Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Institute for Quality and Efficiency in Health Care

Dokumentation und Würdigung der Stellungnahmen zum „Entwurf der Allgemeinen Methoden Version 4.0 vom 09.03.2011“

Version 1.0 vom 23.09.2011

Kontakt:

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)

Dillenburger Straße 27

D-51105 Köln

Tel: +49-221 / 35685-0

Fax: +49-221 / 35685-1

E-Mail: methoden@iqwig.de

Vorwort

Im Stellungnahmeverfahren zum Entwurf der Allgemeinen Methoden Version 4.0 vom 09.03.2011 gingen fristgerecht bis zum 08.04.2011 Stellungnahmen von 14 Organisationen, Institutionen, Firmen und Einzelpersonen ein. Es gab weitere Stellungnahmen, die nicht fristgerecht eingingen. Alle Stellungnahmen wurden in der Überarbeitung des Methodenpapiers berücksichtigt. In dem hier vorliegenden Dokument werden jedoch nur die fristgerecht eingereichten Stellungnahmen gewürdigt.

Inhaltsverzeichnis

1 Würdigung allgemeiner Stellungnahmen

1.1 Strukturen des Instituts

Eine Stellungnahme schlägt vor, die Strukturen des Instituts nicht in der Präambel, sondern wie in der vorigen Version in einem eigenen Kapitel zu behandeln.

Es erscheint generell nicht sinnvoll, Strukturen des Instituts, die sich aus der Satzung ergeben oder z. B. im Organigramm beschrieben sind und auch Änderungen unterliegen, zusätzlich in dem Methodenpapier zu beschreiben. Der Hinweis wird zum Anlass genommen, diese Teile zu streichen.

1.2 Darstellung der „bestverfügbaren Evidenz“

In einer Stellungnahme wird festgestellt, das Methodenpapier „suggeriert (durchgehend), der Ansatz von Sackett und Kollegen sei ‚lückenhaft oder nicht zuverlässig‘ (S. 5)“.

Das Institut geht davon aus, dass es sich bei diesem Kommentar wohl um ein Versehen handelt. Der zitierte Satz heißt vollständig und korrekt: „Oft erweist sich die „bestverfügbare Evidenz“ aber als lückenhaft oder nicht zuverlässig“ und ist somit unmissverständlich und offensichtlich zutreffend.

Eine Stellungnahme stellt fest, dass die „Darstellung der bestverfügbaren Evidenz auch in den künftigen Berichten des Instituts zumeist unterbleiben soll“. Sie unterstellt dem Institut unter Bezugnahme auf § 139a (3) weiter, dass „eine Angleichung der Methoden an den gesetzlichen Auftrag weiterhin nicht beabsichtigt ist. Dies schränkt die Verwertbarkeit der Berichte im G-BA unnötig ein.“

Weder dem Methodenpapier noch den bisherigen Berichten des Instituts ist zu entnehmen, dass die Darstellung von Evidenz unterbleiben soll und unterbleibt. Diejenige Evidenz, die geeignet ist, Informationen zum Nutzen zu liefern, wurde und wird auch weiterhin dargestellt. Der Kommentar unterstellt, dass das Institut sich nicht gesetzeskonform verhält. Abgesehen von einer unvollständigen Wiedergabe des Gesetzestextes (s. u.) findet sich kein Anhaltspunkt, worauf sich dieser Vorwurf bezieht. Dem Institut ist keine einzige Äußerung des G-BA bekannt, die den genannten Kritikpunkt unterstützen würde. Die Stellungnahme verkennt schließlich, dass die Aufträge des G-BA meist Aufträge zur Bewertung des Nutzens sind, für die das Institut gemäß § 139a (4) „zu gewährleisten (hat), dass die Bewertung des medizinischen Nutzens nach den international anerkannten Standards der evidenzbasierten Medizin … erfolgt“. Es ist daher naheliegend, dass die Evidenz insoweit dargestellt wird, als sie für die Bewertung des Nutzens geeignet ist. Dies ist kongruent mit dem Vorgehen vieler anderer, insbesondere internationaler Institutionen.

1.3 Forderung nach RCT

Mehrere Stellungnahmen thematisieren die im Methodenpapier herausgestellte besondere Bedeutung von RCT. Sie kritisieren dabei die „unbedingte Forderung" nach RCT, die im Gegensatz etwa zum Vorgehen des NICE stehe, sowie die „alleinige Berücksichtigung idealer empirischer Bedingungen", die die „Arbeit des Instituts limitiere".

Eine solche unbedingte Forderung gibt es im Methodenpapier nicht. Das Institut ist vielmehr in Methodenpapier und Praxis mit dem NICE auf einer Linie. Auch diese Praxis bedeutet genau nicht die „Berücksichtigung aller anderen Evidenzlevel". Positive Bewertungen (Belege) des Nutzens sind selbstverständlich von möglichst guten, d. h. ergebnissicheren Studiendaten abhängig zu machen. Wie das Methodenpapier sowie die bisherige Praxis deutlich machen, können für die Arbeit des Instituts aber keineswegs „lediglich" idealtypische Studien herangezogen werden, sondern, insbesondere beim Fehlen solcher Studien, auch weitere, etwa nicht randomisierte vergleichende Studien, sofern sie grundsätzlich geeignet sind, ausreichende Ergebnissicherheit zu liefern.

Durch die Konzentration auf geeignete Studien wird die Arbeit des Instituts in keiner Weise limitiert, sondern im Gegenteil erst möglich gemacht.

1.4 Zusammenfassung von Nichtunterlegenheitsstudien

In einer Stellungnahme wird ausgeführt, dass eine klare Aussage wünschenswert sei, „dass es bei der Zusammenfassung von Nichtunterlegenheitsstudien nicht das Ziel ist, die Studien durch eine metaanalytische Zusammenfassung als eine große Nichtunterlegenheitsstudie zu analysieren, sondern eine Abwägung zwischen dem (impliziten) Zusatznutzen und einem möglichen Wirksamkeitsverlust vorzunehmen".

Dieser Hinweis (Abwägung zwischen dem [impliziten] Zusatznutzen und einem möglichen Wirksamkeitsverlust) entspricht im Prinzip der Position und dem Vorgehen des Instituts (z. B. Abschlussbericht N04-01 [34]), wobei es sich dabei schon im Grundsatz (weiterhin) um eine Nicht-Unterlegenheitsfragestellung handelt.

1.5 Geschlechtergerechte Sprache

In zwei Stellungnahmen wird der Gebrauch einer nicht geschlechtergerechten Sprache kritisiert.

Das Methodenpapier wurde entsprechend überarbeitet. Die Allgemeinen Methoden Version 4.0 erscheinen in geschlechtergerechter Sprache.

2 Würdigung der Stellungnahmen zu Kapitel 1

Alle Argumente der Stellungnahmen zu Kapitel 1 des Methodenpapiers werden an anderer Stelle gewürdigt.

3 Würdigung der Stellungnahmen zu Kapitel 2

3.1 Konkretisierung von Aufträgen

Eine Stellungnahme erklärt, dass es wünschenswert sei, dass nach der Auftragserteilung durch den G-BA oder das BMG die Präzisierung und Formulierung der wissenschaftlichen Fragestellungen zu den verschiedenen Produkten des Instituts (Berichte, Rapid Reports etc.) in Abstimmung mit dem Auftraggeber vorgenommen werden.

Bisher und in Zukunft erfolgt(e) selbstverständlich die „Präzisierung und Formulierung" eines Auftrags in Abstimmung mit dem Auftraggeber. Genauer gesagt, erfolgt die Konkretisierung durch den Auftraggeber in Abstimmung mit dem Institut. Dieser Prozessschritt erfolgt jedoch vor Auftragserteilung und ist daher nicht Gegenstand des Methodenpapiers.

3.2 Stellungnahme und Fristen

Eine Stellungnahme äußert die Erwartung, dass das „endgültige, dem Auftraggeber zur Verfügung zu stellende Produkt … sowohl die Bewertungen durch das IQWiG als auch die der Stellungnehmenden widerspiegeln" sollte.

Es ist unklar, was mit dieser Stellungnahme ausgedrückt werden soll. Zu allen Produkten des Instituts, die einem öffentlichen Stellungnahmeverfahren unterzogen werden (vorläufige Berichtspläne und Vorberichte), wird eine Dokumentation und Würdigung der Stellungnahmen erstellt, die Teil der Auftragsdokumentation ist. Somit ist die zitierte Forderung der Stellungnehmenden genau erfüllt.

Eine Stellungnahme stellt fest, es sei „auffallend, dass das IQWiG betont, zunehmend Produkte zu erstellen, für die Stellungnahmeverfahren (nach § 139a Abs. 5 SGB V) zu allen wichtigen Erstellungsabschnitten nicht vorzusehen sind".

Eine besondere Betonung ist weder formuliert noch intendiert. Neben den gesetzlich neu vorgesehenen Dossierbewertungen besteht die einzige Änderung in der Ergänzung von „Addenda" (in der Entwurfsversion noch als „Stellungnahmen des IQWiG" bezeichnet). Diese sind auf eine sehr spezifische Situation mit dem Ziel, die Transparenz zu fördern, beschränkt.

Allen drei Produkttypen, für die kein externes Stellungnahmeverfahren im Rahmen der Produkterstellung im Institut vorgesehen ist (Dossierbewertungen, Rapid Reports, Addenda), ist gemein, dass sie in möglichst kurzer, teilweise gesetzlich vorgegebener Zeitfrist erarbeitet und dem Auftraggeber zur Verfügung gestellt werden sollen bzw. müssen. Die für vorläufige Berichtspläne und Vorberichte etablierten Stellungnahmeverfahren erfordern in der Regel eine Zeitdauer, die bereits der o. g. Frist entspricht und somit nicht praktikabel ist. Rapid Reports und Addenda werden jedoch vor

Veröffentlichung dem Kuratorium des Instituts, in dem alle gesellschaftlich relevanten Gruppen vertreten sind, übersandt, sodass hier die Möglichkeit zur Stellungnahme besteht. Für die Dossierbewertungen gibt es die Stellungnahmemöglichkeit beim Gemeinsamen Bundesausschuss.

Mehrere Stellungnahmen kritisieren die zu kurzen Fristen für Stellungnahmen, sowohl für die Berichte als auch für das Methodenpapier selbst. Solange den nach § 139a SGB V Berechtigten nicht ausreichend Zeit für Stellungnahmen eingeräumt werde, könne auch nicht von deren Billigung ausgegangen werden.

Die vom Institut gesetzten Fristen entsprechen üblicher nationaler und internationaler Praxis. Sie können auch deshalb als ausreichend angesehen werden, da die Aufträge und ihre Inhalte lange vorher öffentlich bekannt sind (Auftrag und Berichtsplan) und eine Fristverlängerung zu weiteren Verzögerungen in den zeitlichen Abläufen führen würde. Im Übrigen beschreibt das Methodenpapier den Regelfall, von dem Ausnahmen möglich sind, was in der Vergangenheit bereits praktiziert wurde. Es gibt somit keinen hinreichend begründeten Anlass, die Fristen für Stellungnahmen zu den Produkten des Instituts generell zu verlängern. Mit einem Stellungnahmeverfahren ist im Übrigen eine Billigung weder vorgesehen noch möglich. Die Fristen für Stellungnahmen zu Grundsatzpapieren wie dem Methodenpapier selbst werden allerdings zukünftig verlängert.

Eine Stellungnahme weist darauf hin, dass in den Kreis der Gremien, an die die Produkte vor Veröffentlichung weitergeleitet werden, auch der Stiftungsrat des Instituts einbezogen werden sollte.

Der Hinweis ist berechtigt und wird umgesetzt.

3.3 Produkte des Instituts

Eine Stellungnahme kritisiert, dass das Methodenpapier selbst in dem Kapitel „Produkte des Instituts“ weder berücksichtigt noch beschrieben wird.

Die „Allgemeinen Methoden“ des Instituts sind nicht als Produkt im Sinne der erwähnten Systematik anzusehen. Der Charakter dieses Papiers wird in der Präambel beschrieben.

3.4 Ablauf der Berichterstellung

Eine Stellungnahme weist darauf hin, dass die Abbildung 1 des Methodenpapiers nicht mehr in allen Punkten den Erläuterungen im Text entspreche. So fehlten die (in der entsprechenden Abbildung der Version 3.0 noch vorhandenen) Textfelder, wonach der vorläufige Berichtsplan und der Vorbericht beim Auftraggeber/Kuratorium/Stiftungsvorstand vorgelegt werden.

Im Entwurf für die Allgemeinen Methoden Version 4.0 wurde im Vergleich zur Version 3.0 dieser Prozessschritt aus der Abbildung aus Gründen der Übersichtlichkeit entfernt. Im Text wird – wie von

den Stellungnehmenden zutreffend konstatiert – weiterhin die Vorlage beim Auftraggeber, Kuratorium und Stiftungsvorstand (und zukünftig auch Stiftungsrat) adressiert. Aus Sicht des Instituts erscheint das ausreichend, sodass sich aus diesem Hinweis kein Änderungsbedarf ergibt.

3.5 Grundlage der Dossierbewertungen

In einer Stellungnahme wird der Verwunderung darüber Ausdruck verliehen, dass bei den Dossierbewertungen diese Nutzenbewertung in der Regel allein aufgrund von Nachweisen des pharmazeutischen Unternehmers (Dossiers) und nur optional unterstützt durch zusätzliche Evidenzrecherchen erfolgen soll. Hierbei käme dann einer Anhörung von Vertreterinnen und Vertretern aus Wissenschaft und Praxis als zusätzliche Informationsquelle für die Abwägung von potenziellem Nutzen und Schaden des neuen Arzneimittels große Bedeutung zu.

Die Grundlagen für die Dossierbewertungen – also insbesondere die Bewertung eines vom Hersteller einzureichenden Dossiers– sind im SGB V und in der AM-NutzenV festgelegt und durch das Institut nicht zu beeinflussen. Das Institut gibt zwar eine Bewertung und Empfehlung ab, der Beschluss zum (Zusatz-)Nutzen wird jedoch durch den Gemeinsamen Bundesausschuss gefällt. Zusätzliche Informationen von Vertreterinnen und Vertretern aus Wissenschaft und Praxis können also den Gemeinsamen Bundesausschuss im Hinblick auf die Abwägung von potenziellem Nutzen und Schaden durch das dort angesiedelte Stellungnahmeverfahren noch erreichen und gehen nicht verloren. Davon unberührt ist im SGB V ein Weg beschrieben, der für den Fall des Vorliegens neuer Erkenntnisse wieder in eine Dossierbewertung mündet. Insgesamt ergibt sich aus den vorgebrachten Argumenten kein Änderungsbedarf.

3.6 Beteiligung externer Sachverständiger an der Erstellung des vorläufigen Berichtsplans

In einer Stellungnahme wird angenommen, das Institut wolle die externen Sachverständigen nicht mehr an der Erstellung des vorläufigen Berichtsplans beteiligen.

Dieses Missverständnis mag auf der Streichung einer entsprechenden, allerdings redundanten Passage im Zusammenhang mit der Beschreibung des Begriffs „Berichtsplans“ im Abschnitt 2.1.1 des Methodenpapiers entstanden sein. Diese Passage wurde gestrichen, weil es bereits zu Beginn dieses Abschnitts unmissverständlich heißt: „Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt. Dabei werden regelhaft externe Sachverständige beteiligt.“ Ein Änderungsbedarf ergibt sich für diesen Aspekt also nicht.

3.7 Weiterleitung von Stellungnahmen zu einem Rapid Report nach Publikation

In einer Stellungnahme wird angeraten, jegliche Stellungnahmen, die nach Veröffentlichung eines Rapid Reports im Institut eingehen, an den Auftraggeber weiterzuleiten.

Im Methodenpapier ist vorgesehen: „Sollten Stellungnahmen zu Rapid Reports eingehen, die substanzielle nicht berücksichtigte Evidenz enthalten, oder erlangt das Institut auf andere Weise Kenntnis von solcher Evidenz, wird dem Auftraggeber begründet mitgeteilt, ob eine Neubeauftragung zu dem Thema (ggf. Aktualisierung des Rapid Reports) aus Sicht des Instituts erforderlich erscheint." Eine Ausweitung dieser Regelung auf jegliche Stellungnahmen ist aus Sicht des Instituts nicht sachgerecht. Stellungnahmen, die dem Auftraggeber zur Kenntnis gelangen sollen, können und sollten direkt an den Auftraggeber adressiert werden. Es ergibt sich somit kein Änderungsbedarf.

3.8 Auswahl externer Sachverständiger (AMNOG)

Eine Stellungnahme schlägt vor, die ausführlichere Beschreibung des Prozederes, die auf den Internetseiten des Instituts verfügbar ist, auch in den „Allgemeinen Methoden" darzustellen. Eine weitere Stellungnahme hinterfragt die Zusammensetzung und Pflege der institutseigenen Sachverständigendatenbank und wünscht die Publikation der Kriterienliste (für die Auswahl).

Das Methodenpapier beschreibt allgemeine Methoden als Rahmen für die Institutsarbeit. Einzelne Konkretisierungen, die auch mitunter einem schnell(er)en Wandel unterworfen sind, würden den Rahmen eines allgemeinen Methodenpapiers sprengen; diese finden sich, soweit erforderlich, auf den Internetseiten des Instituts und können dort eingesehen werden. Dazu zählt auch die Kriterienliste, sodass die gewünschte Veröffentlichung gegeben ist. Konkrete Maßnahmen der internen Qualitätssicherung sind in den institutsinternen Arbeitsabläufen beschrieben. Ein Änderungsbedarf aus den Stellungnahmen zu den genannten Aspekten lässt sich daher nicht ableiten.

3.9 Auswahl der externen Reviewer

In einer Stellungnahme werden (an mehreren Stellen) die Auswahlkriterien für die externen Reviewer erfragt.

Dies ist nicht ganz nachvollziehbar, weil die Auswahlkriterien und der Auswahlprozess unmissverständlich in Abschnitt 2.1.1 des Methodenpapiers („… einem oder mehreren externen Gutachtern mit ausgewiesener methodischer und / oder fachlicher Kompetenz vorgelegt") und Abschnitt 2.2.3 des Methodenpapiers beschrieben sind: „Die Identifikation und die Auswahl potenzieller externer Reviewer sind abhängig vom Umfang des beauftragten Reviews. Sehr umfangreiche externe Reviews können auch als wissenschaftliche Forschungsaufträge vergeben werden. Für diese gelten dann die in Abschnitt 2.2.1 genannten Bedingungen. Ansonsten kann die

Identifikation externer Reviewer durch eine entsprechende Recherche, durch die Kenntnis der Projektgruppe, durch das Ansprechen von Fachgesellschaften, durch eine Bewerbung im Rahmen der Ausschreibung für die Auftragsbearbeitung usw. erfolgen. Eine Darlegung potenzieller Interessenkonflikte muss aber in jedem Fall erfolgen. Die Auswahl der externen Reviewer erfolgt durch das Institut." Ein Änderungsbedarf ergibt sich also zu diesem Punkt nicht.

In dieser Stellungnahme wird an anderer Stelle der Auswahlprozess kritisiert und auf das Peer-Review-Verfahren von wissenschaftlichen Zeitschriften verwiesen.

Offensichtlich handelt es sich hier um ein Miss- oder Unverständnis. HTA-Berichte (auch die des Instituts) unterlaufen gemäß internationalen Standards zahlreiche Qualitätssicherungsmaßnahmen. Hierzu gehören im Institut auch das externe Review, aber auch zahlreiche andere Reviewschritte davor und danach (siehe auch nachfolgenden Punkt). Das Peer-Review-Verfahren von wissenschaftlichen Zeitschriften verfolgt einen völlig anderen Zweck, nämlich aus einer Vielzahl von eingereichten Beiträgen die am besten geeigneten auszuwählen, wobei neben Qualität noch eine Reihe anderer Faktoren (z. B. Originalität) eine Rolle spielen. Aus dem Kritikpunkt resultiert also kein Änderungsbedarf.

3.10 Mindestens 2 Gutachterinnen/Gutachter

In einer Stellungnahme wird im Zusammenhang mit dem externen Review gefordert, es seien „bis auf Weiteres Stellungnahmen von mindestens zwei unabhängig voneinander agierenden Gutachtern zu fordern". In den nachfolgenden Erläuterungen, Behauptungen und Forderungen der Stellungnehmenden wird allerdings dieser Aspekt auf die externen Sachverständigen erweitert, zudem wird der Eindruck erweckt, dass während der Berichterstellung keine unabhängigen Bewertungsschritte zweier Bearbeitender und kein Einbezug medizinisch-inhaltlicher Expertise erfolge.

Zunächst ist klarzustellen, dass das Institut gemäß internationaler Standards der evidenzbasierten Medizin Schritte zur Identifikation und Bewertung wissenschaftlicher Studien immer durch mindestens 2 Personen durchführen lässt. Darüber hinaus wird ein Institutsprodukt in der Folge zahlreichen Reviewschritten unterworfen, u. a. auch einem externen Review, welches zusätzlicher Bestandteil der (internen) Qualitätssicherung ist. Die in diesem Zusammenhang von den Stellungnehmenden vorgebrachte Behauptung, es werde „oftmals [durch die an der Berichterstellung unmittelbar beteiligten Wissenschaftlerinnen und Wissenschaftler] nur methodische Kompetenz [abgedeckt]", entbehrt jeglicher Grundlage.

3.11 Namensnennung von externen Sachverständigen und externen Reviewern

Im Hinblick auf eine im Methodenpapier formulierte Ausnahmeregelung („... Auf ausdrücklichen Wunsch der externen Sachverständigen, auf Aufforderung des Auftraggebers oder aufgrund anderer wichtiger Umstände ist es möglich, die Namen externer Sachverständiger zur Gewährleistung ihrer Unabhängigkeit und zur Vermeidung einer interessenbedingten Einflussnahme nicht zu veröffentlichen ....") wird in einer Stellungnahme gefordert, die Sachverständigendatenbank des Instituts als auch die Namen der externen Sachverständigen der jeweiligen Institutsprodukte (offenbar in jedem Fall) öffentlich zu machen. Dies sei erforderlich, um zu prüfen, ob „Sachverständige beteiligt waren und die diesbezüglichen Regeln eingehalten wurden". An anderer Stelle wird in dieser Stellungnahme zusätzlich gefordert, (auch) die Namen der externen Reviewer (obligatorisch) zu veröffentlichen.

Das Institut setzt sich für eine möglichst umfassende Transparenz seiner Prozesse ein. Bis auf wenige Ausnahmefälle wurden daher in der Vergangenheit die Namen der an der Berichterstellung beteiligten externen Sachverständigen und der externen Reviewer in den Abschlussberichten bzw. Rapid Reports öffentlich gemacht. Schon deshalb ist die implizite Befürchtung bzw. gar Unterstellung, das Institut würde sich nicht an die (gesetzlichen) Regeln halten, nicht begründet. Die Veröffentlichung der Namen von externen Sachverständigen des Instituts bzw. potenziellen Interessentinnen und Interessenten und der externen Reviewer muss andererseits mit der einschlägigen Datenschutzgesetzgebung in Einklang stehen. Allein aus diesen Gründen ist z. B. die Veröffentlichung der Sachverständigendatenbank, die lediglich Interessentinnen und Interessenten für die Mitarbeit an Dossierbewertungen gemäß § 35a SGB V enthält, ausgeschlossen. Selbstverständlich beachtet das Institut daher alle rechtlichen Verpflichtungen.

3.12 Aufgabenbereiche der externen Sachverständigen und der externen Reviewer

In einer Stellungnahme wird angemerkt, dass die Aufgabenbereiche „der Sachverständigen, der externen Gutachter und der externen Reviewer nicht klar voneinander abgegrenzt" seien.

Hierzu ist anzumerken, dass (bisher) im Methodenpapier die Begriffe Reviewer und Gutachterin bzw. Gutachter im Zusammenhang mit der Produkterstellung synonym verwendet wurden. Da dies offenbar zu Missverständnissen geführt hat, wurde dies nun für die finale Version vereinheitlicht. Grundsätzlich bezieht das Institut gemäß § 139b Abs. 3 SGB V externe Sachverständige in die Berichterstellung ein. Darüber hinaus werden Vorberichte und Rapid Reports einem externen Review als Bestandteil der (internen) Qualitätssicherung unterzogen. Externe Sachverständige, die an der Berichterstellung mitgewirkt haben, können entsprechend wissenschaftlichen Gepflogenheiten nicht die Funktion eines externen Reviewers für diesen Bericht übernehmen.

3.13 Anzahl externer Sachverständiger für die frühe Nutzenbewertung nach § 35a SGB V

In mehreren Stellungnahmen wird befürchtet, dass durch den Einbezug von jeweils nur einem externen Experten unterschiedliche fachliche Aussagen von externen Experten von vornherein – verfahrensbedingt – ausgeschlossen seien. Es wird deshalb in einer Stellungnahme angeregt, nach Auswertung der ersten Erfahrungen mit der Einbindung der externen Expertise die Beteiligung von regelhaft zwei externen Sachverständigen pro Bewertungsverfahren vorzusehen.

Die Befürchtung der Stellungnehmenden wird vom Institut (zunächst) so nicht geteilt, da ja im Institut selbst (auch medizinisch-fachliche) Expertise vorgehalten wird und somit eine Zweitmeinung gewährleistet ist. Wie in einer der Stellungnahmen angeregt, wird es jedoch nach einer angemessenen Frist eine Überprüfung des Auswahlprozesses geben, die ggf. zu einer Modifikation führen kann. Dies kann allerdings nur Einfluss auf die nächste Methodenversion haben, sodass zum jetzigen Zeitpunkt kein Änderungsbedarf besteht.

3.14 Einbezug von pharmazeutischen Unternehmen in die Erstellung von Addenda

In einer Stellungnahme wird angeregt, in die Erstellung von Addenda (im Entwurf noch als „Stellungnahmen des IQWiG“ bezeichnet) zu Arzneimittelbewertungen den entsprechenden pharmazeutischen Unternehmer in die Erarbeitung einzubeziehen.

Die direkte Einbindung des pharmazeutischen Unternehmers in die Erstellung von Produkten des Instituts widerspricht dem Unabhängigkeitsgebot nach § 139a und § 139b des SGB V und ist dort so nicht vorgesehen. Somit ergibt sich hieraus keine Änderungsmöglichkeit.

3.15 Einbezug externer Expertise in die Erstellung von Rapid Reports

In einer Stellungnahme wurde angeregt, den Einbezug „medizinischer Fachexperten“ in die Erstellung von Rapid Reports nicht nur optional, sondern regelhaft vorzusehen.

Rapid Reports dienen der Information des Auftraggebers in möglichst kurzer Zeitfrist. Um eine möglichst rasche Erarbeitung zu gewährleisten, müssen die Arbeitsabläufe entsprechend angepasst werden. Bei Themen, für die im Institut ausreichend fachlich-inhaltliche (auch medizinische) Expertise vorhanden ist, ist es ausreichend, durch das externe Review im Rahmen der (internen) Qualitätssicherung eine (weitere) unabhängige Bewertung des Produkts zu erhalten. Anderenfalls (wenn keine ausreichende fachlich-inhaltliche Expertise im Institut vorhanden ist) werden – wie in der Vergangenheit zumeist – externe Sachverständige in die Erstellung des Rapid Reports einbezogen.

Insofern ist die Optionsregelung angemessen, und es ergibt sich aufgrund der Stellungnahme kein Änderungsbedarf.

3.16 Einbeziehung von Patientinnen und Patienten

In zwei Stellungnahmen wird die Frage gestellt, wie Patientinnen und Patienten bzw. Patientenvertreter für die Beteiligung an Institutsprodukten ausgewählt werden und ob dabei ein „methodisches Vorgehen, welches Repräsentativität anstreben sollte, regelhaft verfolgt wird".

Die Beteiligten werden vorrangig über die im Bundesausschuss vertretenen Patientenorganisationen ausgewählt. Die Auswahl erfolgt also durch die Betroffenen selbst. Repräsentativität wird derzeit nicht angestrebt und ist auch nicht realistisch. Wenn deutlicher würde, was „Repräsentativität" in diesem Zusammenhang bedeutet, ob sie hier ein relevantes Ziel ist und wie sie ggf. herzustellen wäre, könnte ein entsprechendes Vorgehen erwogen werden.

3.17 Wissenschaftlicher Beirat

Eine Stellungnahme hinterfragt den Begriff „Notwendigkeit" im Hinblick auf die Einbeziehung des wissenschaftlichen Beirats des Instituts.

Ein Beirat wird üblicherweise dann einbezogen, wenn die zu beratende Institution einen Bedarf zur Beratung hat. Der Begriff wird dementsprechend in „Bedarf" geändert.

3.18 Spezifizierung der Anhörungspraxis

In einer Stellungnahme wird gefordert, die Anhörungspraxis näher zu spezifizieren. Dabei wird der Vorwurf erhoben, mündliche Erörterungen (im Institut) hätten in der Vergangenheit einen Tribunal-Charakter gehabt und seien nicht ergebnisoffen gewesen. Insbesondere die Tonbandaufnahmen bei zusätzlicher Wortprotokollierung durch einen vereidigten Protokollführer ließen Misstrauen und Argwohn vermuten und lösten Befremden aus.

Bei diesem Kommentar bleibt dem Institut unklar, was bei der Anhörungspraxis im Detail näher spezifiziert werden soll und woraus sich der erhobene Vorwurf begründet. Das Institut ist und war zu jedem Zeitpunkt im Rahmen der Anhörungen wissenschaftlichen Argumenten gegenüber offen. Das Vorgehen zur Protokollerstellung der mündlichen Erörterung wird vom Institut als allen Beteiligten dienender Prozessschritt empfunden, der die notwendige Qualität des (Protokoll-)Ergebnisses sicherstellen soll. Insgesamt kann aus dem vorgebrachten Argument kein Änderungsbedarf abgeleitet werden.

3.19 Qualitätssicherungsprozess beschreiben

In mehreren Stellungnahmen wird darauf hingewiesen, dass die (internen) Qualitätssicherungsverfahren nicht im Methodenpapier beschrieben seien.

Das Methodenpapier beschreibt – wie bereits an anderer Stelle erwähnt – allgemeine Methoden als Rahmen für die Institutsarbeit. Einzelne Konkretisierungen, die auch mitunter einem schnell(er)en Wandel unterworfen sind, würden den Rahmen eines allgemeinen Methodenpapiers sprengen. Das Institut gehört zu den wenigen HTA-Institutionen (und vergleichbaren Einrichtungen – sowohl national als auch international), die bereits unmittelbar nach Gründung ein solches Methodenpapier erstellt und publiziert haben. Es ist national und international nicht üblich, interne Qualitätsmanagement- und Qualitätssicherungsdokumente öffentlich zu machen. Ein Änderungsbedarf ergibt sich aus Sicht des Instituts aus dem vorgebrachten Punkt nicht.

4 Würdigung der Stellungnahmen zu Kapitel 3

4.1 Hierarchisierung von Endpunkten

In mehreren Stellungnahmen wurde im Abschnitt zur Definition des patientenrelevanten Nutzens darauf hingewiesen, dass die Hierarchisierung von Endpunkten ein komplexer Prozess sei, für den Kriterien angegeben werden sollten. Zudem sei eine solche Wertung Aufgabe des G-BA. Darüber hinaus wurde der in diesem Zusammenhang bemühte Verweis auf ein Urteil des Bundessozialgerichts kritisch hinterfragt, insbesondere in Anbetracht des wissenschaftlichen Auftrags des Instituts.

Zunächst erfolgt die Adressierung der Hierarchisierung von Endpunkten in diesem Abschnitt des Methodenpapiers bei der Aufzählung von Nutzenkriterien. Ein Teil dieser Nutzenkriterien (Mortalität, Morbidität und Lebensqualität) finden sich als Bewertungsmaßstab explizit im SGB V, ein anderer im Methodenpapier als nachrangig gekennzeichneter Teil (interventions- und erkrankungsbezogener Aufwand, Zufriedenheit der Patienten) nicht. Auch wenn die Aufzählung im SGB V nicht als abschließend betrachtet werden kann, ergeben sich im Zusammenhang mit den Ausführungen zur Quantifizierung des Zusatznutzens in der Arzneimittel-Nutzenbewertungsverordnung (AM-NutzenV) eindeutige Anhaltspunkte für eine Hierarchisierung von Endpunkten: Bestimmte Endpunkte (z. B. „Überlebensdauer“) finden sich prioritär in den beiden höchsten Nutzenkategorien, andere Endpunkte (z. B. „nicht schwerwiegende Symptome der Erkrankung“) nur in der untersten Nutzenkategorie.

Es wird ergänzt, dass für alle „Zielgrößen gilt, dass es erforderlich sein kann, diese jeweils nur im Zusammenhang mit Informationen zur vergleichenden Beeinflussung anderer Zielgrößen zu bewerten“. Dies ist im Zusammenhang mit der Zulassung seit Jahrzehnten geübte Praxis, insbesondere im Zusammenhang mit sogenannten Nicht-Unterlegenheitsstudien, wo es eben nicht allein ausreicht, den Nachweis der Überlegenheit für eine Zielgröße (implizit oder explizit) zu führen, sondern gleichzeitig nachgewiesen werden muss, dass eine andere Zielgröße nicht in einem nicht mehr akzeptablen Ausmaß ungünstig beeinflusst wird. Auch hier ist eine weitere Präzisierung überflüssig. Warum es wegen des wissenschaftlichen Auftrags des Instituts hierbei unangemessen sein soll, auf höchstrichterliche Rechtsprechung im Sinne einer Normgebung zu verweisen, wurde in der Stellungnahme nicht näher ausgeführt und erschließt sich nicht. Schließlich wird an anderen Stellen des Methodenpapiers auf z. B. ein Urteil des Bundesverfassungsgerichts oder auf das SGB V rekurriert, ohne dass diesbezüglich eine entsprechende Kritik von den Stellungnehmenden formuliert wurde.

Schließlich wurde im Hinblick auf die Aussage im Entwurf zur Version 4.0 des Methodenpapiers, dass „sowohl Nutzen- als auch Schadenaspekte eine unterschiedliche Wichtigkeit für die Betroffenen haben [können]…“, in einzelnen Stellungnahmen erneut auf fehlende Kriterien dafür verwiesen.

Auch dieser Kritikpunkt kann nicht recht nachvollzogen werden, da die Kriterien im unmittelbaren Anschluss genannt werden („… die sich ggf. durch qualitative Erhebungen oder bereits bei der Beratung durch Betroffene, Patientenvertretungs- und/oder Verbraucherorganisationen im Zusammenhang mit der Definition patientenrelevanter Endpunkte abzeichnet“), nämlich die Sicht der Betroffenen. Dennoch wird dieser Kritikpunkt aufgegriffen und an dieser Stelle ein Verweis auf entsprechende Ansätze, die später im Methodenpapier bei der Frage nach Methoden zur Aggregation von Endpunkten genannt werden, ergänzt (siehe auch nachfolgenden Punkt).

4.2 Aggregation von Endpunkten

In einer Stellungnahme wurde bemängelt, dass das Institut in der Entwurfsversion für die Allgemeinen Methoden 4.0 eine Ausführung, wie Endpunkte aggregiert werden können, schuldig geblieben sei. Weiterhin wurde in diesem Zusammenhang darauf hingewiesen, dass dieser Abwägungsprozess genuin dem Aufgabenbereich des G-BA zuzuordnen sei.

Es ist sicherlich zutreffend, dass es in die letztendliche Zuständigkeit des G-BA fällt, Endpunkte in einer Erstattungsentscheidung gegeneinander abzuwägen. In der Methode, wie man zu einer Gewichtung kommen kann, kann sich der Entscheidungsträger allerdings beraten lassen bzw. eine Empfehlung entgegennehmen. Im Übrigen stellen viele allgemein akzeptierte Endpunkte bereits implizit eine Aggregation von Sub-Endpunkten dar, allerdings zumeist mit einer simplen Gleichgewichtung. Sollte es für einen Auftrag an das Institut erforderlich sein, Endpunkte zu aggregieren, so werden die dafür angewendeten Methoden im Berichtsplan präzisiert. Dennoch wird die Anregung der Stellungnehmenden aufgegriffen. Entsprechende Ansätze werden im Methodenpapier dargestellt.

4.3 Patientenrelevanz

In einer Stellungnahme wurde festgestellt, dass die im Abschnitt 3.1.1 des Entwurfs für die Version 4.0 des Methodenpapiers genannten Zielgrößen (Mortalität, Morbidität und gesundheitsbezogene Lebensqualität sowie der interventions- und erkrankungsbezogene Aufwand) vor allem bei der Bewertung onkologischer Therapien zu undifferenziert seien.

Da das Methodenpapier nur einen allgemeinen Rahmen beschreiben soll und kann, ist es nicht sinnvoll (und auch nicht möglich), hier für einzelne Indikationen die jeweils spezifischen patientenrelevanten Zielgrößen zu definieren. Eine solche Spezifizierung erfolgt projektbezogen im Berichtsplan.

Weiterhin wurde in der Stellungnahme angeregt, in der Onkologie neben Gesamtüberleben und (gesundheitsbezogener) Lebensqualität „alternative patientenrelevante Effektmaße, z. B. die Zeit des krankheitsfreien Überlebens (Progression-Free Survival, PFS) oder die Zeit bis zum Progress einer

Erkrankung (Time-to-Progression, TTP), als Nutzenmaße für die isolierte Nutzenbewertung festzulegen, um den Patientenbedürfnissen angemessen Rechnung zu tragen“.

Da auch die Morbidität als zentrale, übergeordnete Zielgröße in dem betreffenden Abschnitt aufgeführt wird, ist klar, dass das Institut neben Gesamtüberleben und (gesundheitsbezogener) Lebensqualität auch andere Zielgrößen in Nutzenbewertungen betrachtet. Einzige Anforderung ist, dass diese Zielgrößen Patientenrelevanz besitzen müssen, sodass auch die von den Stellungnehmenden explizit aufgeführten Effektmaße Berücksichtigung finden, sofern sich in ihrer (weiteren) Operationalisierung patientenrelevante Aspekte niederschlagen (z. B. die Symptomatik der Patientinnen und Patienten, siehe auch die weiter unten geführte Diskussion zu Surrogaten). Insgesamt ergibt sich aus der Stellungnahme kein Änderungsbedarf, da beide Forderungen bereits im Methodenpapier umgesetzt sind.

4.4 Nutzennachweis

In einer Stellungnahme wurde folgende Aussage im Zusammenhang mit der Nutzenbewertung von Arzneimitteln bemängelt: „Aufgrund der Zielsetzung der Nutzenbewertung durch das Institut werden in die jeweilige Bewertung nur Studien einer Evidenzstufe eingeschlossen, die zum Nachweis des Nutzens grundsätzlich geeignet ist. Studien, die lediglich Hypothesen generieren können, sind deshalb im Allgemeinen für die Nutzenbewertung nicht relevant. Die Frage, ob eine Studie einen Nachweis eines Nutzens erbringen kann, hängt im Wesentlichen von der Ergebnissicherheit der erhobenen Daten ab.“ Damit werde laut Stellungnehmenden suggeriert, „dass grundsätzlich nur RCTs für den Nachweis eines Nutzens geeignet sind, andere Studientypen somit lediglich hypothesengenerierend sein können“. Dies entspreche nicht dem internationalen Standard der evidenzbasierten Medizin.

In dem kritisierten einleitenden Abschnitt kommt der Begriff „Randomisierung“ nicht vor. Der von den Stellungnehmenden in diesem Zusammenhang zitierte Satz „Die Frage, ob eine Studie einen Nachweis eines Nutzens erbringen kann, hängt im Wesentlichen von der Ergebnissicherheit der erhobenen Daten ab“ wird offenbar von ihnen selbst dahin gehend interpretiert, dass „grundsätzlich nur RCTs“ eine ausreichende Ergebnissicherheit für einen Nutzennachweis haben. Wie an anderen Stellen des Methodenpapiers ausführlich erläutert, ist es gerade das Kennzeichen der evidenzbasierten Medizin, die (Nutzen-)Nachweisfähigkeit von Studien an deren Ergebnissicherheit zu messen (siehe Abschnitt 1.2.5 des Methodenpapiers). Insofern ergibt sich kein Widerspruch zu den internationalen Standards der evidenzbasierten Medizin und kein Änderungsbedarf.

Im Anschluss wurde in dieser Stellungnahme formuliert, dass „auch zum Wirksamkeitsnachweis für Interventionen entsprechender Erkenntnisgewinn aus Beobachtungsstudien ableitbar ist“.

Abgesehen davon, dass es nicht Aufgabe des Instituts ist, Wirksamkeitsnachweise zu bewerten, sondern Nutzennachweise, steht die Aussage nicht im Widerspruch zu den Ausführungen des Entwurfs für die Version 4.0 des Methodenpapiers. An verschiedenen Stellen des Methodenpapiers wird hervorgehoben, dass unter bestimmten Voraussetzungen auch andere Studientypen als RCTs (und damit auch Beobachtungsstudien) in eine Bewertung eingeschlossen werden und damit für einen Nutzennachweis geeignet sein können (siehe z. B. Abschnitt 3.2.3 des Methodenpapiers). Es besteht somit kein Änderungsbedarf.

In einer anderen Stellungnahme wurde kritisiert, dass im Entwurf für die Version 4.0 des Methodenpapiers im Vergleich zu früheren Versionen „im Kern als Regelanforderung der Beleg eines statistisch signifikanten Effekts durch eine Metaanalyse von Studien mit endpunktbezogen geringer Ergebnisunsicherheit oder durch mindestens zwei voneinander unabhängig durchgeführte Studien mit endpunktbezogen geringer Ergebnisunsicherheit bestehen“ bleibe. Dies sei schon zu früheren Zeiten als zu rigide eingestuft worden.

Dieser Kritikpunkt kann nicht nachvollzogen werden. Von den Stellungnehmenden wird erstens der entsprechende Passus nicht vollständig wiedergegeben. Dort heißt es: „In der Regel wird an die Aussage eines Belegs die Anforderung zu stellen sein, dass eine Meta-Analyse von Studien, die endpunktbezogen **in der Mehrheit eine hohe Ergebnissicherheit aufweisen**, einen entsprechenden statistisch signifikanten Effekt zeigt.“ Damit bleiben die Anforderungen des Methodenpapiers im Prinzip hinter den Anforderungen von Zulassungsbehörden, z. B. der EMA, zurück. In dem in diesem Zusammenhang zitierten „Points-to-Consider“-Papier der EMA wird der Stellenwert von Meta-Analysen eher kritisch eingeschätzt (im Sinne als ggf. nicht ausreichend für einen Wirksamkeitsnachweis): “The use of a meta-analysis to provide the pivotal evidence in an application will always be problematic. One really robust trial supported by smaller trials is stronger than either a meta-analysis of studies, none of which is convincing on its own or a meta-analysis of seemingly conflicting results” [13]. Dadurch, dass im Entwurf für die Version 4.0 des Methodenpapiers sogar Studien geringerer Ergebnissicherheit für den Einbezug in eine entsprechende Meta-Analyse als akzeptabel angesehen werden (nur eben nicht in der Mehrheit), ist das ein deutlich liberaleres Vorgehen. Zweitens ist das Erfordernis der Replikation nahezu eine wissenschaftliche Conditio sine qua non, die im Grundsatz keiner näheren Begründung bedarf. Dennoch sei noch mal das o. g. Dokument der EMA zitiert: “There is a general demand for replication of scientific results” [13]. Davon unberührt, wird im Methodenpapier des Instituts in Anlehnung an das Vorgehen der Zulassungsbehörden formuliert: „Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und deren Ergebnisse besondere Anforderungen zu stellen.“ Aus dem vorgebrachten Kritikpunkt lässt sich somit kein Änderungsbedarf ableiten.

