AMNOG: Anforderungen und Herausforderungen bei der Analyse und Bewertung von Lebensqualitätsdaten

Es braucht neue Ansätze für Lebensqualitätsdaten

Seit Inkrafttreten des Arzneimittelmarktneuordnungsgesetzes (AMNOG) im Jahr 2011 in Deutschland ist der Stellenwert von Lebensqualitätsdaten in der Nutzenbewertung von Arzneimitteln deutlich gestiegen. Das Gleiche gilt jedoch auch für die methodischen und konzeptionellen Anforderungen, die hinsichtlich ihrer Erhebung und Auswertung im Rahmen klinischer Studien gestellt werden. Aktuelle Problemfelder umfassen unter anderem die Validierung der zur Erhebung konzipierten Messinstrumente, die Etablierung akzeptierter Relevanzschwellen für die Abbildung eines klinisch bedeutsamen Effekts, den Umgang mit ungenügendem Rücklauf der Fragebögen und Konzepte zur statistischen Auswertung von Lebensqualitätsdaten im Rahmen von Studien mit Überlebenszeitendpunkten. Ansätze zur Lösung dieser Probleme werden in einem Feld, in dem die unterschiedlichen Interessen von akademischer Forschung, den für die Bewertungen zuständigen Institutionen und den Arzneimittelherstellern aufeinandertreffen, diskutiert. Ein konstruktiver Dialog aller Beteiligten ist für die Schaffung tragfähiger Konzepte erforderlich.

>> Seit Inkrafttreten des Arzneimittelmarktneuordnungsgesetzes (AMNOG) im Jahr 2011 sind pharmazeutische Unternehmen verpflichtet, den Zusatznutzen eines neu zugelassenen Arzneimittels, einer neuen Kombination oder eines bereits zugelassenen Arzneimittels, dessen Zulassung erweitert wurde, gegenüber einer oder mehrerer Standardtherapien nachzuweisen. Das Ergebnis dieser Nutzenbewertung (NB) nach § 35a SGB V bildet die Basis für die anschließend stattfindenden Verhandlungen zum Erstattungsbetrag mit dem Spitzenverband der Gesetzlichen Krankenkassen.

Zentrales Element der NB ist das vom pharmazeutischen Unternehmer (pU) einzureichende Nutzendossier, in dem der Zusatznutzen gegenüber der sogenannten zweckmäßigen Vergleichstherapie – die vom Komparator in den pivotalen Zulassungsstudien abweichen kann – anhand patientenrelevanter Endpunkte zu belegen ist. Die NB wird vom Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) oder vom Gemeinsamen Bundesausschuss (G-BA) basierend auf dem Nutzendossier des pU durchgeführt. Der G-BA beschließt anschließend über die Anerkennung eines Zusatznutzens. Die methodischen Anforderungen an die NB werden in der Verfahrensordnung des G-BA (Gemeinsamer Bundesausschuss 2018) und in den Allgemeinen Methoden des IQWiG (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020) beschrieben.

Der Stellenwert von Lebensqualitätsdaten im Rahmen der NB

Neben Mortalität und Morbidität ist die gesundheitsbezogene Lebensqualität (health-related quality of life, HRQoL) eine maßgebliche patientenrelevante Zielgröße für die Beurteilung des therapeutischen Nutzens eines Arzneimittels (Gemeinsamer Bundesausschuss 2018; Bundesministerium für Gesundheit 2010). Der Stellenwert wird auch vom IQWiG hervorgehoben: „Der gesundheitsbezogenen Lebensqualität wird die gleiche Bedeutung beigemessen wie schwerwiegenden (bzw. schweren) Symptomen, Folgekomplikationen und Nebenwirkungen." (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020: 204).

Die HRQoL hat seit Inkrafttreten des AMNOG einen immer größeren Stellenwert in der Bewertung des Zusatznutzens neuer Arzneimittel eingenommen. Der Vorsitzende des G-BA, Prof. Josef Hecken, betont, wie wichtig die Lebensqualität (LQ) bei der Abwägungsentscheidung zum Zusatznutzen sei (Deutscher Bundestag 2014; Staeck 2016). So werte er „die Lebensqualität von Patienten in der letzten Lebensphase höher als eine marginale Verlängerung der Lebenszeit" (Staeck 2016). Fehlende adäquate Daten zur HRQoL könnten, weil sie bspw. nicht erhoben wurden oder weil die verwendeten Messinstrumente oder deren klinische Relevanzschwellen nicht validiert wurden, folglich auch zu Konsequenzen in der NB in Form eines nicht hinreichend belegten Zusatznutzens oder einer Befristung des Beschlusses des G-BA führen (Staeck 2016; Gemeinsamer Bundesausschuss 2018).

Im Folgenden werden methodische Anforderungen bei der Analyse und Bewertung von LQ-Daten im Kontext der NB sowie aktuelle Herausforderungen beschrieben und diskutiert.

