



**IQWiG im Dialog 2021**

# **Umgang mit Skalen in der Nutzenbewertung**

**Abstracts der Präsentationen**

## **Benutzung von Skalen für patientenberichtete Endpunkte: Herausforderungen bei der Entwicklung, Validierung, Anwendung und Interpretation**

**Susanne Singer**

*Abt. Epidemiologie und Versorgungsforschung, Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsmedizin der Johannes Gutenberg-Universität Mainz*

Instrumente zur Erfassung von patientenberichteten Endpunkten in klinischen Studien müssen vielen Anforderungen genügen. Sie sollen möglichst alles erfassen, was für diese Patienten-gruppe relevant ist und gleichzeitig sollen sie möglichst kurz sein, um den Aufwand bei Ausfüllen gering zu halten. Die Items müssen so formuliert sein, dass die Antworten eindeutig interpretierbar sind. Eine Zusammenfassung der Items zu sinnvollen und reliablen Skalen ist häufig gewünscht, jedoch müssen Redundanzen in den Inhalten vermieden werden (selbst wenn das die interne Konsistenz verbessern würde). Einerseits soll die Test-Retest-Reliabilität hoch sein, andererseits auch die Änderungssensitivität. Nötig sind Skalen, die weder Boden- noch Deckeneffekte, eine gute Varianz und möglichst wenige Fehlwerte aufweisen. Die Items müssen zudem eindeutig übersetzbar sein, so dass in multinationalen Studien die Daten der verschiedenen Länder bzw. Sprachen problemlos zusammengefasst werden können. Nicht zuletzt sollen die Fragebögen auch (Neben)wirkungen aktueller Behandlungsmethoden erfassen können (Stichwort Flexibilität), was im Widerspruch stehen kann zu dem Wunsch, Daten über lange Zeiträume hinweg vergleichen zu können (Stichwort Stabilität). Eine weitere Herausforderung stellt die Interpretierbarkeit der Skalen dar: Minimal Important Differences, Minimal Important Changes und Schwellenwerte können dabei zwar helfen, aber diese differieren häufig zwischen Verbesserung und Verschlechterung sowie zwischen verschiedenen Patientengruppen. Bei der Entwicklung und Validierung von Instrumenten kann man versuchen, all diesen Anforderungen so weit wie möglich zu entsprechen, ganz erreichen wird man sie selten. Die Frage ist: ab wann ist ein „not perfect“ immerhin ein „good enough“, und wann nicht?

## **Die Messung der Lebensqualität und spezieller patientenberichteter Endpunkte – elementare Komponenten der Nutzenbewertung**

**Martin Danner**

*Bundesarbeitsgemeinschaft Selbsthilfe von Menschen mit Behinderung, chronischer Erkrankungen und ihren Angehörigen e.V. (BAG SELBSTHILFE), Düsseldorf*

Der Bezugsrahmen der Nutzenbewertung hängt von der jeweiligen Zwecksetzung der Bewertung ab. So ist es ein Unterschied, ob eine Nutzenbewertung zum Zwecke der konkreten Optimierung einer bestimmten Behandlung oder zum Zwecke des Health Technology Assessments vorgenommen werden soll. Zumindest im letzteren Fall geht es eigentlich um eine Zusatznutzenbewertung, da mehrere therapeutische oder

diagnostische Optionen miteinander verglichen werden sollen. Nach den gesetzlichen Vorgaben des § 35a SGB V wird der Nutzenbegriff bei der Nutzenbewertung von Arzneimitteln zwingend an patientenrelevanten Effekten eines Arzneimittels ausgerichtet. Unter anderem wird als ein solcher patientenrelevanter Effekt die „Verbesserung der Lebensqualität“ ausgewiesen. Gerade wegen der generellen Vorgabe der Patientenrelevanz des therapeutischen Effekts ist aber ebenso akzeptiert, dass auch hinsichtlich der übrigen Endpunkte der Verbesserung des Gesundheitszustandes, der Verkürzung der Krankheitsdauer oder der Verringerung der Nebenwirkungen patientenorientierte Effekte maßgeblich für das Assessment sein sollen.