4.5 Frühe Nutzenbewertung (Dossierbewertung)

In einigen der Stellungnahmen wurden die Methoden der frühen Nutzenbewertung (Dossierbewertung) adressiert. Dabei wurden insbesondere folgende Punkte angesprochen:

1. *Die Anforderungen für Nutzenbelege (Abstufung Beleg, Hinweis etc., Abschnitt 3.1.4 der Methoden) sind nach Ansicht eines Stellungnehmenden für die Situation der frühen Nutzenbewertung ungeeignet, da der Zusatznutzen anhand dieser Anforderungen nicht angemessen dargestellt werden könne. Es sei daher u. a. sinnvoll, eine andere Evidenzklassifizierung für diese Situation zu verwenden, und zwar diejenige von Grimes 2002 [20].*
2. *Bei den Ausführungen zu Surrogatendpunkten (Abschnitt 3.1.2 der Methoden) ist nach Ansicht eines Stellungnehmenden der Hinweis, dass in Zulassungsverfahren nicht immer der patientenrelevante Nutzen untersucht werde, zu ändern. Richtig sei, dass nicht immer der patientenrelevante Zusatznutzen untersucht werde.*
3. *Bei den Ausführungen zur Bewertung der Kosten im Rahmen der frühen Nutzenbewertung (Abschnitt 3.3.3 der Methoden) schlägt ein Stellungnehmender vor, dass ein Verweis auf die Methoden zur Kosten-Nutzen-Bewertung des Instituts ergänzt werde.*
4. *Nach Ansicht eines Stellungnehmenden ist der Hinweis aus der Arzneimittel-Nutzen-Verordnung (AM-NutzenV [7]), dass zwischen Therapien regelhaft Unterschiede gemäß der Fach- und Gebrauchsinformation bestehen könnten, nicht auf die Inanspruchnahme von Leistungen zu beziehen, sondern auf die zulassungsgemäße Anwendung der Arzneimittel. Die Bewertung der Kosten im Rahmen der frühen Nutzenbewertung (Abschnitt 3.3.3 der Methoden) sei daher zu ändern.*
5. *Eine Prüfung der Zweckmäßigkeit einer vom pharmazeutischen Unternehmer im Dossier gewählten Vergleichstherapie (Abschnitt 3.3.3 der Methoden) kann vom Institut nach Ansicht eines Stellungnehmenden nur dann vorgenommen werden, wenn der G-BA keine Festlegung der zweckmäßigen Vergleichstherapie im Rahmen einer Beratung vorgenommen hat.*

Zu 1.: Das Institut teilt die Ansicht des Stellungnehmenden nicht, dass die Abstufung in Beleg, Hinweis und Anhaltspunkt für die frühe Nutzenbewertung ungeeignet ist. Zunächst ist es auch bei der frühen Nutzenbewertung sinnvoll, die Ergebnissicherheit der vorliegenden Evidenz zu beschreiben. Dies ist auch in der AM-NutzenV vorgesehen, laut der die „Wahrscheinlichkeit“ für das Vorliegen eines Zusatznutzens dargestellt werden soll [7]. Dies wird mittels der o. g. Abstufung operationalisiert. Es ist des Weiteren auch durchaus möglich, dass bereits zum Zeitpunkt der Zulassung ein Beleg für einen Zusatznutzen festgestellt werden kann. Hierfür müssen z. B. 2 ergebnissichere Studien vorliegen, in denen eine Überlegenheit gegenüber der geeigneten (zweckmäßigen) Vergleichstherapie

bezüglich patientenrelevanter Zielkriterien nachgewiesen wird. Werden 2 für die Zulassung notwendige pivotale Studien entsprechend ausgerichtet und zeigen diese ein entsprechendes Ergebnis, kann auch ein Nutzenbeleg aus solchen Studien abgeleitet werden. Schließlich kann dem Vorschlag der Verwendung einer anderen Evidenzklassifizierung nicht gefolgt werden, da die Evidenzklassifizierung durch die AM-NutzenV vorgegeben ist [7]. Davon abgesehen ist die vorgeschlagene Evidenzklassifizierung für die frühe Nutzenbewertung ungeeignet, da sie auf die Bewertung einzelner Studien ausgerichtet ist, nicht aber auf die Zusammenführung der Ergebnisse mehrerer Studien im Rahmen einer systematischen Übersicht, wie dies bei der frühen Nutzenbewertung der Fall ist [20]. Zusammenfassend ergaben sich durch diesen Aspekt der Stellungnahme keine Änderungen an den Methoden des Instituts.

Zu 2: Das Institut teilt die Auffassung des Stellungnehmenden nicht, dass im Rahmen der Zulassung immer der patientenrelevante Nutzen untersucht wird. Zwar ist dies je nach Indikation durchaus möglich, wenn die für die Zulassung verwendeten Wirksamkeitskriterien auch Nutzenkriterien darstellen. Dies ist z. B. in der Regel bei Thrombozytenaggregationshemmern der Fall, für deren Zulassung in den letzten Jahren Studien mit patientenrelevanten Zielgrößen (kardio- und zerebrovaskuläre Morbidität) durchgeführt wurden. Entsprechend wurde z. B. im Abschlussbericht A04-01B [33] des Instituts für Clopidogrel ein Nutzenbeleg auf Basis der Zulassungsstudie (CURE-Studie) ausgesprochen [33]. Im Gegensatz dazu ist jedoch z. B. bei oralen Antidiabetika der Wirksamkeitsnachweis in Form einer Blutzuckersenkung (sowie der Nachweis von Qualität und Unbedenklichkeit) für die Zulassung ausreichend. Die Blutzuckersenkung allein ist jedoch kein geeigneter Nutzenendpunkt für die Bewertung einer Therapie bei Typ 2 Diabetes mellitus, wie zuletzt das Beispiel Rosiglitazon gezeigt hat. Bei Typ 2 Diabetes mellitus ist die Verhinderung makrovaskulärer Folgekomplikationen ein wesentliches Nutzenkriterium. Trotz Blutzuckersenkung hat Rosiglitazon aber das Risiko für Herzinfarkte nicht reduziert, sondern offenbar erhöht, weshalb die Zulassung für Rosiglitazon von der europäischen Zulassungsbehörde zurückgezogen wurde [15]. Für die Wirkstoffgruppe der Glinide ergab sich auch fast 10 Jahre nach der Zulassung kein Hinweis auf einen Nutzen, da nicht nur die Zulassungsstudien den Nutzen nicht untersucht haben, sondern auch in Folge keine solchen Studien durchgeführt wurden [35]. Insgesamt ergab sich aus der Stellungnahme bezüglich dieses Punktes keine inhaltliche Änderung der Methoden des Instituts. Zur Klarstellung, dass sich der geschilderte Sachverhalt nicht nur auf den patientenrelevanten Nutzen, sondern auch auf den patientenrelevanten Zusatznutzen bezieht, wird der Verweis auf den patientenrelevanten Zusatznutzen in Abschnitt 3.1.2 der Methoden ergänzt.

Zu 3: Ein Verweis auf die Methoden zur Kosten-Nutzen-Bewertung bei der Darstellung der Methoden der frühen Nutzenbewertung erscheint nicht sinnvoll, da bei der frühen Nutzenbewertung keine Kosten-Nutzen-Bewertung gemäß den gesetzlichen Vorgaben durchgeführt wird, sondern lediglich

eine separate Bewertung der Kosten nach einem in der AM-NutzenV vorgegebenen Schema. Ein Änderungsbedarf für die Methoden des Instituts ergab sich diesbezüglich demnach nicht.

Zu 4: Das Institut teilt die oben skizzierte Auffassung des Stellungnehmenden nicht. Dem widerspricht auch die Umsetzung der AM-NutzenV durch den GBA in seiner Verfahrensordnung einschließlich der zugehörigen Anhänge (Dossiervorlagen) [7,18]. Hier ist klar beschrieben, dass bei der Kostenermittlung die zusätzliche Inanspruchnahme von Leistungen auf Basis der Ausführungen in der Fach- oder Gebrauchsinformation zu berücksichtigen ist. Die Verfahrensordnung des GBA wurde nicht beanstandet. Es ergab sich aus der Stellungnahme daher in dieser Hinsicht ebenfalls kein Änderungsbedarf für die Methoden des Instituts.

Zu 5: Es ist richtig, dass der G-BA nach den gesetzlichen Vorgaben die zweckmäßige Vergleichstherapie festlegt. Hierzu kann der pharmazeutische Unternehmer eine Beratung durch den G-BA beantragen. Der pharmazeutische Unternehmer kann in dem von ihm vorgelegten Dossier jedoch von der in der Beratung vorgenommenen Festlegung der zweckmäßigen Vergleichstherapie begründet abweichen (siehe Anhang zur Verfahrensordnung des G-BA [18]). Darüber hinaus können sich neue Erkenntnisse ergeben, die ggf. zu einer anderen zweckmäßigen Vergleichstherapie führen, insbesondere dann, wenn die Beratung mehrere Jahre vor der Dossiereinreichung durchgeführt wurde. Wird das Institut mit der Bewertung solcher Dossiers beauftragt, ist auch eine Bewertung, ob die Auswahl der zweckmäßigen Vergleichstherapie begründet ist, Gegenstand der Dossierbewertung durch das Institut. Die Festlegung der zweckmäßigen Vergleichstherapie bleibt aber auch in solchen Fällen Aufgabe des G-BA, der die Bewertung durch das Institut dabei berücksichtigen kann. Insgesamt ergab sich aus der Stellungnahme auch hier kein Änderungsbedarf für die Methoden des Instituts.

4.6 Dramatischer Effekt

In einer Stellungnahme wurde die Adressierung sogenannter dramatischer Effekte begrüßt, allerdings die beispielhaft angeführte Operationalisierung (relatives Risiko > 10) als ggf. zu strikt angesehen. Alternativ wurde auf die Diskussion zu Kriterien für die Anerkennung einer Berufskrankheit verwiesen, bei der bereits eine Risikoverdoppelung (relatives Risiko > 2) offenbar als wegweisend angesehen werde.

In der Tat ist die Risikoverdopplung ein verbreitetes Kriterium in der Arbeitsmedizin im Hinblick auf die Anerkennung einer Berufskrankheit, um einen von zwei notwendigen Kausalitätsnachweisen zu führen, nämlich den der sogenannten haftungsausfüllenden Kausalität [56]. Damit gemeint ist, ob ein *Individuum* einer bestimmten Schwelle einer Expositionsdosis ausgesetzt worden ist, um bei diesem *Einzelfall* von einem ursächlichen Zusammenhang zwischen Exposition und Krankheit ausgehen zu

können. Wesentlich dabei ist allerdings, dass *populationsbezogene Kausalität* bereits (epidemiologisch) nachgewiesen, also Voraussetzung ist. Hier kann z. B. der Nachweis einer Dosis-Wirkungs-Beziehung eine entscheidende Rolle spielen.

Bei der Frage nach einem dramatischen Effekt im Methodenpapier des Instituts geht es aber zunächst einmal genau um diese populationsbezogene Kausalität. Anders ausgedrückt: Ab welcher Größenordnung kann ein beobachteter Unterschied mit ausreichender Sicherheit nicht mehr allein durch Verzerrung verursacht worden sein, also Ausdruck eines echten Effekts sein. Das *Risikoverdopplungs-Kriterium* für die Anerkennung einer Berufskrankheit speist sich demgegenüber aus einer Interpretation des relativen Risikos als attributables Risiko [49]: „Denn wenn das sogenannte Verdopplungsrisiko überschritten ist (im Sinne eines Relativen Risikos > 2), liegt die Wahrscheinlichkeit der beruflichen Verursachung oberhalb 50 % (gleichbedeutend mit einem attributablen Risiko unter den Exponierten > 50 %). Aus dieser unmittelbaren Evidenz erklärt sich die steigende Bedeutung des Verdopplungsrisikos für die Einführung neuer sowie für die Umsetzung bestehender Berufskrankheiten.“ Das Kriterium an sich und diese Interpretation waren und sind offenbar immer noch Gegenstand einer Kontroverse [50]. Davon unberührt spielt das Problem, dass entsprechende Schätzwerte verzerrt sein könnten, in dieser Diskussion praktisch keine Rolle.

Dass jedoch die alleinige Beobachtung einer Risikoverdopplung bzw. umgekehrt Risikohalbierung aus nicht randomisierten Studien vermutlich nicht ausreicht, um nicht allein durch Verzerrung erklärt werden zu können, zeigen prominente Beispiele aus der jüngeren Vergangenheit (z. B. zu vermeintlich nützlichen Effekten von Beta-Karoten bezüglich Lungenkrebs bzw. der Hormonersatztherapie bezüglich Herzinfarkten), bei denen annähernd Risikohalbierungen in nicht randomisierten Studien beobachtet wurden, sich dann aber in randomisierten Studien nicht nur nicht diese Ergebnisse reproduzieren ließen, sondern sogar schädliche Effekte beobachtet wurden [19].

Die im Methodenpapier im Zusammenhang mit dramatischen Effekten zitierte Arbeit von Glasziou et al. [19] hat den Vorteil, dass die Empfehlung eines Grenzwertes für das relative Risiko von 10 auf Simulationsstudien basiert, nach denen ein beobachtetes relatives Risiko von 5-10 nicht mehr plausibel allein durch Störgrößeneinflüsse erklärt werden kann [19]. In einer Publikation der GRADE-Arbeitsgruppe findet sich eine ähnliche Empfehlung: Für den Fall „sehr großer Effekte mit einem RR [relativen Risiko] > 5 oder < 0,2 ohne plausible Confounder“ kann die „Qualität der Evidenz“ um 2 Stufen als Ausdruck eines überzeugenden Anstiegs der Ergebnissicherheit angehoben werden [36]. In dieser Empfehlung wird – anders als bei Glasziou et al. [19] – nicht noch eine zusätzliche Anforderung an die Präzision der Schätzung (explizit) formuliert. Eine solche Anforderung erscheint allerdings sinnvoll und angemessen.

Insgesamt wirkt es weiterhin angemessen, bei der Einordnung eines beobachteten Unterschieds als „dramatischen Effekt“ sich an der Empfehlung von Glasziou et al. [19] zu orientieren. Dessen ungeachtet ist es aber sinnvoll, im Methodenpapier noch deutlicher zu machen, dass eine solche Grenze nicht starr, sondern auch von den Begleitumständen (u. a. Qualität der zugrunde liegenden Studien) abhängig ist. Damit wird auch ein weiterer Aspekt in der Stellungnahme berücksichtigt, indem hinterfragt wird, warum nur auf „Fallserien“ Bezug genommen würde.

Darüber hinaus wird in der Stellungnahme im Zusammenhang mit der Frage nach dramatischen Effekten gefordert, dass „in jedem Fall eine Darstellung entsprechender Studien erfolgt, auch wenn der Effekt kleiner ausfallen sollte“.

Dem wird insofern Rechnung getragen, als die Aufnahme entsprechender Studien nicht vom Ergebnis dieser Studien abhängig gemacht wird, sondern vom Ergebnis von orientierenden Vorabrecherchen oder auch entsprechenden Hinweisen an das Institut (z. B. im Rahmen der Stellungnahmeverfahren). Diese Möglichkeit ist in dem entsprechenden Abschnitt des Methodenpapiers, aber auch anderen Stellen erwähnt (z. B. Abschnitt 2.1.1 des Methodenpapiers: Vorlage von Unterlagen im Rahmen der Anhörung), sodass hieraus keine Notwendigkeit einer Änderung resultiert.

4.7 Studiendauer

Einer der Stellungnehmenden kritisierte, dass im Abschnitt 3.2.4 der Methoden des Instituts („Studiendauer“) nicht auf Probleme bestimmter Designkomponenten in Langzeitstudien eingegangen wurde.

Im genannten Abschnitt 3.2.4 der Methoden wird beschrieben, dass für verschiedene Studienziele unterschiedliche Studiendauern sinnvoll sind. Es ist nicht Gegenstand dieses Abschnitts, spezifische Probleme bei der Durchführung von Langzeitstudien zu adressieren, zumal derartige Hindernisse je nach Indikation unterschiedlich stark vorhanden sein können und deshalb in der konkreten Nutzenbewertung zu adressieren sind. Die in diesem Zusammenhang explizit angesprochene Randomisierung ist im Übrigen kein Problem der Studiendauer, da Strukturgleichheit mittels Randomisierung ohnehin nur zum Zeitpunkt der Randomisierung, also dem Studienbeginn, realisiert werden kann [38]. Darüber hinaus heißt es an anderer Stelle: „Probleme, die bei langjährigen Studien auftreten (z. B. mit der Zeit zunehmende Ausfallraten von Patienten), treffen wiederum nicht ausschließlich auf RCTs zu. Sie treten hier aufgrund von etablierten Empfehlungen zur Publikationspraxis [48] nur deutlicher in Erscheinung, während sie sich in nichtrandomisierten Studien hinter der u. U. selektiven und möglicherweise verzerrten Auswahl von Patienten, für die Langzeitdaten vorliegen, verstecken können.“ [39] Es ergab sich kein Änderungsbedarf für die Methoden.

4.8 Patientenberichtete Endpunkte

In einer der Stellungnahmen wurde angeregt, dass im Abschnitt 3.2.5 der Methoden des Instituts („Patientenberichtete Endpunkte“) ausführlichere Angaben bezüglich der die Validität solcher Endpunkte betreffenden Probleme unter genauer Benennung des Verzerrungspotenzials getroffen werden. Überdies läge eine Inkonsistenz bei der Bewertung lokaler Verfahren zur Behandlung des benignen Prostatasyndroms (Auftrag N04-01 [34]) vor, da die Bewertung nahezu ausschließlich auf patientenberichtete Endpunkte gestützt worden sei, während objektive Endpunkte wie die maximale Harnflussrate als Surrogat unklarer Validität gewertet worden seien.

Im genannten Abschnitt 3.2.5 der Methoden werden allgemeine Aspekte bei der Bewertung von patientenberichteten Endpunkten dargestellt. Dazu gehört auch die Notwendigkeit, dass die verwendeten Instrumente validiert wurden. Darauf wird auch im Abschnitt 3.2.5 des Methodenpapiers verwiesen (Verwendung geeigneter Instrumente). Ist ein Instrument nicht valide bzw. ist seine Validität ungeklärt, bezieht das Institut die Ergebnisse, die auf Basis solcher Instrumente erhoben wurden, in der Regel nicht in die Bewertung ein. Von der Validität abzugrenzen sind Studienaspekte, die das Verzerrungspotenzial beeinflussen, z. B. ein offenes Studiendesign. Dies ist bereits im Abschnitt 3.2.5 der Methoden dargestellt. Die erstgenannte Anregung aus den Stellungnahmen führte daher nicht zu einer Änderung der Methoden des Instituts.

Bezüglich der vermeintlichen Inkonsistenz ist anzumerken, dass das primäre Kriterium für die Eignung eines Endpunktes für die Nutzenbewertung nicht ist, ob er objektiv erhoben werden kann, sondern ob er patientenrelevant ist bzw. patientenrelevante Zielkriterien abbildet. Dies ist für die Harnflussrate nicht der Fall, jedoch für die im Bericht N04-01 verwendeten patientenberichteten Endpunkte [34]. Eine Inkonsistenz liegt demnach nicht vor, und ein Änderungsbedarf für die Methoden des Instituts ergab sich auch aus dieser Anmerkung nicht.

4.9 Bewertung des Schadens

In einer der Stellungnahmen wurde kritisiert, dass das Institut neben dem Nutzen auch den Schaden medizinischer Interventionen bewertet, da die Bewertung des Schadens im Rahmen der Zulassung einer Maßnahme geprüft werde und, im Falle einer Arzneimittelbewertung, die Bewertung des Instituts der Zulassung nicht widersprechen dürfe. Überdies wurde die Verwendung des Begriffs „Schaden“ kritisiert, und zwar sowohl unter Verweis auf SGB V, in dem nur der Begriff „Nutzen“ verwendet werde, als auch unter Verweis auf das Arzneimittelgesetz (AMG), wo der Begriff „Risiko“ anstelle des Begriffs „Schaden“ verwendet werde. Schließlich wurde besonders hervorgehoben, dass das Institut bei der Bewertung des Schadens offenbar andere Ansprüche an die Evidenz stelle und sogar z. B. tierexperimentelle Studien heranziehe.

Zunächst ist anzumerken, dass nicht jede medizinische Maßnahme einem Zulassungsverfahren unterzogen wird und damit der logische Schluss des Stellungnehmenden, dass der Schaden im Rahmen der Zulassung einer Maßnahme geprüft werde, nicht immer richtig ist. Dessen ungeachtet ist es gemäß SGB V auch bei Interventionen, zu denen ein Zulassungsverfahren existiert, Aufgabe des Instituts, auch negative Effekte einer medizinischen Intervention zu bewerten, unabhängig davon, wie diese negativen Effekte benannt werden. So werden z. B in der AM-NutzenV für die frühe Nutzenbewertung von Arzneimitteln unter den patientenrelevanten Endpunkten auch Nebenwirkungen aufgeführt [7]. Dem widerspricht auch die Anmerkung des Stellungnehmenden, dass die im Rahmen der Arzneimittelzulassung geprüften Kriterien von Qualität, Wirksamkeit und Unbedenklichkeit nicht abweichend von der Bewertung der Zulassungsbehörde bewertet werden dürfen, nicht. Zum einen stellt die Feststellung, dass ein Arzneimittel neben einem positiven Effekt (Nutzen) auch einen negativen Effekt (Schaden) hat, die Unbedenklichkeit des Arzneimittels nicht infrage: Es ist allgemein bekannt und üblich, dass Arzneimittel auch Nebenwirkungen (Schaden) verursachen. Zum anderen ist ein wesentlicher Bestandteil der Nutzenbewertung auch der Vergleich verschiedener Therapieoptionen. Das Ergebnis kann dabei sein, dass von 2 grundsätzlich „geeigneten“ (weil mehr Nutzen als Schaden verursachenden) Therapien eine Therapie mehr Schäden als die andere verursacht und im Vergleich ein ungünstigeres Nutzen-Schaden-Verhältnis aufweist. Dies ist eine wichtige Information für Ärztinnen und Ärzte, Patientinnen und Patienten und die Entscheidungsträger und Entscheidungsträgerinnen und damit zwingend Bestandteil einer Nutzenbewertung.

Die Verwendung des Begriffs „Schaden“ (englisch „Harm“) für die Bewertung negativer Effekte medizinischer Interventionen ist international üblich, auch deshalb, weil der Begriff „Risiko“ sprachlich uneindeutig ist, da er zum einen für negative Effekte, zum anderen aber auch für die Wahrscheinlichkeit des Eintritts eines Effekts verwendet wird [5]. Das British Medical Journal (BMJ) hat diesem Thema vor einigen Jahren eine Spezialausgabe gewidmet und darin für die Verwendung des Begriffs „Harm“ anstelle von „Risk“ geworben [51]. Ein weiteres Beispiel ist das aktuelle Projekt der US-amerikanischen HTA-Agentur AHRQ zur Nutzen-Schaden-Abwägung (Methods for Balancing Harms and Benefits in Systematic Reviews) [3].

Abschließend ist anzumerken, dass die Vermutung des Stellungnehmenden, dass das Institut bei der Bewertung des Schadens auch tierexperimentelle Studien heranziehe, auf einem Missverständnis beruht. Im Abschnitt 3.1.3 der Methoden („Ermittlung des Schadens medizinischer Interventionen“) ist beschrieben, dass vor einer Bewertung potenziell relevante unerwünschte Wirkungen, die für die Entscheidungsfindung einen wesentlichen Stellenwert haben können, zusammengestellt werden. Bei der zugehörigen Vorrecherche können dabei auch tierexperimentelle Studien herangezogen werden. In die Bewertung selbst werden solche Studien nicht einbezogen.

Insgesamt ergab sich aus den genannten Stellungnahmen kein inhaltlicher Änderungsbedarf für die Methoden des Instituts. Aufgrund einer entsprechenden Empfehlung des wissenschaftlichen Beirats des Instituts wird die Benennung der Abschnitte 3.1 („Patientenrelevanter medizinischer Nutzen“) und 3.1.1 („Definition des patientenrelevanten medizinischen Nutzens“) jeweils um den Begriff „Schaden“ ergänzt.

4.10 Ableitung von Aussagen zur Beleglage

In einer Stellungnahme wurde das rein textlich beschriebene Vorgehen zur Ableitung von Aussagen zur Beleglage bemängelt.

Zur besseren Strukturierung und Erfassung dieses Vorgehens wurde der betreffende Abschnitt um eine Tabelle erweitert und der Text um zwei Listen ergänzt.

4.11 Ergebnissicherheit

In einer Stellungnahme wurde der erste Satz des Abschnitts 3.2.1 des Methodenpapiers („Ergebnissicherheit“) kritisiert. Es sei nicht richtig, dass jedes Studienergebnis unsicher ist.

Der betreffende Satz wurde entsprechend dem Vorschlag der Stellungnahme um den wesentlichen Punkt der Prüfung auf Ergebnissicherheit ergänzt. Die Inhalte des Abschnitts 3.2.1 des Methodenpapierentwurfs vom 09.03.2011 wurden verschoben in den Abschnitt 3.1.4 („Zusammenfassende Bewertung“) des Methodenpapiers.

4.12 Benachteiligung von Psychotherapiestudien

Im Zusammenhang mit der Adressierung von Problemen bestimmter Designkomponenten in Langzeitstudien aus der bereits oben angesprochenen Stellungnahme (siehe Punkt 4.2) wurde zusätzlich bemerkt, dass durch die fast ausschließlich den RCT-Studien zuerkannte hohe Ergebnissicherheit Psychotherapiestudien benachteiligt werden könnten. Dies wurde in der Stellungnahme allerdings nicht näher begründet.

Das Argument kann nicht ganz nachvollzogen werden, da für die Psychotherapieforschung durch eine große Zahl von randomisierten Studien, die in wiederum zahlreichen Meta-Analysen und systematischen Übersichten eingeschlossen wurden, gezeigt ist, dass auch hier sehr wohl randomisierte Studien durchführbar sind, durchgeführt wurden und durchgeführt werden (z. B. [21,22,26,37,41,42]). Auch der Wissenschaftliche Beirat Psychotherapie nach § 11 PsychThG beschreibt in seinem Methodenpapier auf einer 3-stufigen Skala für RCTs die höchste Ergebnissicherheit. Für eine (noch) ausreichende interne Validität muss zumindest eine „Parallelisierung“ oder „teilweise“ Randomisierung oder „Quasi-Randomisierung“ gegeben sein [55].

Dies kann – wenngleich eventuell mit einer etwas anderen Gewichtung – als durchaus vergleichbar mit dem Methodenpapier des Instituts angesehen werden, das ebenfalls gerade für den nichtmedikamentösen Bereich die Option für nicht randomisierte Studien offenhält. Und dies wurde in der Vergangenheit auch bereits in Berichten des Instituts umgesetzt (z. B. [27-30]). Entsprechend dem Methodenpapier des Wissenschaftlichen Beirats Psychotherapie sind für den Einbezug von nicht randomisierten Studien Qualitätsanforderungen zu stellen, insbesondere bzgl. der Maßnahmen zur Sicherung der Strukturgleichheit. Insgesamt ergibt sich somit aus diesem Aspekt der Stellungnahme kein Änderungsbedarf.

4.13 Bewertung komplexer Erkrankungen und Interventionsstrategien

In einer Stellungnahme wurde darauf hingewiesen, dass dichotome Entscheidungen im Sinne der Ableitung eines Nutzens oder Zusatznutzens bei komplexen Erkrankungen oft nicht getroffen werden können. Darüber hinaus wird darauf verwiesen, dass der Begriff Zusatznutzen im SGB V nur im Rahmen der Bewertung von Arzneimitteln verwendet wird.

In SGB V § 139a Absatz 3 wird dem Institut explizit die Aufgabe der „Bewertung“ des aktuellen medizinischen Wissenstands zu diagnostischen und therapeutischen Verfahren zugewiesen. Diese hat sich an den international anerkannten Standards der evidenzbasierten Medizin zu orientieren (SGB V, § 139a, Absatz 4). Im internationalen Kontext bedeutet dies, dass die Ergebnisse von verschiedenen Interventionen (Prüf- und Vergleichsbehandlung) miteinander verglichen werden und die Qualität und Ergebnissicherheit der entsprechenden Studien bewertet werden. Als Vergleichsbehandlung stehen mehrere Alternativen zur Verfügung: aktive Komparatoren, „keine“ Behandlung oder – wie vielfach in der Onkologie eingesetzt – „Best Supportive Care“. Die Bewertung dieser verschiedenen Vergleiche und die daraus resultierenden Empfehlungen werden vom Institut mit den Begriffen Nutzen bzw. Zusatznutzen operationalisiert. Somit ist das Vorgehen des Instituts rechtskonform mit SGB V § 139a, auch wenn der Begriff „Zusatznutzen“ im Gesetzestext nur explizit im Rahmen der Arzneimittelbewertung verwendet wird. Eine „Dichotomisierung“ der Empfehlung – denn das Institut formuliert Empfehlungen und trifft keine Entscheidungen – ist bereits in der Vergangenheit insbesondere im nichtmedikamentösen Bereich nicht erfolgt, da im Hinblick auf den (Zusatz-)Nutzen bzw. Schaden 3 Abstufungen vorgesehen waren, im Entwurf zur Methodenversion 4.0 sind es sogar 4 Abstufungen. Davon unberührt sind Entscheidungen im Wortsinn immer dichotom. Ein Änderungsbedarf aus diesem Hinweis der Stellungnehmenden besteht somit nicht.

4.14 Surrogate des patientenrelevanten medizinischen Nutzens

In verschiedenen Stellungnahmen wurden die geforderten Kriterien für die Validitätseinschätzung von Surrogatendpunkten als zu streng bemängelt. Die Stellungnehmenden forderten zudem, dass etablierte

Surrogatendpunkte, die u. a. von den Zulassungsbehörden gefordert würden, auch in der Nutzenbewertung zu akzeptieren seien. Insbesondere bei schwerwiegenden oder seltenen Erkrankungen sei die Validierung eines Surrogatendpunkts nach den im Methodenpapier geforderten Methoden nicht oder nur sehr schwer durchführbar.

Laut gesetzlichem Auftrag soll das Institut unter anderem den patientenrelevanten Nutzen von Interventionen untersuchen. Dazu soll es patientenrelevante Endpunkte wie Mortalität, Morbidität oder gesundheitsbezogene Lebensqualität berücksichtigen. Wenn valide Ergebnisse zu solchen Endpunkten (noch) nicht vorliegen (können), kann auf Surrogatendpunkte zurückgegriffen werden. Die Ergebnisse auf Basis solcher Surrogatendpunkte können aber irreführend sein. Es gab in der Vergangenheit zahlreiche Beispiele, bei denen es durch Schlussfolgerungen aufgrund von Surrogatendpunkten zu Schäden bei Patientinnen und Patienten gekommen ist. Ein Beispiel hierfür ist der Endpunkt ventrikuläre Extrasystolen bei der Bewertung von Medikamenten zur Behandlung von Arrhythmien. So hat sich in der CAST-Studie zwar ein Vorteil bezüglich des Auftretens von ventikulären Extrasystolen unter der Behandlung mit Antiarrhythmika gezeigt, auf der anderen Seite stand aber eine erhöhte Sterblichkeit [8,9]. Ein weiteres Beispiel betrifft das orale Antidiabetikum Rosiglitazon, das aufgrund von Effekten hinsichtlich des HbA1c-Wertes zugelassen wurde. Dabei ist zu beachten, dass der HbA1c-Wert ein von den Zulassungsbehörden akzeptierter Endpunkt ist, der auch als Surrogatendpunkt für mikro- und makrovaskuläre Folgekomplikationen bei Diabetes mellitus angesehen wird. Unter Rosiglitazon zeigte sich zwar eine Verringerung des HbA1c-Wertes, auf der anderen Seite traten aber mehr Herzinfarkte auf. Unter anderem deshalb wurde die Zulassung von Rosiglitazon ausgesetzt [14]. Dieses Beispiel zeigt also, dass auch bei allgemein akzeptierten Surrogatendpunkten eine Validierung notwendig ist.

Grundsätzlich ist zu beachten, dass, außer im Falle einer nachgewiesenen Validität, die Aussage zum (Zusatz-)Nutzen hinsichtlich eines patientenrelevanten Endpunkts auf Basis eines Surrogatendpunkts in der Regel mit einer höheren Unsicherheit behaftet ist, als wenn der patientenrelevante Endpunkt direkt erhoben wurde. Weiterhin gilt, dass die Unsicherheit mit sinkender Höhe der Korrelation der Effekte auf den Surrogat- und den zugehörigen patientenrelevanten Endpunkt in der Validierungsstudie zunimmt. Im Rahmen einer Nutzenbewertung muss zunächst anhand der im Methodenpapier beschriebenen Methoden die Validität eines Surrogatendpunkts bewertet werden. Liegt kein Nachweis einer Validität vor, muss mit der erhöhten Unsicherheit insofern umgegangen werden, als die Ergebnisse aus Studien auf Basis eines Surrogatendpunkts ggf. nicht zu einem Beleg für einen (Zusatz-)Nutzen für eine Intervention führen.

Im Übrigen teilt das Institut die Ansicht nicht, dass die Anforderungen an die Validierung zu streng oder gar unerfüllbar seien. Letztlich sind hier sowohl die Hersteller als auch Studiengruppen gefordert,

ihre Daten aus zahlreichen RCTs zu einer Fragestellung in geeigneter Weise und im Sinne einer systematischen Übersicht zusammenzuführen, um daraus Erkenntnisse zur Validierung von Surrogat-Endpunkten zu gewinnen. Entsprechende Protokolle und die gewonnenen Erkenntnisse müssen dann öffentlich zugänglich gemacht werden.

Das Argument, dass gerade bei seltenen Erkrankungen die Validierung sehr schwer durchführbar ist, ist zwar nachvollziehbar, aber die Unsicherheit beim Ziehen von wissenschaftlichen Schlussfolgerungen aufgrund von Surrogatendpunkten ist unabhängig von der Prävalenz oder vom Schweregrad der Erkrankung. Die Aufgabe des Instituts ist es, auch in diesen Situationen die Unsicherheit, die durch die Verwendung von Surrogatendpunkten entsteht, bezogen auf den patientenrelevanten Nutzen zu beschreiben. Insgesamt ergibt sich somit kein Änderungsbedarf am Abschnitt 3.1.2 des Methodenpapiers.

Weiterhin wurde in einer Stellungnahme mit Bezug auf den vom Institut veröffentlichten Rapid Report zur Aussagekraft von Surrogatparametern in der Onkologie [32] kritisiert, dass nicht diskutiert wurde, ob Endpunkte wie die Zeit bis zur Tumorprogression, der als Surrogatendpunkt für den patientenrelevanten Endpunkt Gesamtüberleben dient, von Patientinnen und Patienten sehr wohl als relevant angesehen würden.

Dies ist so im Grundsatz nicht zutreffend. Im o. g. Rapid Report heißt es im Kapitel 1 („Hintergrund“): „Tumorwachstum bzw. die Progression einer onkologischen Erkrankung kann sich auch in einer Symptomatik oder in einer Veränderung der gesundheitsbezogenen Lebensqualität der Patientinnen und Patienten niederschlagen. Die Symptomatik bzw. die gesundheitsbezogene Lebensqualität ist ein direkt patientenrelevantes Maß für die Progression der Erkrankung. So grenzt auch die FDA z. B. den Endpunkt ‚Zeit bis zur Progression von Symptomen‘ als direktes Maß des klinischen Nutzens von der ‚Zeit bis zur Progression des Tumors‘ als potenzielles Surrogat ab [2].“ Die Fragestellung des Rapid Reports war darüber hinaus nicht, was patientenrelevante Endpunkte in der Onkologie sind, sondern vielmehr Empfehlungen für die Validierung von Surrogatendpunkten in der Onkologie darzustellen.

Dennoch ist der Hinweis hilfreich, da dadurch verdeutlicht werden kann, dass der Begriff „Surrogat“ relativ zu verstehen ist. Er bedeutet im Wortsinn zunächst einmal „Ersatz“. Und somit kann durchaus ein patientenrelevanter Endpunkt in diesem Sinn auch ein Surrogat für einen anderen patientenrelevanten Endpunkt darstellen. Um dies klarzustellen, wird ein entsprechender Absatz im Abschnitt 3.1.2 des Methodenpapiers ergänzt.

In einer Stellungnahme wurde eine Differenzierung zwischen intermediären und Proxy-Zielgrößen verlangt.

Im Gegensatz zu intermediären Endpunkten ist zunächst der Begriff der Proxy-Zielgröße nicht allgemein definiert. Sofern hierunter ein Messinstrument für einen Endpunkt in Form eines Konstrukts (z. B. das Messinstrument des ACR [45] für den Endpunkt „Krankheitsaktivität der rheumatoiden Arthritis“) gemeint ist, ist das Messinstrument nicht als Surrogat für den Konstrukt-Endpunkt, sondern als Messinstrument des Endpunktes selbst zu betrachten. Der durch die Proxy-Zielgröße gemessene Effekt besäße nur dann die Eigenschaft eines Surrogates, wenn damit Schlussfolgerungen auf einen anderen Endpunkt gezogen werden, der nicht der eigentliche Konstrukt-Endpunkt ist. Ähnlich ist ein intermediärer Endpunkt nur dann auch Surrogatendpunkt, wenn er verwendet wird, um damit Schlussfolgerungen auf einen anderen, zeitlich später messbaren Endpunkt zu erlauben. Sofern also intermediäre Endpunkte oder Proxy-Zielgrößen als Surrogatendpunkte im obigen Sinne eingesetzt werden, ist eine methodische Differenzierung zwischen ihnen nicht notwendig.

Schließlich wurde in mehreren Stellungnahmen kritisiert, dass die Ausführungen im Abschnitt 3.1.2 des Methodenpapiers zu allgemein gehalten seien und deshalb das Vorgehen konkretisiert werden solle.

Die der Validierung zugrunde liegenden Methoden sind ausführlich in der zitierten Literatur beschrieben. Dennoch wurden die Ausführungen in Abschnitt 3.1.2 des Methodenpapiers an einigen Stellen zur weiteren Verdeutlichung konkretisiert.

4.15 Seltene Erkrankungen

In einer Stellungnahme wurde bemängelt, dass im Abschnitt 3.2.6 („Nutzen in kleinen Populationen“) des Entwurfs für die Version 4.0 des Methodenpapiers keine definitorische Abgrenzung von seltenen, sehr seltenen und extrem seltenen Erkrankungen vorgenommen werde.

International existiert keine einheitliche Definition dessen, was unter einer „seltenen“ Erkrankung zu verstehen ist [54]. Das beginnt schon damit, dass teilweise von „Rare Diseases“ (also tatsächlich im Wortsinn von „selten“), teilweise von „Orphan Diseases“ (eher im Sinne von „verwaisten“ bzw. „vernachlässigten“ Erkrankungen, was nicht zwangsläufig etwas mit der Häufigkeit der Erkrankung zu tun hat) die Rede ist. Entsprechende Angaben reichen von 1 pro 50 Tausend Einwohner (hier allerdings als „Ultra Orphan“ bezeichnet) bis hin zu 7,5 pro 10 Tausend Einwohner [54]. Im Zusammenhang mit „extrem seltenen“ Erkrankungen wird im Methodenpapier auf die zuvor genannte Quelle verwiesen, in der als erste Näherung < 100 Fälle EU-weit vorgeschlagen wird [54]. Der Begriff „selten“ ist bei den zuvor genannten Definitionen sicherlich relativ zu sehen, wenn beispielsweise die häufig bemühte EU-Definition für eine „verwaiste Erkrankung“ (5 pro 10 Tausend Einwohner [10]) bedeutet, dass EU-weit bis zu 230.000 Patientinnen und Patienten an einer bestimmten Erkrankung leiden.

Um dem zuvor diskutierten Umstand Rechnung zu tragen, wird im Methodenpapier ergänzt, dass es keine einheitliche Definition für das Vorliegen einer seltenen Erkrankung gibt. Darüber hinaus wird die (nicht beabsichtigte) Kategorie „sehr selten“ entfernt.

4.16 Bewertung nichtmedikamentöser Verfahren

Zum Abschnitt 3.4 (und auch 3.2) des Methodenpapiers wird in den Stellungnahmen angemerkt, dass zu nichtmedikamentösen Verfahren randomisierte Studien nur sehr erschwert oder gar nicht durchführbar seien. Dies führe zu einer Benachteiligung der nichtmedikamentösen gegenüber den medikamentösen Interventionen. Es wird daher gefordert, präziser festzulegen, wann und wie weit die Berichterstellung auch nicht randomisierte Studien verschiedenster Evidenzstufen berücksichtigt.

Wie in Abschnitt 1.2.6 des Methodenpapiers ausführlich begründet, sind i. d. R. randomisierte Studien notwendig, um eine Kausalbeziehung zwischen einer Intervention und einem klinischen Ergebnis ableiten zu können. Dieses Grundprinzip gilt in gleicher Weise für nichtmedikamentöse und medikamentöse Interventionen, auch wenn in Abschnitt 3.4 des Methodenpapiers auf die Möglichkeit hingewiesen wird, auch nicht randomisierte Studien in die Bewertung einzubeziehen. Da die Gründe für das Einbeziehen nicht randomisierter Evidenz sehr unterschiedlich sein können, ist es nicht möglich, ein für alle Einzelfälle gültiges Vorgehen festzulegen. In diesem Zusammenhang muss auch auf die G-BA-Verfahrensordnung verwiesen werden, die ein solches Vorgehen ebenfalls von individuellen Begründungen abhängig macht.