Methodische Anforderungen an LQ-Daten in der Nutzenbewertung

Aus Sicht des IQWiG sind bei der Analyse von LQ-Daten die gleichen methodischen Anforderungen an Studiendesign und Analyse und Bewertung der Daten zu erfüllen wie bei anderen patientenrelevanten Endpunkten (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020); dass der G-BA sich dieser Sichtweise anschließt, ist aus den bereits über 560 abgeschlossenen Bewertungsverfahren (Stand: April 2021) erkennbar. Eine methodisch fundierte Bewertung der LQ erfordert psychometrisch validierte Messinstrumente (z. B. Fragebögen), was sowohl krankheitsspezifische als auch generische Instrumente einschließt. Die Validierung solcher Instrumente muss anhand von Daten dokumentiert werden, welche die strukturelle Integrität des Messinstruments in der Zielpopulation und den Vorhersagewert des Messinstruments für den patientenrelevanten Endpunkt belegen. Generische Messinstrumente allein sind häufig nicht geeignet, um therapiespezifische Effekte zu erfassen, da sie nicht hinreichend sensitiv sind, um Unterschiede in klinischen Populationen mit einer häufig niedrigen LQ abzubilden (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020). Zudem müssen sie auch im betreffenden Krankheitsgebiet validiert worden sein.

Neben validierten Messinstrumenten sind auch akzeptierte Relevanzschwellen, z. B. in Form einer Minimal Important Difference (MID), erforderlich, um auf Basis von LQ-Daten einen Zusatznutzen ableiten zu können. Anhand dieser Relevanzschwellen, deren Validität dabei für die aktuelle bzw. verwendete Version des Messinstruments zu belegen ist, werden individuell Kriterien für das Vorliegen einer Therapieantwort („Response") definiert (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020). Eine Bewertung der Relevanz kann auf Basis von Responderanalysen als auch standardisierter

Mittelwertdifferenzen (SMDs, in Form von Hedges' g) durchgeführt werden. Kritik an Responderanalysen auf Basis einer individuellen MID wurde in der aktuellen Version 6.0 der Allgemeinen Methoden des IQWiG hinsichtlich Methodik und weiterer Kontextfaktoren geübt: Wurde die MID eines validierten Messinstruments für eine spezifische Entität (noch) nicht empirisch ermittelt, was bei generischen Instrumenten häufiger der Fall ist (Klakow-Franck 2014), kann diese nicht als Relevanzschwelle verwendet werden. Es kann nicht davon ausgegangen werden, dass eine für eine Entität gültige MID auf eine andere Patientengruppe übertragbar ist.

Entsprechend der Methodik des IQWiG sind für LQ-Daten Responderanalysen unter Verwendung eines validen Responsekriteriums – z.B. der MID – durchzuführen (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020). Bei Vorliegen patientenindividuell unterschiedlich langer Beobachtungszeiten, wie sie bspw. in onkologischen Studien sehr häufig zu finden sind, werden Ereigniszeitanalysen, die die Zeit bis zum Auftreten eines Ereignisses mit berücksichtigen (Unkel et al. 2019), als die adäquate Methodik angesehen.

Als Ereignis wird bei einem Patienten je nach Entität und Krankheitsstadium bspw. die Verbesserung oder Verschlechterung einer Skala des Messinstruments, definiert als hinreichend sicher für die Patienten spürbare Veränderung gegenüber dem Ausgangswert bei Studieneinschluss, gewertet. Einen etablierten Standard, mit dem die Aussagekraft der herangezogenen MIDs abgeschätzt werden kann, gibt es bislang nicht. Für einen im Jahr 2020 veröffentlichten ersten Vorschlag für ein Instrument zur Qualitätsbewertung von Studien zur Ermittlung von MIDs und deren Aussagekraft fehlen bisher noch umfassende Ergebnisse (Devji et al. 2020) und die dort angewandten Kriterien zur Bewertung der Qualität von Ankern zur Festlegung von Schwellenwerten werden von Experten nicht unkritisch gesehen (Peipert 2020). Das IQWiG definiert aktuell – statt bisher eine Veränderung um mindestens die validierte MID – einen Wert von 15% der Skalenspannweite als plausiblen Schwellenwert für eine zwar kleine, aber hinreichend sicher spürbare Veränderung (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020).

Liegt das präspezifizierte Responsekriterium im Sinne einer MID unterhalb von 15% der Skalenspannweite, können neben einer post-hoc Definition des Responsekriteriums von genau 15% der Skalenspannweite auch stetige Modelle (z. B. Gemischte Modelle für wiederholte Messungen) unter Hinzunahme von normativen Effektschätzern wie Hedges' g verwendet werden. Unter der Kenntnis der zeitlichen Stabilität der Skala in der Zielpopulation können solche normativen Effektschätzer prinzipiell genutzt werden, um die Verteilung von Veränderungsmaßen zu schätzen (Morris & DeShon 2002), mit denen auch Responseraten auf Basis möglicher MID-Szenarien diskutiert werden können.