Blickt man auf die Historie der Nutzenbewertung beim Gemeinsamen Bundesausschuss zurück, dann ist zunächst einmal festzustellen, dass insbesondere der aggregierte Endpunkt der „Verbesserung der Lebensqualität“ eine beachtliche Anerkennungskarriere in der Methodendiskussion hinter sich gebracht hat. War dieser Endpunkt anfangs nur als Konstrukt mit eher subjektivem Einschlag angesehen worden, hat sich die Lage angesichts einer Stabilisierung der Regularien zur Validierung von Lebensqualitätsfragebögen, zu Rücklauf-quoten etc. doch sehr verändert. Auch hinsichtlich der spezifischen PRO's ist eine Stabilisierung der Akzeptanz zu verzeichnen. Diese Entwicklung ist aus Patientensicht sehr zu begrüßen.

Gegenstand der aktuellen wissenschaftlichen Diskussion ist nun offenbar die Frage, wie die etablierten Kriterien zur Beurteilung der methodischen Qualität von Lebensqualitäts-instrumenten noch verfeinert werden können. Insbesondere das Kriterium der Interpretierbarkeit von Lebensqualitätsmessungen ist in der Diskussion. Die Frage ist, ob die eingesetzten Instrumente derart differenzierte Ergebnisse ermöglichen, so dass die erhobenen Messwerte auch tatsächlich bedeutsame Veränderungen anzeigen. Die Thematik der MID wird daher intensiv im gesundheitspolitischen Raum diskutiert. Aus einer wissenschaftlichen Perspektive heraus betrachtet, sind solche Diskussionen äußerst fruchtbar, da sie auf eine Verbesserung der zur Verfügung stehenden Instrumente abzielen.

Leider geht es in der Regel aber nicht um eine rein wissenschaftliche Diskussion, sondern um ein „Alles oder Nichts“: Sollen die Ergebnisse von Befragungen in einem konkreten Assessment berücksichtigt werden oder nicht. Da das Assessment seinerseits wieder genutzt wird, um die Frage der Finanzierung und der Verfügbarkeit therapeutischer Optionen zu klären, kann eine sachliche Diskussion zur wissenschaftlichen Weiterentwicklung von Instrumenten gar nicht als solche geführt werden. Dies ist aus Patientensicht sehr zu beklagen.

Auch hinsichtlich anderer Kriterien gibt es nämlich wissenschaftlichen Weiterentwicklungsbedarf. Gerade bei der Bewertung therapeutischer Optionen bei progredienten Erkrankungen wäre es schön, wenn noch weiter an der sog. Responsiveness der Instrumente gearbeitet würde. Das Verdikt, dass Instrumente bislang in vielen Fällen nicht hinreichend änderungssensitiv sind, wäre auch hier eher destruktiv. Stattdessen sollte generell an derartige Fragen mit dem wissenschaftlichen Elan, Verbesserungen erzielen zu wollen, herangegangen werden.

## **Patientenberichtete Endpunkte: Wie können aussagekräftige minimale patientenrelevante Unterschiede hergeleitet werden?**

**Jens-Otto Andreas**

*UCB BIOSCIENCES GmbH, Monheim am Rhein*

Patientenberichtete Endpunkte sind ein wichtiges Instrument zur Beurteilung des Einflusses einer Behandlung auf die Lebensqualität des Patienten. Neben der Validierung der Messinstrumente sind auch akzeptierte (minimale) Relevanzschwellen (minimal important difference, MID) erforderlich, um auf Basis von Lebensqualitätsdaten einen quantifizierbaren Zusatznutzen (bzw. Schaden) ableiten zu können. Anhand einer MID werden dann Kriterien für das Vorliegen eines individuellen Therapieansprechens (Response) definiert. Dabei spielt die adäquate Validierung der Instrument-spezifischen MID eine entscheidende Rolle.

In diesem Vortrag werden verschiedene Vorgehensweisen zur Validierung einer MID/Response-schwelle vorgestellt und diskutiert. Dabei wird auch die Methodik des IQWiGs, die eine Responseschwelle von  $\geq 15\%$  der Skalenspannweite eines Messinstrumentes als Relevanzschwelle vorsieht, mit einbezogen.