4.17 Bewertung diagnostischer Verfahren

Eine Stellungnahme fordert, dass in der Bewertung diagnostischer Verfahren (Abschnitt 3.5 des Methodenpapiers) auch der Selbstwert einer diagnostischen Erkenntnis für Arzt und Patient berücksichtigt werden solle, selbst wenn sich aus der diagnostischen Erkenntnis keine Änderung in der Behandlung ergebe. Es wird ausgeführt, dass die Betrachtung von Diagnose und Therapie als gemeinsame Einheit nur in sehr aufwendigen Studien gelingen könne. Auch entspreche die Bewertung von Testgütestudien nicht den internationalen Standards, die z. B. von der Cochrane Diagnostic Test Accuracy Working Group vorgegeben würden.

Dass in der Bewertung diagnostischer Verfahren primär randomisiert-kontrollierte Studien zu fordern und zu berücksichtigen sind, entspricht dem internationalen Standard. Gerade die führenden Wissenschaftler der Cochrane Diagnostic Test Accuracy Working Group (z. B. J. Reitsma und P. Bossuyt) haben hier in den vergangenen Jahren erhebliche methodische Fortschritte erzielt, die auch im Methodenpapier des Instituts Eingang fanden. So schreibt J. Reitsma auf der Website der Cochrane-Gruppe: “Randomized trials of test-treatment combinations provide the best direct

evidence." Auch der GFR-IQWiG-Workshop „Diagnostische Studien im Fokus" am 04.02.2011 zeigte, dass diagnostische Methoden idealerweise als diagnostisch-therapeutische Einheit betrachtet und evaluiert werden sollen. Wenn es in Testgütestudien unmöglich ist, einen Referenzstandard (z. B. Biopsie) zur Überprüfung der Testergebnisse einzusetzen, gibt es diverse Möglichkeiten, dieses Problem im Studiendesign oder in der Analyse der Studie zu lösen [46].

Ob eine diagnostische Information, die keinerlei therapeutische Konsequenz hat, für die Patientin oder den Patienten dennoch einen Nutzen besitzt, ist sehr zweifelhaft. So wird z. B. die Information, an einer unheilbaren Krankheit zu leiden (bzw. nicht an dieser Krankheit zu leiden), die allgemeine Lebensqualität von Patientinnen und Patienten beeinflussen. Hierbei ist aber wichtig, zwischen der gesundheitsbezogenen und der allgemeinen Lebensqualität zu unterscheiden. Wenn eine diagnostische Information lediglich dazu dient, bei der weiteren privaten Lebensplanung (z. B. Familienplanung, Karriere, Urlaub, Hauskauf, Lebensversicherung etc.) Berücksichtigung zu finden, darf bezweifelt werden, dass es sich hierbei um einen medizinischen Nutzen im Sinne des SGB V handelt.

4.18 Einteilung diagnostischer Evidenz nach Fryback und Thornbury

Eine Stellungnahme hält die Einteilung diagnostischer Evidenz nach Fryback und Thornbury [17] für nicht sinnvoll, da diese Einteilung in der G-BA-Verfahrensordnung nicht verwendet wird und ihr zum Teil (nämlich bei Evidenzstufe Ic) widerspreche.

Dass die Einteilung diagnostischer Studien nach Fryback und Thornbury [17] im Methodenpapier des Instituts, nicht aber in der G-BA-Verfahrensordnung genannt wird, bedeutet keinen Widerspruch. Konkret wird die mögliche Rolle „anderer Interventionsstudien" als Evidenzstufe Ic laut G-BA-Verfahrensordnung angesprochen. Diese Evidenzstufe Ic kann sich aber nicht auf diagnostische Interventionsstudien mit schlechter oder fehlender Kontrollgruppe beziehen, weil diese Graduierung der Evidenz dann der Graduierung therapeutischer Studien grob widersprechen würde. Ein systematischer Vergleich verschiedener Stufenschemata zur Bewertung diagnostischer Evidenz hat im Übrigen gezeigt, dass fast alle vorhandenen Schemata im Grundsatz dieselbe Einteilung verwenden [44]. Die Mehrzahl der neueren Schemata beruht sogar direkt auf der ursprünglichen Einteilung von Fryback und Thornbury [17].

4.19 Machbarkeit randomisierter Studien zu diagnostischen Verfahren

Nach Aussage verschiedener Stellungnehmender sind randomisierte Studien zu diagnostischen Verfahren besonders aufwendig, da zum Teil besondere Regularien (z. B. Strahlenschutzverordnung) zu beachten seien. Auch erfordere die Betrachtung der diagnostisch-therapeutischen Einheit hohe Fallzahlen von Patientinnen und Patienten.

Die gesetzlichen Rahmenbedingungen der klinischen Forschung dienen primär dem Schutz der Patientin oder des Patienten. Sie zu bewerten ist nicht Aufgabe des Instituts. Die Literatur zeigt jedoch, dass es durchaus machbar ist, in randomisierten Studien (teilweise sogar monozentrisch) diagnostische Verfahren, wie z. B. die Positronen-Emissionstomografie (PET), zu untersuchen. Im Übrigen wären auch bei reinen Therapiestudien je nach Art der Therapie natürlich Regularien wie die Strahlenschutzverordnung zu beachten.

Die notwendige Patientenzahl randomisierter Diagnostikstudien hängt von mehreren Faktoren ab, nämlich der Testgüte des diagnostischen Verfahrens, dem Anteil testpositiver und testnegativer Patientinnen und Patienten und der Therapieeffektivität in testpositiven bzw. testnegativen Patienten. Wenn nur eine Subgruppe (z. B. die Test-Positiven) von der Therapie profitieren, verringert sich in einer Studie die Patientenzahl für diese Analyse um den Anteil testnegativer Patientinnen und Patienten. Gleichzeitig aber verbessert sich üblicherweise die Therapieeffektivität in dieser Subgruppe, sodass eher geringere Patientenzahlen benötigt werden, wie z. B. Meta-Analysen zu neueren onkologischen Substanzen (sog. „Targeted Therapies“) zeigen [4]. Damit bestimmen zwei gegenläufige Mechanismen die notwendige Fallzahl einer Studie, und es ist nicht erkennbar, dass randomisierte Diagnostikstudien nicht mit moderatem Aufwand durchführbar sind.

4.20 Änderungen im klinischen Management als Endpunkt

Eine allgemeine Stellungnahme merkte an, dass Änderungen im klinischen Management als Folge diagnostischer Interventionen durchaus geeignet seien, den Nutzen der diagnostischen Intervention zu belegen. In einer Stellungnahme zur Evaluation diagnostischer Verfahren wird kritisiert, dass eine beispielhaft zitierte Studie von van Tinteren et al. [52] zur PET bei Lungenkarzinom einen primären Endpunkt – nämlich die Rate unnützer Thorakotomien – verwende, der nur fragliche Patienten-relevanz besitze und daher als Surrogatendpunkt eingestuft werden müsse.

In der van-Tinteren-Studie [52] wurde eine Thorakotomie in verschiedenen Befundkonstellationen als unnütz bewertet, z. B. wenn das Lungenkarzinom so weit fortgeschritten war, dass eine vollständige Resektion unmöglich war. Es ist in der Tat denkbar, dass eine Tumorresektion einen patienten-relevanten Nutzen hat, selbst wenn es nicht gelingt, den Krebs vollständig zu entfernen (R0-Resektion). Die Stellungnahme übersieht aber, dass in der Studie auch die 1-Jahres-Rezidivrate und -Mortalität untersucht wurden. Da sich hier ein tendenzieller Vorteil zugunsten der Interventionsgruppe zeigte (0 vs. 9 lungenkrebsbedingte Todesfälle), kann man davon ausgehen, dass keine der vermiedenen Thorakotomien im Interventionsarm fälschlicherweise als unnötig klassifiziert wurde.

Allgemein können Änderungen im therapeutischen Management allein kaum als Nutzenbeleg verwendet werden, weil im Regelfall unklar bleibt, ob die Änderungen im therapeutischen Management tatsächlich für die Patientin oder den Patienten einen Vorteil i. S. Mortalität, Morbidität oder Lebensqualität bedingen. Oft ist auch das Gegenteil vorstellbar, wie gerade das Beispiel einer nur scheinbar unnötigen Thorakotomie zeigen kann.

4.21 „Zulassungsstatus“ bei diagnostischen Verfahren

Eine Stellungnahme weist darauf hin, dass in der Bewertung diagnostischer (und auch nicht-medikamentöser) Verfahren zu unterscheiden sei zwischen der arzneimittelrechtlichen Zulassung und dem Konformitätsbewertungsverfahren gemäß Medizinproduktegesetz (MPG).

Diagnostische Methoden beinhalten im Regelfall eher Medizinprodukte (z. B. Laboranalytik, Großgeräte zur Bildgebung etc.) als Medikamente (z. B. Augentropfen). Da prinzipiell sowohl die Zulassung gemäß AMG als auch die Konformitätsbewertung mit CE-Zertifizierung gemäß MPG zu beachten sind, werden beide regulatorischen Voraussetzungen sowohl am Ende von Abschnitt 3.4 als auch in Abschnitt 3.5 des Methodenpapiers präzisiert.

4.22 Direkter Nutzen diagnostischer Verfahren

Zu Abschnitt 3.5 des Methodenpapiers wird angemerkt, dass hier die diagnostisch-therapeutische Kette zu stark im Vordergrund stehe. Es gebe auch einen direkten Nutzen, wenn eine neue weniger oder nicht invasive diagnostische Methode (z. B. bildgebende Verfahren oder molekulare Marker) die bisherige invasive Diagnostik ersetze, wobei jedoch gleiche Testgüte vorausgesetzt werde. Hierbei sei es nicht notwendig, die therapeutischen Konsequenzen mit zu betrachten. Auch wird kritisiert, dass die Bewertung der diagnostischen Kette als Gegensatz zu randomisierten Studien dargestellt werde.

Dass diagnostische Verfahren ihren Nutzen nur im Verbund mit einer Therapie entfalten können, wird bereits im 2. Absatz betont: Ein patientenrelevanter Nutzen diagnostischer Verfahren sei „durch die Vermeidung risikobehafteter(er) bzw. komplikationsträchtiger(er) Interventionen oder durch den gezielt(er)en Einsatz von Interventionen“ zu erzielen. Der Begriff der „diagnostischen Kette“ taucht erst dort auf, wo beide Teile der Kette (Diagnostik und Therapie) separat betrachtet werden (Linked Evidence).

Eine grundsätzliche Unterscheidung zwischen der Bewertung einer diagnostisch-therapeutischen Kette und eines diagnostischen Tests alleine ist zum Teil logisch und didaktisch hilfreich, erscheint aber für die Darstellung hier wenig zielführend, weil je nach klinisch-inhaltlichem Hintergrund sowohl die eine als auch die andere Sichtweise dominieren. Um klarer hervorzuheben, dass sich auch aus dem

diagnostischen Test direkte Effekte auf patientenrelevante Zielkriterien ergeben können, wird ein entsprechender Satz eingefügt.

4.23 Randomisierung der Testreihenfolge bei diagnostischen Verfahren

In einer Stellungnahme wird kritisiert, dass im Abschnitt 3.5 des Methodenpapiers zwei Studiendesigns zum Vergleich der Testgüte als gleichwertig dargestellt werden:

1. *Studien mit Durchführung beider Tests in zufälliger Reihenfolge bei allen Patientinnen und Patienten und*
2. *Studien mit zufälliger Zuordnung aller Patientinnen und Patienten zu je einem der beiden Tests.*

Wie in der Stellungnahme treffend ausgeführt wird, erlaubt das erste Studiendesign auch einen direkten intraindividuellen Vergleich der Testergebnisse, sodass auch diskrepante Befunde erkennbar werden. Dagegen kann das zweite Studiendesign nur einen Vergleich der Gesamttestgüte liefern, jedoch zum Informationsgewinn aus der Kombination beider Tests keine Aussage machen.

Im Methodenpapier werden die beiden Studiendesigns jedoch nicht als gleichwertig dargestellt. Es wird lediglich erklärt, dass die beiden Studiendesigns die „höchste Ergebnissicherheit“ erzielen. Auch wenn dem Stellungnehmenden grundsätzlich zuzustimmen ist, hat in der Praxis das Studiendesign mit randomisierter Testreihenfolge oft Probleme in der Verblindung der Tests untereinander, sodass dieses Design nicht in jedem Fall als besser zu bewerten ist.

5 Würdigung der Stellungnahmen zu Kapitel 4

5.1 Rechtliche Vorgaben bzw. Fehlen des gesetzlichen Auftrags des Instituts

In 3 Stellungnahmen wird angemerkt, dass die beschriebenen Methoden nicht den konkreten Aufgaben des Instituts laut den rechtlichen Vorgaben entsprechen würden.

Die Argumentation, die vorliegende ausführliche Beschreibung der Methoden der Leitlinienbewertung und der Versorgungsanalyse würde nicht dem gesetzlichen Auftrag entsprechen, kann hier nicht nachvollzogen werden. Die Aufgaben des Instituts begründen sich auf den § 139a SGB V, der neben der Bewertung von Leitlinien und der Abgabe von Empfehlungen zu DMP im Absatz 3 Nr. 2, eben auch besagt: „Erstellung von wissenschaftlichen Ausarbeitungen, Gutachten, Stellungnahmen zu Fragen der Qualität und Wirtschaftlichkeit der im Rahmen der gesetzlichen Krankenversicherung erbrachten Leistungen unter Berücksichtigung alter-, geschlechts- und lebenslagenspezifischer Besonderheiten“. Dieses entspricht unter anderem der im Kapitel 4 verwendeten Definition einer Versorgungsanalyse, die insbesondere die Fragen zur Versorgungsqualität durch Analyse und Bewertung von Versorgungsaspekten einer definierten Bevölkerungsgruppe zu einer konkreten medizinischen und systembezogenen Fragestellung mit einschließt. Dies spiegelt sich auch im Generalauftrag wider: „Der G-BA geht bei diesem Auftrag davon aus, dass das Institut auf den ihm gemäß § 139a Abs. 3 SGB V übertragenen Arbeitsfeldern nicht nur Einzelaufträge des G-BA bearbeitet, sondern aus der eigenverantwortlichen wissenschaftlichen Arbeit heraus dem G-BA für dessen gesetzliche Aufgaben notwendige Informationen über versorgungsrelevante Entwicklungen in der Medizin zur Verfügung stellt und konkrete Vorschläge für Einzelaufträge erarbeitet, die aus Sicht des Instituts vor dem Hintergrund dieser Informationen relevant sind.“ Versorgungsanalysen wurden bereits vom G-BA beauftragt (z. B. V06-01[31]).

5.2 Fehlender Bezug zu Leitlinien und DMP

Wie schon unter Punkt 5.1 beschrieben, sprechen 2 der eingegangenen Stellungnahmen einen ihrer Meinung nach fehlenden Bezug zu der eigentlichen Aufgabe des Ressorts Versorgungsqualität an. Diese sehen die Stellungnehmenden lediglich in der Bewertung von Leitlinien und der Abgabe von Empfehlungen zu Disease-Management-Programmen.

Die in Kapitel 4 des Methodenpapiers ausführlich beschriebenen Methoden zur Leitlinienbewertung sind jedoch die grundlegenden Methoden, die unter anderem zur Abgabe von Empfehlungen zu DMP verwendet werden. Zum besseren Verständnis wurde ein neuer Abschnitt (4.3) eingefügt, der die Nutzung der unter 4.2. beschriebenen Methoden für Empfehlungen zu Disease-Management-Programmen erläutert.

5.3 Verwendung von AGREE und DELBI

Zwei der Stellungnehmenden merkten an, dass das in Deutschland entwickelte DELBI-Instrument nicht vom Institut verwendet wird, obwohl dieses doch hauptsächlich zur Bewertung von Leitlinien in Deutschland verwendet würde.

Das AGREE-Instrument ist das zurzeit einzig validierte Bewertungsinstrument von Leitlinien. Das Institut verwendet bevorzugt validierte Instrumente zur Bewertung. Das Institut ist an der Weiterentwicklung von DELBI beteiligt. Wünschenswert wäre die Nutzung eines validierten und auf die deutschen Belange abgestimmten Instruments zur Leitlinienbewertung.

5.4 Inhaltliche Leitlinienbewertung

In 3 Stellungnahmen wird angemerkt, dass die inhaltliche Leitlinienbewertung unter praktischen Aspekten zu umfangreich für die zur Verfügung stehenden Ressourcen des Instituts sei. Es wird die Frage aufgeworfen, warum eine methodische Bewertung von Leitlinien nicht ausreiche.

Instrumente zur methodischen Bewertung gibt es (z. B. AGREE, DELBI). Sie prüfen jedoch nicht inhaltlich, ob die existierende relevante Evidenzbasis vollständig identifiziert ist und – entsprechend den Methoden der evidenzbasierten Medizin – adäquat (insbesondere im Hinblick auf die Ergebnissicherheit) bewertet wurde. Das vom Institut entwickelte Manual zur Prüfung der internen Validität von Leitlinienempfehlungen geht daher über die rein methodische Bewertung hinaus. Gegenstand ist nicht die gesamte Leitlinie. Ziel ist vielmehr eine Methode vorzuhalten, mit der bei Bedarf, z. B. bei Nachfrage des G-BA, einzelne Empfehlungen gezielt geprüft werden können.

Eine weitere Stellungnahme befürchtet die Priorisierung der internen vor der externen Validität.

Im Methodenpapier des Instituts wird die interne Validität als eine Basis für die externe Validität gesehen. Eine Analyse der externen Validität ist bislang nicht vorgesehen.

6 Würdigung der Stellungnahmen zu Kapitel 5

6.1 Zielpersonen der Gesundheitsinformation

Ein Stellungnehmender weist darauf hin, dass die Unterscheidung der Zielpersonen der Gesundheitsinformationen in „Patienten und Bürger“ impliziere, dass Patienten keine Bürger seien.

Das Kapitel 5 des Methodenpapiers wurde diesbezüglich überarbeitet.

6.2 Rechtliche Grundlage für die Erstellung von Gesundheitsinformationen

In einer Stellungnahme wird kritisiert, bei einigen vom Ressort Gesundheitsinformation erstellten Patienteninformationen würde teilweise nicht erkennbar, wie sie durch die gesetzlichen Vorgaben nach SGB V § 139a respektive durch den Generalauftrag abgedeckt seien. Beispielhaft sind die Bell’sche Parese, das Töpfchen-Training, der Krampfaderbruch und das PCO-Syndrom aufgeführt.

In § 139a, Abs. 3 Nr. 6 findet sich neben dem Kriterium der „epidemiologischen Bedeutung“ auch der Aspekt „verständliche allgemeine Informationen zur Qualität und Effizienz in der Gesundheitsversorgung“ als weitere Grundlage für die Erstellung von Gesundheitsinformationen durch das Institut. In der Begründung wird zudem weiter ausgeführt: „Um die Bürgerinnen und Bürger über die Erkenntnisse und Arbeitsergebnisse des Instituts zu informieren und deren Autonomie zu stärken, regelt die Vorschrift nach Nummer 6 die Verpflichtung des Instituts, diese gemäß erteilten Aufträgen über Qualität und Effizienz in der ambulanten und stationären Versorgung zu informieren.“

Zudem sei bezüglich des Begriffes „epidemiologische Bedeutung“ auf die entsprechenden Ausführungen in den Methoden verwiesen. Hier wird ausführlich dargestellt, dass sich keine allgemein akzeptierte Definition des Begriffes „Krankheit mit erheblicher epidemiologischer Bedeutung“ findet. Daher werden bei der Auswahl der Themen Faktoren berücksichtigt, die Hinweis auf die Krankheitslast und damit auch auf die epidemiologische Bedeutung geben können. Hierzu gehören unter anderem neben Kennzahlen der Morbidität und Mortalität auch Behandlungskosten, Inanspruchnahme medizinischer Leistungen und Lebensqualität. Auch die Schwere einer Erkrankung, die Krankheitsdauer, eine mögliche Chronifizierung und das Haupterkrankungsalter und identifizierte Informationslücken können Aspekte sein, die bei der Themenauswahl und -priorisierung Berücksichtigung finden.

Zusammenfassend sieht das Institut keinen Anlass für den implizierten Vorwurf, das Institut würde sich bei der Erstellung von Gesundheitsinformationen nicht gesetzeskonform verhalten.

6.3 Doppelung von Informationen

Ein Stellungnehmender kritisiert die Doppelung von Gesundheitsinformationen mit denen anderer Institutionen und empfiehlt, solche Doppelungen oder gar konkurrierende Informationsmaterialien zu vermeiden.

Es bleibt unklar, warum eine Doppelung von Informationen per se als kritisch angesehen wird, zumal sich die Zielgruppe und/oder der inhaltliche Fokus und damit auch die Aufbereitung eines Themas durch verschiedene Anbieter durchaus unterscheiden können. Zudem werden themenspezifisch maßgebliche Institutionen aktiv in die Begutachtung der entsprechenden Information eingebunden.

6.4 Qualitätsindikatoren für Gesundheitsinformationen

Ein Stellungnehmer hält die Aussage, dass es kein Instrument gibt, das sich als zuverlässiger Indikator für die Qualität von Gesundheitsinformationen oder -websites erwiesen hat, für überprüfungswürdig und verweist auf die Kriterien zur Linkpolicy auf dem gemeinsamen Portal Patienteninformation der Bundesärztekammer und der Kassenärztlichen Bundesvereinigung.

Die dort aufgeführten Punkte orientieren sich hauptsächlich an den bekannten formalen Kriterien nach HON und DISCERN und stellen kein Instrument zur umfassenden Beurteilung inhaltlicher Aspekte dar. Da der Stellungnehmer keine weiteren wissenschaftlichen Referenzen zu ausreichend validierten Instrumenten vorlegt, halten wir die Aussage daher weiter für nicht widerlegt.

6.5 Beteiligung von Bürgerinnen und Bürgern

Ein Stellungnehmer kritisiert die mangelnde Umsetzung der Empfehlung des WHO-Gutachtens zur externen Nutzertestung. Die Kriterien und das Vorgehen seien trotz der im Gutachten angemerkten mangelnden Repräsentativität nicht erkennbar aufgearbeitet worden und nicht nachvollziehbar.

Dazu heißt es im WHO-Gutachten auf Seite 22 f: „Anfängliche Bedenken, dass die Patientengruppen zu klein sein könnten, wurden mehr oder weniger aufgelöst. Die Universität führt fast jeden Test mit neuen Patienten durch und versucht gegebenenfalls, Patienten aus speziellen Zielgruppen einzubeziehen. Die Mitarbeiter des Ressorts gehen aktiv auf Patientengruppen zu und pflegen Kontakte, um sie für eine Beteiligung zu gewinnen und sicherzustellen, dass ihre Belange berücksichtigt werden.“ Weiter heißt es unter Empfehlungen: „Das Institut sollte sich weiterhin gemeinsam mit der Patientenuniversität Hannover darum bemühen, dass die Testpersonen die Zielpopulation so gut wie möglich repräsentieren, …“

Wie sich aus diesen Ausführungen der Gutachterinnen und Gutachter oben genannte Kritik ableiten lässt, bleibt unklar. Letztlich stellt sich hier auch die Frage, wie man in diesem Kontext als Gegensatz

zu mangelnder Repräsentativität denn ausreichende oder gar vollständige Repräsentativität operationalisieren will und sich diese in Folge erreichen lässt.

6.6 Identifikation kultureller Unterschiede

Ein Stellungnehmer fordert, die Entscheidungswege zur Identifikation kultureller Unterschiede auszuführen.

Es bleibt unklar, warum gerade dieser Begriff aus seinem Kontext gelöst wurde. Im Originaltext wird ausgeführt, dass ein Ziel bei der Erstellung unserer Informationen ist, „Sensibilität und Respekt vor dem Wissen, den Wertvorstellungen und Sorgen der Nutzerinnen und Nutzer, vor ihrer Autonomie, ihren kulturellen Unterschieden sowie gegenüber geschlechts-, alters- und behindertenspezifischen Belangen zu zeigen“. Wie das Ressort versucht, diese Aspekte zu erfassen, ist unseres Erachtens in Kapitel 5 ausführlich beschrieben. Sollten sich beispielsweise im Rahmen der qualitativen Forschung oder der Einbindung von Betroffenen in den Erstellungsprozess Hinweise auf unterschiedliche kulturelle Aspekte möglicher Nutzerinnen und Nutzer einer Information zeigen, werden wir versuchen, diese in der finalen Version adäquat zu berücksichtigen. Wir behaupten nicht, dass dies bei der großen Zahl und Diversität der potenziellen Leserinnen und Leser immer vollumfänglich gelingen kann.

6.7 Bewertungsinstrumente

Ein Stellungnehmer schlägt vor, neben dem Oxman-Guyatt-Score auch die Bewertungsskala nach Leichsenring und Rüger[43] anzuwenden, um auch nicht streng kontrollierte oder naturalistische Studien in ihrer Evidenz zu beurteilen.

Bei der Bewertungsskala nach Leichsenring und Rüger [43] handelt es sich um ein Instrument zur Beurteilung von Primärstudien und nicht von systematischen Übersichten. Letztere bilden jedoch die Basis für die Gesundheitsinformationen des Instituts. Der Vorschlag bietet damit methodisch bedingt keine verwendbare Ergänzung für die Arbeit im Ressort.

6.8 Berücksichtigung herstellergesponserter Studien

Ein Stellungnehmer kritisiert die fehlende Berücksichtigung herstellergesponserter Studien.

Das Ressort Gesundheitsinformation hat die entsprechenden systematischen Reviews wegen möglicher Verzerrungen bisher nicht berücksichtigt. In Zukunft werden diese Reviews nicht per se ausgeschlossen, vielmehr ist die Überprüfung eines möglichen Funding Bias zukünftig ein Aspekt der Bewertung des Verzerrungspotenzials einer systematischen Übersicht. Bei fehlendem Hinweis auf einen entsprechenden Bias wird das Review nicht grundsätzlich ausgeschlossen.

7 Würdigung der Stellungnahmen zu Kapitel 6

7.1 Berücksichtigung unpublizierter Daten

In Kapitel 6 zur Informationsbeschaffung wird beschrieben, dass Daten, die dem Institut übermittelt werden, aber nicht publiziert werden dürfen, nicht inhaltlich in die Bewertung des Instituts einfließen können, da dies dem Transparenzgebot widerspricht.

Eine Stellungnahme sieht darin ein Fehlverständnis des Instituts, da kein Normkonflikt zwischen dem Transparenzgebot in § 139a und Bewertungsentscheidungen existiere. Aus Sicht der Stellungnehmenden sind auch Daten, in deren Publikation nicht eingewilligt wurde, auf Basis von § 20 Abs. 2 SGB X zu berücksichtigen.

Der genannte Paragraf ist für das Institut nicht einschlägig, da er Untersuchungsgrundsätze für Behörden regelt. Die Notwendigkeit der Transparenz von Informationen aus klinischen Studien (Studienmethodik und Studienergebnisse) ist international unumstritten. Entsprechende gesetzliche Regelungen wurden in den USA [16], in Europa [11,12] und in Deutschland [1 §42b,2] verabschiedet. Die Forderung der Stellungnehmenden, Daten aus klinischen Studien in die Bewertung einzubeziehen, trotzdem aber als Betriebs- und Geschäftsgeheimnisse geheim zu halten, bleibt damit weit hinter internationalen, aber auch deutschen Standards zurück.

7.2 Vollständigkeit von Primärstudien

In Kapitel 6 zur Informationsbeschaffung steht, dass bei einer Nutzenbewertung auf Grundlage systematischer Übersichten keine Vollständigkeit im Sinne einer Berücksichtigung aller verfügbaren Primärstudien angestrebt wird.

Eine Stellungnahme geht davon aus, dass dies nicht für die Nutzenbewertungen von Arzneimitteln gilt, da das Institut die vollständige Einreichung von Studien von den Firmen verlangt.

Eine Nutzenbewertung von Arzneimitteln kann – sofern bestimmte Voraussetzungen erfüllt sind – auf Grundlage systematischer Übersichten erfolgen. Dies ist nicht der Fall, wenn eine relevante Menge unpublizierter Daten zu erwarten ist. In der finalen Version der Allgemeinen Methoden 4.0 wird noch einmal etwas deutlicher formuliert, wann eine Nutzenbewertung nicht auf Grundlage von systematischen Übersichten durchgeführt werden kann.

7.3 Systematische Recherche nach Leitlinien

In Kapitel 6 zur Informationsbeschaffung wird beschrieben, dass innerhalb der Nutzenbewertung Leitlinien als Informationsquelle nicht grundsätzlich ausgeschlossen werden. Es erfolgt jedoch i. d. R. keine systematische Recherche nach Leitlinien.

Eine Stellungnahme sieht durch einen Verzicht auf eine systematische Recherche nach Leitlinien eine wichtige Evidenzquelle außer Acht gelassen.

Bei der Nutzenbewertung auf Grundlage von Primärstudien werden systematische Übersichten recherchiert, um diese nach weiteren relevanten Primärpublikationen zu durchsuchen. Eine Vollständigkeit der systematischen Übersichten wird daher nicht angestrebt und macht deshalb auch eine gezielte Suche nach Leitlinien nicht notwendig.

Das von den Stellungnehmenden zitierte Gutachten des Sachverständigenrates [47] weist explizit darauf hin, dass ein Großteil der derzeit vorliegenden ausländischen und deutschen Leitlinien nicht den international geforderten Qualitätskriterien einer evidenzbasierten Leitlinienentwicklung entspricht. Es mag sein, dass die Erstellung von S-3-Leitlinien auf der Grundlage von systematischen Übersichten erfolgt. Jedoch genügt die Dokumentation des Vorgehens häufig nicht den Anforderungen für den Einschluss in einen Bericht des Instituts. So wurde für das PET-Sammelprojekt (D06-01E-K) in sieben Indikationen eine systematische Recherche nach Leitlinien durchgeführt. Dabei konnte keine Leitlinie identifiziert werden, die den Kriterien einer hochwertigen systematischen Übersicht entsprochen hätte.

Findet eine Nutzenbewertung auf Grundlage von systematischen Übersichten statt, wird deshalb über eine zusätzliche gezielte Suche nach Leitlinien projektspezifisch entschieden. Sollte im Zuge der Anhörung eine hochwertige Leitlinie genannt werden, die den Einschlusskriterien des Berichtsplans entspricht, wird diese berücksichtigt.

8 Würdigung der Stellungnahmen zu Kapitel 7

8.1 Einschluss von Studien / Bedeutung der Zulassung

In einer der Stellungnahmen wurde gefordert, dass im Abschnitt 7.1.1 des Methodenpapiers die Kriterien zum Einschluss von Studien, die die Einschlusskriterien nicht vollständig erfüllt haben, dahin gehend ergänzt werden, dass Studien, auf denen eine Zulassung oder Zulassungserweiterung basiert, immer in die Nutzenbewertung eingeschlossen werden. Des Weiteren wurde angeregt, dass im Abschnitt 7.1.1 aus Konsistenzgründen ein Hinweis aus Abschnitt 3.3.1 der Methoden („Stellenwert des Zulassungsstatus") aufgenommen wird, dass Studien, die das Einschlusskriterium „Population" nicht vollständig erfüllen, auch dann aufgenommen werden können, wenn hinreichend sicher plausibel ist, dass die Effektschätzer patientenrelevanter Endpunkte nicht wesentlich durch das betreffende Merkmal (nicht erfülltes Einschlusskriterium) beeinflusst werden.

Ob eine Studie in eine Nutzenbewertung eingeschlossen wird, hängt nicht davon ab, ob sie zur Zulassung geführt hat, sondern ob sie die Fragestellung der Nutzenbewertung beantwortet. Beispielsweise könnte die Zulassung mit erheblichen Einschränkungen verbunden worden sein (z. B. Zweitlinientherapie), sodass ein Teil der Zulassungsstudien größtenteils oder sogar vollständig außerhalb des erteilten Zulassungsstatus durchgeführt wurde (z. B. zur Erstlinientherapie). Bei solchen Studien kann dann nicht regelhaft davon ausgegangen werden, dass sie zuverlässige Aussagen über die Fragestellung der Nutzenbewertung (Anwendung des Arzneimittels gemäß Zulassung) erlauben. Es ist daher nicht sinnvoll, Zulassungsstudien grundsätzlich in die Nutzenbewertung einzuschließen. Es ist allerdings möglich, dass das nicht erfüllte Einschlusskriterium ein Merkmal darstellt, das die Effekte patientenrelevanter Endpunkte nicht beeinflusst. Im Abschnitt 3.3.1 der Methoden ist für solche Fälle vorgesehen, dass diese Studien, die aufgrund dieses Merkmals ganz oder teilweise außerhalb der Zulassung durchgeführt wurden, in die Nutzenbewertung eingeschlossen werden. Wie vom Stellungnehmenden vorgeschlagen, ist es sinnvoll, diese Möglichkeit für das Merkmal „Population" unabhängig von der Frage der Zulassung zu eröffnen und im Abschnitt 7.1.1 zu ergänzen. Die Methoden wurden entsprechend geändert.

8.2 Aspekte der Bewertung des Verzerrungspotenzials

In einer Stellungnahme wird darauf hingewiesen, dass der in Abschnitt 7.1.4 verwendete Begriff „Behandler" durch seine Verwendung im Dritten Reich als Diskriminierung jüdischer Ärzte belastet sei.

Der Hinweis ist berechtigt. Der Begriff wird durch „behandelnde Personen" ersetzt.

Eine Stellungnahme wirft die Frage auf, warum das Institut nur eine zweistufige anstelle einer mehrstufigen Bewertung des Verzerrungspotenzials entsprechend der Methodik der Cochrane Collaboration durchführt.

Das Institut verwendet in der Tat zwei Stufen (niedrig, hoch). Bei der von der Cochrane Collaboration vorgeschlagenen Methodik von einem „mehrstufigen" Verfahren zu sprechen, ist irreführend. Auch im Cochrane Handbook finden sich primär zwei Ausprägungen, nämlich „High" und „Low Risk of Bias". Dort ist lediglich beschrieben, dass bei fehlenden Informationen zur Bewertung des Verzerrungspotenzials die Ausprägung „Unclear" zu verwenden ist. Leider ist im Cochrane Handbook nicht klar beschrieben, wie diese Ausprägung bei der Interpretation der Ergebnisse und bei Analysen zu berücksichtigen ist. Es findet sich die Angabe, dass in Analysen Studien mit „Low" und „Unclear Risk of Bias" nicht kombiniert werden sollten. Das deckt sich mit der Verfahrensweise des Instituts, da bei fehlenden Informationen zur Bewertung des Verzerrungspotenzials von einem hohen Verzerrungspotenzial ausgegangen wird.

Eine Stellungnahme fordert die Streichung des Satzes im Abschnitt 7.1.4, dass für nicht randomisierte vergleichende Studien in der Regel keine zusammenfassende Bewertung der Verzerrungsaspekte durchgeführt wird. Es widerspräche der Verfahrensordnung des G-BA, die auch Meta-Analysen von „Nicht-RCTs" vorsieht.

Hier liegt offensichtlich ein Missverständnis vor, da in diesem Satz nicht die Zusammenfassung einzelner Studienergebnisse, sondern die Zusammenfassung einzelner Verzerrungsaspekte zu einer Gesamtbewertung des Verzerrungspotenzials adressiert wird.

Es wird in einer Stellungnahme gewünscht, die Methoden zur Minimierung des Verzerrungspotenzials bei nicht randomisierten Studien zu ergänzen.

Die grundsätzlichen Punkte zur Vermeidung / Minimierung von Verzerrungen sind im Abschnitt 7.1.4 des Methodenpapiers sowohl für randomisierte als auch für nicht randomisierte Studien aufgeführt. Des Weiteren werden in diesem Abschnitt diverse Richtlinien zu diesem Thema zitiert. Es ergibt sich somit kein Änderungsbedarf.

8.3 Rückfragen bei Autorinnen und Autoren

Eine Anmerkung bez. Abschnitt 7.1 besagt, dass Rückfragen bei Autorinnen und Autoren nicht explizit erwähnt werden.

Dabei wurde offensichtlich übersehen, dass genau diesem Punkt ein eigener Abschnitt (Abschnitt 7.3.10 des Methodenpapiers) gewidmet ist. Eine Änderung der Methoden ist deshalb nicht erforderlich.

8.4 Nutzenbewertung auf Basis systematischer Übersichten

Eine Stellungnahme kritisiert die regelhafte Anforderung des Instituts, dass als Grundlage für eine Nutzenbewertung mindestens zwei systematische Übersichten vorliegen sollten, da eine hochwertige systematische Übersicht bereits „das höchste Evidenzniveau" bedeute.

Zum einen wird diese Anforderung im betreffenden Abschnitt begründet. Nur so lässt sich nämlich die Konsistenz der Ergebnisse der systematischen Übersichten überprüfen. Zum anderen wird der Fall des Vorliegens lediglich einer systematischen Übersicht explizit beschrieben. Es ist durchaus mit Begründung möglich, zur Nutzenbewertung eine einzelne systematische Übersicht heranzuziehen. Eine solche Begründung kann z. B. genau darin liegen, dass diese systematische Übersicht von besonderer Güte ist.

Des Weiteren wird angemerkt, dass der in diesem Abschnitt verwendete Begriff „Sekundärliteratur" missverständlich sei.

Die Anmerkung wurde aufgegriffen und dieser Begriff durch „systematische Übersicht" ersetzt.

Darüber hinaus wird gefordert, die „wichtigsten Kernelemente" einer systematischen Übersicht aufzuführen.

Diese Forderung verwundert etwas, da die in der Stellungnahme genannten Merkmale alle im vorherigen Abschnitt 7.2.1 genannt sind. Darüber hinaus wird in diesem Abschnitt explizit auf zwei Bewertungsinstrumente (Oxman-Guyatt-Index, AMSTAR-Instrument) hingewiesen.

In einer weiteren Stellungnahme wird die in Abschnitt 7.2.2 formulierte Forderung, dass für den Einbezug systematischer Übersichten qualitativ hochwertige Primärstudien enthalten sein müssen, hinterfragt.

Diese Forderung ist in der Tat falsch. Systematische Übersichten, die transparent und nachvollziehbar darstellen, dass keine qualitativ hochwertigen Primärstudien existieren, können sehr wohl zur Nutzenbewertung herangezogen werden. Die betreffende Textpassage wurde korrigiert.

8.5 Beurteilung statistischer Signifikanz

In einer Stellungnahme wird gefragt, wie der Diskurs zum in Abschnitt 7.3.2 des Methodenpapiers formulierten Sachverhalt „Beide Aspekte – ob eine Hypothese ein- oder zweiseitig zu formulieren ist und ob für multiples Testen adjustiert werden muss – werden in der wissenschaftlichen Literatur immer wieder kontrovers diskutiert" aussehe.

Zur Klärung wurden zwei Artikel zitiert.

8.6 Beurteilung klinischer Relevanz

In einer Stellungnahme werden die generellen Erläuterungen zur Relevanzproblematik in Kapitel 7.3.3. hinterfragt, speziell die Unterscheidung einer Systemebene von einer individuellen Ebene sowie die Verbindung mit ökonomischen Aspekten.

Die Hinweise bzgl. Unklarheiten sind nachvollziehbar. Die Erläuterungen wurden überarbeitet.

In einer Stellungnahme wird darauf hingewiesen, dass die Bewertung der Relevanz im Sinne des Ausmaßes eines Effektes eine zusätzliche Hürde darstelle, die zur Folge habe, dass in Indikationsgebieten, bei denen Fortschritte nur in vielen kleinen Schritten erzielt werden können (z. B. in der Onkologie), längerfristig der Status quo zementiert wird, unabhängig davon, ob diese Fortschritte kosteneffizient sind oder nicht.

Die internationale wissenschaftliche Diskussion beklagt seit vielen Jahren, dass bei der Interpretation von Studienergebnissen die Bewertung der Relevanz zu kurz kommt. Dies gilt umso mehr, wenn durch die Zusammenfassung von Studien mit resultierenden großen Patientenzahlen auch sehr kleine Effekte statistisch signifikant werden können. Das Institut sieht es vor diesem Hintergrund als seine Aufgabe an, bei seinen Empfehlungen an den Auftraggeber auch das Ausmaß eines Nutzens oder Zusatznutzens zu bewerten. Dies steht in Einklang mit gesetzlichen Regelungen, ganz explizit durch Regelungen im AMNOG inkl. der zugehörigen Rechtsverordnung. Natürlich kann die Bewertung der Relevanz dazu führen (nach gesetzgeberischer Zielsetzung soll sie offenbar dazu führen) festzustellen, dass die Verbesserung einer Therapie in einem sehr kleinen Schritt keinen derartigen therapierelevanten (Zusatz-)Nutzen darstellt, der eine breite Anwendung und eine solidarische Finanzierung rechtfertigt.

In der Stellungnahme wird weiter ausgeführt, dass solche Therapien für bestimmte Patientengruppen wertvolle Therapiealternativen darstellen, wenn z. B. Alternativen versagt haben oder nicht verträglich waren.

Diese Stellungnahme kann nur so verstanden werden, dass es für eine neue Therapie zwar insgesamt nur eine geringe Verbesserung geben könnte, für bestimmte (Sub-)Gruppen jedoch einen Zusatznutzen. Dies ist richtig, hat aber selbstverständlich die Konsequenz, den Zusatznutzen in bestimmten Patientengruppen zu belegen. In den Methoden des Instituts ist eine solche Situation abgebildet.

In den Stellungnahmen nimmt die Kritik an der auf 0,2 Standardabweichungen festgelegten Irrelevanzschwelle für die standardisierte Mittelwertdifferenz im Falle des Fehlens validierter bzw. etablierter Kriterien zur Relevanzbewertung einen großen Raum ein. Eine solche fixe Schwelle sei willkürlich und ohne Begründung gewählt. In die Relevanzbewertung flössen Werturteile ein, die nicht

durch einen rein statistischen Ansatz zu lösen seien. Die Anwendung einer fixen Schwelle berücksichtige die spezifischen klinischen Situationen nicht. Daher sei eine solche Schwelle kontextabhängig festzulegen.

Grundsätzlich ist festzustellen, dass in den Stellungnahmen zu diesem Thema zwar umfangreiche Kritik an dem beschriebenen Vorgehen geübt wird, mögliche konkrete Lösungsansätze neben der allgemeinen Forderung, kontextabhängige Gegebenheiten zu berücksichtigen, jedoch nicht formuliert wurden.

In den Stellungnahmen bleibt weitgehend unberücksichtigt, dass in der betreffenden Textpassage zur Relevanzbewertung verschiedene Ansätze zur Ableitung eines relevanten Effekts formuliert sind. Der Ansatz mit Verwendung einer auf 0,2 Standardabweichungen festgelegten Irrelevanzschwelle ist lediglich als letzte Möglichkeit aufgeführt. Falls validierte bzw. etablierte Kriterien vorliegen, werden diese zur Relevanzbewertung herangezogen. Dies können auf der betrachteten Skala vorliegende Schwellen für Gruppenunterschiede oder auch Responsekriterien für individuelle Veränderungen sein. Damit ist sichergestellt, dass bei Vorliegen von kontextspezifischen Informationen der auf der festen Schwelle beruhende Ansatz nicht zur Anwendung kommt. Lediglich in Situationen ohne weitere Informationen stellt dieser Ansatz aus Sicht des Instituts eine weitere Alternative dar, relevante Unterschiede aufzudecken. Die Implementierung auf unterster Ebene stellt somit eine Erweiterung und keine Beschränkung dar.