Der Rücklauf der Fragebögen sollte zur Verwertbarkeit der Daten hinreichend groß sein. In der Regel wird hier eine Schwelle von 70% angesetzt, d.h. die Ergebnisse sollen auf mindestens 70% der eligiblen Studienteilnehmenden basieren. Das IQWiG differenziert aktuell in einem zweiten Schritt die Follow-up-Verluste und beschreibt, dass diese ggf. durch adäquate Ersetzungsstrategien ausgeglichen werden können (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020).

Die statistische Auswertung von LQ-Daten im Rahmen von Studien mit Überlebenszeitendpunkt stellt, unter anderem durch unterschiedlich lange Nachbeobachtungszeiten und Zensierungen, besondere Anforderungen an die statistische Analyse.

Die im Folgenden näher beschriebenen Aspekte der Analyse und Bewertung von LQ-Daten stellen eine Auswahl wichtiger und aktuell meist diskutierter Teilbereiche im Kontext der NB dar. Weitere relevante Aspekte, wie die Validität der Messinstrumente, Boden- und Deckeneffekte, Größenordnung des Effekts (jenseits der Relevanzschwelle) werden hier nicht näher betrachtet, sollten jedoch im Rahmen einer weiterführenden methodischen Auseinandersetzung berücksichtigt werden.

Kommentar zu Responseschwellen und MID

Im Rahmen der Bewertung der Relevanz beobachteter Veränderungen auf einer LQ-Skala ist nicht nur deren statistische Signifikanz, sondern insbesondere auch deren klinische Relevanz zu berücksichtigen. Hierzu ist die Definition eines Schwellenwertes, allgemein als Responseschwelle bezeichnet, als Kriterium für eine relevante Veränderung auf der LQ-Skala erforderlich.

Während eine Responseschwelle nicht zwingend eine minimale Schwelle für relevante Veränderungen darstellt, handelt es sich bei der MID um den Wert, der den kleinsten klinischen relevanten Unterschied zwischen Zeitpunkten definiert. Die Festlegung der MID ist jedoch mit Herausforderungen verbunden.

Ein erstes Problem besteht bereits in der uneinheitlichen Verwendung des Begriffes der MID. In randomisierten klinischen Studien, in denen Patientengruppen (= die Studienarme) miteinander verglichen werden, wäre der kleinste klinisch relevante Unterschied zwischen Gruppen zu bestimmen. Die „Relevanz" des Unterschiedes ist jedoch schwierig zu bewerten. Bezieht man sich hierbei auf externe Kriterien wie bspw. die Funktionsfähigkeit oder Toxizitätsprofile, vergleicht man verschiedene Konstrukte. Die inhaltliche Zulässigkeit ist fraglich, da es sich bei externen Kriterien um Fremdeinschätzungen durch den Arzt handelt, LQ jedoch per definitionem ein subjektives Konstrukt ist. Ein Heranziehen der Fremdeinschätzung des Arztes als externes Kriterium für die LQ wäre ausschließlich zulässig, wenn belegt werden könnte, dass diese mit der Selbsteinschätzung der Patienten übereinstimmt, was in der Regel nicht der Fall ist. Daher wird meist ein zweiter Weg gewählt: Die LQ-Scores von Patienten zu zwei Zeitpunkten werden miteinander verglichen. Zusätzlich werden die Patienten befragt, inwieweit ihre LQ besser oder schlechter geworden ist. Die mittlere Differenz in den LQ-Scores bei den Patienten, die minimale Veränderungen in der LQ angegeben haben, ist die MID. Hier prüft man also den kleinsten klinisch relevanten Unterschied zwischen Zeitpunkten. Da der Vergleich zwischen Studienarmen und der zwischen Zeitpunkten sich konzeptionell grundlegend unterscheidet, sollte man dies auch in der Begrifflichkeit unterscheiden. Crosby et al. (2003) schlugen für ersteres den Begriff MID vor, für letzteres den Begriff „Clinically Meaningful Change" (CMC) (Crosby et al. 2003).

Weitere Probleme sind zum einen, dass sich empirisch gezeigt hat, dass sich der CMC bei Verschlechterungen versus Verbesserungen der LQ unterscheidet (Cocks et al. 2012). Zum anderen müssen MID und CMC, die für eine Patientengruppe ermittelt wurden, nicht automatisch auch bei einer anderen Gruppe gelten.