## **Methodische Aspekte bei der Analyse von Daten zur Lebensqualität in Nutzenbewertungen**

**Christoph Schürmann**

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

Die gesundheitsbezogene Lebensqualität ist nicht nur wegen der gesetzlichen Vorgabe, sondern vor allem wegen ihres besonderen Bezugs zur persönlich empfundenen Krankheitslast von Patientinnen und Patienten seit jeher eine relevante Endpunktkategorie bei der Nutzenbewertung therapeutischer Behandlungen durch das IQWiG. Eine zentrale Anforderung für die Messung ist zunächst die Verwendung geeigneter, valider Instrumente bzw. Skalen sowie die Festlegung von wiederholten Erhebungszeitpunkten. Für die Bewertung ist anschließend die Wahl aussagekräftiger biometrischer Verfahren entscheidend, insbesondere vor dem Hintergrund der verfügbaren Datenlage. Die Auswahl einer jeweiligen Analyse muss vor allem in Abhängigkeit des Ziels der Bewertung erfolgen, aus Sicht des IQWiG ist dies der Vergleich der Lebensqualität zwischen den jeweiligen Behandlungsgruppen im gesamten Beobachtungszeitraum. Das Effektmaß und der Umgang mit fehlenden Beobachtungen sind so zu wählen, dass die entsprechenden Analysen diesem Ziel möglichst nahekommen. Sofern bei der Erhebung oder Auswertung Responsekriterien zum Einsatz kommen, müssen diese verlässlich interpretierbar sein. Hierzu hat das IQWiG in der letzten Version seines Methodenpapiers (Version 6.0) erstmals einen neuen Ansatz vorgestellt, um ein praktikables Vorgehen zu ermöglichen, das gleichzeitig eine ergebnisgesteuerte Berichterstattung vermeidet. Der Vortrag gibt eine Übersicht über diese methodischen Faktoren und erläutert ihre Bedeutung für die Nutzenbewertung an Beispielen.

## Die Entwicklung und Verwendung von Schwellenwerten für klinische Relevanz zur Unterstützung der Interpretation von patientenberichteten Endpunkten

**Johannes Giesinger**

*Universitätsklinik für Psychiatrie II, Medizinische Universität Innsbruck*

Patientenberichtete Endpunkte (*Patient-reported Outcomes*) werden in klinischen Studien verwendet um die Patientenperspektive auf die Folgen von Erkrankungen und Behandlungen zu evaluieren. Für die Erfassung solcher Endpunkte ist eine Vielzahl von umfangreich validierten und international weit verbreiteten Instrumenten verfügbar. Diese werden in klinischen Studien und der klinischen Routine für die systematische und standardisierte Erhebung von Parametern wie beispielsweise Lebensqualität, Symptombelastung, und Behandlungsnebenwirkungen eingesetzt.

Derartige Instrumente beschreiben die erhobenen Parameter zumeist mit Punktwerten. Zum Beispiel der EORTC QLQ-C30, das in der Onkologie am häufigsten verwendete Instrument zur Erfassung von Lebensqualität und Symptombelastung, beschreibt die Ausprägung der einzelnen Parameter (z.B. Schweregrad von Fatigue oder Atemnot) mit Punktwerten zwischen 0 und 100. Die Interpretation wird allerdings dadurch erschwert, dass diese minimalen und maximalen Werte meist nur die Grenzen des Messbereichs eines Instruments widerspiegeln und nicht unbedingt die minimal oder maximal mögliche Ausprägung eines Symptoms o.ä. beschreiben. Um die Interpretierbarkeit der Punktwerte von einzelnen PatientInnen oder Patientengruppen zu verbessern, werden daher Normwerte (beispielsweise basierend auf der Allgemeinbevölkerung), sogenannte kleinste relevante Differenzen (für die Beurteilung von Unterschieden zwischen Patienten oder Zeitpunkten) oder Schwellenwerte verwendet.

Schwellenwerte zur Beurteilung der klinischen Relevanz von Punktwerten (auch als *Thresholds for Clinical Importance* bezeichnet) beziehen sich auf absolute Punktwerte einzelner PatientInnen zu einem konkreten Zeitpunkt. Sie ermöglichen das Screening nach Symptomen und Beeinträchtigungen in der klinischen Routine und die Einteilung von PatientInnen in *Treatment Responder* bzw. *Non-Responder* in klinischen Studie (wenn das Behandlungsziel beispielsweise Symptombefreiheit ist). Schwellenwerte basieren zumeist auf externen Kriterien, wie Patientenbeurteilungen von Gesundheitszuständen oder Experten-einschätzungen, und werden mit Hilfe von unterschiedlichen statistischen Methoden definiert (zum Beispiel mittels *Receiver Operating Characteristic* Analysen oder logistischer Regressionsmodelle). Die zu Grunde liegende Methodik bestimmt dabei im Wesentlichen die Bedeutung der Schwellenwerte und ihre Anwendbarkeit im jeweiligen Kontext.

Zusammenfassend: Schwellenwerte können die Interpretation von patientenberichteten Endpunkten erleichtern und dadurch einen wichtigen Beitrag zur Etablierung und weitreichenden Verwendung von patientenberichteten Endpunkten in der klinischen Forschung und Praxis leisten.