Da das hierarchische Vorgehen des Instituts im betreffenden Abschnitt offensichtlich nicht verständlich formuliert war, wurde der betreffende Text überarbeitet. Insbesondere die verschiedenen Hierarchiestufen wurden deutlicher hervorgehoben. Darüber hinaus wurden die möglichen Ansätze zur Relevanzbewertung erweitert. Liegen zwar keine skalenspezifischen Irrelevanzschwellen vor, jedoch validierte, etablierte oder anderweitig gut begründete Relevanzschwellen (z. B. aus Fallzahlplanungen), werden diese Relevanzschwellen zur Ableitung von Irrelevanzschwellen herangezogen.

In mehreren Stellungnahmen wird kritisiert, dass das beschriebene Vorgehen international nicht empfohlen wird.

Ein klarer Standard zur Relevanzbewertung existiert bisher weder international noch national. Dies wird auch dadurch deutlich, dass sich in den Stellungnahmen keine Verweise auf solche Standards finden.

Eine Stellungnahme kritisiert mit Verweis auf das von Röhmel verfasste und der Stellungnahme beigefügte Diskussionspapier, dass sowohl die Anwendung einer verschobenen Nullhypothese anstelle

des Vergleichs des Effektschätzers mit der Irrelevanzschwelle als auch die Schwelle 0,2 „wissenschaftlich nicht begründbar" sei.

Bei der Aussage zur fehlenden wissenschaftlichen Begründung der Anwendung einer verschobenen Nullhypothese auf Röhmel zu verweisen, ist bemerkenswert, da dieser Ansatz von Röhmel selbst und anderen in der Vergangenheit unterbreitet wurde. Im Indikationsgebiet der pAVK wurde seinerzeit von Heidrich et al. [23,24] vorgeschlagen, eine verschobene Nullhypothese zum Nachweis der Relevanz zu testen. Stattdessen zu fordern, dass lediglich der Effektschätzer die Irrelevanzschwelle überschreitet, missachtet selbst die Ausführungen im Diskussionspapier von Röhmel. Hierbei wird übersehen, dass eine Irrelevanzschwelle nicht einer Relevanzschwelle gleichzusetzen ist [53], was im Diskussionspapier von Röhmel auch klar beschrieben ist. Doch selbst wenn der Effektschätzer mit der Relevanzschwelle verglichen würde, um Relevanz zu attestieren, besteht das Problem, dass die Wahrscheinlichkeit für den Fehler 1. Art nicht kontrolliert werden kann. Darüber hinaus ist dieser Ansatz weniger effizient, da mit dessen Anwendung im Falle eines geschätzten Effekts zwischen Irrelevanz- und Relevanzschwelle kein Relevanznachweis möglich ist. Dem gegenüber lässt der Ansatz mit der verschobenen Hypothese auch dann bei entsprechender Präzision (schmales Konfidenzintervall) die Ableitung einer Relevanz zu.

Die Kritik von Röhmel richtet sich auch gar nicht an die prinzipielle Verwendung von verschobenen Hypothesen. Röhmel schreibt vielmehr, man könne „der Annahme der Gruppe [des Instituts] zustimmen, dass es Sinn macht, von einer bekannten Relevanzschranke δ_r auszugehen", um eine Irrelevanzschwelle als verschobene Hypothesengrenze abzuleiten. Röhmels Kritik bezieht sich ausschließlich auf die Quantität der Schwelle 0,2. Er führt an, dass bei üblichen Annahmen zur Fallzahlplanung in klinischen Studien die Power für den Nachweis der Relevanz anhand der Schwelle 0,2 bei einer Studie nur 50 % beträgt. Für eine übliche Power von 90 % müsste die Fallzahl um den Faktor 2,7 erhöht werden. Diese Berechnungen sind richtig. Es ist jedoch prinzipienbedingt klar, dass der Nachweis einer „höheren" Hürde (Relevanz) nicht ohne Inkaufnahme einer erhöhten Fallzahl gegenüber dem „einfacheren" Nachweis eines Unterschieds einhergehen kann. Der Schlussfolgerung, dass der Mehraufwand „zu hoch" sei, folgt das Institut nicht. In Analogie zu den Berechnungen von Röhmel ist die Power zum Nachweis der Relevanz, falls zwei Studien vorliegen, bereits etwa 80 %. Hierbei ist auch zu beachten, dass für Fragestellungen des Instituts häufig mehrere Studien vorliegen, die metaanalytisch zusammengefasst werden können.

In einer Stellungnahme wird gefordert, Responderanalysen gegenüber der Verwendung von Effektgrößen zur Bewertung der Relevanz primär heranzuziehen.

Sowohl in der ursprünglichen Fassung, auf die sich die Stellungnahme bezieht, als auch in der aktuellen überarbeiteten Fassung waren bzw. werden Responderanalysen gegenüber Ergebnissen zu standardisierten Effektgrößen vorgezogen.

In zwei Stellungnahmen wird kritisiert, dass durch die Verwendung einer Irrelevanzgrenze von 0,2 Standardabweichungen bei einem „kleinen, aber nachweisbaren Effekt“ von 0,2 Standardabweichungen niemals ein Nutzenbeleg erbracht werden kann, da die untere Grenze des Konfidenzintervalls auch bei großen Studien immer kleiner als 0,2 sein muss.

Dieses Argument ist mehrfach irreführend. Auch aus solchen Studien und bei solchen Effektgrößen kann ein Nutzen abgeleitet werden, z. B. durch entsprechende Responderanalysen. Es ist zu beachten, dass die Irrelevanzgrenze von 0,2 Standardabweichungen nur auf der letzten Stufe der Hierarchie der Relevanzbewertung herangezogen wird. Es wird bei diesem Argument auch noch ein weiterer Sachverhalt ignoriert. Eine Irrelevanzgrenze von 0,2 bedeutet nicht, dass Effekte oberhalb von 0,2 zwingend als relevant anzusehen sind. Dazu muss ein Effekt oberhalb der Relevanzgrenze liegen. Dass ein Graubereich zwischen Irrelevanzgrenze und Relevanzgrenze existiert [53], ist auch außerhalb des Instituts anerkannt und wurde z. B. im Diskussionspapier von Röhmel klar so dargestellt. Die Unterscheidung zwischen Irrelevanz- und Relevanzgrenze ist in diesem Zusammenhang wichtig (siehe auch folgenden Absatz).

Eine Stellungnahme weist darauf hin, dass das in Abschnitt 7.3.3 formulierte Kriterium (das Konfidenzintervall liegt vollständig oberhalb der Irrelevanzgrenze) noch nicht dem Vorliegen eines relevanten Effekts entspräche, sondern eines nicht sicher irrelevanten Effekts. Dadurch entstünde eine nicht definierte Grauzone.

In der Tat kann in diesem Fall (wie auch im betreffenden Absatz beschrieben) nur davon ausgegangen werden, dass die Effektstärke nicht in einem sicher irrelevanten Bereich liegt. Das Institut wertet dies als ausreichend für den Nachweis eines relevanten Effekts, da sich in diesem Fall die beobachteten Effekte in der Regel deutlich oberhalb der Irrelevanzschwelle realisieren. Zur Klarstellung wurde der betreffende Absatz ergänzt.

8.7 Bewertung subjektiver Endpunkte bei offenen Studiendesigns

In verschiedenen Stellungnahmen wird die in Abschnitt 7.3.4 beschriebene Methodik zum Umgang mit Verzerrungspotenzialen, die durch die Erhebung subjektiver Endpunkte in offenen Studien entstehen, kritisiert. Die Kritik bezieht sich auf die beschriebene Möglichkeit, durch eine adjustierte Entscheidungsgrenze im Sinne einer verschobenen Nullhypothese dem Verzerrungspotenzial zu begegnen. Eine Stellungnahme merkt an, die Ausführungen suggerierten den Versuch, ein qualitatives Problem quantitativ zu lösen.

Das Institut fasst die in diesem Abschnitt beschriebene mögliche Verzerrung nicht als qualitatives Problem auf. Es ist empirisch belegt, dass Effekte bezüglich subjektiver Endpunkte bei offenen Studien im Mittel größer ausfallen als in Studien mit Verblindung [57]. Dies bedeutet eine auf Verzerrung beruhende quantitative Verschiebung der Effektschätzer, der aus Sicht des Instituts auch quantitativ begegnet werden kann.

Eine weitere Stellungnahme sieht ein Risiko für eine zu variable Berücksichtigung von Studien, da die Beschreibung der Methodik nur vage sei.

Es gibt zwei Gründe dafür, dass der Umgang mit diesbezüglichen Verzerrungspotenzialen in diesem Abschnitt nur in Ansätzen beschrieben ist. Zum einen existiert bisher kein allgemein akzeptierter oder etablierter Umgang mit diesem Problem. Bisher wird dieses Problem im Rahmen von systematischen Übersichten entweder schlicht ignoriert oder aus den Ergebnissen keine belastbaren Aussagen abgeleitet. Die Anwendung einer adjustierten Entscheidungsgrenze wird in diesem Abschnitt als „eine Möglichkeit“ genannt, dem Problem jenseits dieser Extreme zu begegnen. Erst dadurch wird es möglich, belastbare Aussagen zu treffen, ohne das Problem zu ignorieren. Zum anderen ist die Wahl der Grenze im Kontext, d. h. projektspezifisch,zu wählen. Daher ist die konkrete Methodik so weit als möglich und sinnvoll in den jeweiligen projektspezifischen Berichtsplänen zu verorten. Das Risiko einer zu variablen Berücksichtigung von Studien wird vom Institut nicht gesehen. Der Ansatz einer adjustierten Entscheidungsgrenze ermöglicht überhaupt erst einen transparenten und operationalisierten Umgang mit dem Problem.

Eine weitere Stellungnahme kritisiert die Verwendung einer festen Entscheidungsgrenze und verweist auf „multivariate Verfahren“ zur „Bereinigung von Effektschätzern“ bzw. „Bayesianische Verfahren zur Effektschätzer-Modifikation (Shrinkage)“, ohne jedoch hierzu auf Literatur zu verweisen. Die Stellungnehmenden weisen selbst darauf hin, dass diese Ansätze in der Praxis noch nicht erprobt sind bzw. für deren Anwendung benötigte Parameter bisher fehlen.

Wie oben bereits geschrieben existiert bisher kein etabliertes Verfahren zur angemessenen Berücksichtigung des hier betrachteten Verzerrungspotenzials. Daher wird die Anwendung einer adjustierten Entscheidungsgrenze im Methodenpapier auch nur als „eine Möglichkeit“ genannt, eine „feste Entscheidungsgrenze“ schon gar nicht vorgegeben. Auch das Institut sieht hier weiteren Forschungsbedarf, weist jedoch darauf hin, dass die Anwendung einer adjustierten Entscheidungsgrenze nicht dazu führt, dass Evidenz unberücksichtigt bleibt. Vielmehr ermöglicht dieser Ansatz unter der Prämisse, dass das Verzerrungspotenzial mitberücksichtigt wird, überhaupt erst die Ableitung von Nutzenaussagen.

8.8 Nachweis der Gleichheit

Eine Stellungnahme zum Abschnitt 7.3.6 sieht in der „Forderung nach Äquivalenzstudien zum Nachweis der Gleichheit" ein „Knock-Out"-Kriterium für eine Vielzahl von RCT auf höchstem Evidenzniveau und infolgedessen die Aberkennung des Nutzens vieler medizinischer Verfahren. Dieses Konzept fordere höchstes methodisches Niveau. Äquivalenzstudien seien selten in der Literatur zu finden, was dazu führe, dass viele Studienergebnisse aus RCT nicht zu verwerten wären. Außerdem gäbe es Machbarkeits- und Finanzierungsprobleme von Äquivalenzstudien aufgrund der großen Fallzahlen.

Grundsätzlich ist das methodische Niveau bei Äquivalenz- bzw. Nicht-Unterlegenheitsfragestellungen nicht höher als bei Fragestellungen zum Nachweis von Unterschieden. Dass bei Äquivalenzfragestellungen spezielle Sachverhalte zu beachten sind (z. B. die Festlegung einer Äquivalenzgrenze), ist gerade die Intention des kritisierten Abschnitts. Dass Äquivalenz- bzw. Nicht-Unterlegenheitsstudien selten in der Literatur zu finden seien, kann nicht nachvollzogen werden, wie systematische Übersichten zu diesem Thema zeigen [40]. Die Kritik, dass mangels Verfügbarkeit von Äquivalenzstudien viele Studienergebnisse aus RCT nicht zu verwerten wären, beruht vermutlich auf einem Missverständnis. Im Rahmen von Nutzenbewertungen werden regelhaft sowohl Überlegenheitsstudien als auch Äquivalenz- bzw. Nichtunterlegenheitsstudien einbezogen und wenn möglich per Meta-Analysen zu einem Gesamtergebnis zusammengeführt. Für die Meta-Analysen ist es im Idealfall unerheblich, ob der geschätzte Effekt und Standardfehler aus einer Überlegenheits- oder Äquivalenzstudie stammt, wobei die speziellen Unterschiede im Studiendesign ggf. beachtet werden müssen (z. B. unterschiedliche ITT-Strategien). Das Argument, Äquivalenzstudien seien aufgrund der großen Fallzahlen schwer durchführbar, ist unverständlich. Solche Studien mit einer abgeschwächten Fragestellung der Nichtunterlegenheit werden z. T. gerade deshalb durchgeführt, da die Fallzahl für die „strengere" Fragestellung der Überlegenheit zu groß erscheint.

Die Sorge der Stellungnehmenden, dass die „Forderung nach Äquivalenzstudien" zur Nichtberücksichtigung vieler Studienergebnisse führe, wird nicht geteilt. Der Abschnitt 7.3.6 des Methodenpapiers enthält gar keine Forderung nach Äquivalenzstudien. Vielmehr wird dort beschrieben, welche Punkte beim Nachweis einer Gleichheit zu beachten sind. Es muss hier auch zwischen der intendierten Fragestellung der einzelnen Studien (Überlegenheit, Nichtunterlegenheit, Äquivalenz) und der Fragestellung im Rahmen der Nutzenbewertungen des Instituts unterschieden werden. Grundsätzlich werden für die Nutzenbewertungen alle Studien unabhängig von ihrer jeweiligen speziellen Fragestellung einbezogen.

8.9 Meta-Analysen

In einer Stellungnahme wird mit Bezug auf Abschnitt 7.3.8 angeregt, zu dem Fall, dass abseits von Meta-Analysen Ergebnisse mehrerer Studien auch qualitativ zusammengefasst werden können, im Methodenpapier Stellung zu nehmen.

Diese Anregung verwundert etwas, da diese Möglichkeit im betreffenden Abschnitt (siehe Punkt B – Heterogenität) explizit Erwähnung findet. Auch im Abschnitt 3.1.4 des Methodenpapiers wurde dieser Punkt adressiert.

Des Weiteren wird in der Stellungnahme die primäre Verwendung von Modellen mit zufälligen Effekten kritisiert. Dieses Vorgehen sei konservativer als das der Cochrane Collaboration. Es sei zu berücksichtigen, dass sich häufig auch homogene Datensätze in Reviews fänden, sodass „in jedem Fall Fixed-Effects-Modelle herangezogen werden sollten“.

Das Vorgehen des Instituts ist keineswegs konservativer als das der Cochrane Collaboration. Im Cochrane Handbook wird überhaupt keine primäre Strategie zum Poolen von Ergebnissen vorgeschlagen, sondern die gängigen Verfahren werden vergleichend beschrieben. Der Verwendung von Modellen mit zufälligen Effekten im Falle von Heterogenität ist im Cochrane Handbook [25 Kapitel 9] ein eigener Abschnitt (9.5.4), der in der Stellungnahme auch zitiert wird, gewidmet. Des Weiteren ist zu beachten, dass bei homogener Ergebnislage die gepoolten Resultate unter Anwendung eines Modells mit zufälligen bzw. festen Effekten identisch sind.

Dieselbe Stellungnahme kritisiert, dass im Teil B (Heterogenität) des Abschnitts 7.3.8 keine inhaltlichen Kriterien zum Umgang mit Heterogenität angeführt würden und stattdessen statistische Betrachtungen zu großen Raum einnähmen. Es bedürfe einer differenzierten Betrachtung.

In diesem Teil der Stellungnahme finden sich allgemeine Aussagen zum Umgang mit Heterogenität, die zum Teil auch statistischer Natur sind, die sich inhaltlich gleichartig im betreffenden Abschnitt des Methodenpapiers finden. Das Methodenpapier beschreibt die allgemeinen, für alle Projekte gleichermaßen geltenden methodischen Grundsätze. Welche klinischen und methodischen Faktoren in einem Projekt konkret zur Untersuchung von Heterogenität herangezogen werden, ist in den projektspezifischen Berichtsplänen festgehalten.

In einer weiteren Stellungnahme wird vorgeschlagen zu präzisieren, wann aufgrund zu großer Heterogenität auf das Poolen der Ergebnisse mittels einer Meta-Analyse verzichtet wird.

Der Vorschlag wurde aufgegriffen. Im Abschnitt 7.3.8 des Methodenpapiers wurde das regelhafte Kriterium (p-Wert für Heterogenitätstest $< 0{,}2$) mit dem Zusatz ergänzt, dass es aufgrund der Kontextabhängigkeit in den Berichtsplänen der Projekte spezifiziert wird.

In dieser Stellungnahme wird des Weiteren darauf hingewiesen, dass eine Nutzenbewertung aufgrund einer zu hohen Heterogenität nicht ausbleiben dürfe.

Im Teil B (Heterogenität) des Abschnitts 7.3.8 des Methodenpapiers wird explizit geschrieben, dass bei gleichgerichteten Effekten ggf. auch ohne quantitative Zusammenfassung (positive) Nutzenaussagen getroffen werden können. Zum besseren Verständnis und zur Präzisierung des Vorgehens der Ableitung von Aussagen im Falle gleichgerichteter Effekte wurde der Abschnitt 3.1.4 des Methodenpapiers („Zusammenfassende Bewertung“) überarbeitet.

8.10 Indirekte Vergleiche

Eine Stellungnahme zum Abschnitt 7.3.9 bemängelt die unkonkrete Beschreibung dazu, welche Methoden für indirekte Vergleiche vom Institut akzeptiert werden.

Eine solche Beschreibung findet sich in dem zitierten Methodenpapier zur Bewertung von Verhältnissen zwischen Kosten und Nutzen. Diese Kriterien wurden übernommen und durch aktuelle Literatur ergänzt. Es wurde außerdem die Aussage ergänzt, dass neben den komplexen MTC Meta-Analysen auch das einfache Verfahren nach Bucher [6] akzeptabel sein kann.

Literaturverzeichnis

1. Gesetz über den Verkehr mit Arzneimitteln (Arzneimittelgesetz - AMG) [online]. 05.2011 [Zugriff: 29.06.2011]. URL: http://bundesrecht.juris.de/bundesrecht/amg_1976/gesamt.pdf.

2. Verordnung über die Nutzenbewertung von Arzneimitteln nach § 35a Absatz 1 SGB V für die Erstattungsvereinbarungen nach § 130b SGB V (Arzneimittel-Nutzenbewertungsverordnung -AM-NutzenV). Bundesgesetzblatt 2010; 1(68): 2324-2328.

3. Agency for Healthcare Research and Quality (AHRQ). Mehods fo balacing harms and benefits in systematic reviews [online]. 03.03.2011 [Zugriff: 15.07.2011]. URL: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=638.

4. Amir E, Seruga B, Martinez-Lopez J, Kwong R, Pandiella A, Tannock IF et al. Oncogenic targets, magnitude of benefit, and market pricing of antineoplastic drugs. J Clin Oncol 2011; 29(18): 2543-2549.

5. Aronson J. Balancing benefits and harms in health care. BMJ 2004; 329: 30.

6. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997; 50(6): 683-691.

7. 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). Bundesgesetzblatt Teil 1 2010; (68): 2324-2328.

8. Cardiac Arrhythmia Suppression Trial. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 1989; 321(6): 406-412.

9. Cardiac Arrhythmia Suppression Trial. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. N Engl J Med 1991; 324(12): 781-788.

10. Europäisches Parlament. Verordnung (EG) Nr. 141/2000 des Europäischen Parlaments und des Rates vom 16. Dezember 1999 über Arzneimittel für seltene Leiden. Amtsblatt der Europäischen Gemeinschaften 2000; 43(L18): 1-5.

11. Europäisches Parlament, Rat der Europäischen Union. Verordnung (EG) Nr. 726/2004 des Europäischen Parlaments und des Rates vom 31.03.2004 zur Festlegung von Gemeinschaftsverfahren für die Genehmigung und Überwachung von Human- und Tierarzneimitteln und zur Errichtung einer Europäischen Arzneimittel-Agentur. Amtsblatt der Europäischen Union 2004; 47(L136): 1-33.

12. Europäisches Parlament, Rat der Europäischen Union. Verordnung (EG) Nr. 1901/2006 des Europäischen Parlaments und des Rates vom 12.12.2006 über Kinderarzneimittel und zur Änderung der Verordnung (EWG) Nr. 1768/92, der Richtlinien 2001/20/EG und 2001/83/EG sowie der Verordnung (EG) Nr. 726/2004. Amtsblatt der Europäischen Union 2006; 49(L378): 1-19.

13. European Medicines Agency (EMA). Points to consider on application with: 1. Meta-analyses; 2. One pivotal study [online]. 31.05.2001 [Zugriff: 22.09.2010]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.

14. European Medicines Agency (EMA). Assessment report for AVANDIA: International non-proprietary name: Rosiglitazone [online]. 03.12.2010 [Zugriff: 15.07.2011]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Assessment_Report_-_Variation/human/000268/WC500100757.pdf.

15. European Medicines Agency (EMA). Questions and answers on the suspension of rosiglitazone-containing medicines (Avandia, Avandament and Avaglim) [online]. 23.09.2010 [Zugriff: 14.07.2011]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Medicine_QA/2010/09/WC500097003.pdf.

16. Food and Drug Administration (FDA). FDA Amendments Act (FDAAA) of 2007: Public Law No. 110-85 §801 [online]. 09.2007 [Zugriff: 29.06.2011].

17. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11(2): 88-94.

18. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses [online]. 20.01.2011 [Zugriff: 02.09.2011]. URL: http://www.g-ba.de/downloads/62-492-548/VerfO_2011-08-04.pdf.

19. Glasziou PP, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334(7589): 349-351.

20. Grimes DA, Schulz K. An overview of clinical research: The lay of the land. Lancet 2002; 359(9300): 57-61.

21. Hartmann A, Weber S, Herpertz S, Zeeck A. Psychological treatment for anorexia nervosa: A meta-analysis of standardized mean change. Psychother Psychosom 2011; 80(4): 216-226.

22. Hay PP, Bacaltchuk J, Stefano S, Kashyap P. Psychological treatments for bulimia nervosa and binging. Cochrane Database Syst Rev 2009; (4): CD000562.

23. Heidrich H, Cachovan M, Creutzig A, Rieger H, Trampisch HJ. Guidelines for therapeutic studies in Fontaine's stages II-IV peripheral arterial occlusive disease. German Society of Angiology. VASA 1995; 24(2): 107-119.

24. Heidrich H, Trampisch HJ, Roehmel J. Stellungnahme zu den Prüfrichtlinien für Therapiestudien im Fontaine Stadium II-IV bei PAVK. VASA 1996; 25(1): 73-75.

25. Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008.

26. Hunot V, Churchill R, Silva de Lima M, Teixeira V. Psychological therapies for generalised anxiety disorder. Cochrane Database Syst Rev 2007; (1): CD001848.

27. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Vakuumversiegelungstherapie von Wunden: Abschlussbericht; Auftrag N04-03 [online]. 13.03.2006 [Zugriff: 12.08.2011]. (IQWiG Berichte; Band 4). URL: https://www.iqwig.de/download/N04-03_Abschlussbericht_Vakuumversiegelungstherapie_zur_Behandlung_von_Wunden..pdf.

28. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Früherkennungsuntersuchung von Hörstörung bei Neugeborenen: Abschlussbericht; Auftrag S05-01 [online]. 28.02.2007 [Zugriff: 12.08.2011]. (IQWiG Berichte; Band 19). URL: https://www.iqwig.de/download/S05-01_Abschlussbericht_Frueherkennungsuntersuchung_von_Hoerstoerungen_bei_Neugeborenen.pdf.

29. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Interstitielle Brachytherapie beim lokal begrenzten Prostatakarzinom: Abschlussbericht; Auftrag N04-02 [online]. 17.01.2007 [Zugriff: 12.08.2011]. (IQWiG Berichte; Band 15). URL: https://www.iqwig.de/download/N04-02_Abschlussbericht_Brachytherapie.pdf.

30. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Stammzelltransplantation bei den Indikationen Akute lymphatische Leukämie (ALL) und Akute myeloische Leukämie (AML) bei Erwachsenen: Abschlussbericht; Auftrag N05-03A [online]. 30.03.2007 (IQWiG Berichte; Band 21). URL: https://www.iqwig.de/download/N05-03A_Abschlussbericht_Stammzelltransplantation_be_ALL_und_AML.pdf.

31. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Qualität der pädiatrisch-hämatologisch-onkologischen Versorgung: Abschlussbericht; Auftrag V06-01 [online]. 13.08.2009 [Zugriff: 12.08.2011]. (IQWiG Berichte; Band 62). URL: https://www.iqwig.de/download/V06-01_Abschlussbericht_Qualitaet_der_paediatrisch_haematologisch_onkologischen_Versorgung.pdf.

32. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Aussagekraft von Surrogatendpunkten in der Onkologie: Rapid Report; Auftrag A10-05 [online]. 31.01.2011 [Zugriff: 15.07.2011]. (Band 80). URL: https://www.iqwig.de/download/A10-05_Rapid_Report_Surrogatendpunkte_in_der_Onkologie..pdf.

33. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Clopidrogel plus Acetylsalicylsäure bei akutem Koronarsyndrom: Abschlussbericht; Auftrag A04-01B [online]. 28.01.2009 [Zugriff: 14.07.2011]. (IQWiG Berichte; Band 43). URL: https://www.iqwig.de/download/A04-01B_AB_Clopidogrel_plus_ASS_bei_akutem_Koronarsyndrom.pdf.

34. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Nichtmedikamentöse lokale Verfahren zur Behandlung der benignen Prostatahyperplasie: Abschlussbericht; Auftrag N04-01 [online]. 02.06.2008 [Zugriff: 14.07.2011]. (Band 33). URL: https://www.iqwig.de/download/N04-01_Abschlussbericht_Nichtmedikamentoese_lokale_Verfahren_zur_Behandlung_de_BPH.pdf.

35. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Glinide zur Behandlung des Diabetes mellitus Typ 2: Abschlussbericht; Auftrag A05-05C [online]. 06.04.2009 [Zugriff: 14.07.2011]. (Band 48). URL: https://www.iqwig.de/download/A05-05C_Abschlussbericht_Glinide_zur_Behandlung_des_Diabetes_mellitus_Typ_2.pdf.

36. Kunz R, Lelgemann M, Guyatt GH, Antes G, Falck-Ytter Y, Schünemann H. Von der Evidenz zur Empfehlung. In: Kunz R, Ollenschläger G, Raspe H, Jonitz G, Donner-Banzhoff N (Ed). Lehrbuch Evidenzbasierte Medizin in Klinik und Praxis. Köln: Deutscher-Ärzte-Verlag; 2007. S. 231-248.

37. Lackner JM, Mesmer C, Morley S, Dowzer C, Hamilton S. Psychological treatments for irritable bowel syndrome: a systematic review and meta-analysis. J Consult Clin Psychol 2004; 72(6): 1100-1113.

38. Lange S. The all randomized/full analysis set (ICH E9) - may patients be excluded from the analysis? Drug Inf J 2001; 35: 881-891.

39. Lange S. Die Rolle randomisierter kontrollierter Studien bei der medizinischen Bewertung von Routineverfahren. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2006; 49(3): 272-277.

40. Lange S, Freitag G. Choice of delta: Requirements and reality - results of a systematic review. Biom J 2005; 47(1): 12-27.

41. Leichsenring F. Are psychodynamic and psychoanalytic therapies effective? A review of empirical data. Int J Psychoanal 2005; 86(Pt 3): 841-868.

42. Leichsenring F, Rabung S. Effectiveness of long-term psychodynamic psychotherapy: A meta-analysis. JAMA 2008; 300(13): 1551-1565.

43. Leichsenring F, Rüger U. Psychotherapeutische Behandlungsverfahren auf dem Prüfstand der Evidence Based Medicine (EBM). Randomisierte kontrollierte Studien vs. naturalistische Studien - Gibt es nur einen Goldstandard? Z Psychosom Med Psychother 2004; 50(2): 203-217.

44. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009; 29(5): E13-21.

45. Pincus T. The American College of Rheumatology (ACR) Core Data Set and derivative "patient only" indices to assess rheumatoid arthritis. Clin Exp Rheumatol 2005; 23(5 Suppl 39): S109-113.

46. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009; 62(8): 797-806.

47. Sachverständigenrat für die Konzentrierte Aktion im Gesundheitswesen. Bedarfsgerechtigkeit und Wirtschaftlichkeit; Band 1: Zielbildung, Prävention, Nutzerorientierung und Partizipation; Band 2: Qualitätsentwicklung in Medizin und Pflege; Gutachten 2000-2001; Kurzfassung [online]. [Zugriff: 08.07.2011]. URL: http://www.svr-gesundheit.de/Gutachten/Gutacht00/kurzf-de00.pdf.

48. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c332.

49. Seidler A. Anerkennung von Berufskrankheiten: Anwendung der risiokoverdopplung und alternativer Kriterien. Zbl Arbeitsmed 2001; 51: 286-295.

50. Seidler A, Nienhaus A. Gibt es bei häufigen Krankheiten ein Verdopplungsrisiko? Arbeitsepidemiologische anmerkungen zur Frage der Gruppentypik der Berufskrankheit Nr. 2108 BKV. Zbl Arbeitsmed 1999; 49: 74-79.

51. Smith R. Think harm always. BMJ 2004; 329.

52. van Tinteren H, Hoekstra OS, Smit EF, van den Bergh JHAM, Schreurs AJM, Stallaert RALM et al. Effectiveness of positron emission tomography in the preoperative assessment of patients wirh suspected non-small-cell lung cancer. Lancet 2002; 359(9315): 1388-1392.

53. Victor N. On clinically relevant differences and shifted null hypotheses. Methods Inf Med 1987; 26(3): 109-116.

54. Windeler J, Lange S. Nutzenbewertung in besonderen Situationen - Seltene Erkrankungen. Z Evid Fortbild Qual Gesundhwes 2008; 102(1): 25-30.

55. Wissenschaftlicher Beirat Psychotherapie (WBP). Verfahrensregeln zur Beurteilung der wissenschaftlichen Anerkennung von Methoden und Verfahren der Psychotherapie [online]. 20.09.2010 [Zugriff: 15.07.2011]. URL: http://www.wbpsychotherapie.de/downloads/Methodenpapier28.pdf.

56. Woitowitz H-J. Ermittlung der Exposition als Grundlage der Begutachtung: Quantifizierung der Expedition aus arbeitsmedizinischer Sicht. Med Sach 2002; 98(3): 86-92.

57. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. BMJ 2008; 336(7644): 601-605.

Anhang: Dokumentation der Stellungnahmen

Inhaltsverzeichnis

A 1 Stellungnahmen von Organisationen, Institutionen und Firmen

A 1.1 Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V. (AWMF)

Autor:

Selbmann, Konrad

Arbeitsgemeinschaft der
Wissenschaftlichen
Medizinischen
Fachgesellschaften e.V.

Association of the
Scientific
Medical
Societies in Germany

Düsseldorf, 08.04.2011

Stellungnahme der AWMF zum
Entwurf des Methodenpapiers Version 4 des IQWiG vom 08.03.2011
zum Aspekt: Produktspezifische Verfahrensabläufe

Die Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) begrüßt die regelmäßige Fortschreibung des Methodenpapiers des IQWiG, das nun in vierter Version als Entwurf zur Stellungnahme vorgelegt wurde. Begrüßenswert ist auch der neue eingeführte Abschnitt „Was ist neu?“, wobei wir aus Kapazitätsgründen ohne Prüfung davon ausgehen, dass in den dort nicht erwähnten Kapiteln keine Änderungen gegenüber der Version 3 erfolgt sind. Dies lässt sich zu einem späteren Zeitpunkt immer noch verifizieren.

Ein wichtiges Anliegen der AWMF ist die frühzeitige Einbindung des Sachverstands von Vertretern der wissenschaftlichen medizinischen Fachgesellschaften und von Vertretern der Patientenorganisationen in die Verfahren zur Nutzenbewertung und zur Kosten-Nutzenbewertung des IQWiG[1,2,3]. Die vorliegende Stellungnahme der AWMF geht daher primär auf die in Kapitel 2.1 dargelegten produktspezifischen Arbeitsabläufe ein. Zur Umsetzung der Rahmenvorgaben des Methodenpapiers im konkreten Einzelfall wird die AWMF im Bedarfsfall gesondert Stellung nehmen.

Im Kapitel 2.1. ist auffallend, dass das IQWiG betont, zunehmend Produkte zu erstellen, für die Stellungnahmeverfahren (nach §139a Abs. 5 SGB V) zu allen wichtigen Erstellungsabschnitten nicht vorzusehen sind. Dazu gehören Rapid Reports, Dossierbewertungen, Arbeitspapiere und in Grenzen IQWiG-Stellungnahmen und Gesundheitsinformationen. Solange den nach §139a SGB V Berechtigten nicht ausreichend Zeit für Stellungnahmen eingeräumt wird, kann auch nicht von deren Billigung ausgegangen werden. Produkte ohne Stellungnahmeverfahren nach §139a SGB V können nicht den gleichen Stellenwert haben wie etwa die ausführlichen Berichte mit Empfehlungen zu den im SGB V beschriebenen Aufgaben.

[1] Stellungnahme der AWMF zum überarbeiteten Kapitel „2.1 Nutzenbewertung in der Medizin“ des Methodenpapiers Version 2 des IQWiG vom 28.9.06. Düsseldorf, im Oktober 2006
[2] Möglichkeiten der Beteiligung der Fachgesellschaften an der Erstellung von Berichten des IQWiG für den Gemeinsamen Bundesausschuss. Vorschläge der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) zur Verbesserung des IQWiG-Verfahrens zur Berichterstellung. Düsseldorf, 13.12.07
[3] Stellungnahme der AWMF zum „Entwurf einer Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung-Version 2.0“ des IQWiG. Düsseldorf, 20.04.2009

Neu gegenüber der Version 3.0 des Methodenpapiers ist insbesondere eine Präzisierung der Unterscheidung der Verfahrensabläufe für die Erstellung von ausführlichen Berichten und Rapid Reports (Schnellberichten). Dabei ist in der Tat kritisch zu hinterfragen, warum im Ablauf für Rapid Reports explizit keine Möglichkeit zur Stellungnahme vorgesehen ist. Gemäß §139a hat das IQWiG in allen wichtigen Abschnitten der Bewertungsverfahren Sachverständigen aus Wissenschaft und Praxis und Vertretern der Patienten Gelegenheit zur Stellungnahme zu geben und deren Stellungnahmen sind in die Entscheidung einzubeziehen. Die Unterscheidung zwischen ausführlichen Berichten und Rapid Reports ist in diesem Zusammenhang nicht einsichtig, da, wie im Entwurf auch dargelegt, Grundlage für die Erstellung beider Produkte die in § 139a SGB V beschriebenen Aufgaben des Instituts sind. Der ergänzte Hinweis zur Zielsetzung der Rapid Reports als zeitnahe Empfehlungen „zu denen aus Sicht des Auftraggebers keine Anhörungen durch das Institut erforderlich sind“ ist unklar und wäre durch die Auftraggeber (G-BA und BMG) zu prüfen.

Neu in die Produktpalette des IQWiG aufgenommen wurden Dossierbewertungen, die der Unterstützung der frühen Nutzenbewertung neu zugelassener Arzneimittel durch den Gemeinsamen Bundesausschuss (G-BA) vor dem Hintergrund des am 1. Januar 2011 in Kraft getretenen Gesetz zur Neuordnung des Arzneimittelmarktes (AMNOG) dienen sollen[4]. Grundsätzlich ist eine solche, das Zulassungsverfahren ergänzende Nutzenwertung zu begrüßen. Es verwundert jedoch, dass diese Nutzenbewertung in der Regel allein aufgrund von Nachweisen des pharmazeutischen Unternehmers (Dossiers) und nur optional unterstützt durch zusätzliche Evidenzrecherchen erfolgen soll. Da zum Zeitpunkt der Zulassung in der Regel nur wenige praxisrelevante Studien vorliegen, kommt der Anhörung von Vertretern aus Wissenschaft und Praxis als zusätzliche Informationsquelle für die Abwägung von potentiellem Nutzen und Schaden des neuen Arzneimittels große Bedeutung zu. Beispielhaft wären klinische Forschergruppen, die an der Entwicklung beteiligt waren und erste Anwender zu nennen. Insofern ist zu bedauern, dass bei den Dossierbewertungen keine Anhörung durch das IQWiG im Rahmen der Berichterstellung vorgesehen ist, sondern erst nach Fertigstellung der Berichte im Rahmen der Bewertung durch den G-BA [5].

Die Zielsetzung einer zeitnahen Fertigstellung von Rapid Reports und Dossierbewertungen erfordert ein pragmatisches Vorgehen. Insbesondere sollte bei den eher als vorläufig anzusehenden Dossierbewertungen im Falle neuer Erkenntnisse eine strukturierte Fortschreibung erarbeitet und in das Methodenpapier aufgenommen werden.

[4] § 35a SGB V. Bewertung des Nutzens von Arzneimitteln mit neuen Wirkstoffen. Verfügbar: http://www.gesetze-im-internet.de/sgb_5/__35a.html
[5] Verfahrensordnung des G-BA. 5. Kapitel: Bewertung des Nutzens von Arzneimitteln nach § 35a SGB V in der Fassung vom 20. Januar 2011, §19: Gesetzliches Stellungnahmeverfahren. Verfügbar: http://www.g-ba.de/downloads/17-98-3025/2011_Kapitel 5 VerfO_zweiseitig.pdf

Insgesamt sind aus Sicht der AWMF für alle Produkte des IQWiG die Kernelemente der systematischen Recherche nach der bestverfügbaren Evidenz und die strukturierte Einbeziehung des Sachverstands von Vertretern der wissenschaftlichen medizinischen Fachgesellschaften sowie von Vertretern der Patientenorganisationen unverzichtbar.

Am Anfang des Kapitels 2.1 wird bei den Gesundheitsinformationen kein Stellungnahmeverfahren erwähnt, im Abschnitt 2.1.6 jedoch ausdrücklich darauf aufmerksam gemacht, dass der endgültige Entwurf eines Informationsprodukts im Rahmen einer einmonatigen Beratungsperiode zur begrenzten Stellungnahme an den Auftraggeber und die Gremien des IQWiG u.a. dem Kuratorium, in dem die AWMF vertreten ist, verschickt wird. Die Fachgesellschaften haben sich wiederholt zu den vorgelegten Texten geäußert und aus der Sicht der AWMF nützliche Hinweise zu Verbesserungen gegeben. Daher halten sie sowohl eine begleitende als auch eine externe schließende Evaluation des Gesamtprodukts „Gesundheitsinformation“ nach dem gegenwärtigen Stand der Wissenschaft, beide transparent für die oben genannten Stellungnahmeberechtigten für erforderlich.

Zusätzlich verweisen wir auf das ausführliche Gutachten von Prof. Rüger, Institut für Statistik an der Universität München, das von der Deutschen Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie (DGPT) in Auftrag gegeben war.[6] Das Gutachten liegt dem IQWiG bereits vor und ist hier nur der Vollständigkeit halber der AWMF-Stellungnahme noch einmal beigelegt. Viele der dort aufgeführten methodischen Gesichtspunkte aus der Sicht der psychotherapeutischen Verfahren lassen sich auch auf andere nicht-medikamentöse Verfahren übertragen.

[6] Prof. Dr. B. Rüger: Stellungnahme zum Entwurf „Allgemeine Methoden“ des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) in der Version 4.0 vom 09.03.2011. Erstellt für die Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie (DGPT) e.V.

A 1.2 Bundesärztekammer (BÄK)

Autor:
Zorn, Ulrich

Stellungnahme der Bundesärztekammer

zu den Allgemeinen Methoden des IQWiG Entwurfsversion 4.0 in der Fassung vom 09.03.2011

Berlin, 08.04.2011

[redacted]

Zu den Allgemeinen Methoden des IQWiG in der Entwurfsversion 4.0 in der Fassung vom 09.03.2011 nimmt die Bundesärztekammer wie folgt Stellung:

Die Bundesärztekammer begrüßt die regelmäßige Überprüfung und Überarbeitung der methodischen Grundlagen des IQWiG. Die Bundesärztekammer hatte hierzu bereits mehrfach in Stellungnahmen Hinweise gegeben. Zu der jetzt vorgelegten Entwurfsversion der „Allgemeinen Methoden" zur Nutzenbewertung in der Version 4.0 nimmt die Bundesärztekammer wie folgt Stellung:

Präambel

In der überarbeiteten Präambel werden jetzt auch Gremien und Organe des IQWiG vorgestellt (Stiftungsrat und Stiftungsvorstand). Hierzu ist anzumerken, dass

- ► Strukturen des Instituts nicht in der Präambel abgehandelt werden sollten, sondern, wie in Version 3.0, in dem dafür geeigneteren Kapitel 1,
- ► als weitere zu nennende Gremien der Beirat und das Kuratorium des Instituts fehlen,
- ► die Zusammensetzung des Stiftungsrats auf Seiten der „Leistungserbringer" die Bundesärztekammer vermissen lässt.

Löschen des ehemaligen Abschnitts 1.2 „Struktur"

Die Erläuterung der Struktur des Instituts mag nicht unmittelbar Bestandteil seiner Methoden sein, die Kenntnisse der Struktur können jedoch das Verständnis für die Arbeitsweise erleichtern. Insofern gehen durch die Streichung dieses Abschnitts möglicherweise nützliche Informationen verloren. Es sollte daher wenigstens ein Hinweis in dem Text eingefügt werden, dass Informationen zu Struktur und Organisation des Instituts auf der Internetseite des IQWiG verfügbar sind.