Um die Variabilität einer MID eines Erhebungsinstrumentes hinsichtlich Art und Schwere einer Erkrankung, der untersuchten Richtung einer Veränderung oder der Methodik abzubilden, hat das IQWiG in seiner aktuellen Methodik einen neuen Ansatz entwickelt, der auf pragmatischem Weg die Ermittlung von Schwellen ermöglichen soll,

die hinreichend sicher einen klinisch bedeutsamen Bereich abgrenzen. Konkret legt das Institut fest, dass es künftig Responderanalysen ab einem Grenzwert von mindestens 15% der Skalenspannweite des verwendeten Erhebungsinstruments für die Bewertung heranzieht (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2019). Begründet wird der gewählte Schwellenwert damit, dass eine Sichtung von systematischen Übersichtsarbeiten zu MIDs ergeben habe, dass sich MIDs in vielen Fällen zwischen 10 und 20% der Spannweite der Scores eines Erhebungsinstruments bewegen.

Durch das Festlegen einer Schwelle von 15% soll nach Aussage des IQWiG erreicht werden, dass eine hinreichend sichere, für die Patienten spürbare Veränderung abgebildet werden kann und das Risiko einer ergebnisgesteuerten Berichterstattung durch eine beliebige Auswahl eines von vielen möglichen MIDs minimiert wird. Dieser neue Ansatz ermöglicht eine Vorhersehbarkeit der Akzeptanz von Responseschwellen und bietet zudem einen Weg, relevante Änderungen der LQ für Erhebungsinstrumente nachzuweisen, für die keine validierte MID in der entsprechenden Indikation vorliegt.

Dieser neue pragmatische Ansatz des IQWiG unterliegt jedoch auch einigen Einschränkungen. So gibt es Erhebungsinstrumente, die keine klar begrenzte Skala haben und die Punktzahlen ihrer Ergebnisse zunächst über einen T-Score umgerechnet werden müssen, um die Veränderungen einordnen zu können (z.B. SF-36). Zudem herrscht kein Konsens darüber, wie ein Schwellenwert von 15% ermittelt werden soll. So können zur Definition des Bereiches die maximal und minimal berechenbaren Scores oder externe (das heißt empirische) Werte verwendet werden. Alternativ kann der Bereich basierend auf einer festgelegten Anzahl von Standardabweichungen vom Mittelwert definiert werden.

Eine weitere Einschränkung besteht darin, dass in die Validierung des Schwellenwertes von 15% keine Bewertung seitens der Patienten einfließt. Eine Möglichkeit, die Sicht der Patienten bezüglich der Effektivität einer Intervention hinsichtlich der LQ einfließen zu lassen, wäre eine Einschätzung der Patienten die Schwelle, ab der von einer „ausreichend guten" oder auch einer „relevant schlechten" LQ gesprochen werden kann, selbst anzugeben. Das Konzept solcher Schwellenwerte (Cut-Offs) ist in anderen Bereichen, in denen patientenberichtete Endpunkte (patient-reported outcomes, PRO) dargestellt werden, durchaus etabliert (Zigmond & Snaith 1983). Im Falle der LQ scheint wichtig zu sein, dass nur Personen, die die Erkrankung und damit die mögliche LQ-Einschränkung selbst erlebt haben, diese Schwelle sinnvoll einschätzen können. In Pilotstudien hat sich dieses Vorgehen als durchführbar gezeigt (Fahsl et al. 2012).

Erste NB-Verfahren nach neuer IQWiG-Methodik zeigen, dass während das IQWiG bereits für die Anerkennung von Responderanalysen strikt die 15%-Grenze festschreibt, der G-BA im Augenblick zusätzlich auch noch Auswertungen nach alter Methodik anerkennt. Im NB-Verfahren zu Secucinumab in der Indikation Psoriasis Arthritis bspw. wurde der Fragebogen HAQ-DI zur Erfassung der Behinderung durch eine entzündlich-rheumatische Gelenkerkrankung herangezogen. Das IQWiG berücksichtigte die initial vom pU dargestellte MID von ≥3,5 unter Verweis auf die neuen Methoden nicht für die NB und zog lediglich die vom pU im Rahmen der Stellungnahme nachgereichte Auswertung mit einer Responseschwelle von 15% (entspricht einer MID von ≥0,45) heran.

Der G-BA hatte in der vorliegenden Indikation bereits die MID von ≥0,35 Punkten als klinisch relevante Veränderung beim HAQ-DI anerkannt und zog vor dem Hintergrund der aktuellen methodischen Diskussion beide Responderanalysen zur Bewertung des Zusatznutzens heran. Gleiches gilt für den Fragebogen SF-36 im selben NB-Verfahren zur Messung der HRQoL in klinischen Studien (Gemeinsamer Bundesausschuss 2020).

Im Rahmen von Beratungsgesprächen und auf der auf seiner Website veröffentlichten FAQs (Gemeinsamer Bundesausschuss 2020, 2021a, 2021b) verweist der G-BA zur Umsetzung und Auswertung der klinischen Relevanzschwellen derzeit auf die „Allgemeinen Methoden" 6.0 des IQWiG. In einer Übergangszeit, die bis zum Inkrafttreten einer angepassten Vorlage für Modul 4 des Nutzendossiers gelte, sind die pU dazu angehalten, für Responderanalysen grundsätzlich sowohl Auswertungen unter Verwendung einer bereits durch den G-BA in der Vergangenheit akzeptierten klinischen Relevanzschwelle (z. B. eine MID) als auch der 15%-Schwelle darzustellen (Gemeinsamer Bundesausschuss 2021a).

