender Priorität könnten auch Einschränkungen der externen Validität hingenommen werden, z. B. durch den Einbezug von Daten aus ähnlichen Indikationsgebieten oder durch den Einsatz von etablierten Surrogatendpunkten innerhalb kombinierter Endpunkte. Auch der Einsatz von adaptiven Designs wird in der Regel mit Einschränkungen der externen Validität verbunden sein. Immerhin bliebe aber wenigstens die interne Validität erhalten.

Die Aufgabe der internen Validität durch Verzicht auf eine Randomisierung stünde in einem solchen hierarchischen Vorgehen aus logischen Gründen an letzter Stelle. Wesentliche Voraussetzung, um sie für (regulatorische) Entscheidungen nutzen zu können, wäre, dass die zugrunde liegenden Daten aus einem Krankheitsregister mit den o. g. Qualitätskriterien stammen oder dass ein derartig großer beobachteter Unterschied vorliegt, der nicht mehr allein durch Bias erklärt werden kann.

## **Ergebnisse der empirischen Untersuchung zur Studiengrundlage für die Zulassung von Orphan Drugs in Europa**

Es wurden 85 Arzneimittel mit europäischer Orphan Drug Designation und europäischer Marktzulassung von 2001 bis 2013 identifiziert. Die Zulassung der 85 Arzneimittel stützte sich auf 125 Hauptstudien (ohne 6 Zulassungen, die auf Literatur-Reviews basierten); darunter waren 82 RCTs.

### **Auswertungen auf Arzneimittelzebene**

#### *Alle Zulassungen*

58 der identifizierten 85 Arzneimittel (68 %) dienen der Behandlung seltener, 27 Arzneimittel (32 %) der Behandlung sehr seltener Erkrankungen. Bei 59 Arzneimitteln (69 %) basierte die Zulassung auf RCTs (55 ausschließlich auf RCTs und 4 auf RCTs in Kombination mit Non-RCTs). Bei 20 Arzneimitteln (24 %) dienten Non-RCTs als Grundlage für die Zulassung.

### *Zulassungen ohne Literatur-Reviews*

Die Zulassung der 79 Orphan Drugs ohne Literatur-Reviews beruhte auf jeweils 1 bis 5 Hauptstudien. Bei sehr seltenen Erkrankungen waren maximal 3 Studien Basis für die Zulassung. Der Anteil von Zulassungen basierend auf Daten aus RCTs betrug etwa 75 %. Die Patientenzahl der Zulassungsstudien pro Arzneimittel betrug zwischen 27 und 2961 Patienten (Median 165). Etwa 70 % aller Patienten wurden in RCTs behandelt. Bei 66 Arzneimitteln (84 %) wurden patientenrelevante Endpunkte in den Studien berücksichtigt (bei 31 Orphan Drugs als primärer, bei 57 als sekundärer Endpunkt). Der Großteil der Studien wurde multizentrisch, multinational und multikontinental durchgeführt. In der Mehrzahl der Fälle zeigten sich keine auffälligen Unterschiede zwischen seltenen und sehr seltenen Erkrankungen in den dargestellten Studiencharakteristika.

### **Auswertungen auf Studienebene**

Grundlage der deskriptiven Auswertungen zu Charakteristika und Methodik der Zulassungsstudien sind alle 82 RCTs (66 %) unter den 125 Hauptstudien (ohne Zulassungen basierend auf Literatur-Reviews) der zugelassenen Orphan Drugs (1 bis 3 RCTs pro Arzneimittel). In den RCTs wurden zwischen 8 und 769 Patienten (Median 160,5) eingeschlossen.

Der Anteil doppelblinder RCTs liegt bei 74 %. Bei 28 % der Studien wurde ein aktiver Komparator eingesetzt. Nichtunterlegenheits- bzw. Äquivalenzstudien kamen mit 6 % der Studien nur in Ausnahmefällen zum Einsatz. Es zeigen sich keine auffälligen Unterschiede zwischen den Studien zu seltenen und sehr seltenen Erkrankungen.

In 52 % der RCTs kamen Methoden zur Kontrolle von Störgrößen zum Einsatz. Cross-over-Designs und sequenzielle Verfahren wurden in 5 % bzw. 12 % der Studien verwendet. Eine Erhöhung des Irrtumsniveaus wurde in 1 Studie vorgenommen. Eine adaptive Randomisierung, Bayes'sche Verfahren oder sonstige spezielle randomisierte Designs wurden in keiner der Zulassungsstudien eingesetzt. Insgesamt wurden in den Studien zu sehr seltenen Erkrankungen weniger häufig spezielle Auswertungsmethoden eingesetzt, sequenzielle Verfahren kamen gar nicht zur Anwendung.

### **Fazit**

Eine Begründung für eine unterschiedliche Herangehensweise bei der Bewertung von medizinischen Interventionen für seltene gegenüber nicht seltenen Erkrankungen kann wissenschaftlich nicht abgeleitet werden. Umgekehrt existieren auch keine spezifischen Designs und statistischen Methoden für seltene Erkrankungen, die nicht auch relevant für häufige(re) Erkrankungen sein könnten. Dies gilt in gleicher Weise für medikamentöse wie für nichtmedikamentöse Interventionen.

Zulassungen und Zulassungsstudien für Orphan Drugs basieren zu einem großen Teil, auch bei sehr seltenen Erkrankungen, auf konventionellen (randomisierten) Designs, sodass die grundsätzliche Machbarkeit nicht infrage steht.

Im Kontext der Bewertung medizinischer Interventionen bei seltenen, insbesondere sehr seltenen Erkrankungen kann es dennoch notwendig oder auch politisch gewünscht sein, Kompromisse bei der Aussagesicherheit einzugehen. Solche Kompromisse sind grundsätzlich auf 3 Ebenen denkbar:



# Health Technology Assessment im Auftrag des

Das (statistische) Irrtumsniveau könnte über den üblichen Wert von (zweiseitig) 5 % angehoben werden (Kompromiss bei der geforderten Präzision).

In absteigender Priorität könnten auch Einschränkungen der externen Validität hingenommen werden, z. B. durch den Einbezug von Daten aus ähnlichen Indikationsgebieten oder durch den Einsatz von etablierten Surrogatendpunkten innerhalb kombinierter Endpunkte.

Die Aufgabe der internen Validität durch Verzicht auf eine Randomisierung stünde in einem hierarchischen Vorgehen aus logischen Gründen an letzter Stelle. Wesentliche Voraussetzung, um sie für (regulatorische) Entscheidungen nutzen zu können, wäre, dass die zugrunde liegenden Daten aus einem Krankheitsregister mit exzellenter Qualität im Hinblick auf Vollständigkeit und Vollzähligkeit stammen.

**Schlagwörter:** Seltene Krankheiten, Nutzenbewertung

**Keywords:** Rare Diseases, Benefit Assessment

**Der deutsche Volltext ist erhältlich unter**

**[https://www.iqwig.de/download/MB13-01\\_Rapid-Report\\_Studien-bei-seltenen-Erkrankungen.pdf](https://www.iqwig.de/download/MB13-01_Rapid-Report_Studien-bei-seltenen-Erkrankungen.pdf)**