Überarbeiteter Abschnitt 1.2.6

Im Zusammenhang mit der Erläuterung der Eigenschaften klinischer Studien werden die Begriffe „Verzerrung" und „Concealment" benutzt, dabei wird aber nur der Begriff der „Verzerrung" erläutert. Auf den Klammerzusatz mit „Concealment" sollte entweder an dieser Stelle verzichtet werden, oder es sollte ein gleichfalls deutschsprachiger Begriff mit Erläuterungen ergänzt werden (Verblindung, verdeckte Zuordnung, …).

Unter Nr. 1 ist vom **Nachweis** von Kausalität die Rede, unter Nr. 2 vom **Beweis** der Kausalität - hier sollte eine einheitliche Begriffsverwendung angestrebt werden, vorzugsweise zugunsten des Begriffs Nachweis, da eine Beweisführung im mathematischen oder naturwissenschaftlich-physikalischen Sinne hier wohl nicht gemeint ist.

Hauptanliegen des Kapitels ist nach wie vor, den hohen Stellenwert von RCTs für die Beurteilung des Nutzens einer Intervention oder Maßnahme darzulegen. Dabei wird erläutert, warum sich hohe Ergebnissicherheit und Alltagsnähe nicht ausschließen müssen, um offenbar wiederkehrenden Forderungen nach Berücksichtigung von Studien auch ohne Randomisierung entgegenzutreten. Zu vermissen ist bei diesen Ausführungen je-

doch die Würdigung des Umstands, dass für viele Fragestellungen gar keine RCTs zur Verfügung stehen, zumindest nicht in solchen Zeiträumen, wie sie der G-BA zur Beantwortung von Fragen zu Leistungsausschluss oder -gewährung innerhalb der GKV benötigt und zur Klärung als Auftrag an das Institut vergibt. Insofern gehen die ausgiebigen Ausführungen nicht auf alle der in den in den letzten Jahren zu beobachtenden Probleme der Entscheidungsfindung bei der Nutzenbewertung ein.

Es sollte außerdem bedacht werden, dass der neu eingefügte Abschnitt 7.3.4 (Bewertung subjektiver Endpunkte bei offenen Studiendesigns) zwar nicht unmittelbar im Widerspruch zur herausgehobenen Bedeutung von RCTs gesehen werden muss, diese jedoch relativiert. Damit bleibt die Frage einer verlässlichen und reproduzierbaren Bewertungsgrundlage bzw. ein möglicher ad-hoc-Umgang des Instituts in Bewertungssituationen vakant.

Überarbeiteter Abschnitt 2.1

Abbildung 1 „Ablauf der Berichterstellung“ entspricht nicht mehr in allen Punkten den Erläuterungen im Text. So fehlen die (in der entsprechenden Abbildung der Version 3.0 noch vorhandenen) Textfelder, wonach der vorläufige Berichtplan und der Vorbericht beim Auftraggeber/Kuratorium/Stiftungsvorstand vorgelegt werden.

Die Auswahl externer Sachverständiger mithilfe einer Sachverständigendatenbank erscheint als pragmatischer Weg zur Bewältigung der durch kurze Fristen gekennzeichneten Nutzenbewertungen nach § 35a SGB V. Die Auswahl von Sachverständigen soll anhand einer Kriterienliste geschehen, die auf Selbstauskünften der Bewerber basiert. Ein Verifizierungsverfahren dieser Angaben durch das Institut ist nicht regelhaft vorgesehen - ob hierzu eine Notwendigkeit besteht, wird sich vermutlich erst nach ersten Erfahrungen mit dem Verfahren erweisen. Es sollte überlegt werden, die ausführlichere Beschreibung des Prozederes, die auf den Internetseiten des Instituts verfügbar ist, auch in den „Allgemeinen Methoden“ darzustellen.

Überarbeiteter Abschnitt 3.1.1

Die stärkere Bezugnahme auf den Aspekt der klinischen Wirksamkeit bei der Definition des patientenrelevanten medizinischen Nutzens von Behandlungsmaßnahmen ist zu begrüßen (die Bundesärztekammer hatte dies in ihrer Stellungnahme vom 13.12.2007 zur Version 3.0 des Methodenpapiers angeregt). Die in diesem Zusammenhang zitierte und als „höchstrichterliche Rechtsprechung“ zusammengefasste Entscheidung des Bundessozialgerichts vom 6. Mai 2009 (Protonentherapie bei Brustkrebs), dass „bestimmte (Nutzen-)Aspekte erst dann notwendigerweise zu bewerten sind, wenn die therapeutische Wirksamkeit hinreichend belegt ist“, eignet sich allerdings nur bedingt als Referenz für einen differenzierten Umgang mit diesen Begriffen bzw. der Problematik. Dies ergibt sich allein schon aus einzelnen Formulierungen der BSG-Entscheidung, die eine Unschärfe bei der Trennung zwischen Nutzen und Wirksamkeit erkennen lassen: „Danach ist eine Methode nur dann entsprechend dem allgemein anerkannten Stand der medizinischen Erkenntnisse erforderlich, wenn ihr **Nutzen** und die medizinische Notwendigkeit

der Methode belegt sind. Fehlen solche Belege für die therapeutische **Wirksamkeit** eines Behandlungskonzepts, so …“ (BSG v. 6.5.2009, B 6 A 1/08 R).

Die Möglichkeit, einen „endpunktbezogen »geringeren Schaden« (im Sinne einer Verringerung von Nebenwirkungen) bei Betrachtung der Effekte auf alle anderen Endpunkte in die abwägende Feststellung eines »Zusatznutzens«“ einzubeziehen, ist zu begrüßen. Dieses Vorgehen sollte nicht nur als „möglich“ im Sinne einer Option dargestellt werden, sondern regelmäßig Anwendung finden.

Überarbeiteter Abschnitt 3.1.2

Die stärkere Ausdifferenzierung der Bewertung des Umgangs mit Surrogatparametern ist zu begrüßen. Die in der letzten Stellungnahme der Bundesärztekammer vom 13.12.2007 vorgeschlagene Differenzierung zwischen intermediären und Proxy-Zielgrößen ist an dieser Stelle jedoch zu wiederholen.

Überarbeiteter Abschnitt 3.1.3

In diesem nur geringfügig geänderten Abschnitt ist aus der Gegenüberstellung von Schaden und Nutzen medizinischer Interventionen nach wie vor die Tendenz erkennbar, in Zweifelsfällen eher ablehnenden bzw. dilatorischen Entscheidungen den Weg zu bahnen (vergleiche die Stellungnahme der Bundesärztekammer vom 13.12.2007).

Überarbeiteter Abschnitt 3.1.4

Trotz der passagenweisen Überarbeitung dieses Abschnitts bleibt im Kern als Regelanforderung der Beleg eines statistisch signifikanten Effekts durch eine Metaanalyse von Studien mit endpunktbezogen geringer Ergebnisunsicherheit oder durch mindestens zwei voneinander unabhängig durchgeführte Studien mit endpunktbezogen geringer Ergebnisunsicherheit bestehen. Die Bundesärztekammer hatte dies bereits in ihrer Stellungnahme vom 13.12.2007 als zu rigide eingestuft, zumindest bezogen auf lebensbedrohliche Erkrankungen. Positiv ist anzumerken, dass nunmehr für die Entscheidungsperspektive des Instituts auf das Prinzip der Risikovorsorge unter Annahme eines Schadenpotenzials verzichtet werden soll.

Neuer Abschnitt 3.3.3

In diesem Abschnitt werden überwiegend die gesetzlichen Vorgaben der Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V wiedergegeben. Mit Blick auf die Nutzenbewertung wird auf die „im vorliegenden Methodenpapier beschriebenen Standards der evidenzbasierten Medizin“ verwiesen; die Bewertung der Kosten soll „auf Basis der Standards der Gesundheitsökonomie“ erfolgen. Zumindest der letztgenannte Aspekt erscheint sehr pauschal, da die Gesundheitsökonomie, wie jeder andere wissenschaftliche Fachbereich auch, von unterschiedlichen Lehrmeinungen und entsprechenden Diskursen geprägt ist. Wünschenswert an dieser Stelle wäre der Verweis auf die institutseigenen Me-

thoden zur Bewertung von Kosten-Nutzen-Verhältnissen, die hierzu allerdings einer Aktualisierung bedürften.

Abschnitt 3.5

Unverändert sollen für die Bewertung diagnostischer Verfahren „dieselben Nutzenkategorien wie bei der Bewertung therapeutischer Interventionen“ gelten, nämlich „Mortalität, Morbidität und gesundheitsbezogene Lebensqualität“. Dies erscheint nach wie vor problematisch, da einerseits die Studiensituation bei diagnostischen Verfahren in aller Regel ungünstiger ist als bei therapeutischen Verfahren, andererseits die ärztliche Perspektive, die für Therapieentscheidungen maßgeblich ist, ausgeblendet wird. Komplett negiert wird der Selbstwert einer diagnostischen Erkenntnis für Arzt und Patient, die unabhängig von der Zielstellung einer therapeutischen Konsequenz existiert.

Überarbeiteter Abschnitt 4

Der neue Titel dieses Abschnitts „Methoden der Versorgungsqualität“ ist, im Gegensatz zum früheren Titel „Leitlinien und Disease-Management-Programme“, nicht selbsterklärend. Es ist anhand des Titels nicht erkennbar, dass es in diesem Abschnitt primär um das Thema Leitlinien geht.

Der Versuch einer eigenen Definition von Leitlinien zu Beginn des Abschnitts wirkt angesichts mittlerweile seit vielen Jahren auf nationaler und europäischer Ebene vorliegender und weithin akzeptierter Formulierungen (siehe etwa unter auf den Webseiten des ÄZQ unter http://www.leitlinien.de) redundant.

Die Aussage, dass das „Ressort Versorgungsqualität des IQWiG seinen Aufgabenbereich in der wissenschaftlichen Identifikation und Beschreibung von Versorgungsstandards auf Basis hochwertiger Leitlinien sieht“, gehört eher in einen Geschäftsbericht, weniger in ein Methodenpapier. Die Aufgabe des IQWiG sollte sich aus seiner gesetzlichen Grundlage (vgl. § 139a SGB V) ergeben und nicht durch die Sichtweise einzelner Ressorts bestimmt werden. Mit Blick auf Leitlinien lautet hier der klare gesetzliche Auftrag „Bewertungen evidenzbasierter Leitlinien für die epidemiologisch wichtigsten Krankheiten“ sowie „Abgabe von Empfehlungen zu Disease-Management-Programmen“ (§ 139a SGB V Abs. 3 Nr. 3 SGB V). Die in Aussicht gestellte Identifikation von Versorgungsstandards auf der Basis hochwertiger Leitlinien ist etwas völlig anderes, indem Leitlinien nur noch Mittel zum Zweck sein sollen und nicht mehr der eigentliche Gegenstand der Befassung. Die Aufgabe, die sich das Ressort Versorgungsqualität des IQWiG hier selber erteilt, erinnert an im Zuge des GKV-Gesundheitsreformgesetzes 2000 in einem neuen § 137e Abs. 3 SGB V formulierte Aufgabe für den damaligen Koordinierungsausschuss, „auf der Grundlage evidenzbasierter Leitlinien die Kriterien für eine im Hinblick auf das diagnostische und therapeutische Ziel ausgerichtete zweckmäßige und wirtschaftliche Leistungserbringung für mindestens 10 Krankheiten je Jahr [zu] beschließen, bei denen Hinweise auf unzureichende, fehlerhafte oder übermäßige Versorgung bestehen und deren Beseitigung die Morbidität und Mortalität der Bevölkerung nachhaltig beeinflussen kann“. Diese, dem Vorläufergremium des G-BA zugedachte Aufgabe verschwand schon nach kurzer Zeit wieder durch das GKV-Modernisierungsgesetz. Die

Gründe für seinen raschen Sinneswandel sind seitens des Gesetzgebers nicht dokumentiert worden, es liegt aber die Vermutung nahe, dass nicht nur die Quantität der Aufgabe (10 Krankheiten pro Jahr) an methodische Grenzen stieß.

Von dieser grundsätzlichen Überlegung abgesehen bedeutet die beabsichtigte Verknüpfung des Untersuchungsgebiets an das Vorhandensein „hochwertiger Leitlinien“ eine Limitierung wider besseres Wissen. Auch hier hatte sich bekanntlich der Gesetzgeber in einer ähnlich gelagerten (Fehl-)Vorstellung selber korrigieren müssen, indem er - ebenfalls per GKV-Modernisierungsgesetz - seinen Anspruch revidieren musste, Patienten in strukturierten Behandlungsprogrammen nach „evidenzbasierten Leitlinien“ behandeln lassen zu wollen (§ 137f Abs. 2 Nr. 1 SGB V alt, Stand 2002-2004), anstatt wie seither und bis heute gültig „nach dem aktuellen Stand der medizinischen Wissenschaft unter Berücksichtigung von evidenzbasierten Leitlinien oder nach der jeweils besten, verfügbaren Evidenz…“

Dass „hohe methodische Qualität von Leitlinien nicht notwendigerweise mit der inhaltlichen Qualität der darin enthaltenen Empfehlungen korreliert“ (siehe Abschnitt 4.3.1), ist zutreffend, die Schlussfolgerung, die Parameter daher als vollständig voneinander entkoppelt betrachten zu müssen und komplett neue Methoden zur Bewertung von Leitlinieninhalten zu entwickeln, erscheint aber ambitioniert. Unter praktischen Aspekten erscheint es fraglich, mit welchen Ressourcen das Institut die inhaltliche Qualität der Empfehlungen prüfen möchte. Nicht umsonst ist in der Version 3.0 der IQWiG-Methoden festgestellt worden: „Die inhaltliche Prüfung umfasst neben der Bewertung von Vollständigkeit und Aktualität der einbezogenen Literatur auch die Interpretation und Bewertung der Studienergebnisse. Da dieses Verfahren sehr aufwendig ist, muss sich die inhaltliche Prüfung auf die in Auftrag gegebenen Fragestellungen des G-BA bzw. auf die wesentlichen Kernempfehlungen der Leitlinie beschränken.“

Insgesamt bietet der neu konzipierte Abschnitt 4 zwar eine Reihe wertvoller Überlegungen zum anspruchsvollen Thema „Critical Appraisal“ (bezeichnenderweise taucht dieser Begriff - außer im Literaturverzeichnis - im Methodenpapier gar nicht auf), hat sich aber vom eigentlichen gesetzlichen Auftrag des Instituts für den Bereich Leitlinien entfernt. Das Thema „Abgabe von Empfehlungen zu Disease-Management-Programmen“ ist bei der Neukonzeption des Abschnitts ebenfalls aus dem Blickfeld geraten.

Überarbeiteter Abschnitt 5.2.1

Es sei darauf hingewiesen, dass die kritische Einschätzung des Instituts, wonach in den Gesundheitsinformationen für Bürger und Patienten „die durchgehende Verwendung maskuliner Personenbezeichnungen (generisches Maskulinum) zu einer gedanklichen Unterrepräsentation von Frauen führt, worin eine sprachliche Benachteiligung von Frauen zu sehen ist“, offenbar nicht für die Zielgruppe seiner Methodentexte gilt (Auszug aus der Präambel des vorliegenden Methoden-Entwurfs 4.0: „In diesem Dokument wird bei der Angabe von Personenbezeichnungen jeweils die männliche Form angewandt. Dies erfolgt mit dem Ziel einer besseren Lesbarkeit“).

Überarbeiteter Abschnitt 5.3.4

Die stärker differenzierte Darstellung des Aktualisierungsprozesses der vom Institut verfassten Gesundheitsinformationen wird begrüßt, insbesondere der Verzicht auf den in der Version 3.0 noch verwendeten groben Hinweis, wonach „sich das Institut überdies über methodologische Entwicklungen auf dem Laufenden“ halte, „um entscheiden zu können, wann die Evidenz für bestimmte Resultate so überzeugend geworden ist, dass es keiner weiteren Aktualisierungen“ bedürfe.

Überarbeiteter Abschnitt 5.5.3

Als Beitrag zur Evaluation der Gesundheitsinformationen des Instituts wird auf einen von der Universität Bielefeld zur Evaluierung von Online-Modellprojekten entwickelten Online-Fragebogen verwiesen. Unverändert gegenüber den Methoden 3.0 ist von der „Auswertung der ersten 2561 ausgefüllten Fragebögen“ die Rede. Die Formulierung „ersten“ erweckt den Eindruck, als wollte das Institut mit den Auswertungen weiterer Fragebögen fortfahren, dies ist aber offenbar nicht gemeint oder zwischenzeitlich nicht der Fall gewesen.

Des weiteren erscheint auch die Aussage, wonach es bisher kein Instrument gäbe, „das sich als zuverlässiger Indikator für die Qualität von Gesundheitsinformationen oder -websites erwiesen hat“, insofern überprüfungswürdig, als die zur Unterstützung dieser Feststellung zitierte Literatur aus den Jahren 1998, 2000, 2002 und 2004 stammt und damit möglicherweise nicht mehr den neuesten Stand auf diesem sehr dynamischen Gebiet berücksichtigt. Verwiesen sei in diesem Zusammenhang auf das gemeinsame Portal Patienten-Information.de von Bundesärztekammer und Kassenärztlicher Bundesvereinigung, bei dem unter dem Stichwort „Linkpolicy“ Kriterien aufgelistet werden, welche zur Bewertung der Qualität von Gesundheitsinformationen bzw. diesbezüglichen Internetangeboten herangezogen werden können.

Überarbeiteter Abschnitt 7.1.4

In diesem Abschnitt wird an mehreren Stellen die Bezeichnung „Behandler“ verwendet, wo „Arzt/Ärztin“ sinngemäß zutreffend wäre, siehe z. B. „Weiterhin ist es erforderlich, dass die Zielvariable unabhängig vom (unverblindeten) Behandler ist bzw. unabhängig vom Behandler verblindet erhoben wird …“. Der Begriff „Behandler“ anstelle von Arzt/Ärztin ist durch seine Verwendung im Dritten Reich als Diskriminierung jüdischer Ärzte belastet und sollte dementsprechend nicht mehr verwendet werden. Im September 1938 war es zum Entzug der Approbationen für jüdische Ärzte und einer anschließenden Bezeichnung als „Behandler“ gekommen. Vertiefende Informationen zu diesem Thema sind z. B. im Internet unter http://www.jahrestag-approbationsentzug.de/ zu finden.

Neuer Abschnitt 7.3.4

Infolge der in diesem Abschnitt verwendeten Definition von subjektiven Endpunkten, darunter der Erhebung und Einschätzung „von den behandelnden bzw. endpunkterhebenden Personen“, die auf eine Stufe gestellt werden soll mit „patientenberichteten Endpunkten“, wird ein beträchtlicher Anteil von RCTs (und darauf fußender Metaanalysen) in den

Verdacht „hochgradiger“ Verzerrung geraten. Die prospektive, projektspezifische Festlegung adjustierter Entscheidungsgrenzen als methodischer Lösungsansatz des Instituts dürfte dabei großen Spielraum für die Bewertung bieten. Hierin könnte ein Risiko für eine zu variable Berücksichtigung von Studien durch das Institut liegen. Unklar erscheint, auf welche empirischen Daten das Institut sich für die Wahl eines Grenzwertes projektspezifisch stützen möchte; der Verweis auf Daten, „wie sie z. B. Wood et al. liefern“ [die zentrale Referenzliteraturstelle für den in diesem Abschnitt thematisierten Zweifel an der Qualität von RCTs], beschreibt dies nur vage.

Es sollte außerdem aus systematischen Gründen überlegt werden, die Ausführungen zum Umgang mit Verzerrung in den Abschnitt 7.3.11 (Darstellung von Verzerrungsarten) einzugliedern.

Fazit

Die Bundesärztekammer erkennt weitere Verbesserungen in den Allgemeinen Methoden im Vergleich zu den Vorversionen. Dies betrifft etwa

- die stärkere Ausdifferenzierung der Bewertung des Umgangs mit Surrogatparametern,

- die stärkere Bezugnahme auf den Aspekt der klinischen Wirksamkeit bei der Definition des patientenrelevanten medizinischen Nutzens von Behandlungsmaßnahmen,

- die Absicht, die Verringerung von Nebenwirkungen stärker in die abwägende Feststellung eines Zusatznutzens einer therapeutischen Methode einzubeziehen,

- die differenziertere Darstellung des Aktualisierungsprozesses der vom Institut verfassten Gesundheitsinformationen.

Andere Inhalte der Methoden erscheinen nach wie vor problematisch, etwa

- das Festhalten an vergleichsweise rigiden Bewertungsmechanismen unter Verwendung von RCTs und Metaanalysen trotz erfahrungsgemäß unbefriedigenden Studiensituationen. Hinsichtlich der Beurteilung zur Ergebnissicherheit und der klinischen Relevanz von Studienergebnissen auf Basis der gegenwärtig besten Evidenz sollten unter Würdigung der jeweiligen Zielgrößen prinzipiell Studien aller Evidenzgrade beim „Assessment“ berücksichtigt und dargestellt werden, um eine möglichst vollständige Information des Auftraggebers G-BA zu gewährleisten und diesem ein möglichst umfassendes „Appraisal“ zu ermöglichen.

- die zu wenig differenzierte Übertragung der Ansprüche der Bewertung des Nutzens therapeutischer auf diagnostische Maßnahmen,

- der Gegenüberstellung von Schaden und Nutzen medizinischer Interventionen mit der Tendenz, im Zweifel eher ablehnenden oder dilatorischen Entscheidungen den Weg zu bahnen,

- das neu formulierte Kapitel zu Leitlinien weist viele interessante Überlegungen zur inhaltlichen Bewertung von Leitlinien auf, hat sich dabei aber vom gesetzlichen Auftrag entfernt,

- eine noch geringe und damit erweiterungsfähige Detailtiefe weisen die im Zuge des Arzneimittelmarkt-Neuordnungsgesetzes (AMNOG) neu formulierten Abschnitte zur Nutzenbewertung von Arzneimitteln nach § 35a SGB V auf. Die Bundesärztekammer geht von einer zeitnahen ausführlichen methodischen Auseinandersetzung des IQWiG auch mit Fragen gesundheitsökonomischer Analysen und Kosten-Nutzen-Bewertungen aus.

Berlin, 08.04.2011

gez.
Dr. rer. nat. Ulrich Zorn, MPH
Bereichsleiter im Dezernat 3

A 1.3 Bundesverband der Pharmazeutischen Industrie e.V. (BPI)

Autor:
Lietz, Christiane

Stellungnahme

Bundesverband der Pharmazeutischen Industrie e.V.

BPI

Leben ist Vielfalt

Stellungnahme zum Entwurf der Allgemeinen Methoden, Version 4.0

Der Bundesverband der Pharmazeutischen Industrie e.V. (BPI) nimmt zu dem vom IQWiG vorgelegten Entwurf der Allgemeinen Methoden, Version 4.0 wie folgt Stellung:

Zu Kapitel 2.1.1 Berichte (S. 13 ff):

Zur adäquaten Vorbereitung von Stellungnahmen wäre es wünschenswert, wenn die Zeit zur Kommentierung der vorläufigen Berichtspläne und Vorberichte durch die Öffentlichkeit - gerade auch aufgrund des erheblichen Umfangs der Berichtsentwürfe - länger als vier Wochen betragen würde.

Zu Kapitel 3.1.1 Definition des patientenrelevanten medizinischen Nutzens (S. 31 ff.):

Nach wie vor will das IQWiG patientenrelevanten Nutzen anhand der Bewertung der Beeinflussung folgender patientenrelevanter Zielgrößen zur Feststellung krankheits- und behandlungsbedingter Veränderungen erlauben:

1. Mortalität,
2. Morbidität (Beschwerden und Komplikationen),
3. gesundheitsbezogene Lebensqualität.

Ergänzend sollen interventions- und erkrankungsbezogener Aufwand und die Zufriedenheit der Patienten mit der Behandlung berücksichtigt werden können.

Diese Zielgrößen erweisen sich vor allem bei der Bewertung onkologischer Therapien als zu undifferenziert. Eine Abbildung des Gesamtüberlebens (overall survival) ist in Studien zu onkologischen Interventionen häufig schwierig, da bei Einsetzen einer Progression weitere Therapien zum Einsatz kommen. Abgesehen von supportiven und palliativen Therapien ist häufig auch der Nachweis einer Verbesserung der Lebensqualität im Studienzeitraum dann schwierig, wenn die Lebensqualität zunächst unter Therapie sinkt und erst zu einem deutlich späteren Zeitpunkt wieder ansteigt. Es erscheint daher sachgerecht, in der Onkologie bei Nutzenbewertungen alternative patientenrelevante Effektmaße, z. B. die Zeit des krankheitsfreien Überlebens (progression-free survival, PFS) oder die Zeit bis zum Progress einer Erkrankung (time-to-progression, TTP) als Nutzenmaße für die isolierte Nutzenbewertung festzulegen, um den

entenbedürfnissen angemessen Rechnung zu tragen [vgl. Aidelsburger/Wasem: Kosten-Nutzen-Bewertungen von onkologischen Therapien, Juli 2008].

Dies entspricht auch dem gesetzlichen Rahmen des § 35 b Abs. 1 SGB V, der gerade keine abschließende Aufzählung für Kriterien des Patientennutzens enthält („insbesondere“).

Zu Kapitel 3.1.4. Zusammenfassende Bewertung (S. 37 ff.):

Die Aussagen zur Belegbarkeit des (Zusatz-)Nutzens sollen jetzt in vier Abstufungen getroffen werden:

- Beleg
- Hinweis
- Anhaltspunkt oder
- keine dieser drei Situationen

Diese sollen anhand definierter Evidenzen festgestellt werden:

Beleg:

- Meta-Analyse von Studien, die endpunktbezogen in der Mehrheit eine hohe Ergebnissicherheit aufweisen, einen entsprechenden statistisch signifikanten Effekt zeigt oder

- mindestens zwei voneinander unabhängig durchgeführte Studien mit endpunktbezogen hoher Ergebnissicherheit und entsprechendem statistisch signifikanten Effekt

Hinweis:

- Meta-Analysen und qualitative Zusammenfassungen von Studien mit endpunktbezogen mäßiger Ergebnissicherheit oder
- Einzelstudienergebnisse mit hoher Ergebnissicherheit

Anhaltspunkt:

- Meta-Analysen und qualitativen Zusammenfassungen von Studien mit geringer Ergebnissicherheit oder
- Einzelstudienergebnissen mit mäßiger Ergebnissicherheit

Diese Anforderungen sind ungeeignet für die Situation der frühen Nutzenbewertung, da der Zusatznutzen anhand dieser Anforderungen nicht angemessen dargestellt werden kann. Die geforderte Evidenz liegt zum Zeitpunkt der Nutzenbewertung nach § 35a SGB V, also kurz nach der Zulassung, in der Regel nicht vor. Damit ist es kaum möglich, in der Frühbewertung nach § 35 a SGB V Belege für einen Zusatznutzen festzustellen. In der Entscheidungspraxis des G-BA wurden bislang selbst Hinweise als nicht ausreichend für die Anerkennung eines Zusatznutzens erachtet. Die Evidenzanforderungen sollten sich daher stärker an der Realität kurz nach der

Zulassung eines Arzneimittels orientieren. Für die frühe Nutzenbewertung wird daher die Berücksichtigung von Evidenzstufen entsprechend dem Standard der internationalen Literatur [Grimes & Schulz, Lancet 2002; 359: 57–61 (14)] vorgeschlagen:

I Evidenz aus mindestens einer korrekt randomisierten kontrollierten klinischen Studie
II-1 Evidenz aus gut angelegten kontrollierten Studien ohne Randomisierung
II-2 Evidenz aus gut angelegten Kohorten- oder Fall-Kontroll-Studien, vorzugsweise von mehr als einem Zentrum oder einer Forschungsgruppe
II-3 Evidenz aus multiplen Zeitserien mit oder ohne Intervention
III Meinungen von angesehenen Experten, gestützt auf klinische Erfahrung, deskriptive Studien oder Berichte von Expertenkommissionen.

Zu Kapitel 7.3.3. Beurteilung klinischer Relevanz (S. 129ff):

Im Kapitel 7.3.3. wird zunächst das Konzept der klinischen Relevanz auf „Systemebene“ dargestellt. Zur besseren Verständlichkeit sollte hier konkretisiert werden, was unter „Systemebene“ zu verstehen ist. In diesem Zusammenhang sollte auch dargelegt werden, warum und mit welcher Konsequenz „ökonomische Überlegungen“ beim Konzept der klinischen Relevanz „auf Systemebene“ eine Rolle spielen. Welcher Zusammenhang besteht zwischen Ressourcen und der Beurteilung klinischer Relevanz? Da dieses Konzept auf „Systemebene“ ausweislich der Ausführungen im Methodenentwurf für die Bewertungen des Instituts von Bedeutung ist, sind hierzu unbedingt nähere Ausführungen erforderlich.

Abgesehen davon sollte das IQWiG Bewertungen nach wissenschaftlichen und nicht systemischen Konzepten durchführen und daher klinische Relevanz unabhängig von ökonomischen Überlegungen beurteilen.

Des Weiteren ist anzumerken, dass die Festsetzung einer fixen Irrelevanzschwelle von 0,2 SMD bei Fehlen validierter bzw. etablierter Relevanzkriterien willkürlich erscheint, da sie nicht näher begründet wird.

Beim Vorschlag der biometrischen Maße für Effektstärken hat Cohen gleichzeitig darauf hingewiesen, dass diese Grenzen relativ sind, vom jeweiligen Forschungsgebiet abhängen und dass die zugrunde gelegten Werte ausschließlich intuitiv gewählt wurden [Cohen 1988]. Die Nachteile der vom IQWiG vorgeschlagenen Methode wurden auch von anderen Autoren beleuchtet [Molnar 2009, S. 537, 539].

Außerdem soll nach der vom IQWiG vorgeschlagenen Methodik für einen Nutzenbeleg die untere Grenze des Konfidenzintervalls oberhalb der Grenze von 0,2 SMD liegen.

Für eine Substanz mit einem kleinen aber nachweisbaren Effekt von 0,2 SMD kann so niemals ein Nutzenbeleg erbracht werden, da die untere Grenze des Konfidenzintervalls selbst bei sehr großen Patientenzahlen immer unter 0,2 liegen wird.

Bundesverband der Pharmazeutischen Industrie e.V.

BPI

Leben ist Vielfalt

Die rein mathematische Ableitung des Nutzens auf der Basis einer berechneten Effektstärke und willkürlich gesetzter Relevanzgrenzen hat einen entscheidenden Einfluss auf die Leistungserstattung. Dies betrifft damit auch Therapiemöglichkeiten in ethisch sensiblen Bereichen wie z.B. in der Krebstherapie. Es bestehen daher erhebliche Zweifel, ob ein starrer Schwellenwert, der methodisch noch nicht einmal international anerkannt ist, in Deutschland rechtlich und gesellschaftlich akzeptabel ist. Dies sollte daher unbedingt korrigiert werden. Die Festlegung von Schwellenwerten sollte vielmehr im Einzelfall Rahmen eines scoping-Workshops unter Einbindung weiterer Experten erfolgen, um ethische Gesichtspunkte angemessen zu berücksichtigen.

Berlin, den 08.04.2011

A 1.4 Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie e.V. (DGPT)

Autoren:

Springer, Anne A.

Gutachter:

Rüger, Bernhard,

DGPT

Deutsche Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie e.V.

Geschäftsführender Vorstand:
Dipl.-Psych. Anne A. Springer (Vorsitzende)
*
Dr.med. Bernhard Janta
Klinik Wittgenstein
Dr.med. Dipl.-Psych. Karsten Münch
Dr.med. Gabriele Friedrich-Meyer
Dr.rer.nat. Dipl.-Psych. Dietrich Munz

Geschäftsstelle:
Dr. rer. pol. Felix Hoffmann, Geschäftsführer
RAin Birgitta Lochner, Justitiarin

Institut für Qualität und Wirtschaftlichkeit
im Gesundheitswesen (IQWiG)
Dillenburger Straße 27

51105 Köln

per E-Mail: methoden@iqwig.de

Allgemeine Methoden
Entwurf für Version 4.0 vom 09.03.2011

Berlin, den 08.04.2011

Sehr geehrter Herr Professor Windeler,
sehr geehrte Damen und Herren,

wir danken für die Zusendung des Entwurfes der Allgemeinen Methoden 4.0, der uns über die AWMF erreicht hat. Sie erhalten anbei eine Stellung, die Professor Rüger, München, im Auftrage der Deutschen Gesellschaft für Psychoanalyse, Psychotherapie, Psychosomatik und Tiefenpsychologie (DGPT) erarbeitet hat.

Wir bitten Sie, diese Stellungnahme bei Ihren weiteren Beratungen zu berück-sichtigen und würden uns über eine Rückmeldung freuen.

Mit freundlichem Gruß

Anne A. Springer
Vorsitzende

Prof. Dr. Bernhard Rüger
Institut für Statistik an der Universität München

Stellungnahme zum Entwurf „Allgemeine Methoden“ des Institutes für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) in der Version 4.0 vom 09.03.2011

Die vorliegende Stellungnahme ist nach den Kapiteln des IQWiG-Entwurfes gegliedert und mündet in der am Schluss gegebenen zusammenfassenden Beurteilung. Die Stellungnahme wurde von der DGPT in Auftrag gegeben; vereinbarungsgemäß sollte darin der IQWiG-Entwurf in erster Linie unter dem Aspekt der Psychotherapieforschung betrachtet werden mit einem besonderen Gewicht auf der Frage, ob durch die einzelnen Methoden des IQWiG-Konzeptes Benachteiligungen der Psychotherapieforschung impliziert werden.

Kapitel 1: Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

In Abschnitt 1.1 stellt sich das Institut mit seinem gesetzlich geregelten Aufgabenbereich vor. Der Auftrag des Gesetzgebers ist weit gespannt: Nach §139a SGB V wird das IQWiG zu Fragen von grundsätzlicher Bedeutung für die Qualität und Wirtschaftlichkeit medizinischer Leistungen tätig, soweit diese innerhalb der gesetzlichen Krankenversicherung erbracht werden. Nach Absatz (4) dieses Paragraphen hat das Institut zu gewährleisten, dass die Bewertung des medizinischen Nutzens nach den international anerkannten Standards der evidenzbasierten Medizin und die ökonomische Bewertung nach den hierfür maßgeblichen international anerkannten Standards, insbesondere der Gesundheitsökonomie erfolgt. Diese gesetzliche Grundlage kann nicht Gegenstand meiner Stellungnahme sein, wohl aber die folgende (wie ich meine ganz entscheidende) Ausweitung des gesetzlichen Rahmens durch das IQWiG:

In der Formulierung von Absatz (4) sind Nutzen und Kosten relativ streng getrennte Bereiche, die nicht unbedingt im Sinne einer Kosten-Nutzen-Analyse zu verstehen sind. Vielmehr hat der Gesetzgeber nach §§35a und b (sowie auch 139a Absatz (3) Punkt 5) die Durchführung von Kosten-Nutzen-Analysen explizit nur für Arzneimittel vorgesehen. Gleichwohl dehnt das IQWiG in seinen „Allgemeinen Methoden“ unter Berufung auf §35a,b den Auftrag von Kosten-Nutzen-Analysen auf „medizinische Technologien“ oder „Gesundheitstechnologien“ allgemeiner Art aus und lässt dabei offen, wie weit gefasst diese Begriffe sind. (Siehe dazu auch die IQWiG-Richtlinien „Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten“ aus dem Jahr 2008.) Hier entsteht der Eindruck, dass das IQWiG eine sehr offene, unbestimmten Ausweitung des gesetzgeberischen Auftrages vornimmt und

Kosten-Nutzen-Analysen auch für solche Therapien durchzuführen beabsichtigt, für die das wegen mangelnder quantitativer Bewertungen sehr viel schwerer möglich ist als auf dem Arzneimittelsektor. Besonders betroffen davon wären dann das Gebiet der Psychotherapie und auch die Psychotherapieforschung.

Auf die evidenzbasierte Medizin gehen die „Allgemeinen Methoden“ des IQWiG in Abschnitt 1.2 ein. Hier habe ich einen kritischen Einwand vorzutragen, der aus den Teilabschnitten 1.2.5 und 1.2.6 resultiert.

Eine von allen Forschern gewünschte hohe Ergebnissicherheit (interne Validität) von Studien wird fast nur randomisierten kontrollierten Studien (RCT-Studien) zugebilligt; an einer Stelle (S. 8/9) heißt es: „Andere Studientypen als randomisierte kontrollierte Studien (Randomized Controlled Trials, RCTs) sind in der Regel für einen Kausalitätsbeweis nicht geeignet. In nicht randomisierten vergleichenden Studien kann grundsätzlich nicht von einer Strukturgleichheit der Gruppen ausgegangen werden. Sie liefern daher immer ein potenziell verzerrtes Ergebnis und können zumeist die maßgebliche Frage nicht hinreichend sicher beantworten, ob ein Unterschied auf der geprüften Intervention beruht.“

Auch die vergleichende Bewertung von interner und externer Validität in Teilabschnitt 1.2.6 fällt eindeutig zugunsten der internen Validität und damit zur Höchstbewertung von RCTs aus. Dazu die folgenden Zitate aus diesem Abschnitt (Seite 9): „Es trifft zu, dass viele randomisierte Studien Aspekte des Alltags der Patientenversorgung nicht abbilden z. B. Patienten mit Begleiterkrankungen ausschließen, die im Alltag häufig sind. Dies ist jedoch keine Konsequenz der Randomisierung, sondern die Folge anderer Faktoren wie z. B. der Festlegung enger Ein- und Ausschlusskriterien der Studie.“ „Aber selbst wenn sich Patientengruppen eines RCTs von Patienten des Versorgungsalltags unterscheiden, muss das die externe Validität des Ergebnisses nicht infrage stellen. Entscheidend ist vielmehr die Frage, ob zu erwarten ist, dass der in einer Population festgestellte Therapieeffekt in einer anderen Population anders ist.“

Diese absolute Auszeichnung von RCTs seitens des IQWiG kann zu Benachteiligungen von Psychotherapiestudien führen, weil bekanntermaßen in der Psychotherapieforschung randomisierte Studien nur selten und wenn, dann mit hohem Aufwand durchführbar sind, vor allem dann, wenn Langzeitpsychotherapien untersucht werden sollen. Darauf gehe ich noch an mehreren Stellen ein.

Kapitel 2:
Produkte des Instituts

Hier werden die verschiedenen Produkte des Institutes ausführlich und klar beschrieben, auch nach ihrem Ablauf und ihrer Erstellung, es ergeben sich keine kritischen Punkte, aus meiner Sicht erübrigt sich eine Stellungnahme zu diesem Kapitel.
.

Kapitel 3:
Nutzen- und Schadenbewertung medizinischer Interventionen

Das Thema dieses relativ umfangreichen Kapitels (es enthält acht Abschnitte) ist zentral und wichtig und in den Abschnitten 3.1, 3.2 und 3.4 auch für Psychotherapiestudien sehr bedeutungsvoll. Nur auf diese Abschnitte gehe ich hier ein.

Der erste Abschnitt ist dem Problem einer Nutzen- und Schadenmessung medizinischer Interventionen gewidmet. Angesichts der Größe und Tragweite dieses Themas fällt der Abschnitt relativ offen und unverbindlich aus. Wahrscheinlich lassen sich aber auch keine allgemein verbindlichen Kriterien zur Messung von Nutzen und Schaden entwickeln, so dass nur derart allgemeine und notgedrungen vage Behandlung des Themas möglich ist. Immerhin wird aber ausführlich auf die Messung durch Surrogate und deren Validierung eingegangen. Eine Benachteiligung der Psychotherapieforschung kann ich (gerade wegen der großen Offenheit der Darstellung) nicht erkennen.

Im zweiten Abschnitt wird auf spezielle Aspekte der Nutzenbewertung eingegangen. Hier möchte ich besonders den Teilabschnitt 3.2.4 über die Studiendauer hervorheben: Das Thema Langzeittherapie wird zwar in Hinblick auf die erforderliche Studiendauer behandelt (was eigentlich selbstverständlich ist), nicht aber bezüglich des geplanten Designs einer Studie. Hier wird die Chance vertan, auf das Problem bestimmter Designkomponenten (Verblindung, Randomisierung u.a.) in Langzeitstudien einzugehen und auf eine damit verbundene Relativierung „harter“ Anforderungen an die Qualität einer Studie. Außerdem wird in Teilabschnitt 3.2.1 eine hohe Ergebnissicherheit einer Nutzenbewertung fast ausschließlich den RCT-Studien zuerkannt und dadurch eine mögliche Benachteiligung von Langzeit- und Psychotherapiestudien geschaffen. Der neu aufgenommene Teilabschnitt 3.2.2 („Auswirkung nicht publizierter Studienergebnisse“) erscheint mir der Sache und Darstellung nach gut und zutreffend gelungen.

In Abschnitt 3.4 wird (endlich einmal!) auf das Gebiet nichtmedikamentöser therapeutischer Interventionen eigens und näher eigegangen: Hier wird einerseits zu recht auf das Problem hingewiesen, dass auf diesem Gebiet Studien mit Randomisierungen oder Verblindungen nur sehr erschwert oder gar nicht durchführbar sind, leider aber daraus keine Konsequenzen gezogen in Hinblick auf eine „natürliche“ Benachteiligung in der Qualitätsbeurteilung von Studien auf diesem Forschungsfeld. Auch hier scheint mir eine Chance vertan zu sein.

Kapitel 4:
Methoden der Versorgungsqualität

Dieses Kapitel ist komplett neu gefasst worden. Es ist wesentlich umfangreicher als sein Vorgänger (Kapitel 4 in Version 2), gleichwohl aber wie er sehr allgemein und nur selten verbindlich gehalten. Es enthält oft Sätze (s. S. 67) wie „ die übergeordneten Ziele einer Versorgungsanalyse sind die Beurteilung der Versorgungsqualität und die Identifizierung möglicher relevanter Lücken in der Versorgung“ oder „eine Versorgungsanalyse wird in

Übereinstimmung mit den Methoden der Versorgungsforschung durch ihre versorgungsrelevante Fragestellung definiert.“ Diese sehr allgemeinen Aussagen finden sich vor allem in den Abschnitten 4.1 und 4.2 und größtenteils auch in 4.4. Sie enthalten sehr einleuchtende und völlig gerechtfertigte Qualitätsanforderungen an die Aufstellung von Leitlinien und Empfehlungen für die öffentliche Gesundheitsversorgung. Wegen der hier vorherrschenden sehr allgemein und offen formulierten Richtlinien können keine tendenziellen Schwerpunkte mit Vor- oder Nachteilen für bestimmte therapeutische Fachrichtungen entstehen. Der Abschnitt 4.3 „Validität von Leitlinienempfehlungen“ ist etwas konkreter gefasst.