Am 16. Dezember 2021 hat der G-BA in seiner öffentlichen Sitzung die Änderung der Modulvorlagen in der Anlage II zum 5. Kapitel der Verfahrensordnung beschlossen (Gemeinsamer Bundesausschuss 2021b). Hiermit werden Änderungen vorgenommen, die die neuen Vorgaben des IQWiG zu MID-basierten Responderanalysen bei PRO aufgreifen.

Kommentar zu Rücklaufquoten

Ein ausreichender Rücklauf von Fragebögen stellt sicher, dass die befragte Stichprobe repräsentativ für die Studienpopulation ist. Bei zu hohen Ablehnerquoten ist mit einem erhöhten Selektionsbias zu rechnen.

Hier müssen die Begrifflichkeiten sauber getrennt werden: Ablehnerquote definiert den Anteil der Patienten, die bereits bei Baseline eine Teilnahme an der Erhebung ablehnen. Abbruchquote definiert den Teil der Patienten, die initial an der Studie teilnehmen, an Follow-up-Zeitpunkten jedoch nicht mehr. Hierdurch kann ein Selektionsbias entstehen, wenn Patienten mit schlechterer LQ häufiger aus der Studie ausscheiden, was aber eher die Regel als die Ausnahme darstellt. Deshalb ist es wichtig, möglichst geringe Ablehner- und Abbruchquoten zu erzielen. Aber was genau heißt „möglichst gering"?

Das IQWiG beschreibt in seinem Methodenpapier 6.0 die Bewertung fehlender Werte in klinischen Studien – hier fehlende Fragebögen – in einem zweistufigen Prozess. Zunächst werden Studienteilnehmer betrachtet, die aus der Analyse vollständig ausgeschlossen wurden (= Ablehnerquote). Ergebnisse sollen auf mindestens 70% der in die Auswertung eingeschlossenen Patienten basieren, um für die NB berücksichtigt zu werden. In der nachfolgenden Betrachtung der Follow-up-Verluste (= Abbruchquote) legt das IQWiG keine festen Grenzen fest, sondern richtet sich kontextabhängig nach folgenden Faktoren: 1. Anzahl und Zeitpunkt der Follow-up-Verluste und 2. Gründe für die Follow-up-Verluste (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2020).

Jedoch zeigen die Erfahrungen aus einer Vielzahl von NB-Verfahren, dass die Bezugsgröße eine wesentliche Frage und Schwierigkeit bei der Ermittlung und Bewertung der Abbruchquoten ist. IQWiG und G-BA berücksichtigen zur Ermittlung der Abbruchquoten alle randomisierten Patienten im Nenner und ziehen dort lediglich verstorbene Patienten ab, was bedeutet, dass Patienten, die gemäß Studienprotokoll keinen Fragebogen mehr erhalten, weil sie die Behandlung (vorab) beendet haben und die LQ nur bis zum Behandlungsende erfasst wird, als „Abbrecher" gewertet werden. Dies steht im Kontrast zur

Auffassung der pU, dass sich Patienten in dieser Konstellation nach Behandlungsende nicht mehr im Risikoset befinden und folglich nicht als „Abbrecher" zu werten sind. Sowohl Ablehner- als auch Abbruchquoten sollten auf die zum jeweiligen Zeitpunkt im Analyseset verfügbaren Patienten bezogen werden. Es könnte zudem informativ und relevant sein, diesen Kennwert den bislang von IQWiG und G-BA verwendeten gegenüberzustellen.

Kommentar zur statistischen Auswertung von LQ-Daten im Rahmen von Studien mit Überlebenszeitendpunkt/unterschiedlich langen Beobachtungszeiten

Insbesondere in der Onkologie werden im Rahmen von Überlebenszeitstudien LQ-Daten erhoben. Studien mit einem Überlebenszeitendpunkt werden aufgrund unterschiedlich langer Nachbeobachtungszeiten und Zensierungen mittels statistischer Verfahren der Überlebenszeitanalyse (v.a. Kaplan-Meier, Cox' Proportionales Hazards Modell) ausgewertet; dies erschwert auch die Analyse der LQ. Mittels Techniken der Überlebenszeitanalyse wird häufig die Zeit bis zur Verschlechterung analysiert, wobei Zensierungsgründe konkret auch Progression oder Tod sein können, welche konkurrierende Risiken darstellen. Zensierung mittels eines konkurrierenden Risikos ist subtil: Das Cox