Hier werden Instrumente zur Erreichung einer hohen Qualität von Leitlinien und Empfehlungen auch selbstverpflichtend vorgeschlagen. Außerdem werden die für Therapiestunden geläufigen Kriterien der internen und externen Validität erstmals auf das Gebiet der Leitlinien und Empfehlungen übertragen. Dieser neue Ansatz ist sicherlich wichtig und interessant; gleichwohl aber auch problematisch. Wie können auf dem Gebiet von Leitlinienempfehlungen diese beiden Kriterien statistisch überprüft werden? Mir scheint das (auch begrifflich) problematisch zu sein. Auch eine Zurückführung dieser Überprüfung auf die den Leitlinien zugrunde liegenden Studien ist nicht gerade einfach. Darüber hinaus wird in Teilabschnitt 4.3.2 der internen Validität eine Priorität vor der externen zugeordnet, obwohl auf dem Gebiet von Studien längst bekannt ist, dass zwischen diesen beiden Kriterien eher eine Konkurrenz besteht und in vielen Fällen eine hohe interne Validität die externe Validität eher senkt und gerade für Leitlinienempfehlungen eine hohe externe Validität gesichert sein sollte. Mit dieser Priorisierung der internen vor der externen Validität entsteht (wie schon in anderen Kapiteln) der Eindruck, dass das IQWiG den RCT-Studien mit ihrer bekanntlich sehr hohen internen Validität den ersten Platz unter den verschiedenen Studienarten beimisst und damit solche Therapien indirekt benachteiligt, die RCT-Studien nur schwer zugänglich sind.

Kapitel 5:
Evidenzbasierte Gesundheitsinformation für Bürger und Patienten

Zu Beginn des Kapitels heißt es: „Dem Institut wurde vom Gesetzgeber die Aufgabe zugewiesen, allgemeine Gesundheitsinformationen für Bürger und Patienten zu erstellen. Ziel des Instituts ist die Verbesserung der Gesundheit und Patientenautonomie durch die Bereitstellung von Gesundheitsinformationen, die die allgemeine Gesundheitskompetenz und die Wissenschaftskenntnisse fördern sollen.“ In dem Kapitel wird klar und ausführlich beschrieben, wie das Institut diesen Auftrag durchführt bzw. durchführen soll. Der Umfang des Auftrages wird vom IQWiG sehr umfassend ausgelegt, ausgenommen wird nur das Gebiet der Einzelberatung. In dem Kapitel treten kaum fachtherapeutischen Gesichtspunkte oder Bewertungen auf, vorherrschend ist eine neutrale Betrachtungsweise, das ganze Kapitel (insbesondere Abschnitt 5.3) ist sehr gut (differenziert und anwendungsnah) gelungen. Gleichwohl habe ich zwei Kritikpunkte vorzutragen:

In Teilabschnitt 5.3.3 auf Seite 90 sollte neben dem Oxman-Guyatt-Index zur Bewertung der Evidenz auch noch die Skala nach Leichsenring und Rüger einbezogen werden, durch die auch nicht streng kontrollierte bis hin zu naturalistischen Studien in ihrer Evidenz beurteilbar werden. (Falk Leichsenring und Ulrich Rüger[2004] in: Zeitschrift für Psychosomatische Medizin und Psychotherapie, 50, 203-217.) Dazu dann mehr in meiner Stellungnahme zu Kapitel 7 weiter unten.

In Teilabschnitt 5.4.1 auf den Seiten 96/97 erscheint mir die Auswahl der medizinischen Bereiche für die Themen der Berichterstattung sehr willkürlich; das Gebiet psychischer Erkrankungen fehlt vollständig, Psychotherapie taucht auch nicht unter den therapeutischen Optionen auf, auch wenn am Ende von psychosozialen Aspekten gesprochen wird, ist nur das soziale Umfeld und seine Kommunikation gemeint. Zwar sind alle diese Themen beispielhaft gemeint, wenn aber schon aufgelistet wird, sollte eine gewisse Ausgewogenheit vorliegen.

Kapitel 6:
Informationsbeschaffung

Hier werden Ziele, Wege und Komponenten einer sorgfältigen Informationsbeschaffung (bis hin zum Ablauf einer Literaturrecherche) ausführlich beschrieben, es ergeben sich keine kritischen Punkte, aus meiner Sicht erübrigt sich eine Stellungnahme zu diesem Kapitel.

Kapitel 7:
Informationsbewertung

Dieses wichtige, umfangreiche und heterogene Kapitel kann ich am besten an Hand seiner einzelnen Abschnitte beurteilen.

Abschnitt 7.1: Qualitätsbewertung von Einzelstudien

Hier werden die zentralen Richtlinien zur Beurteilung klinischer Studien aufgestellt. Meine Stellungnahme dazu fällt nicht ganz eindeutig aus.

Einerseits wird eine objektive, wissenschaftlich fundierte und gleichwohl sehr weit gefasste Kriterienbildung zur Beurteilung der Qualität klinischer Studien entwickelt mit einer (durchaus auch kritisch vorgetragenen) Bewertung und Rangbildung der verschiedenen Studienarten und ihrer Evidenzgrade (Teilabschnitte 7.1.2 bis 7.1.4). Der Teilabschnitt 7.1.3 enthält eine Rangordnung für die verschiedenen Studienarten mit sieben Stufen: RCT-Studien und systematische Übersichten über RCT-Studien haben höchste Evidenz, danach kommen (in dieser Reihenfolge) nichtrandomisierte Interventionsstudien, prospektive und retrospektive Beobachtungsstudien, nicht experimentelle Studien sowie Expertenmeinungen. Diese in der internationalen Literatur mehr oder weniger klar festgelegte Rangordnung dient dem IQWiG als Orientierung, nicht aber als streng einzuhaltendes Regelwerk. Dass der hier angedeutete Spielraum seitens des IQWiG ernst genommen und verantwortungsvoll genutzt werden soll, wird an verschiedenen Stellen der „Allgemeinen Methoden“ deutlich, nicht zuletzt auch an den in Teilabschnitt 7.1.4 vorgeschlagenen elf „Standards“, unter denen sich auch solche für nichtrandomisierte und nicht experimentelle Studien befinden. Unter Wahrung dieses Spielraumes sind keine Benachteiligungen psychotherapeutischer Studie zu erwarten.

Andererseits werden in Teilabschnitt 7.1.4 für nicht verblindbare Verfahren, die in Studien über Psychotherapien wohl eher die Regel als die Ausnahme bilden, strenge Forderungen zur Vermeidung von (auch subjektiv bedingten) Verzerrungen aufgestellt: Erstens eine vor Beginn der Studie unabhängig vom Behandler festgelegte Zielvariable (die verblindet erhoben oder aber mit einem „harten“ Endpunkt objektiv markiert wird) und zweitens eine randomisierte oder doch wenigstens verdeckte Zuweisung der Patienten auf die in der Studie zu vergleichenden Behandlungsgruppen. Die erste Forderung ist in Psychotherapiestudien

wohl (gerade noch) erfüllbar, die zweite aber nur sehr schwer oder gar nicht mehr: Allein schon wegen der üblichen und dem Patienten rechtlich zustehenden Vorgespräche vor der Aufnahme einer psychotherapeutischen Behandlung ist eine verdeckte Zuweisung kaum möglich. Hierin könnte eine Ursache liegen für eine Benachteiligung psychotherapeutischer Studien gegenüber pharmakologischen und vielen medizinisch-somatischen Studien.

Ähnliches gilt auch für die weiter oben schon beschriebene Priorität, die das IQWiG der internen vor der externen Validität einräumt und der damit verbunden Höchstbewertung von RCT-Studien. Dadurch wird der von mir positiv bewertete Spielraum in der Bewertung von Studien wieder etwas eingeschränkt in Richtung auf eine Benachteiligung (im Sinne einer niedrigeren Evidenzzuweisung) für Psychotherapiestudien wegen ihrer Schwierigkeiten bei der Durchführung von RCT-Studien.

Trotz dieser Einschränkungen überwiegt aber hier (und vor allem auch in Teilbschnitt 7.1.4) eine Offenheit und Ausgewogenheit der vom IQWiG vorgeschlagenen Bewertungskriterien und eine damit verbundene wertfreie Neutralität gegenüber Psychotherapiestudien. Das gilt auch für die drei abschließenden und überzeugenden Teilabschnitte 7.1.5 bis 7.1.7.

Zusammenfassend ist festzustellen: Trotz einer deutlichen Priorität (mit einer erkennbaren Überbewertung) von RCT-Studien werden im IQWiG-Konzept auch andere, weniger strenge Studiendesigns hoch eingeschätzt, darunter auch solche, die in der Psychotherapieforschung besonders wichtig sind und häufig eingesetzt werden. Hier ist festzuhalten, dass das IQWiG für die Beurteilung der Qualität klinischer Studien einerseits (völlig zu Recht) strenge Regeln aufstellt, andererseits aber auch Lockerungen und Abweichungen von diesen Regeln zulässt, für die es dann aber die Kompetenz für eine Evidenzbeurteilung beansprucht.

Abschnitt 7.2: Berücksichtigung systematischer Übersichten

Der Abschnitt ist sehr allgemein und neutral gehalten, er enthält keinerlei Benachteiligungen gegenüber Psychotherapiestudien. Allerdings wird die in Teilabschnitt 7.2.2 behandelte Nutzenbewertung nicht sehr genau ausgearbeitet. Wenn hier auch eine Nutzenbewertung von Therapien gemeint ist, treffen hier implizit die kritischen Einwände aus meinen Anmerkungen zur IQWiG-Methodik für die Bewertung von Nutzen und Kosten aus dem Jahr 2009 zu. Die in Abschnitt 7.2 dargestellte Auswertung systematischer Übersichten ist davon natürlich nur mittelbar betroffen.

Abschnitt 7.3: Spezielle biometrische Aspekte

Die biometrisch statistischen Methoden, die hier vom IQWiG zugrunde gelegt werden, stellen ein aktuelles und ausgewogenes Methodenkonzept dar ohne besondere Schwerpunktsetzung auf irgendwelche medizinischen, pharmakologischen oder psychotherapeutischen Gebiete. Es entstehen daher auch keine Benachteiligungen psychotherapeutischer Forschungsstudien. Der Abschnitt ist, wie ich meine, gut gelungen; erfreulich sind die ausführliche Behandlung meta-analytischer Verfahren und die klare und ausführliche Gegenüberstellung von statistischer Signifikanz und klinischer Relevanz. (Etwas dürftig geraten ist lediglich der Teilabschnitt 7.3.7.)

Abschnitt 7.4: Qualitative Methoden

Ein sehr begrüßenswertes Thema in diesem (berechtigterweise nur) den quantitativen Methoden gewidmeten Arbeitskonzept. Das Thema ist hier als ergänzender Ausblick einbezogen worden. Natürlich könnten sich hier bei einer ausführlicheren Darstellung (falls diese überhaupt schon möglich ist) ganz besondere Aspekte für die Psychotherapie ergeben.

Zusammenfassende Beurteilung:

Die neue Version der „Allgemeinen Methoden“ des IQWiG stellt eine umfassende, wissenschaftlich gut fundierte Rahmenbildung für den vom Gesetzgeber vorgegebenen Aufgabenbereich des Institutes dar.

In der Darstellungsform des IQWiG-Konzeptes fällt eine gewisse Heterogenität bis hin zu manchen Unstimmigkeiten auf, so als ob mehrere Autoren mit deutlich unterschiedlichen Gesichtspunkten das Konzept erarbeitet hätten. Dabei können zwei Argumentationsebenen unterschieden werden: Eine, auf der klare und verbindliche Kriterien und Methoden festgelegt werden, und eine andere, die sehr weit gefasste, allgemeine und oft etwas vage Regeln und Zielvorstellungen enthält.

Nur auf der ersten Ebene gibt es die oben beschriebenen Punkte möglicher Benachteiligungen der Psychotherapieforschung mit Konsequenzen für Umfang und Gewicht der Psychotherapie in der Gesundheitsversorgung. Auf der zweiten Ebene sind solche Punkte naturgemäß nicht zu finden. Hier beansprucht das IQWiG in den Fällen, in denen keine scharfen verbindlichen Kriterien vorliegen (können), eine eigene Entscheidungskompetenz mit der Entwicklung projektspezifischer Methoden.

Generell entsteht der Eindruck, dass sich das IQWiG zu einer sehr eigenständigen, um nicht zu sagen eigenmächtigen Institution entwickelt hat. Dafür spricht auch die eingangs unter Kapitel 1 dargelegte, drch das IQWiG vollzogene Ausweitung des gesetzgeberischen Auftrages. Wahrscheinlich lässt sich diese Entwicklung nicht mehr korrigieren.

Gleichwohl kann sich bei erforderlicher Durchsetzungsfähigkeit die Psychotherapieforschung im Rahmen des IQWiG-Konzeptes vertreten finden. Wie gut das gelingt, hängt einerseits von den Psychotherapieforschern und ihren Studien ab, andererseits aber auch vom IQWiG und seinen seinen zukünftigen Entwicklungen der projektspezifischen Methoden und ihrer Anwendungsfähigkeit auf psychotherapeutische Forschungsfragen.

Ein weiteres wichtiges Kriterium für eine adäquate Vertretung der Psychotherapieforschung im Rahmen des IQWiG-Konzeptes ist das Gewicht, das die vom IQWiG entwickelten „Allgemeinen Methoden zur Bewertung von Verhältnissen zwischen Nutzen und Kosten“ innerhalb des IQWiG-Konzeptes besitzen werden. Die darin entwickelte Analyse von Effizienzgrenzen habe ich in einer eigenen Stellungnahme kritisiert mit dem Ergebnis, dass diese Analyse auf dem Gebiet der Psychotherapie kaum einsetzbar ist. Sollte die Analyse von Effizienzgrenzen zur Nutzen-Kosten-Bewertung wesentlicher Bestandteil des IQWiG-Konzeptes werden, worauf in der Präambel der „Allgemeinen Methoden“ des IQWiG eigens verwiesen wird, so wird eine adäquate Erfassung der Psychotherapie durch das IQWiG-Konzept ernsthaft in Frage destellt..

A 1.5 Deutsche Krankenhausgesellschaft (DKG)

Autoren:
Schlottmann, Nicole

DEUTSCHE
KRANKENHAUS
GESELLSCHAFT

Bundesverband
der Krankenhausträger
in der Bundesrepublik
Deutschland

Herrn
Prof. Dr. J. Windeler
Institut für Qualität und Wirtschaftlichkeit im
Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln

Unser Zeichen	Telefon Durchwahl	Telefax	Datum
V/Dr. Schl/wu	[redacted]	[redacted]	07.04.2011

Stellungnahme der Deutschen Krankenhausgesellschaft zum Entwurf der „Allgemeinen Methoden" Version 4.0 des Institutes für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)

Sehr geehrter Herr Professor Windeler,

in der **Anlage** übersenden wir Ihnen die Stellungnahme der DKG zu den oben genannten allgemeinen Methoden des IQWIG. Nach den zahlreichen intensiven Diskussionen zur Methodenbewertung im Kontext des Gemeinsamen Bundesausschusses sowie des IQWiG freuen wir uns, dass nach drei Jahren nochmals ein überarbeiteter Entwurf vorgelegt wird. Wir begrüßen ausdrücklich, die allgemeinen Methoden regelhaft zu überprüfen und gegebenenfalls zu aktualisieren sowie ein Stellungnahmeverfahren durchzuführen.

Uns hat jedoch die Tatsache irritiert, dass die Veröffentlichung des Entwurfes faktisch ohne jede Vorankündigung innerhalb der Selbstverwaltung erfolgte und gleichzeitig eine Frist von lediglich vier Wochen für die Stellungnahme eingeräumt wurde. Bei aller Wertschätzung vermittelt dieses Vorgehen bei immerhin über 500 Literaturverweisen den Eindruck, dass der Aufwand für eine fundierte Stellungnahme weit unterschätzt oder diese vielleicht nicht ernsthaft erwünscht wird. Unklar bleibt überdies der weitere Fortgang in Bezug auf den Entwurf, beispielsweise ob eine öffentliche Beratung dazu stattfinden wird. Entsprechend den sonstigen Transparenzmaßstäben des IQWiG gehen wir davon aus, dass es auf Grundlage der schriftlichen Stellungnahmen eine mündliche Anhörung und eine öffentliche schriftliche Würdigung der Stellungnahmen geben wird.

Bei der Lektüre entsteht bedauerlicherweise der Eindruck, dass die vielfach geäußerte Kritik an den Methoden des IQWiG keinen Eingang in den Entwurf gefunden hat. Vielmehr entfernt man sich noch weiter von in der Versorgungsrealität Machbarem in Richtung primärer Methodendominanz. Obwohl das IQWiG sich nunmehr im sechsten Jahr seiner Arbeit befindet und damit auf eine nicht unerhebliche Anzahl erstellter Berichte zurückblicken kann, ist es dem Institut nicht gelungen, eine seiner gesetzlich zugewie-

[redacted]

senen Bedeutung Rechnung tragende Akzeptanz in der medizinischen Fachwelt zu erlangen. Anstatt die Kooperation mit den medizinisch-wissenschaftlichen Experten/Fachgesellschaften und Leistungserbringern und damit gemeinsame praxistaugliche Lösungen zu suchen, erfolgt bei zahlreichen medizinisch-inhaltlichen Fragestellungen (z. B. Bewertung von Leitlinien) eine sehr konfrontative Positionierung des Instituts. Aus unserer Sicht wird hierdurch eine wichtige Chance vertan, die methodische Kompetenz des Institutes durch medizinisch-wissenschaftliche Kompetenz zu erweitern und damit zu sachgerechten Bewertungen zu kommen.

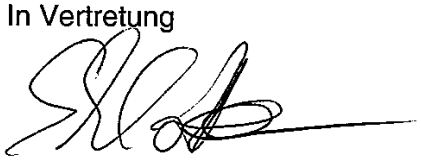
Die Überarbeitung des Methodenpapiers bot aus unserer Sicht überdies eine gute Gelegenheit, auch international abweichende methodische Ansätze im Sinne eines optimalen Ausschöpfens des vielfältigen Methodenrepertoires zu diskutieren und auf dieser Basis das eigene Vorgehen nachvollziehbar zu begründen, was an entscheidenden Stellen leider nicht erfolgt ist.

Bereits in der Vergangenheit hatten wir wiederholt zum Ausdruck gebracht dass das IQWiG seinen Aufgaben im Sinne der im § 139a SGB V festgelegten Grundsätze nicht in angemessener Weise nachkommt. Dort heißt es zu den Aufgaben des IQWIG, die *„Recherche, Darstellung und Bewertung des aktuellen medizinischen Wissenstandes zu diagnostischen und therapeutischen Verfahren bei ausgewählten Krankheiten"* vorzunehmen. Diese Regelung umschreibt sehr deutlich die Darstellung der bestverfügbaren Evidenz (≙ aktueller medizinischer Wissensstand) und deren Bewertung. Eben diese Kenntnis über die bestverfügbare Evidenz, eingebettet in den Gesamtabwägungsprozess, befähigt den G-BA, differenzierte Entscheidungen im Sinne der evidenzbasierten Medizin zu treffen. Leider soll die Darstellung der bestverfügbaren Evidenz auch in den künftigen Berichten des IQWiG zumeist unterbleiben, was sich beispielsweise aus den Vorgaben für die Literaturrecherche ableiten lässt. Hier wird sich in der Regel auf die Suche der „methodisch-theoretisch bestmöglichen" Evidenz (zumeist Stufe 1) beschränkt. Dem vorliegenden Entwurf ist demnach zu entnehmen, dass eine Angleichung der Methoden an den gesetzlichen Auftrag weiterhin nicht beabsichtigt ist. Dies schränkt die Verwertbarkeit der Berichte im G-BA unnötig ein.

Sehr geehrter Herr Professor Windeler, dem IQWiG kommt mit seiner Aufgabe, dem Gemeinsamen Bundesausschuss Grundlagen für seine Entscheidung zur Verfügung zu stellen, eine große Verantwortung im deutschen Gesundheitswesen zu. In jüngerer Zeit sind Entscheidungen des G-BA, wie zum Beispiel zur PET bei malignen Lymphomen nicht zuletzt auch durch Einbeziehung der entsprechenden IQWiG-Berichte erfolgt, die Deutschland medizinisch-wissenschaftlich und leistungsrechtlich international isolieren. Das Methodenpapier in seiner vorliegenden Form lässt befürchten, dass solche Fehlentwicklungen in der Zukunft weiter zunehmen werden. Im Interesse der Patienten können wir Sie daher nur bitten, sich mit unseren Anregungen auseinanderzusetzen.

Abschließend möchten wir noch festhalten, dass wir uns vor dem Hintergrund des Umfanges des Methodenpapiers sowie der nur kurzen Stellungnahmefrist vorbehalten, auch im weiteren Verlauf noch Anmerkungen einzubringen. Zudem weisen wir ausdrücklich darauf hin, dass dieses nicht die Stellungnahme einer Einzelperson, sondern einer Organisation darstellt, die sich im Falle einer Anhörung vorbehält, weitere oder auch andere Personen von Seiten der DKG zu der Anhörung zu entsenden.

Mit freundlichen Grüßen
Der Hauptgeschäftsführer
In Vertretung



Dr. N. Schlottmann
Geschäftsführerin Dezernat Medizin

Anlage

Stellungnahme der Deutschen Krankenhausgesellschaft zum Entwurf der„Allgemeinen Methoden“ Version 4.0 des Institutes für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)

Vorbemerkung:

Ein wesentlicher Aspekt des Entwurfes der Allgemeinen Methoden 4.0 - im Folgenden als „Methodenpapier“ bezeichnet - ist die Umsetzung der Grundsätze der Evidenzbasierten Medizin. Richtungsgebende Äußerungen zu dieser Thematik finden sich in Kapitel 1.2. Bevor weiter unten im Detail auf die einzelnen Kapitel des Methodenpapiers eingegangen wird, möchten wir zu diesen grundsätzlichen Äußerungen Stellung nehmen.

In Kapitel 1.2.1 „Praktische evidenzbasierte Medizin“ heißt es auf S. 4:

„... Evidenzbasierte Medizin ist von der Idee her als Strategie für Ärzte gedacht, die für ihre Patienten unter möglichen Interventionen die vielversprechendsten und deren Bedürfnissen am ehesten entsprechenden Alternativen herausfinden und die Erfolgsaussichten neutral darstellen wollen. Diese Anwendung der evidenzbasierten Medizin in der täglichen Praxis für „individuelle Patienten“ haben im Jahr 1996 David Sackett und Kollegen [118] folgendermaßen definiert: „EbM ist der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen, wissenschaftlichen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten. Die Praxis der EbM bedeutet die Integration individueller klinischer Expertise mit der bestverfügbaren externen Evidenz aus systematischer Forschung.“

Dazu sei Folgendes angemerkt:

- ⇨ In deutlichem Kontrast zu dieser Definition, die bestverfügbare Evidenz abzubilden, werden auch in der überarbeiteten Methodenbeschreibung lediglich idealtypische Studien nach höchsten methodischen Kriterien gefordert, um einer positiven Nutzenbewertung zu entsprechen. Im Methodenpapier wird durchgehend suggeriert, der Ansatz von Sackett und Kollegen sei „lückenhaft oder nicht zuverlässig“ (S. 5). Höchste methodische Anforderungen sind zwar überaus wünschenswert, müssen sich aber in realen Versorgungskontexten bewähren und prüfen lassen. Dabei werden im Methodenpapier durchgängig Ansprüche an die Evidenzgenerierung formuliert, die an entscheidenden Stellen weder ausreichend systematisch referenziert und im Kontext internationaler Literatur diskutiert werden, noch hinreichend die realen Gegebenheiten der Durchführbarkeit und Finanzierbarkeit berücksichtigen. Dazu zählen unter anderem
  - Die unbedingte Forderung nach randomisierten kontrollierten Studien (RCT) und systematischen Reviews über RCT (Evidenzlevel I der Verfahrensordnung des G-BA) bei unzureichender Berücksichtigung aller anderen Evidenzlevel, obwohl international anerkannte Einrichtungen der evidenzbasierten Medizin wie z.B. das britische NICE durchaus weitere Evidenzlevel in ihre Bewertungen einbeziehen, wenn die Studienlage auf höchstem Level unzureichend ist,
  - Die Forderung nach mindestens zwei systematischen Reviews, um diese berücksichtigen zu können, ohne diese Forderung ausreichend durch Referenzen zu belegen,

- Die Forderung nach Äquivalenzstudien, was Fragen zur Festsetzung von Schwellenwerten und zum Aufwand der dazu erforderlichen großen Zahl an Studienteilnehmern aufwirft, um kleine Gruppenunterschiede ausschließen zu können,
- Die Forderung nach Schwellenwerten für die „Standardized Mean Difference" (SMD) von 0,2 samt Konfidenzintervall, die international umstritten ist,
- Die Forderung nach dem primären Einsatz von Random-Effects-Modellen in Meta-Analysen, wodurch die Konfidenzintervalle breiter werden und die Hürde von 0,2 wiederum noch schwieriger zu überwinden ist, obwohl der Einsatz verschiedener Modelle und der Umgang mit Heterogenität international weitaus differenzierter und komplexer beurteilt wird, als auch
- Die lediglich zweistufige Bewertungsskala für Studien, obwohl in der internationalen Literatur mehrstufigere Skalen diskutiert und häufig empfohlen werden, sowie
- Die unzureichende Bereitschaft, Leitlinien regelmäßig zu recherchieren und in ihre Bewertungen einfließen zu lassen. Letzteres gilt offenbar auch für hochwertige Leitlinien mit zugrunde liegenden systematischen Reviews, da keine methodische Differenzierung von z.B. S1- bis S3-Leitlinien nach deutschen Evidenzklassifizierungen vorgenommen wird. Gerade hochwertige Leitlinien haben jedoch durch die enge Verbindung von externer Evidenz mit klinischer Expertise unter Einbeziehung von Experten-Panels hohe Akzeptanz und unmittelbaren Praxisbezug.

⇨ Es sei betont, dass Visionen von idealtypischen Studienkonzepten wichtig sind, um wissenschaftliche Weiterentwicklungen zu befördern; da es sich bei den vom IQWiG vorgenommenen Nutzenbewertungen aber nicht um rein erkenntnisorientierte, zweckfreie Grundlagenforschung, sondern um eine anwendungsorientierte Basis leistungsrechtlicher Bewertungen handelt, die erhebliche Konsequenzen und Leistungseinbußen für die GKV-Versicherten in Deutschland nach sich ziehen können, besteht die Verpflichtung, aktuelle Forschungs- und Versorgungsrealitäten zu berücksichtigen. Die alleinige Berücksichtigung idealer empirischer Bedingungen limitiert die Arbeit des Instituts erheblich, da vorhersagbar ist, dass es kaum Studien geben wird, die all diesen Ansprüchen genügen können, sodass eine positive Nutzenbewertung für viele Verfahren, mit deren Bewertung das Institut beauftragt wird, von vornherein sehr unwahrscheinlich wird. Studien im klinischen Versorgungsalltag bis hin zu Experimenten unter hochkontrollierten Bedingungen können immer wieder Kompromisse zur Durchführbarkeit abverlangen, ohne dass deshalb jeder Erkenntnisgewinn negiert werden muss; vielmehr erfordert dies eine kritische und differenzierte Gewichtung, wie hoch dieser Erkenntnisgewinn einzuschätzen ist. Diese Differenzierung fehlt in weiten Teilen des Methodenpapiers. Erläuterungen dazu finden sich in den detaillierten Kommentierungen und Rückfragen zu den einzelnen Kapiteln.

⇨ Leider erfährt der Leser kaum etwas über den Erstellungsprozess des Methodenpapiers. So wäre es beispielsweise aus Transparenzgründen interessant, wie die einzelnen Ressorts eingebunden wurden und ob eine systematische Literaturrecherche und Bewertung Grundlage dieses

Methodenpapiers darstellen. Der insgesamt narrative Sprachstil lässt in dieser Hinsicht Zweifel aufkommen, zumal an kaum einer Stelle abwägende Ausführungen zur zitierten Literatur zu finden sind.

.

Zu „1 Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen":

Kapitel 1.2 Evidenzbasierte Medizin, S. 4-5

⇨ Die Geschichte der Evidenzbasierten Medizin (EbM) zeichnet sich nicht nur durch die Geschichte der Randomisierung, sondern vielmehr durch die Entwicklung einer transparenten, reliablen und validen Systematik zur Zusammenfassung der aktuellen wissenschaftlichen Erkenntnisse, also der bestverfügbaren Evidenz, aus. Dies führte auch zur Etablierung von Evidenzleveln mit Aussagen auf unterschiedlichem Kausalitätsniveau. Es ist zu empfehlen, diesen wesentlichen Aspekt im Methodenpapier zu ergänzen.

Mehr dazu z.B. unter

- www. cochrane.de
- Egger M, Smith GD. Meta-analysis: Potentials and Promise. BMJ 1997; 315: 1371-1374
- Antes G, Bassler D, Galandi D. Systematische Übersichtsarbeiten. Deutsches Ärzteblatt 1999, Heft 10 (96): 616-622
- Antes G, Koch G. Gesichertes Medizinisches Wissen. Deutsches Ärzteblatt 2000/ Praxis Computer 2: 23-25
- Glass GV. Primary, secondary and meta-analysis of research. Educational Researcher 1976 (5): 3-8
- Schwarzer R. Meta-Analysen: Methodik, Anwendungsbeispiel und Computerprogramm, in: D. Liepmann, G. Mohr, R. Schwarzer (Hrgb.), 1987, Arbeitsberichte des Instituts für Psychologie
- Oxford Centre for Evidence-based Medicine - Levels of Evidence. March 2009, May 2001; Internet: http://www.cebm.net/index.aspx?o=1025

Kapitel 1.2.2 Bedeutung von EbM für das Institut, S. 5

„...Der Auftrag des Instituts zielt also ganz bewusst nicht auf die Behandlung einzelner Patienten mit ihren möglichen Besonderheiten, sondern darauf, für welche Gruppen von Patienten es Belege für einen Nutzen gibt."

⇨ Es handelt sich bei diesen Ausführungen um einen rein utilitaristischen Ansatz, welcher der individuellen medizinischen Versorgung nicht gerecht wird. Es wird dabei die Tatsache, dass dieses Vorgehen gravierende Rückwirkung (z.B. durch Leistungsausschlüsse) auf die Individualversorgung hat, nicht ausreichend beachtet; diese Trennung von Kollektivnutzen und Individualversorgung ist artifiziell. Genau diese Brücke sollte jedoch durch EbM nach Sackett und Kollegen sowie dem Sachverständigenrat für die

konzertierte Aktion im Gesundheitswesen im Versorgungskontext gebaut werden!

Mehr dazu z.B. unter

- Sachverständigenrat für die konzertierte Aktion im Gesundheitswesen: Bedarfsgerechtigkeit und Wirtschaftlichkeit. Band I Zielbildung, Prävention, Nutzerorientierung und Partizipation, Band II Qualitätsentwicklung in Medizin und Pflege. Gutachten 2000/2001. Internet: http://www.svr-gesundheit.de/Gutachten/Gutacht00/kurzf-de00.pdf http://www.svrgesundheit.de/Gutachten/Gutacht00/kurzf-de00.pdf (Zugriff am 03.12.2010)

Kapitel 1.2.4 Die Strategien der EbM, S. 6-7

„3. ... Diese Regeln und Werkzeuge, die international als Standard akzeptiert sind und laufend weiterentwickelt werden, sind die methodische Basis der EbM und der Arbeit des Instituts."

⇨ Bzgl. „internationaler Standards" wäre es hilfreich, zumindest beispielhaft zu definieren, welche diese sind, und einen möglichen Querverweis zu den späteren Kapiteln anzuführen. Wie sind diese definiert, wer ist damit gemeint? Auf internationale Kontexte wird gerade in den ersten Kapiteln häufig verwiesen, ohne dies zu spezifizieren und mit Referenzen zu belegen.

⇨ Es stellt sich dabei auch die grundsätzliche Frage, wie systematisch und umfassend die verschiedenen Kapitel zur Methodik überhaupt recherchiert und internationale Diskurse dargestellt wurden. Um dem Vorwurf der Beliebigkeit vorzubeugen, wären diesbezügliche Erläuterungen geboten.

Kapitel 1.2.6 Die Verbindung von Ergebnissicherheit und Alltagsnähe, S. 8-10

„...Die Randomisierung (zusammen mit einem sorgfältigen Concealment) ist das beste gegenwärtig verfügbare Instrument, um diese Verzerrung zu minimieren. Die zufällige Zuteilung der Teilnehmer zu den Gruppen stellt sicher, dass es keine systematischen Unterschiede zwischen den Gruppen gibt, und zwar weder bezüglich bekannter Eigenschaften (z. B. Alter, Geschlecht, Schwere der Erkrankung) noch bezüglich unbekannter Faktoren...."

⇨ Der Begriff „sicher" sollte durch „wahrscheinlich" ersetzt werden, da es auch zufällig passieren kann, dass die Vergleichsgruppen bzgl. bekannter und unbekannter Confounder nicht gleich verteilt sind. Dann muss man für die nach Baseline-Status ungleich verteilten Variablen adjustieren. Dies ist aber nur für bekannte Confounder möglich, unbekannte können naturgemäß nicht gemessen werden. Auch RCT können keine „Sicherheit" in Anspruch nehmen, da sich auch diese lediglich probabilistischer Modelle und statistischer Aussagen bedienen, die „nur" Wahrscheinlichkeiten berechnen, wenngleich die Aussagekraft im Vergleich zu anderen Studiendesigns die höchste ist – vorausgesetzt, der RCT wurde methodisch korrekt durchgeführt.

„...Andere Studientypen als randomisierte kontrollierte Studien (Randomized Controlled Trials, RCTs) sind in der Regel für einen Kausalitätsbeweis nicht geeignet...."

⇨ Diese Aussage ist problematisch und wirft verschiedene Fragen auf: In der Tat ist das Evidenz- und Kausalitätslevel für RCT und systematische Übersichten von RCT am höchsten. Die Aussagekraft hinsichtlich kausaler Assoziationen

nimmt mit zunehmend höheren Evidenzleveln zu. Allerdings handelt es sich sowohl in der Verfahrensordnung des G-BA als auch in internationalen Evidenzklassifizierungen um eine Vielzahl an kategorialen Abstufungen der Evidenzlevel. So dichotom (RCT: ja versus nein), wie im Methodenpapier an dieser Stelle die Frage kausaler Assoziationen postuliert wird, ist die Sachlage nicht. Die Frage der Kausalität ist auch für andere Studiendesigns überaus bedeutsam! Beobachtende Studien, also keine RCT, können ebenfalls konfirmatorischen Zusammenhanghypothesen nachgehen – dies ist nicht nur RCT vorbehalten (siehe auch S. 44 Methodenpapier IQWIG). Für ätiologische Fragen beispielsweise sind aufgrund ethischer Bedenken keine RCT möglich. So existieren intensive Debatten und auch Standards darüber, ab wann von Zusammenhängen z.B. einer potentiell krebserzeugenden Noxe tatsächlich auszugehen ist oder nicht. Das IQWIG-Postulat berücksichtigt somit einen Großteil epidemiologischer Forschung nicht ausreichend.

Weiteres findet sich dazu in z.B.

- International Agency for Research on Cancer, Lyon, France
- Environmental Protection Agency, USA etc.

⇨ Darüber hinaus existieren auch für Fragen der Wirksamkeit therapeutischer Interventionen international validierte Instrumente zur Qualitätsbewertung von Beobachtungsstudien (Kohorten und Fall-Kontrollstudien: siehe dazu z.B. CASP unter http://www.phru.nhs.uk/casp/casp.htm)

⇨ Der Umstand, dass die Cochrane Collaboration eine „Non-Randomized Methods Group" eingerichtet hat, deutet darauf hin, dass Beobachtungsstudien durchaus zu Fragen kausaler Zusammenhänge/Assoziationen beitragen können.

„... Es trifft zu, dass viele randomisierte Studien Aspekte des Alltags der Patientenversorgung nicht abbilden z. B. Patienten mit Begleiterkrankungen ausschließen, die im Alltag häufig sind. Dies ist jedoch keine Konsequenz der Randomisierung, sondern die Folge anderer Faktoren wie z. B. der Festlegung enger Ein- und Ausschlusskriterien der Studie. Zudem werden Patienten in randomisierten Studien oft anders (intensiver und enger) betreut als im Alltag. Das alles sind aber bewusste Entscheidungen derjenigen, die in einer Studie eine bestimmte Frage beantworten wollen. Durch einen Verzicht auf die Randomisierung werden diese Entscheidungen nicht geändert. Auch in nicht randomisierten Studien findet eine Selektion der Teilnehmer durch Designcharakteristika statt, sodass externe Validität ebenso wenig automatisch besteht wie bei RCTs."

⇨ Bei diesen Aussagen stellt sich die Frage, warum solche Entscheidungen von Untersuchern hinsichtlich eher rigider Ein-/Ausschlusskriterien getroffen werden. Dies unterliegt keiner Beliebigkeit, sondern ist der Tatsache geschuldet, dass sich in homogenen Gruppen Unterschiede durch verschiedene Interventionen einfacher nachweisen lassen und Bias-Quellen dadurch minimiert werden. Bei sehr heterogenen Studienpopulationen können trotz Randomisierung auch andere Effekte als die durch die Interventionen bewirkten zum Tragen kommen, welche nicht immer adjustiert werden können.

„... Aber selbst wenn sich Patientengruppen eines RCTs von Patienten des Versorgungsalltags unterscheiden, muss das die externe Validität des Ergebnisses nicht infrage stellen. Entscheidend ist vielmehr die Frage, ob zu erwarten ist, dass der in einer Population festgestellte Therapieeffekt in einer anderen Population anders ist..."

⇨ Woran wird letzteres festgemacht, empirisch oder theoretisch durch Analogieschluss?

„...Werden beispielsweise sehr breite Einschlusskriterien gewählt, stellt sich im Besonderen die Frage, ob die (Gesamt-)Studienergebnisse auf die gesamte Studienpopulation anwendbar sind [516], was letztlich über adäquate Subgruppenanalysen zumindest ansatzweise zu beantworten wäre."

⇨ Hierzu wären nähere Erläuterungen hilfreich: Subgruppen-Analysen sind oft mit Problemen behaftet, wenn sie nicht a-priori definiert wurden, da Studien primär für die ganze Studienpopulation ausgerichtet sind. Diese Probleme umfassen die Power, Selektionseffekte etc. und reichen bis hin zur Problematik des Data-Mining.

Zu „2 Produkte des Instituts":

⇨ Leider wird in dem Kapitel „Produkte des Instituts" das Produkt *„Allgemeine Methoden"*, also das Methodenpapier selbst, als entscheidende Grundlage der Arbeit des Instituts weder berücksichtigt noch beschrieben:

- Wie sieht das Vorgehen bei der Erstellung des Produkts *„Allgemeine Methoden"* aus? Wie wird die Evidenz berücksichtigt? Welche Systematik liegt dem zugrunde?
- Wie gestalten sich das Stellungnahmeverfahren und die Anhörung zum Methodenpapier?
- Wie sieht das externe Reviewverfahren aus?
- Erfolgt die Erstellung des Dokuments unter Einbeziehung von Sachverständigen, und wenn ja, welchen?

⇨ Generell wäre in Bezug auf dieses Kapitel wünschenswert, dass nach der Auftragserteilung durch den G-BA oder das BMG die Präzisierung und Formulierung der wissenschaftlichen Fragestellungen zu den verschiedenen Produkten des IQWIG (Berichte, Rapid Reports, etc.) in Abstimmung mit dem Auftraggeber vorgenommen wird. Diese Anmerkung begründet sich auch darin, dass im späteren Kapitel unter 6.1.1 (Ablauf der Recherche) auf Seite 107 u.a. angeführt wird, dass die Fragestellung vom IQWiG offenbar selbst konkretisiert wird. Ein Einbeziehen des Auftraggebers erst im Stadium der vorläufigen Version des Berichtsplans oder des fertigen Rapid Reports birgt die Gefahr einer wissenschaftlichen und monetären Ressourcenverschwendung, wenn erst dann Ungenauigkeiten oder Missverständnisse durch die Auftraggeber benannt werden können. Eine Klärung des genauen Vorgehens wäre deshalb sehr hilfreich.

⇨ Darüber hinaus sind zu den Produkten des Instituts die Definitionen und Aufgabenbereiche der Sachverständigen, der externen Gutachter und der externen Reviewer nicht klar voneinander abgegrenzt. So stellt sich die Frage, ob ein Sachverständiger auch ein externer Gutachter werden kann, und die Frage, ob ein externer Gutachter seine Expertise aktiv anbieten kann. Handelt es sich hierbei um unterschiedliche Personengruppen oder gibt es Überschneidungen?

⇨ An mehreren Stellen wird ausgeführt, dass Produkte vor Veröffentlichung zunächst an den Auftraggeber, an das Kuratorium sowie den Vorstand der Stiftung weitergeleitet werden. Es erscheint sinnvoll, bei den Adressaten einer solchen Weiterleitung auch den Stiftungsrat des Instituts mit einzubeziehen.

Kapitel 2.1 Produktspezifische Verfahrensabläufe, S. 11

„...Entsprechend sind Rapid Reports insbesondere für zeitnahe Empfehlungen gedacht, zu denen aus Sicht des Auftraggebers keine Anhörungen durch das Institut erforderlich sind...."

⇨ Aufgrund des fehlenden Anhörungsprozesses bei „Rapid Reports" ist zu erwägen, medizinische Fachexperten nicht nur optional, sondern regelhaft in die Erstellung aller Produkte einzubeziehen.

⇨ In den verschiedenen Kapiteln lässt die *„...Einbeziehung...der Meinung einzelner Betroffener..."* (S.26, Kapitel 2.1.6 Arbeitspapiere), die Berücksichtigung der *„...Patientenperspektive über Patienten bzw. Patientenorganisationen..."* und das Verschicken von Fragebögen an „Patienten bzw. Patientenorganisationen" (S. 19-20, Kapitel 2.1.3 Dossierbewertung) nicht erkennen, ob ein methodisches Vorgehen, welches Repräsentativität anstreben sollte, regelhaft verfolgt wird oder die Auswahl involvierter Patienten bzw. Patientenvertreter nach subjektiven Maßstäben, denen bekanntermaßen ein Selektionsbias innewohnt, erfolgt.

Kapitel 2.1.1 Berichte, S. 13-17

„...Je nach Notwendigkeit wird auch der wissenschaftliche Beirat des Instituts einbezogen...."

⇨ Wie ist diese Notwendigkeit definiert?

„...Regelhaft werden dabei ... entsprechende Patientenorganisationen beteiligt...."

⇨ Aufgrund welcher Kriterien werden Patientenorganisationen, Patienten bzw. Patientenvertreter beim Entstehungsprozess von IQWIG-Produkten wie Berichten, Dossiers, Rapid Reports etc. hinzugezogen? Wie sieht das Auswahlverfahren aus?

„...Zudem wird als weiterer Schritt der Qualitätssicherung der Vorbericht einem oder mehreren externen Gutachtern mit ausgewiesener methodischer und/oder fachlicher Kompetenz vorgelegt."

⇨ Wie erfolgt die Auswahl der Gutachter?

„...Für eine Frist von mindestens vier Wochen wird dann der Öffentlichkeit Gelegenheit zur Stellungnahme (Anhörung) gegeben (schriftliche Stellungnahmen). Die Gelegenheit zur Stellungnahme bezieht sich dabei insbesondere auf das projektspezifische methodische Vorgehen zur Beantwortung der Fragestellung. Die Fragestellung selbst ist i. d. R. durch den Auftrag vorgegeben und nicht Gegenstand des Stellungnahmeverfahrens. Optional kann eine mündliche wissenschaftliche Erörterung mit Stellungnehmenden durchgeführt werden. Diese Erörterung dient der ggf. notwendigen Klärung von Inhalten der schriftlichen Stellungnahmen mit dem Ziel der Verbesserung der wissenschaftlichen Qualität des Berichtsplans..."

⇨ Die Anhörungspraxis sollte näher spezifiziert werden. Zuweilen hatten Anhörungen in der Vergangenheit Tribunal-Charakter und ließen einen ergebnisoffenen Charakter, der wissenschaftlichen Erörterungen

üblicherweise innewohnt, vermissen. Die Tonbandaufnahmen bei zusätzlicher Wortprotokollierung durch einen vereidigten Protokollführer lässt Misstrauen und Argwohn vermuten und löst Befremden aus. Um zu einer konstruktiven Atmosphäre beizutragen, wäre z.B. eine der beiden Protokollierungsvarianten ausreichend.

Kapitel 2.1.2 Rapid Reports, S. 18-19

„...Vor Fertigstellung kann optional ein Entwurf des Rapid Report einem oder mehreren externen Gutachtern mit ausgewiesener methodischer und / oder fachlicher Kompetenz als weiterer Schritt der Qualitätssicherung vorgelegt werden...."

⇨ Wie gestaltet sich dieser weitere Schritt zur Qualitätssicherung genau?

„... Sollten Stellungnahmen zu Rapid Reports eingehen, die substanzielle nicht berücksichtigte Evidenz enthalten, oder erlangt das Institut auf andere Weise Kenntnis von solcher Evidenz, wird dem Auftraggeber begründet mitgeteilt, ob eine Neubeauftragung zu dem Thema (ggf. Aktualisierung des Rapid Reports) aus Sicht des Instituts erforderlich erscheint. Der Auftraggeber entscheidet über die Beauftragung des Instituts..."

⇨ Zudem erscheint es als sinnvoll, alle Stellungnahmen, die nach Veröffentlichung eines Rapid Reports im IQWiG eingehen, an den Auftraggeber weiterzuleiten.

Kapitel 2.1.3 Dossierbewertung, S. 19-20

⇨ Zur Auswahl und Einbindung externer Sachverständiger in die frühe Nutzenbewertung von Arzneimitteln nach § 35a SGB V ist Folgendes anzumerken:

- Im Rahmen der frühen Nutzenbewertung von Arzneimitteln nach § 35a SGB V beabsichtigt das IQWiG, regelhaft externe Sachverständige zu beteiligen. Diese Beteiligung der medizinischen Fachkreise soll mittels eines Fragenkataloges erfolgen, der jeweils zu Beginn des Verfahrens von den externen Sachverständigen auszufüllen ist. Darüber hinaus können die externen Experten im Bedarfsfall zur Klärung spezieller Fragestellungen auch im laufenden Bewertungsverfahren hinzugezogen werden.
- Es ist grundsätzlich zu begrüßen, dass das IQWiG gemäß seinem gesetzlichen Auftrag im Verfahren der frühen Nutzenbewertung von Arzneimitteln die Beteiligung externer Sachverständiger vorsieht. Auch in den Krankenhäusern tätigen Ärzten wird eine Beteiligung als externer Sachverständiger an den Bewertungsverfahren des IQWiG zur frühen Nutzenbewertung empfohlen.
- Das IQWiG sieht jedoch zunächst die Beteiligung von ausschließlich einem externen Experten je Bewertungsverfahren vor. Dies ist angesichts dieses grundlegend neu zu etablierenden Verfahrens der Dossierbewertung sowie der für die Bewertung gesetzlich vorgegebenen engen Bearbeitungsfristen in der Startphase auch sachgerecht. Durch die Beschränkung auf einen externen Sachverständigen werden aber unterschiedliche fachliche Aussagen der externen Experten von vornherein – verfahrensbedingt – ausgeschlossen. Kämen im Falle einer Beteiligung von jeweils zwei

externen Sachverständigen diese zu identischen Aussagen (beispielsweise zur Frage der zweckmäßigen Vergleichstherapie), würde dies die Aussagekraft der Stellungnahmen der externen Experten deutlich erhöhen.

- Es sollte deshalb, nach Auswertung der ersten Erfahrungen mit der Einbindung der externen Expertise, die Beteiligung von regelhaft zwei externen Sachverständigen pro Bewertungsverfahren vorgesehen werden. Dies würde die Aussagekraft der jeweiligen Stellungnahmen der externen Experten und damit die Qualität und den Stellenwert dieser neuen Beteiligungsmöglichkeit für die medizinischen Fachkreise insgesamt deutlich erhöhen.

⇨ Wie gestaltet sich die Zusammensetzung und Pflege der *„institutseigenen Sachverständigendatenbank"*?

Kapitel 2.1.4 Stellungnahme des IQWiG, S. 20

„...Alle Arbeitsschritte werden in Verantwortung des Instituts, bei Bedarf unter Einbeziehung externer Expertise, getätigt...."

⇨ Wie erfolgt die Einbeziehung der externen Expertise und Auswahl der Fachleute genau?

Kapitel 2.1.5 Gesundheitsinformationen, S. 22-24

⇨ Es wird nicht auf die zeitliche Beziehung zwischen IQWiG-Veröffentlichungen und G-BA-Beratungen eingegangen (siehe auch Kap. 5.3). Dieses ist jedoch angesichts der leistungsrechtlichen Entscheidungshoheit des G-BA nicht unerheblich, insbesondere dann, wenn Gesundheitsinformationen möglicherweise anders ausfallen als die Entscheidungen des G-BA getroffen werden. Eine sehr frühzeitige Publikation der Gesundheitsinformationen könnte dann eine öffentliche Beeinflussung unter Vorwegnahme der G-BA-Entscheidung bedeuten.

⇨ Wie gestaltet sich die *„ressortinterne Qualitätssicherung"* für Gesundheitsinformationen genau?

Kapitel 2.2.1 Auswahl externer Sachverständiger, S. 27

„...Für die zu vergebenden Projekte wird anhand einer Kriterienliste der jeweils am besten geeignete Bewerber des entsprechenden Fachgebiets aus der Sachverständigendatenbank ausgewählt und beauftragt...."

⇨ Warum wird die Kriterienliste weder auf der Internetseite noch im Methodenpapier publiziert? Wie ist diese definiert?

Kapitel 2.2.2 Gewährleistung der fachlichen Unabhängigkeit, S. 29:

„...Auf ausdrücklichen Wunsch der externen Sachverständigen, auf Aufforderung des Auftraggebers oder aufgrund anderer wichtiger Umstände ist es möglich, die Namen externer Sachverständiger zur

Gewährleistung ihrer Unabhängigkeit und zur Vermeidung einer interessenbedingten Einflussnahme nicht zu veröffentlichen...."

- ⇨ Werden die Namen der externen Sachverständigen nicht publiziert, kann Transparenz nicht gewahrt werden.
- ⇨ Sowohl die Sachverständigendatenbank des IQWiG als auch die Namen der externen Sachverständigen der jeweiligen IQWIG-Produkte sollten öffentlich gemacht werden. Dies dient der Transparenz. Andernfalls kann nicht geprüft werden, ob Sachverständige beteiligt waren und die diesbezüglichen Regeln eingehalten wurden.
- ⇨ Sollte ein Sachverständiger nicht mit der Veröffentlichung seines Namens einverstanden sein, so müsste er jedoch zumindest einer Offenlegung gegenüber dem Vorstand zustimmen.
- ⇨ Das Hinzuziehen von externen Sachverständigen zur Erstellung des vorläufigen Berichtsplans sollte, wie im vormaligen Methodenpapier festgeschrieben, beibehalten werden. Mögliche sachliche und insbesondere medizinisch-fachliche Vorgehensweisen und Inhalte sollten mit Experten der jeweiligen Fragestellungen abgestimmt sein, um einer wissenschaftlichen und monetären Ressourcenverschwendung vorzubeugen, die entstehen könnte, wenn sich aus dem Review des Vorberichtes grundlegende und zwingende Änderungsvorschläge ergäben.

Kapitel 2.2.3 Begutachtung der Produkte des Instituts, S. 29-30:

„...Alle Produkte einschließlich der jeweiligen Zwischenprodukte unterliegen einem umfangreichen mehrstufigen internen Qualitätssicherungsverfahren. Darüber hinaus wird im Verlauf der Erstellung von Berichten, Rapid Reports und z.T. auch von Gesundheitsinformationen ein externes Reviewverfahren als weiterer Schritt der Qualitätssicherung durchgeführt. Bei Arbeitspapieren und bestimmten Gesundheitsinformationen (siehe Abschnitt 2.1.5) ist das externe Reviewverfahren optional...."

- ⇨ Warum wird der interne Qualitätssicherungsprozess auch hier nicht dargestellt? Wie gestaltet sich dieser? Diesen Prozess transparent zu machen, ist in international anerkannten Einrichtungen durchaus üblich.
- ⇨ Hier besteht ein möglicher Widerspruch in Bezug auf das externe Reviewverfahren. Während im Kapitel 2.2.3 ein externes Reviewverfahren im Verlauf der Erstellung von Rapid Reports als weiterer Schritt der Qualitätssicherung durchgeführt wird, wird die Durchführung dessen im Kapitel 2.1.2 als optional beschrieben.

„...Die Frage, wie effektiv bestimmte Verfahrensweisen beim externen Review sind, wird erst in jüngerer Zeit in gezielten Studien untersucht. Bislang gibt es allerdings nur wenige aussagekräftige Untersuchungen dazu [164,400]...."

- ⇨ Wenn es auch bislang keine belastbare Evidenz für eine bestimmte Anzahl von Gutachern gibt, so sind bis auf weiteres Stellungnahmen von mindestens zwei unabhängig voneinander agierenden Gutachtern zu fordern (siehe auch Qualitätsbewertung der Publikationen einschließlich Mindestanforderungen auf S. 126, Oxman und Guyatt). Von diesen Gutachtern muss sowohl die methodische als auch die fachliche Kompetenz abgedeckt werden. Dies ist umso wichtiger, da die an der Berichterstellung unmittelbar beteiligten

Wissenschaftlerinnen und Wissenschaftler, wie bereits mehrfach von der DKG angemerkt, oftmals nur methodische Kompetenz abdecken. Die gleiche Forderung bezieht sich auch auf den Erstellungsprozess der Rapid Reports. Ferner sollten die Gutachter darüber hinaus bei den Anhörungen anwesend sein, um einen methodischen und gleichermaßen medizinisch-inhaltlichen Diskurs zu ermöglichen und die Argumentation anderer Experten aus erster Hand nachvollziehen zu können. Nach dem Eingang von Stellungnahmen und ggf. Anhörungen sollte der Abschlussbericht ebenfalls durch zwei Gutachter, die insgesamt sowohl die methodische als auch die fachliche Kompetenz abdecken, auf mögliche Veränderungen und deren Implikationen hin überprüft werden.

„...Die Identifikation und die Auswahl potenzieller externer Reviewer sind abhängig vom Umfang des beauftragten Reviews. Sehr umfangreiche externe Reviews können auch als wissenschaftliche Forschungsaufträge vergeben werden. Für diese gelten dann die in Abschnitt 2.2.1 genannten Bedingungen. Ansonsten kann die Identifikation externer Reviewer durch eine entsprechende Recherche, durch die Kenntnis der Projektgruppe, durch das Ansprechen von Fachgesellschaften, durch eine Bewerbung im Rahmen der Ausschreibung für die Auftragsbearbeitung usw. erfolgen. Eine Darlegung potenzieller Interessenkonflikte muss aber in jedem Fall erfolgen. Die Auswahl der externen Reviewer erfolgt durch das Institut. Eine Höchstgrenze von Reviewern gibt es nicht. Die externen Gutachten werden hinsichtlich ihrer Relevanz für das jeweilige Produkt geprüft. Eine Veröffentlichung der externen Gutachten erfolgt nicht...."

- ⇨ Wie genau erfolgt die Auswahl externer Reviewer (Methodiker, medizinische Fachexperten etc.)? Was sind die Kriterien? Die Auswahl und Integration medizinischer Fachexperten ist dabei besonders relevant, da dies in zentraler Weise die Akzeptanz der Berichte und des gesamten Instituts betrifft. Beispiele für Akzeptanz sind die Leitlinienerstellung des ÄZQ oder des NICE mit ihren großen Experten-Panels.
- ⇨ Dadurch, dass das IQWIG für seine eigenen Produkte die externen Reviewer selbst aussucht, ist deren Unabhängigkeit nicht gewährleistet, was zu Verzerrungen führen und die Qualität des Reviewverfahrens in Frage stellen kann. Im Rahmen der Publikationsprozesse in Fachjournalen sind gegenüber dem Autor die Reviewer oftmals anonymisiert. Autoren können zuweilen zwar Wünsche hinsichtlich der Gutachter angeben, die Festlegung erfolgt dann aber vom Journal, nicht vom Autor selbst.
- ⇨ Es könnte auch eine Veröffentlichung der externen Gutachten stattfinden. Dies wird im Rahmen von Peer-Review-Verfahren teilweise auch in internationalen Fachjournalen so gehandhabt.

„..Die Auswahl der internen und externen Gutachter erfolgt primär auf Basis ihrer methodischen und/oder fachlichen Expertise..."

- ⇨ Aufgrund welcher spezifischen Kriterien werden externe Gutachter ausgewählt? Welche Aufgaben haben sie? Handelt es sich um nationale oder internationale Experten? Können Sachverständige externe Gutachter werden?

„...Die Namen der externen Gutachter von Berichten und Rapid Reports werden i. d. R. im Abschlussbericht bzw. Rapid Report veröffentlicht, einschließlich einer Darstellung ihrer potenziellen Interessenkonflikte, analog zur Vorgehensweise bei externen Sachverständigen...."

- ⇨ Werden die Namen von externen Gutachtern nicht publiziert, kann Transparenz nicht gewahrt werden. (siehe auch Kommentare zu Kapitel 2.2.2)

„...Neben dem oben beschriebenen externen Qualitätssicherungsverfahren unter Beteiligung vom Institut ausgewählter und beauftragter Gutachter ist durch die Veröffentlichung der Institutsprodukte und die damit verbundene Möglichkeit zur Stellungnahme ein offenes und unabhängiges Reviewverfahren gewährleistet...."

⇨ Die Beschreibung des Qualitätssicherungsverfahrens fehlt.

Kapitel 2.2.4 Veröffentlichung der Produkte des Instituts, S. 30

„...Zur Wahrung der Unabhängigkeit des Institutes muss ausgeschlossen werden, dass die Auftraggeber oder interessierte Dritte Einfluss auf die Inhalte der Berichte nehmen können. Dies könnte zu einer Vermengung der wissenschaftlichen Ergebnisse mit politischen und/oder wirtschaftlichen Aspekten und/oder Interessen führen. Gleichzeitig muss vermieden werden, dass das Institut seinerseits bestimmte Ergebnisse zurückhält...."

⇨ Die Vermengung ist de facto gar nicht zu verhindern, da sowohl die Trägerorganisationen als auch die Auftraggeber politische Einrichtungen sind.

Zu „3 Nutzen- und Schadenbewertung medizinischer Interventionen":

Kapitel 3.1.1 Definition des patientenrelevanten medizinischen Nutzens

S.31

„...Die Bewertung der Evidenz soll nach Möglichkeit in eine eindeutige Feststellung münden, dass entweder das Vorliegen eines (Zusatz-)Nutzens (bzw. Schadens) einer Maßnahme oder das Fehlen eines (Zusatz-)Nutzens (bzw. Schadens) belegt ist oder das Vorliegen oder Fehlen eines (Zusatz-) Nutzens (bzw. Schadens) nicht belegt und daher unklar ist, ob ein (Zusatz-)Nutzen (bzw. Schaden) durch die Maßnahme erzielt wird...."

⇨ Der Auftrag an das IQWiG besteht darin, die Evidenz hinsichtlich eines Nutzens oder Zusatznutzens einer Maßnahme zu prüfen. Die Bewertung eines durch eine Maßnahme möglicherweise verursachten Schadens wird hingegen im Rahmen der Zulassung einer Maßnahme geprüft. Zudem sollte statt des Begriffs des „Schadens" eher der Terminus „Risiko" verwendet werden, da anders als beim Nutzen, bei dem eine möglichst hohe Eintrittswahrscheinlichkeit angestrebt wird, die Wahrscheinlichkeit des Eintritts eines Schadens naturgemäß möglichst gering ausfallen sollte. Dieses Kontinuum von Wahrscheinlichkeiten und die damit verbundenen Abwägungsprozesse werden durch den Begriff „Risiko" bzw. des „Nutzen-Risiko-Verhältnisses" im Kontext der „Nutzen-Risiko-Abwägung" treffender abgebildet. Auf diese Problematik wird in den Anmerkungen zum Kapitel 3.3 („Nutzenbewertung von Arzneimitteln") ausführlicher eingegangen.

⇨ Die vom G-BA in der Vergangenheit an das IQWiG vergebenen Aufträge im Bereich der nicht-medikamentösen Verfahren betreffen nahezu ausschließlich komplexe Themenfelder wie die Positronen-Emissions-Tomographie (PET) oder die Stammzelltransplantation. Die Entwicklung in diesen Bereichen ist sehr dynamisch, zudem handelt es sich hier wie auch sonst häufig um seltene und lebensbedrohliche Erkrankungen, die naturgemäß eine besonders umsichtige und verantwortungsvolle Bewertung erforderlich machen. Bei diesen Erkrankungen ist es oft schwierig, zu eindeutigen und generalisierbaren Studienergebnissen zu gelangen, u.a. aufgrund der Betrachtung unterschiedlicher Krankheitsstadien und Unterschieden in den jeweiligen Vorbehandlungen sowie wegen der Problematik der

„Standardtherapie", die nicht selten im Wandel ist (deutlich wird dies z.B. am Beispiel des Einsatzes der „novel agents" bei der Behandlung des multiplen Myeloms). U.a. aus diesen Gründen können die an vielen Stellen im Methodenpapier suggerierten dichotomen Entscheidungen oftmals nicht getroffen werden. Auch ist darauf hinzuweisen, dass der Begriff des „Zusatznutzens" im SGB V nur im Kontext der Bewertung von Arzneimitteln gebraucht wird – dies sollte in Bezug auf die nichtmedikamentösen Verfahren rechtskonform deutlicher dargestellt werden.

S.32

⇨ Mit dem Zitat [70] wird Bezug auf das Urteil des Bundessozialgerichtes vom 06.05.2009 genommen. Eine Präzisierung, auf welchen Teil (Randnummern) der Entscheidung Sie rekurrieren, erscheint geboten. Auch ist zu prüfen, inwieweit diese den G-BA betreffenden Ausführungen des Gerichts auf die Arbeit des IQWiG übertragen werden können. Dies ist kritisch zu bewerten und in Anbetracht des wissenschaftlichen Auftrags des Instituts zu empfehlen, Referenzen dieser Art im Methodenpapier nicht zu verwenden.

S.32

„...Sowohl Nutzen- als auch Schadenaspekte können eine unterschiedliche Wichtigkeit für die Betroffenen haben, die sich ggf. durch qualitative Erhebungen oder bereits bei der Beratung durch Betroffene, Patientenvertretungs- und/oder Verbraucherorganisationen im Zusammenhang mit der Definition patientenrelevanter Endpunkte abzeichnet. In einer solchen Situation kann es sinnvoll sein, eine Hierarchisierung von Endpunkten vorzunehmen."

⇨ Die Hierarchisierung von Endpunkten ist nicht trivial und beinhaltet komplexe Werteentscheidungen. Wenn das Institut planen sollte, eine solche Hierarchisierung vorzunehmen, so wäre unter Bezug auf validierte Instrumente darzustellen, nach welchen Kriterien das Institut hierbei vorgehen will.

Kapitel 3.1.2 Surrogate des patientenrelevanten medizinischen Nutzens, S. 33

⇨ Im vorliegenden Entwurf zur Version 4.0 der Allgemeinen Methoden werden die Anforderungen an die Validierung von Surrogatparametern konkretisiert. Dies geschieht laut Pressemitteilung des IQWiG vom 11.03.2011 auf Basis des Rapid Reports A 10-05 „Aussagekraft von Surrogatparametern in der Onkologie". Dieser wurde jedoch erst kurz vor Veröffentlichung des vorliegenden Methodenentwurfs am 28.02.2011 publiziert und deshalb bisher weder in der Fachwelt noch im beauftragenden G-BA selbst ausreichend diskutiert. Hierzu ist anzumerken:

- Insbesondere in der Onkologie basieren viele Zulassungen von Arzneimitteln auf Nachweisen der Beeinflussung von Parametern des Tumoransprechens (i.e. z.B. Verbesserung des krankheitsfreien Überlebens, objektive Ansprechrate, Zeit bis zur Progression, Zeit bis zum Therapieabbruch oder progressionsfreies Überleben). Zum Zeitpunkt der Zulassung können i.d.R. noch keine Langzeitstudien mit Daten zur Mortalität oder Morbidität vorgelegt werden, ohne dass die Zulassung des neuen Arzneimittels sich erheblich verzögern würde. Patienten müssen jedoch insbesondere bei schwerwiegenden Erkrankungen frühzeitig Zugang zu neuen Therapiemöglichkeiten erhalten. Die Verwendung von Surrogatparametern geschieht genau aus diesem Spannungsfeld heraus.

- Bei der frühen Nutzenbewertung von Arzneimitteln nach § 35a SGB V ist bereits unmittelbar nach der Zulassung durch den G-BA bzw. das IQWiG eine Bewertung auf Basis der zu diesem Zeitpunkt vorliegenden Zulassungsstudien vorzunehmen. Dabei ist durch den G-BA auf dieser Grundlage zu entscheiden, ob das neue Arzneimittel einen Zusatznutzen im Vergleich zu den bisherigen Therapiealternativen besitzt oder nicht. Deshalb ist als Entscheidungsgrundlage für die Verfahren der frühen Nutzenbewertung die grundsätzliche Klärung notwendig, ob und in welchem Ausmaß diese Parameter als valide Surrogatendpunkte zur Bewertung bzw. Feststellung eines Zusatznutzens von neu zugelassenen Arzneimitteln geeignet sind.
- Der Gesetzgeber hat für das Verfahren der frühen Nutzenbewertung nach § 35 a SGB V mehrere Vorgaben getroffen: Zum einen wurde ausweislich der Gesetzesbegründung ausdrücklich deutlich gemacht, dass der frühe Zeitpunkt der Bewertung zu berücksichtigen ist. Nach § 2 Absatz 3 der Verordnung über die Nutzenbewertung von Arzneimitteln (AM-NutzenV) wird weiter konkretisiert, dass unter dem Nutzen eines Arzneimittels der patientenrelevante therapeutische Effekt zu verstehen ist, insbesondere hinsichtlich der Verbesserung des Gesundheitszustands, der Verkürzung der Krankheitsdauer, der Verlängerung des Überlebens, der Verringerung von Nebenwirkungen oder einer Verbesserung der Lebensqualität. Weiter wird durch § 5 Absatz 5 der AM-NutzenV ausdrücklich klargestellt, dass für den Fall, dass valide Daten zu patientenrelevanten Endpunkten noch nicht vorliegen können, die Bewertung auf Grundlage der verfügbaren Evidenz unter Berücksichtigung der Studienqualität mit Angabe der Wahrscheinlichkeit für den Beleg eines Zusatznutzen erfolgt. Darüber hinaus kann der G-BA in diesem Falle eine Frist setzen, bis wann valide Daten zu patientenrelevanten Endpunkten vorgelegt werden sollen.

⇨ Der Gesetzgeber hat somit eine Bewertung des Zusatznutzens durch den G-BA bzw. das IQWiG auf Basis von Surrogatendpunkten mit den beschriebenen Restriktionen – Berücksichtigung der Studienqualität und insbesondere Angabe der Wahrscheinlichkeit eines Zusatznutzens – ausdrücklich vorgegeben. Die Ausführungen des IQWiG in Abschnitt 3.1.2 des vorliegenden Methodenentwurfes tragen diesem Anliegen des Gesetz- bzw. Verordnungsgebers jedoch nicht ausreichend Rechnung:

- Zum einen werden zur Validierung von Surrogaten durch das IQWiG Kriterien angeführt, die noch nicht, als „allgemein akzeptiert" bezeichnet werden können (u.a. Meta-Analysen von mehreren randomisierten Studien, die in Patientenkollektiven und Interventionen durchgeführt werden müssten, die Aussagen über das der Nutzenbewertung zugrundeliegende Anwendungsgebiet und die zu bewertende Intervention sowie die Vergleichsintervention erlauben). Zudem sind diese, da ein Surrogat die Kriterien zum Zeitpunkt der frühen Bewertung gerade aufgrund der fehlenden Endpunktstudien nicht erfüllen kann, wenig hilfreich. Weiterhin wird den Besonderheiten schwerwiegender, seltener Erkrankungen nicht genug Rechnung getragen, wie dies noch in der Version 3 der „Allgemeinen Methoden" erfolgt ist. Ein weiterer Aspekt der noch intensiver diskutiert werden muss, ist die Frage der Bestimmung der Patientenrelevanz der klinischen Endpunkte. Im Rapid Report wurden die „Zeit zur Progression einer Erkrankung" als Surrogat bezeichnet. Diesem

mag man zustimmen, wenn man die „Zeit bis zur Progression" als Indikator für das „Gesamtüberleben" heranziehen möchte. Ob die „Zeit zur Progression" aber von den Patienten als relevant angesehen wird, wird damit nicht diskutiert.

- Wie auch durch das IQWiG auf Seite 34 ausgeführt, gibt es bisher weder ein Standardverfahren noch allgemein akzeptierte Kriterien, deren Erfüllung den Nachweis der Validität eines Surrogats bedeuten würde. Die Verwendung eines Surrogatparameters wird vielmehr immer eine gewisse Unsicherheit bergen müssen. Bei einer belegten 100%igen Validität eines Surrogates, würde es sich nicht mehr um ein Surrogat, sondern um einen klinischen Endpunkt handeln. Für Interventionen, die bei Bewertung schon längerfristig verfügbar sind, ist es akzeptabel zu fordern, dass klinische Endpunktstudien für die Bewertung vorliegen müssen.

⇨ Im Ergebnis sind die vom IQWiG geforderten Kriterien zur Validierung eines Surrogats zum Zeitpunkt der frühen Bewertung nicht erfüllbar, da die geforderten Daten zeitlich noch nicht vorliegen können. Es erscheint wenig hilfreich, dass ein Surrogat nur dann als valide akzeptiert wird, und damit nur dann ein „Nutzenbeleg" erreicht werden kann, wenn im Grunde kein Surrogat mehr benötigt wird, da die Daten zu den klinischen Endpunkten bereits vorliegen (da sie zum Beweis der Validität benötigt werden). Im Entwurf des Methodenpapiers wird im Ergebnis lediglich ausgeführt, dass Surrogatendpunkte von nicht gegebener oder nicht adäquater Validierung nicht als Beleg für einen Nachweis des (Zusatz-) Nutzens einer Intervention geeignet sind. Allerdings ist schon jetzt eine abgestufte Einstufung der Wahrscheinlichkeit des Zusatznutzens möglich, und das Methodenpapier sollte dieser Tatsache auch Rechnung tragen.

⇨ In diesem Zusammenhang ist zwar grundsätzlich zu begrüßen, dass das IQWiG die Einführung einer neuen Kategorie zur Verlässlichkeit von Aussagen vorsieht. Dazu sollen die bisherigen Ergebniskategorien „Beleg" (hohe Sicherheit) und „Hinweis" (mittlere Sicherheit) um die Kategorie „Anhaltspunkt" ergänzt werden, womit ausgedrückt werden soll, dass eine Aussage nur mit geringer Sicherheit möglich ist. Leider ist dem vorliegenden Abschnitt zur Verwendung von Surrogaten aber an keiner Stelle zu entnehmen, dass die Einführung der Kategorie „Anhaltspunkt" auch den Zweck hat, „die Verlässlichkeit von Aussagen zu Nutzen und Schaden auf Basis von Surrogatendpunkten kenntlich zu machen"; dazu werden bislang lediglich in der o.g. Pressemitteilung vom 11.03.2011 entsprechende Aussagen gemacht.

⇨ Zudem basiert der vorliegende neue Abschnitt „Surrogate des patientenrelevanten medizinischen Nutzens" nach Aussage des IQWiG, wie bereits dargestellt, auf den Ergebnissen des Rapid Reports „Aussagekraft von Surrogatparametern in der Onkologie". Die nur sehr allgemein gehaltenen diesbezüglichen Ausführungen im vorliegenden Methodenentwurf auf eineinhalb Seiten bilden die zentralen Ergebnisse des Rapid Reports und das darin vorgeschlagene methodische Vorgehen aber nur sehr unzureichend ab und sind deshalb weiter zu konkretisierten. Insgesamt wäre es erforderlich gewesen, das im o.g. Rapid Report entwickelte Methodenkonzept zumindest in Grundzügen im vorliegenden Methodenentwurf darzustellen- dies umso mehr, da es international bislang noch kein Standardverfahren und keine allgemein akzeptierten Kriterien zur Validierung von Surrogaten gibt und den

Fachkreisen bisher noch keine Gelegenheit zur Stellungnahme und fachlichen Diskussion des vom IQWiG vorgeschlagene Methodenkonzepts zur Surrogatvalidierung gegeben wurde.

⇨ Zusammenfassend ist festzustellen, dass der vorliegende Abschnitt das methodische Vorgehen des IQWiG zur Validierung von Surrogatendpunkten nicht ausreichend und nachvollziehbar beschreibt. Die Ausführungen sind weiter zu konkretisieren. Insbesondere ist darzustellen, welche Methoden der Validierung von Surrogaten zugrunde gelegt werden und anhand welcher Kriterien Schlussfolgerungen zur Aussagesicherheit der Ergebnisse getroffen werden – hier insbesondere, welche Anforderungen das IQWiG an die unterschiedlichen Ergebniskategorien „Beleg", „Hinweis" und „Anhaltspunkt" stellt.

Kapitel 3.1.3 Ermittlung des Schadens medizinischer Interventionen

S. 37

„...Die Zusammenstellung erfolgt im Rahmen der Vorrecherche zur jeweiligen Fragestellung insbesondere auf Grundlage der Daten kontrollierter Interventionsstudien, in denen zielgerichtet der Nutzen der Intervention untersucht wurde, sowie ggf. auf Basis vorliegender epidemiologischer Studien (zum Beispiel Kohorten- oder Fall-Kontroll-Studien), von Pharmakovigilanzdaten, Informationen von Zulassungsbehörden etc. Im Einzelfall können hier auch Ergebnisse aus Tierexperimenten sowie aus Experimenten zur Überprüfung eines pathophysiologischen Konstrukts hilfreich sein."

⇨ Es ist als bemerkenswert einzuschätzen, dass in Bezug auf die Bewertung eines „Schadens" (zur Problematik dieses Begriffes siehe unsere voherigen Ausführungen) offensichtlich andere Ansprüche an die Evidenz gestellt werden, da auch die untersten Stufen der „Evidenztreppe" Berücksichtigung finden und tierexperimentelle Studien herangezogen werden sollen. Unstrittig ist, dass – gerade bei seltenen und schweren Komplikationen oder unerwünschten Ereignissen – eine Beschränkung auf vergleichende Studien nicht sinnvoll ist. Für eine methodisch schlüssige Abwägung von „Nutzen" und „Schaden" ist es jedoch nicht nachvollziehbar, dass das Institut beim „Nutzennachweis" eine rigide Beschränkung auf möglichst randomisierte, kontrollierte Studien vornimmt (Evidenzstufe I), in Bezug auf den Schaden jedoch Studien der Evidenzstufe V heranzuziehen bereit ist.

⇨ Grundsätzlich möchten wir – auch in Bezug auf Kapitel 3.1.1 – darauf hinweisen, dass die im Methodenpapier verwendete Terminologie der Bewertung von „Nutzen" und „Schaden" eher durch den international gebräuchlichen Begriff der Nutzen-Risiko-Abwägung ersetzt werden sollte (siehe auch unsere Ausführungen zu Kapitel 3.3).

Kapitel 3.1.4 Zusammenfassende Bewertung

S. 38-39

„...Eine Möglichkeit der gemeinsamen Würdigung von Nutzen und Schaden ist die Gegenüberstellung der endpunktbezogenen Nutzen- und Schadenaspekte. Dabei werden die Effekte auf alle Endpunkte (qualitativ oder semiquantitativ) gegeneinander abgewogen mit dem Ziel zu einer endpunktübergreifenden Aussage zum Nutzen bzw. Zusatznutzen einer Intervention zu kommen. Eine weitere Möglichkeit der gemeinsamen Würdigung besteht darin, die verschiedenen patientenrelevanten Endpunkte zu einem einzigen Maß zu aggregieren. In diesem Fall würden die Aussagen des Instituts für jeden einzelnen patientenrelevanten Endpunkt gewichtet z. B. in einen

Summenscore einfließen. Die genaue Durchführung der gemeinsamen Würdigung von Nutzen und Schaden ist themenspezifisch und sollte – wenn dies prospektiv möglich ist – im Berichtsplan und ansonsten im Vorbericht beschrieben werden. Eine quantitative Gewichtung unter Verwendung von Summenscores sollte prospektiv zum Zeitpunkt der Auswahl der zu untersuchenden Endpunkte erfolgen."

⇨ Bereits bei der Diskussion der IQWiG-Methodik zu Kosten und Nutzen wurde deutlich, wie problematisch die Aggregierung von Endpunkten ist. Das IQWiG hat es auch in dieser Revision der allgemeinen Methoden unterlassen auszuführen, wie dies im Allgemeinen aussehen könnte und auf welche Methodik dabei rekurriert wird. Auch muss berücksichtigt werden, dass dieser Abwägungsprozess dem Aufgabenbereich des G-BA zuzuordnen ist. Es ist daher zu empfehlen, diesen Aspekt in diesem Kapitel zu streichen. Alternativ wäre eine deutlich detailliertere Darstellung der geplanten allgemeinen Methodik notwendig – ein Rekurs auf die Notwendigkeit einer themenspezifischen Ausarbeitung reicht hierbei nicht aus.

Kapitel 3.2 Spezielle Aspekte der Nutzenbewertung

Kapitel 3.2.3 Dramatischer Effekt

S. 42

„...Dabei sollten auch mögliche Schäden der Maßnahme berücksichtigt werden. Glasziou et al. [186] versuchen, die Einordnung als dramatischer Effekt zu operationalisieren. In einer ersten Näherung wird vorgeschlagen, einen beobachteten Effekt dann als nicht mehr allein durch die Einwirkung von Störgrößen erklärbar anzusehen, wenn er auf dem Niveau von 1 % signifikant ist und als relatives Risiko ausgedrückt den Wert 10 übersteigt [186]."

⇨ Es ist zu begrüßen, dass dieser Aspekt thematisiert wird. Es wird ausgeführt, dass – unter Bezugnahme auf Glasziou et al. [186] – ein „dramatischer Effekt" bei einem relativen Risiko (RR) von über 10 als gegeben angesehen wird. Gleichzeitig wird jedoch darauf verwiesen, dass es sich hierbei nur um eine „erste Näherung" handelt. Wir halten es für erforderlich, dass in den Berichten des Instituts in jedem Fall eine Darstellung entsprechender Studien erfolgt, auch wenn der Effekt kleiner ausfallen sollte. Es sollte auch ergänzt werden, weshalb hier nur Bezug auf „Fallserien" genommen wird – die Einbeziehung auch anderer Studientypen (z.B. vergleichende Beobachtungsstudien) erscheint geboten. Der Nachweis dramatischer Effekte wäre zudem bei der im Methodenpapier vorgenommenen Definition nur sehr selten zu erwarten, im Bereich chronischer Erkrankungen dürfte ein Nachweis auf dieser Basis – trotz ggf. hoher Effektstärke - zumeist nicht gelingen. In anderen Kontexten z.B. im Rahmen der Anerkennung von Berufskrankheiten wird eine ähnliche Diskussion mit allerdings weitaus niedrigeren Effektstärken als 10, d.h. lediglich zur Risikoverdopplung mit RR=2 geführt. Eine Einbeziehung entsprechender Arbeiten erscheint deshalb sinnvoll:

- Morfeld, P.; Piekarski, C.: Anerkennung von Berufskrankheiten aus der Sicht der Epidemiologie – Missverständnis und Missbrauch des Kriteriums der Risikoverdopplung. Zbl. Arbeitsmed. 51 (2001), 276-285
- Morfeld, P.; Straif, K.: Zur Anerkennung von Berufskrankheiten. Zbl. Arbeitsmed. 51 (2001), 254-261
- Seidler, A.; Pfahlberg, A.; Hornung, J.; Elsner, G.; Gefaller, O.: Anerkennung von Berufskrankheiten – Anwendung der

Risikoverdopplung und alternativer Kiterien. Zbl. Arbeitsmed. 51 (2001), 286-295

- Fritze J, Mehrhoff F (2008): Die ärztliche Begutachtung. Rechtfragen, Funktionsprüfungen, Beurteilungen, Steinkopff Verlag, Würzburg
- Becker P (2007): Die wesentliche Bedingung – aus juristischer Sicht. In: Med Sach 103 3/2007

Kapitel 3.2.5 Patientenberichtete Endpunkte

S. 43

„...Für die in Abschnitt 3.1.1 beschriebenen patientenrelevanten Nutzendimensionen können auch patientenberichtete Endpunkte (Patient Reported Outcomes, PROs) zum Einsatz kommen. Neben der gesundheitsbezogenen Lebensqualität und der Behandlungszufriedenheit können auch andere Nutzendimensionen mittels PRO erfasst werden, wie z. B. Symptome der Erkrankung. Wie auch für die Erfassung der Lebensqualität und der Behandlungszufriedenheit sind hierfür Instrumente zu fordern, die für den Einsatz in klinischen Studien geeignet sind [148]. Bei der Auswahl der zu berücksichtigenden Evidenz (insbesondere Studientypen) für den Nachweis eines Effekts gelten i. d. R. die gleichen Prinzipien wie bei anderen Endpunkten [167]. Das heißt, dass auch im Falle von PROs einschließlich gesundheitsbezogener Lebensqualität und Behandlungszufriedenheit randomisierte kontrollierte Studien am besten für den Nachweis eines Effekts geeignet sind. Da Angaben zu PROs aufgrund ihrer Natur subjektiv sind, sind offene Studien in diesem Bereich nur von eingeschränkter Validität... Es gibt empirische Evidenz für ein hohes Verzerrungspotenzial bezüglich subjektiver Endpunkte in offenen Studien [521]. Dies ist bei der Interpretation solcher Studien zu beachten (siehe auch Abschnitte 7.1.4 und 7.3.4). Allerdings sind Situationen denkbar, in denen eine Verblindung von Ärzten und Patienten nicht möglich ist. In solchen Situationen sind – sofern möglich – andere Anstrengungen zur Verzerrungsminimierung bzw. -einschätzung (z. B. verblindete Endpunkterhebung und -bewertung) zu fordern..."

⇨ Hier sind ausführlichere Aussagen bzgl. der die Validität von PROs betreffenden Probleme unter genauerer Benennung des Verzerrungspotentials notwendig. Lebensqualität, Zufriedenheit, etc. sind zwangsläufig subjektiv, dennoch existieren diesbezüglich validierte Messverfahren. In diesem Zusammenhang ist darauf hinzuweisen, dass das Institut bei der Bewertung der nichtmedikamentösen lokalen Verfahren zur Behandlung des benignen Prostatasyndroms (N04-01 vom 02.06.2008) seine Bewertung nahezu ausschließlich auf patientenberichtete Endpunkte (v.a. IPSS- bzw. AUA-Score) gestützt hat, während es objektivere Parameter wie die maximale Harnflussrate als Surrogat unklarer Validität gewertet hat. Hier scheint eine klärungsbedürftige Inkonsistenz vorzuliegen.

Kapitel 3.2.6 Nutzen in kleinen Populationen, S. 43-44

⇨ Hier wird Bezug auf Patienten mit seltenen Erkrankungen genommen. Leider bleibt unklar, welche Größenordnungen hier konkret angenommen werden. Es ergibt sich die Frage, wie seltene, sehr seltene und extrem seltene Erkrankungen hier definitorisch voneinander abgegrenzt werden.