Literatur

1. Gemeinsamer Bundesausschuss (G-BA) (2018): Verfahrensordnung des Gemeinsamen Bundesausschusses. Zuletzt geändert am 16. August 2018 - veröffentlicht im Bundesanzeiger BAnz AT 05.03.2019 B2 - in Kraft getreten am 6. März 2019. In: https://www.g-ba.de/downloads/62-492-1777/VerfO_2018-08-16_iK-2019-03-06.pdf. (abgerufen am: 17.06.2021)
2. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) (2020): Allgemeine Methoden. Version 6.0 vom 05.11.2020. In: https://www.iqwig.de/methoden/allgemeine-methoden_version-6-0.pdf?rev=144030. (abgerufen am: 18.01.2021)
3. Bundesministerium für Gesundheit (2010): Arzneimittel-Nutzenbewertungsverordnung - AM-NutzenV. In: http://www.gesetze-im-internet.de/amnutzenv/AM-NutzenV.pdf. (abgerufen am: 17.06.2021)
4. Deutscher Bundestag (2014): Hecken lobt Arzneimittelgesetz. In: Heute im Bundestag 2014, 145. In: https://www.bundestag.de/webarchiv/Presse/hib/2014_03/01-262758. (abgerufen am: 25.08.2021)
5. Staeck, F. (2016): GBA-Chef will Vorfahrt für Lebensqualität. In: Ärzte Zeitung, 03.06.2016, http://www.aerztezeitung.de/politik_gesellschaft/arzneimittelpolitik/nutzenbewertung/article/9 12741/nutzenbewertung-gba-chef-will-vorfahrt-lebensqualitaet.html. (abgerufen am: 12.08.2021)
6. Gemeinsamer Bundesausschuss (G-BA) (2018): Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Arzneimittel-Richtlinie (AM-RL): Anlage XII – Verlängerung der Befristung der Geltungsdauer eines Beschlusses über die Nutzenbewertung von Arzneimitteln mit neuen Wirkstoffen nach § 35a SGB V – Trifluridin/Tipiracil. In: https://www.g-ba.de/downloads/40-268-5105/2018-07-05_AM-RL-XII_Trifluridin-Tipiracil_Verlaengerung-Befristung_D-252_TrG.pdf. (abgerufen am: 17.06.2021)
7. Klakow-Franck, R. (2014): Die Bedeutung von Lebensqualität für die Arbeit des Gemeinsamen Bundesausschusses. In: ZEFQ 2014, 108, 2: 151-6
8. Unkel, S., Amiri, M., Benda, N., Beyersmann, J., Knoerzer, D., Kupas, K., et al. (2019): On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. In: Pharm Stat 2019, 18, 2: 166-83
9. Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N. et al. (2020): Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. In: BMJ 2020, 369: m1714
10. Peipert, J. D. (2020): Re: Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. In: BMJ 2020, 369:m1714
11. Morris, S. B., DeShon, R. P. (2002): Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. In: Psychological Methods 2002, 7, 1: 105-25
12. Crosby, R. D., Kolotkin, R. L., Williams, G. R. (2003): Defining clinically meaningful change in health-related quality of life. In: J Clin Epidemiol 2003, 56, 5: 395-407
13. Cocks, K., King, M. T., Velikova, G., de Castro, G. Jr., Martyn St-James, M., Fayers, P. M. et al. (2012): Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. In: Eur J Cancer 2012, 48, 11: 1713-21
14. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) (2019): Allgemeine Methoden. Entwurf für Version 6.0 vom 05.12.2019. In: https://www.iqwig.de/download/Allgemeine-Methoden_Entwurf-fuer-Version-6-0.pdf. (abgerufen am: 17.06.2021)
15. Zigmond, A. S., Snaith, R.P. (1983): The hospital anxiety and depression scale. In: Acta Psychiatr Scand 1983, 67, 6: 361-70
16. Fahsl, S., Keszte, J., Boehm, A., Vogel, H. J., Völkel, W., Meister, E. F. et al. (2012): Clinical relevance of quality-of-life data in laryngectomized patients. In Laryngoscope 2012, 122, 7:1532-8
17. Gemeinsamer Bundesausschuss (G-BA) (2020): Nutzenbewertungsverfahren zum Wirkstoff Secukinumab (Neue wissenschaftliche Erkenntnisse (§ 14): Psoriasis-Arthritis). In: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/590/#nutzenbewertung. (abgerufen am: 17.06.2021)
18. Gemeinsamer Bundesausschuss (G-BA) (2021): Antworten auf häufig gestellte Fragen zum Verfahren der Nutzenbewertung. Verfügbar unter: https://www.g-ba.de/themen/arzneimittel/arzneimittel-richtlinie-anlagen/nutzenbewertung-35a/faqs/#wie-soll-vor-dem-hintergrund-der-veroffentlichung-des-methodenpapiers-60-des-iqwig-am-5-november-2020-derzeit-in-der-dossiererstellung-mit-der-bestimmung-von-klinischen-relevanzschwellen-bei-komplexen-skalen-umgegangen-werden (abgerufen 12.08.2021)
19. Gemeinsamer Bundesausschuss (G-BA) (2021): Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Verfahrensordnung: Änderung der Modulvorlage in der Anlage II zum 5. Kapitel vom 16. Dezember 2021. In: https://www.g-ba.de/downloads/39-261-5217/2021-12-16_VerfO_Aenderung-Modulvorlage-Anlage-II-Kap-5.pdf. (abgerufen am: 27.12.2021)
20. Allignol, A., Beyersmann, J., Schmoor, C. (2016): Statistical issues in the analysis of adverse events in time-to-event data. In: Pharm Stat 2016, 15, 4: 297-305
21. Beyersmann, J., Schmoor, C. (2019): The Analysis of Adverse Events in Randomized Clinical Trials. In: Halabi, S., Michiels, S. (Hrsg.) (2019): Textbook of Clinical Trials in Oncology - A Statistical Perspective. 1. Auflage. New York: Chapman and Hall/CRC, 537-557
22. Olschewski, M., Schulgen, G., Schumacher, M., Altman, D. G. (1994): Quality of life assessment in clinical cancer research. In: Br J Cancer 1994, 70, 1: 1-5
23. Olschewski, M., Schumacher, M. (1990): Statistical analysis of quality of life data in cancer clinical trials. In: Statistics in Medicine 1990, 9, 7: 749-63
24. Schumacher, M., Olschewski, M., Schulgen, G. (1991): Assessment of quality of life in clinical trials. In: Statistics in Medicine 1991, 10, 12: 1915-30
25. Gran, J. M., Lie, S.A., Øyeflaten, I., Borgan, Ø., Aalen, O. O. (2015): Causal inference in multi-state models–sickness absence and work for 1145 participants after work rehabilitation. In: BMC Public Health 2015, 15, 1: 1082
26. Rizopoulos, D. (2012): Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. New York: Chapman and Hall/CRC
27. Coens, C., Pe, M., Dueck, A. C., Sloan, J., Basch, E., Calvert, M., et al. (2020): International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. In: Lancet Oncol 2020, 21, 2: e83-e96

Modell liefert ein korrektes Hazard Ratio (HR) für die Zeit bis zur Verschlechterung. Die Zensierung ist jedoch informativ bezüglich Wahrscheinlichkeitsaussagen und muss daher durch die Möglichkeit, mehrere Ereigniszeiten konkurrierender Risiken simultan zu analysieren, ergänzt werden. Auch Kaplan-Meier-Kurven sind nicht sinnvoll, da diese die tatsächlichen Häufigkeiten überschätzen. Stattdessen ist der Aalen-Johansen-Schätzer der sogenannten kumulativen Inzidenzfunktion zu verwenden (Allignol et al. 2016; Beyersmann & Schmoor 2019).

Eine besondere Rolle spielt die „Zensierung durch Progression". Häufig sieht in diesem Fall das Studienprotokoll vor, LQ-Daten nicht weiter zu erheben. Dies ist häufig mit der Kritik verbunden, dass auch Progression eine informative Zensierung bezüglich des HR sei. Dies ist jedoch nicht der Fall für die Analyse des HR im Kontext konkurrierender Risiken und für die Zeitdauer des progressionsfreien Überlebens.

Ein Kritikpunkt an der Zielgröße „Zeit bis zur Verschlechterung" ist, dass diese eine mögliche Besserung der LQ nicht erfasst. Eine Analyse der Dauer einer relativ guten LQ ist jedoch mittels einer Erweiterung konkurrierender Risiken hin zu Mehrstadienmodellen möglich, was bereits Anfang der 90er Jahre vorgeschlagen wurde (Olschewski et al. 1994; Olschewski & Schumacher 1990; Schumacher et al. 1991). Aktuelle Entwicklungen verknüpfen Mehrstadienmodelle, Zensierung bei Behandlungswechsel (wie z. B. nach Progression) mit kausalen Analysen (Gran et al. 2015) und sind Bestandteil der aktuellen Estimand-Debatte (Unkel et al. 2019). Eine Ausweitung des Diskurses auf die LQ-Messung wäre ebenfalls wünschenswert.

Die bisher diskutierten Analysen erfordern, dass LQ kategorisiert wird, was sowohl die statistische Analyse als auch deren Interpretation erleichtert. Um jedoch den longitudinalen Verlauf der LQ direkt zu untersuchen, werden häufig Gemischte Lineare Modelle verwendet. Eine implizite Annahme ist hier, dass nicht vorliegende LQ-Daten „Missing at Random" sind, d.h. der Umstand, ob Daten fehlen oder nicht, ist unabhängig von der Ausprägung der fehlenden Werte. Soll jedoch die longitudinale LQ-Trajektorie als solche ausgewertet werden, bietet sich als Alternative zum Gemischten Linearen Modell ein Joint Model an, in dem der „informative Drop Out" Prozess mitmodelliert wird und daher „Missing Not at Random" sein darf (Rizopoulos 2012). Der Preis hierfür sind typischerweise relativ strikte Modellannahmen sowie nicht selten numerische Schwierigkeiten. Ferner erlaubt ein Joint Model sogar die fragwürdige Extrapolation von LQ nach dem Tod.

Schlussfolgerung/Fazit

Die Erhebung der HRQoL von Patienten im Rahmen klinischer Studien beruht auf subjektiven Beurteilungen durch den Patienten. Ein wichtiges Ziel bei der Auswertung der dadurch gewonnenen Daten ist es, dennoch möglichst verlässliche Erkenntnisse daraus zu ziehen. Spätestens mit der Einführung der NB von Arzneimitteln im Rahmen des AMNOG wird dieses Anliegen noch dringlicher, da hier der Fokus ausschließlich auf patientenrelevanten Endpunkten liegt. Anhand der hier diskutierten aktuellen Beispiele wird die wachsende Bedeutung von LQ-Daten im NB-Verfahren deutlich, jedoch auch herausgearbeitet, welche methodischen und konzeptionellen Probleme und Herausforderungen daraus entstehen und über welche Lösungsansätze aktuell diskutiert wird. Es besteht ein akuter Bedarf an neuen Ansätzen, die möglichst auch alle involvierten Interessen berücksichtigen sollten. Ein stetiger Austausch zwischen akademischer Forschung, den für die Bewertungen zuständigen Institutionen (G-BA; IQWiG) und den pU mit dem Ziel, solche Lösungen zu finden, ist – gerade auch im Sinne der Patienten – mehr als wünschenswert.

Positive Ansätze in dieser Richtung sind vorhanden: Genannt sei hier die Arbeit der Projektgruppe „Analyse unerwünschter Ereignisse bei variablen Beobachtungszeiten", der Arbeitsgruppe „Therapeutische Forschung" (ATF) der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS) und der Arbeitsgruppe „Pharmazeutische Forschung" (APF) der Deutschen Region der Internationalen Biometrischen Gesellschaft (IBS-DR).

Es ist zu begrüßen, dass ein Austausch hinsichtlich der Analyse und Bewertung von PRO einen immer größeren Stellenwert einnehmen und die Entwicklung von Lösungsansätzen dadurch vorangetrieben wird. So waren im Jahr 2020 die Herausforderungen bei der Analyse von PRO bspw. Thema eines Workshops der ATF der GMDS und im Juni 2021 Fokus der Veranstaltungsreihe „IQWiG im Dialog". Das interdisziplinäre Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data (SISAQOL) Consortium veröffentlichte im Februar 2020 seine internationalen Empfehlungen für die Analyse von PRO (Coens et al. 2020), die bei der Entwicklung von Standards bei der Analyse von PRO auch in der NB berücksichtigt werden sollen. <<

von:
Sarah Böhme MSc[1]
Prof. Dr. rer. med. Susanne Singer[2]
Stephanie Kauffmann MSc[3]
Dipl.-Stat. Friedhelm Leverkus[4]

1: Manager Health Technology Assessment bei Pfizer in Deutschland
2: Leiterin der Abteilung Epidemiologie und Versorgungsforschung am Institut für Medizinische Biometrie, Epidemiologie und Informatik der Universitätsmedizin Mainz
3: Junior Manager Health Technology Assessment bei Pfizer in Deutschland
4: Director Health Technology Assessment & Outcomes Research Health & Value bei Pfizer in Deutschland

Autorenerklärung

Sarah Böhme, Friedhelm Leverkus und Stephanie Kauffmann sind Mitarbeiter der Pfizer Pharma GmbH. Dr. Susanne Singer berichtet den Erhalt eines Honorars durch Pfizer im Rahmen der vorliegenden Arbeit und abseits davon Honorare von Bristol-Myers Squibb, Boehringer-Ingelheim und Lilly. Unterstützung beim Medical Writing erhielten die Autoren von AMS Advanced Medical Services GmbH (München). Sponsor der vorliegenden methodischen Analyse ist Pfizer Deutschland GmbH.

Danksagung

Die Autoren bedanken sich bei Jan Beyersmann und Robert Miller für wertvolle und hilfreiche Anmerkungen und für die Unterstützung bei der Erstellung des Manuskripts. Die Autoren danken zudem AMS Advanced Medical Services GmbH (München) für die Unterstützung beim Medical Writing.

Zitationshinweis

Böhme et al.: „Es braucht neue Ansätze für Lebensqualitätsdaten", in „Monitor Versorgungsforschung" (01/22), S. 43-47. http://doi.org/10.24945/MVF.01.22.1866-0533.2373

Korrespondenzadresse

Kontakt: Sarah.Boehme@Pfizer.com