Kapitel 3.3 Nutzenbewertung von Arzneimitteln, S. 44-49

⇨ Zur Verwendung des Begriffs „Schaden" im Kontext der Nutzenbewertung von Arzneimitteln durch das IQWiG ist folgendes anzumerken:

- Das IQWiG kann seit Inkrafttreten des AMNOG durch den G-BA sowohl mit der frühen Nutzenbewertung nach § 35a SGB V als auch weiterhin auf Grundlage des § 139a SGB V mit einer ausführlichen Nutzenbewertung beauftragt werden. Mit § 35b SGB V sowie konkretisierend für das Verfahren der frühen Nutzenbewertung mit § 2 Abs. 3 AM-NutzenV stellt der Gesetz- bzw. Verordnungsgeber klar, dass zur Bewertung des Nutzens insbesondere die Verbesserung des Gesundheitszustands, eine Verkürzung der Krankheitsdauer, eine Verlängerung der Lebensdauer, eine Verringerung der Nebenwirkungen sowie eine Verbesserung der Lebensqualität zu berücksichtigen sind. Das IQWiG hingegen spricht im vorliegenden Methodenpapier im Rahmen der Nutzenbewertung von Arzneimitteln von „positiven" Aspekten (Nutzen) und demgegenüber „negativen" Aspekten (Schaden).
- Die Verwendung des Begriffs des „Schadens" im Rahmen der Nutzenbewertung von Arzneimitteln durch das IQWiG erscheint aufgrund folgender Aspekte nicht angemessen:
    - Im Rahmen der Nutzenbewertung soll nach SGB V wie oben skizziert die Verringerung der Nebenwirkungen ausdrücklich als wesentlicher Aspekt des Nutzens für den Patienten berücksichtigt werden. So ist beim Vergleich zweier Arzneimittel, bei denen ein neues Arzneimittel A weniger Nebenwirkungen hat als Arzneimittel B, dies ein Vorteil oder ein Nutzen für Arzneimittel A. Wenn das neue Arzneimittel im umgekehrten Falle höhere Nebenwirkungen als Arzneimittel B aufweist, ist dies ein Nachteil bzw. geringerer Nutzen von Arzneimittel A gegenüber B.
    - Darüber hinaus wurde durch das AMNOG ausdrücklich klargestellt, dass der G-BA – und damit auch das IQWiG – die im Rahmen der Arzneimittelzulassung geprüften Kriterien der Qualität, Wirksamkeit und Unbedenklichkeit eines Arzneimittel nicht abweichend von der Beurteilung der Zulassungsbehörde bewerten darf. Eine systematische Nutzen-Schaden-Abwägung bei der Nutzenbewertung von Arzneimitteln durch das IQWiG widerspricht somit den sozialgesetzlichen Grundlagen. Damit die Bewertungen und Aussagen des IQWiG auch in Zukunft rechtskonform in die Entscheidungen des G-BA einbezogen werden können, bedarf es in diesem Punkt einer notwendigen Korrektur des vorliegenden Methodenentwurfs. Es muss klargestellt werden, dass das IQWiG im Rahmen der Bewertung von Arzneimitteln auf Grundlage der durch das AMNOG konkretisierten Rechtsgrundlagen tätig wird.
    - Darüber hinaus widerspricht die Verwendung des Begriffs „Schaden" im Rahmen der Arzneimittelnutzenbewertung auch den im Arzneimittelgesetz (AMG) vorgenommenen Begriffsbestimmungen. Im AMG, dessen Ziel die Sicherheit im Verkehr mit Arzneimitteln ist, findet sich schon bei den Begriffsdefinitionen eben nicht der Begriff des „Schadens" sondern der Begriff der Nebenwirkung, für die bei bestimmungsgemäßen Gebrauch auftretende nachteilige

unbeabsichtigte Reaktion (§ 4 Nr. 13 AMG). Weiterhin wird im AMG der Begriff des Nutzen-Risiko-Verhältnisses eingeführt. Hierunter versteht der Gesetzgeber nach § 4 Nr. 28 AMG eine Bewertung der positiven therapeutischen Wirkungen des Arzneimittels im Verhältnis zu jedem Risiko (im Zusammenhang mit der Qualität, Sicherheit oder Wirksamkeit) für die Gesundheit der Patienten oder die öffentliche Gesundheit. Wenn das Nutzen-Risiko-Verhältnis ungünstig ist, muss eine Zulassungsbehörde die Zulassung versagen (§ 25 Abs.2 Nr. 5 AMG) oder zurücknehmen (§ 30 Abs. 1 AMG). Der Begriff „Schaden" taucht in diesem Zusammenhang ebenfalls nicht auf. Auch auf EU-Ebene wird analog von „benefit-risk-assessment" gesprochen. Bei der Bewertung des Nutzen-Risiko-Verhältnisses berücksichtigen die Zulassungsbehörden neben den gemessenen Vor- und Nachteilen, den Unsicherheiten und Risiken auch verschiedene Patienten- und Krankheitscharakteristika und die zur Verfügung stehenden Therapiealternativen.

- Da der Gesetzgeber im Rahmen des SGB V und des AMG den Begriff des „Schadens" bewusst vermeidet, sollte zur Umsetzung dieser gesetzlichen Vorgaben auch das IQWiG in Zukunft im Rahmen der Nutzenbewertung von Arzneimitteln von der Verwendung dieses negativ besetzten Begriffes absehen. Stattdessen sollte von **Nebenwirkungen** und Nachteilen, die sich in einem geringeren Nutzen ausdrücken **und** - um die Unsicherheiten in Bezug auf die Relevanz einer Nebenwirkung zu berücksichtigen - von **Risikopotenzial** oder Nutzen-**Risiko**-Abwägung gesprochen werden, wie dies auch im Kontext der Arzneimittelzulassung üblich ist.

S. 44

„...Aufgrund der Zielsetzung der Nutzenbewertung durch das Institut werden in die jeweilige Bewertung nur Studien einer Evidenzstufe eingeschlossen, die zum Nachweis des Nutzens grundsätzlich geeignet ist. Studien, die lediglich Hypothesen generieren können, sind deshalb im Allgemeinen für die Nutzenbewertung nicht relevant. Die Frage, ob eine Studie einen Nachweis eines Nutzens erbringen kann, hängt im Wesentlichen von der Ergebnissicherheit der erhobenen Daten ab...."

⇨ Durch diese Aussagen wird suggeriert, dass grundsätzlich nur RCTs für den Nachweis eines Nutzens geeignet sind, andere Studientypen somit lediglich hypothesengenerierend sein können. Dies entspricht nicht den „international anerkannten Standards der evidenzbasierten Medizin", denen das Institut gemäß § 139a Abs. 4 SGB V verpflichtet ist, und stellt eine unzulässige Verkürzung dar.

Kapitel 3.3.2 Studien zur Nutzenbewertung von Arzneimitteln
S. 46

„... Das Studiendesign hat insofern erheblichen Einfluss auf die Ergebnissicherheit, als dass mit Beobachtungsstudien, prospektiv oder retrospektiv, ein kausaler Zusammenhang zwischen Intervention und Effekt in der Regel nicht dargestellt werden kann, während die kontrollierte Interventionsstudie grundsätzlich hierfür geeignet ist [198]...."

- ⇨ Der Begriff der Ergebnissicherheit wird im Methodenpapier überstrapaziert. Auch RCTs können nur Wahrscheinlichkeitsaussagen treffen. Als alternative Terminologie könnte treffender von „Aussagekraft der Assoziation" oder des Zusammenhangs im Sinne eines hohen oder weniger hohen Kausalitätsniveaus gesprochen werden.
- ⇨ Bezüglich des Postulats, dass Kausalität nur von RCT begründet werden kann, ist anzumerken, dass im Rahmen epidemiologisch-ätiologischer Fragestellungen sicher andere Maßstäbe gelten (siehe Ausführungen zu Kapitel 1.2.6). Die damit verbundenen grundsätzlichen Überlegungen machen deutlich, dass auch zum Wirksamkeitsnachweis für Interventionen entsprechender Erkenntnisgewinn aus Beobachtungsstudien ableitbar ist:
  - Gordis L (2009): Epidemiology. Saunders Elsevier, Philadelphia
  - Rothmann KJ, Greenland S, Lash TL (2008): Modern Epidemiology. Lippincott Williams & Wilkins, Philadelphia

Kapitel 3.4 Nichtmedikamentöse therapeutische Interventionen, S. 49-50

- ⇨ Umfang und Inhalt dieses Kapitels, das wesentliche Grundlagen der Bewertung nichtmedikamentöser Verfahren des Instituts beinhalten sollte, stehen leider weiterhin in keiner angemessenen Relation zur Relevanz dieser Thematik in Anbetracht der diversen bereits erfolgten und zukünftig geplanten Beauftragungen des Instituts durch den G-BA. Weitere Präzisierungen und Ergänzungen sind erforderlich. In diesem weitgehend propädeutisch formulierten Kapitel fällt auf, dass einige Themen und Problemstellungen durchaus benannt werden, jedoch nicht der aktuelle Stand des wissenschaftlichen Diskurses dargestellt wird. Der Aussage, dass es notwendig sein kann, auch nicht randomisierte Studien einzubeziehen, ist z.B. nachdrücklich zuzustimmen, jedoch finden diese tatsächlich oft genug keinen Eingang in die abschließende Bewertung des Instituts. Auch die Verweise auf die Verfahrensordnung sollten überprüft werden.

Kapitel 3.5 Diagnostische Verfahren, S.50-52

- ⇨ Diesem Kapitel fehlt insgesamt eine Darstellung des aktuellen Standes der wissenschaftlichen Basis der Bewertung von diagnostischen Verfahren. Diese ist insgesamt nicht unstrittig, hierbei sei auch exemplarisch auf die *„Cochrane diagnostic test accuracy working group"* verwiesen. In der Diskussion des IQWiG mit nuklearmedizinischen Experten im Rahmen von Anhörungen zur PET wurden zudem etliche Probleme deutlich, z.B. die Unmöglichkeit der Validierung von Ergebnissen durch einen „Goldstandard" (z.B. Biopsie). Es kann erwartet werden, dass das Institut bei der nunmehr vierten Auflage seines Methodenpapiers solche Erkenntnisse mit einfließen lässt, Unsicherheiten benennt und Vorschläge für einen sachgerechten Umgang hiermit darstellt. In Bezug auf die Verwendung der Klassifikation von *Fryback und Thornbury* ist eine differenziertere Betrachtung der verschiedenen Evidenzstufen einschließlich der diagnostischen und prognostischen Güte erforderlich. In Anbetracht der nachfolgenden leistungsrechtlichen Tragweite der Bewertungen des Instituts muss hier eingedenk der international noch im Fluss befindlichen Bewertungskriterien diagnostischer Verfahren eine

wissenschaftliche Auseinandersetzung mit den kontroversen Punkten erfolgen.

⇨ Die isolierte Betrachtung der Wirkung einzelner diagnostischer Verfahren auf patientenrelevante Endpunkte kann auf dem Hintergrund einer Vielzahl an diagnostischen und therapeutischen Maßnahmen der oft genannten „diagnostisch-therapeutischen Einheit" (z.B. D06-01A, S. vi) sehr schwierig sein und hohe Zahlen an Studienteilnehmern erfordern, welche ihrerseits die Machbarkeit von Studien limitieren.

S. 50

„...Allgemein kann der Prozess zur Evaluierung diagnostischer Verfahren in Analogie zur Arzneimittelprüfung in verschiedene, hierarchisch angeordnete Phasen oder Stufen eingeteilt werden [174,291]. Prospektiv geplante, kontrollierte diagnostische Studien der Phase 4 nach Köbberling et al. [291] bzw. Studien der Stufe 5 nach Fryback und Thornbury [174] mit einer (idealerweise zufälligen) Zuteilung von Patienten zu einer Strategie mit bzw. ohne Anwendung der zu prüfenden diagnostischen Maßnahme oder zu Gruppen mit bzw. ohne Offenlegung der (diagnostischen) Testergebnisse können dabei in Entsprechung zu (Arzneimittel-)Zulassungsstudien der Phase 3 („Wirksamkeitsstudien") gesehen werden. Dementsprechend wird ihnen auch der höchste Evidenzgrad zugesprochen (siehe z. B. Verfahrensordnung des G-BA [181]). Auch für die Zulassung für bestimmte Indikationen von Arzneimitteln und biologischen Produkten, die im Zusammenhang mit bildgebenden diagnostischen Methoden entwickelt werden, werden von der FDA derartige Studien empfohlen [166], und Beispiele zeigen, dass sie – je nach erwartetem Effekt – auch mit vergleichsweise moderatem Aufwand durchführbar sind [493]."

⇨ Der Verweis auf die Verfahrensordnung des G-BA erscheint hier nicht korrekt, da in dieser kein Bezug auf die Klassifikation von Fryback und Thornbury genommen wird. Zudem werden in Kapitel 2 § 11 Abs. 2 auf der höchsten Evidenzstufe unter „Ic" bereits "andere Interventionsstudien" genannt.

⇨ In Bezug auf die Einschätzung des Aufwands der Durchführung klinischer Studien ist Zurückhaltung anzuraten. Es dürfte auch dem Institut bekannt sein, dass sich insbesondere durch die AMG-Novelle stark erhöhte Anforderungen an die Planung und Durchführung von Studien ergeben haben, hierauf haben Stellungnehmende beim IQWiG mehrfach hingewiesen. Insbesondere bei diagnostischen Verfahren, die mit einer Strahlenbelastung der Patienten einhergehen, sind zudem besondere strahlenschutzrechtliche Vorgaben zu beachten und langwierige Prüfungen erforderlich. Im Bereich der PET wurde von Stellungnehmenden z.B. auf die PETAL-Studie hingewiesen, die mit einem finanziellen Aufwand in 7-stelliger Höhe (Eurobeträge) einhergeht – dies als „moderat" zu bezeichnen, erscheint nicht sachgerecht.

S. 51

"...Liegen solche Studien nicht oder in nicht ausreichender Quantität und Qualität vor, kann eine Bewertung der sogenannten diagnostischen Kette erfolgen [345]."

⇨ Die Einbeziehung des Konzeptes der „linked evidence" ist zu begrüßen, hierzu sind jedoch weitere Ausführungen im Methodenpapier erforderlich, wie sie z.B. in dem zitierten Papier des australischen Medical Services Advisory Committee (MSAC) [345] getroffen werden.

S. 51

„...Ggf. können allein solche Studien, auch ohne spezielle Bewertung der diagnostischen Güte, für einen Nutzennachweis eines diagnostischen Verfahrens herangezogen werden [322], wenn sich

daraus mit ausreichender Sicherheit Aussagen zur Wechselwirkung zwischen diagnostischer Information und (zumeist therapeutischer) Konsequenz ableiten lassen."

⇨ Die Bedeutung der diagnostischen und prognostischen Güte als Lebensqualitätsaspekt wird nicht angesprochen.

S. 52

„... Der Stellenwert ... letztlich unkontrollierter Studien im Rahmen von Nutzenbewertungen diagnostischer Verfahren muss als weitgehend unklar angesehen werden. Informationen zu Managementänderungen allein können deshalb nicht für einen Nutzennachweis herangezogen werden, so lange keine Informationen über die patientenrelevanten Auswirkungen solcher Änderungen vorliegen...."

⇨ Auch hier zeigt sich erneut der grundlegende Dissens über das Wesen des Nutzennachweises. Wie in den allgemeinen Anmerkungen zu Kapitel 3.5 bereits vermerkt, kommen andere Institutionen (z.B. die US-amerikanischen Centers for Medicare and Medicaid Services (CMS) auf der Basis des Nachweises von Managemententscheidungen (Stufe 4 nach Fryback and Thornbury)) bereits zu einer positiven Bewertung. Auch wenn dem IQWiG im Gegensatz zu den CMS keine leistungsrechtliche Bewertung obliegt, so erfordern grundlegende Unterschiede zu internationalen Gepflogenheiten eine umfassende und nachvollziehbare wissenschaftliche Begründung.

S. 52

„...Auch bei der Bewertung von Diagnostika kann es notwendig sein, den Zulassungsstatus zu beachten (siehe Abschnitt 3.3.1)."

⇨ Hierbei ist unklar, was in Bezug auf diagnostische Verfahren mit „Zulassungsstatus" gemeint ist. Oftmals dürfte eine Beauftragung des IQWiG durch den G-BA aufgrund von Bewertungsverfahren nach den §§ 135 und §137c SGB V erfolgen – in diesem Kontext von „Zulassung" zu sprechen, erscheint nicht sachgerecht. Die abweichenden gesetzlichen Regelungen sind zu beachten.

Kapitel 3.6 Früherkennung und Screening

⇨ Allgemein ist anzumerken, dass Umfang und Ausführlichkeit dieses Kapitels nicht mit dem Stellenwert dieses Bereiches – auch in Bezug auf bisherige und künftige Beauftragungen durch den G-BA – korrelieren. Die Ausführungen im Methodenpapier erschöpfen sich in allgemeinen propädeutischen Aussagen. Eine Darstellung des aktuellen, ggf. auch kontroversen wissenschaftlichen Diskurses unterbleibt. Hierzu ist exemplarisch Folgendes anzumerken:

- Die Testgenauigkeit ist neben der Wirkung auf die Krankheitslast von großer Bedeutung und sollte nicht nur dann herangezogen werden, wenn der Nutzen in Bezug auf das Morbiditäts- oder Mortalitätsgeschehen nicht hinreichend oder gar nicht untersucht ist.
- Die Testgenauigkeit ist deshalb so bedeutsam, da Screenings definitionsgemäß in der symptomlosen und zumeist auch gesunden Allgemeinbevölkerung durchgeführt werden. So ist die Kenntnis der Wahrscheinlichkeit für falsch-positive Ergebnisse essentiell, da diese zu unnötigen weiteren diagnostischen und therapeutischen Interventionen führen könnten.

- Wie ist „früh“ oder „spät“ zu definieren? Wie wird mit dem Problem der zeitlichen Vorverlagerung umgegangen?

Kapitel 3.7 Prävention

⇨ Auch bei diesem Kapitel fällt – insbesondere in Anbetracht der bereits erfolgten Beauftragung durch das BMG ([P10-01] „Primärprävention und Gesundheitsförderung bei Männern“) – auf, dass eine umfassende wissenschaftlich begründete Darstellung der Methodik des Instituts fehlt. Ergänzungen erscheinen dringend geboten.

Kapitel 3.8 Prognosestudien

⇨ Insgesamt erschließt sich die Notwendigkeit einer – gerade im Vergleich zu den beiden vorgenannten Kapiteln – derart ausführlichen Neufassung eines Abschnittes zu „Prognosestudien“ nicht. Solche Studien werden vornehmlich im Rahmen der Bewertung von diagnostischen Verfahren und von Maßnahmen zur Früherkennung/Screening herangezogen. Eine ausführlichere Darstellung des wissenschaftlichen Diskurses in den beiden vorgenannten Kapiteln ist erforderlich, ggf. könnten Teilaspekte dieses Kapitels dort verortet werden. Exemplarisch hierzu folgende Anmerkung:

S. 54-57
„...Ein prognostisches Merkmal liefert eine Information, die letztlich keinen Selbstzweck haben kann, sondern zu einer Konsequenz führen sollte, die ihrerseits einen prüfbaren Nutzen für den Patienten bedeutet....“

⇨ Ein Informationsgewinn ist nicht nur Selbstzweck, sondern kann erheblichen Einfluss auf die Krankheitsbewältigung und Lebensqualität haben (z.B. bei Tumorpatienten, siehe Ausführungen zu Kapitel 3.6).

Zu „4 Methoden der Versorgungsqualität“:

⇨ Dieses Kapitel stellt scheinbar ein neues Feld der „Versorgungsanalyse“ dar, auf dem das IQWiG offensichtlich tätig werden möchte. Für dieses sehr umfängliche Unterfangen ist jedoch weder einen gesetzlicher Auftrag noch eine Begründung durch den Generalauftrag des G-BA auszumachen. Außerordentlich befremdlich wirkt überdies, dass in den Ausführungen zur „Versorgungsanalyse“ in erheblichen Maße Inhalte übernommen wurden, die im Rahmen eines Modellprojektes des G-BA von der AG „Versorgungsorientierung/Priorisierung“ erarbeitet wurden, ohne dass ein deutlich sichtbarer Verweis auf dieses Modellprojekt erfolgt ist. Nimmt man die Aspekte, welche sich hinter den in diesem Kapitel besonders thematisierten Begriffen „Leitlinien“ und „Qualität“ in Bezug auf die gesetzlichen Aufgaben des Instituts nach § 139 a SGB V verbergen, so wären in diesem Kapitel eindeutig zuzuordnende Ausführungen zur methodischen Herangehensweise zu folgenden Punkten zu erwarten:

- Bewertungen evidenzbasierter Leitlinien für die epidemiologisch wichtigsten Krankheiten

- Abgabe von Empfehlungen zu Disease-Management-Programmen
- Erstellung von wissenschaftlichen Ausarbeitungen, Gutachten und Stellungnahmen zu Fragen der Qualität und Wirtschaftlichkeit der im Rahmen der gesetzlichen Krankenversicherung erbrachten Leistungen unter Berücksichtigung alters-, geschlechts- und lebenslagenspezifischer Besonderheiten

⇨ Auf die Ausführungen zur „Versorgungsanalyse" sollte daher im Methodenpapier verzichtet werden. Ein gesetzlicher oder ein Auftrag des G-BA existiert nicht. Wenngleich auch die Ausführungen zur Versorgungsanalyse (z. B. Problematik Abgrenzung Versorgungsanalyse von Nutzenbewertung) durchaus einer kritischen Kommentierung bedürften, wird vor dem Hintergrund der Empfehlung auf generelle Streichung in dieser Stellungnahme inhaltlich nicht weiter darauf eingegangen.

Kapitel 4.1 Hintergrund, S. 58

„Leitlinien sind wissenschaftlich basierte Instrumente, die den Umgang mit Krankheitsbildern innerhalb eines Gesundheitssystems oftmals über die gesamte Versorgungskette hinweg abbilden. Im Vergleich zur Nutzenbewertung besitzen sie oftmals eine geringere Informationstiefe. Sie können jedoch normativ Standards in allen Bereichen der Versorgungskette beschreiben, sei es Diagnostik, Behandlung, Rehabilitation oder Nachsorge. Versorgungsstandards beinhalten wesentliche Informationen über die in einem Gesundheitssystem angestrebte Versorgungsqualität. Die Bestimmung eines Versorgungsstandards ist eine zentrale Voraussetzung, um Aussagen über die Versorgungsqualität in einem Gesundheitssystem treffen zu können.."

⇨ Die pauschale Aussage einer geringeren Informationstiefe von Leitlinien ist insbesondere in Hinblick auf Leitlinien auf dem Niveau S3 u. E. nicht haltbar. Die Informationstiefe ist gerade dann als sehr hoch einzuschätzen, wenn neben aufwendigen systematischen Reviews zur Evidenzlage praktische Erwägungen im Versorgungskontext und verschiedene Evidenzstufen mit unterschiedlichen Empfehlungsgraden in diesen Leitlinien berücksichtigt werden. Gerade weil die gesamte Versorgungskette (Kliniker etc.), ergänzt durch Methodiker, im Erstellungsprozess eingebunden ist, ist besonders bei diesen Leitlinien die Akzeptanz sehr hoch. Ganz anders verhält es sich dagegen nicht selten mit der Akzeptanz von isolierten Nutzenbetrachtungen einzelner Aspekte der Versorgungskette, insbesondere, wenn sie fast ausschließlich von Methodikern ohne klinischen Alltagsbezug erstellt wurden. Evidenzbasierte Leitlinien hingegen sind der Archetyp für die gelebten Grundgedanken der Evidenzbasierten Medizin, u. a. bestehend aus der Integration von externer und interner Evidenz.

Kapitel 4.2.2 Methodische Bewertung von Leitlinien, S. 59

„...Das von einem Netzwerk von Forschern und Gesundheitspolitikern entwickelte und validierte AGREE-Instrument .... ist international am weitesten verbreitet und liegt mittlerweile in 13 Sprachen vor. Auch das deutschsprachige DELB-Instrument der AWMF und des ÄZQ basiert auf dem Bewertungsinstrument der AGREE Collaboration. Um ggf. einen Vergleich der Ergebnisse der Leitlinienentwicklung des Instituts mit den in anderen Studien veröffentlichten Leitlinienbewertungen zu vereinfachen, wird für die methodische Bewertung von Leitlinien im Institut regelhaft das AGREE verwendet. An der Weiterentwicklung des DELB-Instrumentes arbeitet das Institut aktiv mit. ..."

⇨ Es stellt sich die Frage, ob die Vergleichsmöglichkeit mit anderen Studien wirklich eine ausreichende Begründung darstellt, regelhaft das AGREE-Instrument zu verwenden. Es wirkt unglücklich, dass ein hochwertiges, in Deutschland entwickeltes Bewertungsinstrument wie DELBI, welches im Übrigen auch vom G-BA im Rahmen seiner Beratungen zu den DMP-Programmen für die Bewertung von Leitlinien herangezogen wird, hier nachrangig gesehen wird, zumal sich DELBI auch noch eng an AGREE anlehnt.

Kapitel 4.3. Validität von Leitlinienempfehlungen/ 4.3.1 Hintergrund, S.62-63

„Für viele Fragestellungen reicht die Überprüfung der methodischen Qualität einer Leitlinie daher nicht aus, um die Wertigkeit einzelner Empfehlungen einzuschätzen [204]. Hieraus ergibt sich die Notwendigkeit, dass Inhalte von Leitlinien analysiert und überprüft werden müssen, insbesondere in Bezug auf die Validität der Empfehlungen.
Es existiert eine Vielzahl methodischer Instrumente, die mit dem Ziel entwickelt wurden, eine hohe Qualität von Leitlinien zu erreichen (insbesondere GRADE, ADAPTE, AGREE) [5,6,210]. Keines der Instrumente ist jedoch dafür geeignet, eine systematische inhaltsbezogene Analyse von Leitlinien bzw. deren Empfehlungen durchzuführen [500]. Eine hohe methodische Qualität von Leitlinien korreliert nicht notwendigerweise mit der inhaltlichen Qualität der darin enthaltenen Empfehlungen [506]. Bisher existiert kein Instrument zur systematischen Bewertung der inhaltlichen Qualität von Leitlinien oder Leitlinienempfehlungen. Daher ist es notwendig, Methoden zur Analyse und Bewertung von Leitlinieninhalten zu entwickeln."

⇨ Es stellt sich, auch in Hinblick auf Machbarkeit und Ressourcen, die Frage, warum eine methodische Überprüfung nicht ausreicht. Es hat zudem seine Gründe, die vor allem in der extremen Komplexität eines solchen Unterfanges liegen, dass bislang, auch international, keine belastbaren Instrumente zur inhaltlichen Leitlinienbewertung vorliegen. Letztlich müsste man den gesamten Leitlinienerstellungsprozess in strukturell gleichartiger Zusammensetzung noch mal durchführen, um feststellen zu können, ob man im Einzelfall zu einer anderen Empfehlung gekommen wäre. Die Konsequenzen einer solchen Feststellung wären völlig unklar. Es wird zumeist nur die Möglichkeit eines Hinweises bleiben, dass - soweit überhaupt Unterschiede bestehen – diese wesentlich auf dem Weg von der Evidenz zur Empfehlung entstanden sind und sich damit einer im Sinne einer Dichotomie von „Richtig oder Falsch" gearteten Bewertung zumeist entziehen werden. Durch das Vorgehen, die Erstellung einer Leitlinie für die inhaltliche Bewertung quasi selbst noch einmal zu entwickeln, kommt es überdies zu einer ungewollten Verschmelzung der Position des Bewerters mit dem eines Erstellers, die es im Kontext von Bewertungen möglichst zu vermeiden gilt. Die bisherigen Erfahrungen mit Produkten des IQWiG zu Leitlinien (z. B. Beauftragungen des G-BA im Zusammenhang mit DMP-Programmen) haben zudem gezeigt, dass das IQWiG bereits allein mit der methodischen Bewertung von Leitlinien in hohem Maße ausgelastet ist. Es steht zu befürchten, dass eine Hinzunahme einer inhaltlichen Bewertung von Leitlinien allein in Bezug auf die zeitlichen Rahmenbedingungen, unter denen der G-BA sich mit DMP-Programmen zu beschäftigen hat, eine sinnvolle Zuarbeit des IQWiG für den G-BA nicht mehr möglich macht. Auch mit Blick auf andere Bewertungsverfahren des IQWiG stellt sich hier noch die generelle Frage, welche Tiefe Bewertungen des IQWiG erlangen sollen. Überträgt man die mit der inhaltlichen Leitlinienbewertung im Zusammenhang stehenden Ansprüche beispielsweise auf die Nutzenbewertung, dann müssten im Rahmen der

Nutzenbewertung alle relevanten wissenschaftlichen Arbeiten (einschließlich der Primärstudien) vom IQWiG noch mal selbst durchgeführt werden.

Zu „5 Evidenzbasierte Gesundheitsinformationen für Bürger und Patienten“:

- ⇨ Die Unterscheidung in *„Patienten und Bürger“* im gesamten Kapitel 5, z.B. als Zielpersonen für die Gesundheitsinformationen des IQWiG, sollte unterbleiben, da sie impliziert, dass Patienten keine Bürger seien. Auch Formulierungen wie *„Der Gesetzgeber hat die Informationsaufgaben des Instituts sowohl am Bürger – also am Gesunden- als auch am Patienten orientiert.“* (Zitat Seite 87) sind auf diesem Hintergrund unglücklich. Ferner fällt insbesondere in Kapitel 5, aber auch in der Präambel (Zitat Präambel letzter Satz: *„In diesem Dokument wird bei der Angabe von Personenbezeichnungen jeweils die männliche Form angewandt. Dies erfolgt mit dem Ziel einer besseren Lesbarkeit.“*), der Gebrauch einer nicht-gendersensiblen Sprache auf. Obwohl die Problematik (Lesbarkeit etc.) grundsätzlich bekannt ist, sollte in einem Grundsatzdokument wie dem Methodenpapier überdacht werden, ob man sich in Widerspruch zum Bundesgleichstellungsgesetz § 1 Satz 1 und 2 stellen möchte (siehe dazu www.gender-mainstreaming.net) Dies sollte auch bedacht werden, da das IQWiG im Kapitel 5.2.1 erklärt, wie wichtig die Nutzung einer gendersensiblen Sprache [bei den Gesundheitsinformationen] sei.
- ⇨ Die rechtliche Grundlage der Erstellung von Gesundheitsinformationen durch das IQWiG beruht auf:
  - SGB V § 139a Abs. 3 Punkt 6. „Bereitstellung von für alle Bürgerinnen und Bürger verständlichen allgemeinen Informationen zur Qualität und Effizienz in der Gesundheitsversorgung sowie zu Diagnostik und Therapie von Krankheiten mit erheblicher epidemiologischer Bedeutung“
  - dem Generalauftrag des G-BA vom 13.03.2008 :

    „Die Erfassung und Auswertung des relevanten Schrifttums und die Informationspflicht des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen erstreckt sich auch auf die Bereitstellung von für alle Bürgerinnen und Bürger verständliche allgemeine Informationen zur Qualität und Effizienz in der Gesundheitsversorgung sowie zu Diagnostik und Therapie von Krankheiten mit erheblicher epidemiologischer Bedeutung.“

- ⇨ In der Vergangenheit war bei einer Vielzahl von Gesundheitsinformationen - angeführt seien hier beispielhaft *„Wassergeburt“, „Bell´sche Parese“, „PCO-Syndrom“, „Töpfchen-Training“ und „Krampfaderbruch“* - nicht erkennbar, wie diese durch den o.g. Auftrag abgedeckt waren. Ferner besteht eine Dopplung mit Gesundheitsinformationen anderer Institutionen. Wie bereits im Gutachten der WHO (2010) angemerkt, sollten Dopplungen oder gar konkurrierende Informationen öffentlicher Informationsmaterialien durch Kooperation mit anderen Organisationen und Einrichtungen vermieden werden. Im jetzt vorliegenden Methodenpapier bleibt unklar, wie das IQWiG die genannten Probleme adressieren will. Aus dem Methodenpapier sollte hervorgehen, wie

spezifische fachliche Kompetenz (Beispiel: infektiologische Themen) beschafft werden kann und wie man sich von anderen, etablierten Institutionen, die Gesundheitsinformationen veröffentlichen, abgrenzen will (z.B. Robert Koch-Institut, Bundesinstitut für Arzneimittel und Medizinprodukte, European Medicines Agency etc.).

Kapitel 5.1, Hintergrund und Ziele

S.76-77

⇨ In der Definition der evidenzbasierten Gesundheitsinformation fehlt die Einbeziehung der ärztlichen Sichtweise, ohne die eine Gesundheitsinformation unvollständig bleibt.

S.77

„... Überdies sind viele Nutzer mit evidenzbasierten Informationen nicht vertraut, was besondere Anforderungen an die Kommunikation stellt [145, 189, 464]....".

⇨ Dieser Satz ist schwer nachvollziehbar, zumal eine Reihe von Institutionen, z.B. Bundeseinrichtungen, ebenfalls wissenschaftliche Informationen zur Gesundheit veröffentlicht.

Kapitel 5.2.1, Patientenzentrierte Kommunikation / Kommunikationsstandards, S.80

⇨ Es sollte ausgeführt werden, auf welche Weise die kulturellen Unterschiede in den Gesundheitsinformationen adressiert werden und wie Entscheidungswege aussehen, solche kulturellen Unterschiede zu identifizieren.

Kapitel 5.2.5, Patientenzentrierte Kommunikation / Methode der Patientenwege, S.84

⇨ Es sollte im Methodenpapier erläutert werden, wie wichtigen Problemen des Mediums Internet (z.B. „accessibility"„information-inequity" etc.) begegnet werden soll.

Kapitel 5.3.2, Recherche, S.89

⇨ Aus dem Methodenpapier geht nicht hervor, wie vom Herausgeber von Gesundheitsinformationen möglichen psychologischen und emotionalen Reaktionen der Leserinnen und Leser begegnet werden kann.

Kapitel 5.4.1., Informationsprodukte / Informationsberichte, Merkblätter und Kurzantworten, S.97

⇨ Wie soll sich der externe Begutachtungsprozess der Informationsberichte, Merkblätter und Kurzantworten gestalten und wie findet die Auswahl der Gutachter statt?

Kapitel 5.4.3, Zusätzliche Elemente, S.99

"Zu welchen Gesundheitsinformationen Pressemitteilungen herausgegeben werden, hängt unter anderem davon ab," [...]"ob die jeweiligen Forschungsergebnisse in der Öffentlichkeit bereits bekannt sind", [...] „welche Themen besonders dafür geeignet sind" [...] und was die Öffentlichkeit „möglicherweise wissen möchte."

⇨ Es bleibt unerwähnt, wie die Entscheidungswege aussehen. Es wäre relevant zu erfahren, auf welche Weise das IQWIG in Erfahrung bringen will, was Patientinnen und Patienten möglicherweise wissen wollen.

Kapitel 5.5.1, Überwachung und Auswertung / Routinemonitoring, S.103

⇨ Es ist nicht nachzuvollziehen, wie Punkt 3 auf Seite 103 mit einer Online-Datenschutzerklärung in Einklang stehen kann.

„...Außerdem wird das Institut den Umfang abschätzen, in dem seine Gesundheitsinformationen durch Multiplikatoren übernommen werden. Dazu muss ausgewertet werden, wie viele andere Websites – vor allem die der Krankenkassen – einen Link zu www.gesundheitsinformation.de geschaltet haben. Darüber hinaus wird die Anzahl der gedruckten Versionen der Gesundheitsinformationen des Instituts überwacht, vor allem derjenigen, die durch die Krankenkassen erstellt wurden...."

⇨ Es ist nicht ohne Weiteres davon auszugehen, dass vor allem mit Hilfe der Anzahl der Verlinkungen der Website oder der Informationsausdrucke durch Krankenkassen Aussagen zur Multiplikation getroffen werden könnten; zu berücksichtigen sind in diesem Zusammenhang insbesondere auch Informationen von Patientenorganisationen.

Kapitel 5.2.3. Beteiligung von Bürgern und Patienten, S.83 und S.24, und Kapitel 5.5.3 Evaluation, S.103-104

⇨ Der bereits von den WHO-Gutachtern formulierten Kritik an der Evaluation des Gesundheitsinformationsdienstes des IQWiGs wird durch das Methodenpapier nicht ausreichend Rechnung getragen. Die Kriterien und das Vorgehen bei der *„externen Nutzertestung"* sind trotz der bereits 2010 durch die WHO angemerkten mangelnde Repräsentativität nicht erkennbar auf eine solche hin aufgearbeitet worden und sind auch nicht nachvollziehbar. Einer nicht-repräsentativen Bewertung von Gesundheitsinformationen wohnt die Gefahr einer Verzerrung inne. Ferner ist nicht erkennbar, wie der potentielle Nutzen der Gesundheitsinformationen bestimmt wird. Im Methodenpapier sollte formuliert werden, wie dem Umstand, dass sich durch eine Nutzerbefragung auf einer Website immer nur ein selektierter Kreis an Nutzerinnen und Nutzern zur Abgabe von Kommentaren aufgefordert fühlt, begegnet wird und ob das IQWiG z.B. eine „non-responder/user"-Analyse durchführt. Das Ergebnis einer solchen Analyse sowie das Ergebnis einer Untersuchung zur etwaigen *„schädlichen Wirkung"* (S. 80) der Gesundheitsinformationen sollte im Rahmen einer Ergebnisqualitätsprüfung offengelegt werden.

⇨ Abschließend sei darauf hingewiesen, dass die Ergebnisse der von der WHO vorgenommenen Evaluation des IQWIG im Methodenpapier insgesamt berücksichtigt werden sollten. Dazu zählen u. a. auch die mangelnde Transparenz zur Einbeziehung verschiedener Interessengruppen im Erstellungsprozess der Gesundheitsinformationen oder die Kritik am

prinzipiellen Ausschluss industriegeförderter Reviews/Forschung als eine der Grundlagen für die Gesundheitsinformationen. Weitere Anregungen zu der Thematik findet sich z. B. in

- De Joncheere K., Gartlehner G., Gollogly L., Mustajoki P., Permanand G.: (2010): Health information for patients and the general public produced by the German Institute for Quality and Efficiency in Health Care. A review by the World Health Organization Europe 2008/2009. Internet (Zugang 18.06.2010): http://apps.who.int/medicinedocs/documents/s17062e/s17062e.pdf
- Lee, K: Has the hunt for conflicts of interest gone too far? No. BMJ 2008; 336; 477
- Stossel, TP: Has the hunt for conflicts of interest gone too far? Yes. BMJ 2008; 336; 477
- Yank, V et al.: Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. *BMJ* 2007;335;1202-1205; originally published online 16 Nov 2007

Zu „6 Informationsbeschaffung":

Kapitel 6.1.4 Leitliniendatenbanken, -anbieter, S. 109

„... Innerhalb der Nutzenbewertung werden Leitlinien als Informationsquelle nicht grundsätzlich ausgeschlossen. Es erfolgt jedoch i. d. R. keine systematische Recherche nach Leitlinien."

⇨ Durch einen Verzicht auf eine systematische Recherche nach Leitlinien wird eine wichtige Evidenzquelle außer Acht gelassen, die, bei hochwertigen Leitlinien mit einem systematischen Review (S-3) gekoppelt, unter Aufbereitung dieser Evidenz durch klinische Fachexperten und Methodiker einen besonders hohen Informationsgewinn besitzt. Gerade hier wird externe und interne Evidenz gemäß der EbM-Definition von Sackett und Kollegen und dem Sachverständigenrat in idealer Weise verbunden, die aufgrund der Integration realer klinischer Versorgungskontexte eine hohe Akzeptanz entfalten können. Bei Nichtbeachtung von Leitlinien stellt sich zudem die Frage, in welcher Form das IQWIG die Produkte international bedeutsamer Einrichtungen der evidenzbasierten Medizin wie z.B. die des NICE in Großbritannien berücksichtigt, da deren Produkte in Form von Leitlinien nach der aktuellen Fassung des IQWIG-Methodenpapiers weitgehend bedeutungslos wären. Insgesamt wird damit auch eine erhebliche Missachtung gegenüber den an der Leitlinienerstellung beteiligten Wissenschaftlern und Medizinern zum Ausdruck gebracht, was die Akzeptanz der Arbeiten des IQWiG in der medizinischen Fachwelt nicht befördert.

Mehr dazu unter Anmerkungen zu Kapitel 1 oder z.B. unter

- Sachverständigenrat für die konzertierte Aktion im Gesundheitswesen: Bedarfsgerechtigkeit und Wirtschaftlichkeit. Band I Zielbildung, Prävention, Nutzerorientierung und Partizipation, Band II Qualitätsentwicklung in Medizin und Pflege. Gutachten 2000/2001. Internet: http://www.svr-gesundheit.de/Gutachten/Gutacht00/kurzf-de00.pdf

http://www.svrgesundheit.de/Gutachten/Gutacht00/kurzf-de00.pdf (Zugriff am 03.12.2010

Zu „7 Informationsbewertung":

Kapitel 7.1.2 Zusammenhang zwischen Studientyp/-art und Fragestellung, S. 115

„... Weitere grundlegende klassische Studientypen der Epidemiologie sind Fall-Kontroll-Studien [60] zur Untersuchung des Zusammenhangs zwischen Expositionen und seltenen Erkrankungen und Kohortenstudien [61] zur Erforschung des Effekts von Expositionen im Zeitverlauf...."

⇨ Warum ist der Zeitverlauf an dieser Stelle wichtig? Es wäre hilfreich, dieses näher zu erläutern. Kohortenstudien bekleiden ein höheres Evidenzniveau als Fall-Kontrollstudien (nach z.B. Oxford Centre for Evidence-based Medicine, siehe oben), was dem Umstand geschuldet ist, dass Fall-Kontrollstudien retrospektiv von der Krankheit ausgehend die Exposition nachträglich erfragen, was mit erheblichen Erinnerungsfehlern (Recall Bias) behaftet sein kann. Somit kann eines der wichtigsten und unverzichtbaren Hill-Kriterien für Kausalität – nämlich der Umstand, dass die Exposition der Erkrankung vorausgehen muss – u. U. nicht garantiert werden. Daraus resultieren größere Unsicherheiten als in Kohortenstudien. Demgegenüber gehen Kohortenstudien grundsätzlich in der geforderten zeitlichen Abfolge von der Exposition aus, um deren Wirkung auf die Erkrankung in der Zukunft abzuschätzen.

„...Kohortenstudien sind in diesem Sinne prospektiv angelegt; allerdings gibt es auch retrospektive Kohortenstudien, in denen die Exposition aus der Vergangenheit erfasst wird (häufig in der Arbeits- oder auch Pharmakoepidemiologie anzutreffen). Grundsätzlich sind prospektive Studien retrospektiven Designs vorzuziehen...."

⇨ Der Begriff „retrospektiv" sollte im Zusammenhang mit Kohortenstudien durch „historisch" ersetzt werden. Der Begriff "retrospektiv" ist an dieser Stelle unpräzise. Fall-Kontroll-Studien sind retrospektiv, da sie die Exposition als Auslöser für die Erkrankung von der bereits festgestellten Krankheit nachträglich, also tatsächlich retrospektiv, erfragen. Eine „retrospektive Kohortenstudie" hingegen geht ebenso wie die prospektive Kohortenstudie von der Exposition aus – erstere von der Vergangenheit zum aktuellen Zeitpunkt des Studienbeginns und die zweite vom aktuellen Zeitpunkt des Studienbeginns in die Zukunft. Der präzisere Terminus „historisch" besagt, dass die Daten früher, also vor Studienbeginn z.B. in arbeitsmedizinischen Akten im Rahmen von Vorsorgeuntersuchungen erfasst wurden. Die Erfassung erfolgte demnach zum Zeitpunkt der Exposition, also vor der Erkrankung. Insofern ist der Begriff „retrospektiv" missverständlich. Lediglich die Verwendung dieser arbeitmedizinischen Akten erfolgt erst nach Studienbeginn. In Fall-Kontroll-Studien hingegen erfolgt die Expositionserfassung tatsächlich immer rückblickend, also retrospektiv - nämlich dann, wenn die Krankheit bereits eingetreten ist. So wäre eine differenziertere Darlegung des Sachverhalts im Methodenpapier angezeigt.

Literatur dazu z.B.:

- Checkoway, H.; Pearce, N.; Kriebel, D.: Research methods in occupational epidemiology. New York: Oxford University Press 2004
- Oxford Centre for Evidence-based Medicine - Levels of Evidence. March 2009, May 2001; Internet: http://www.cebm.net/index.aspx?o=1025

Kapitel 7.1.4 Aspekte der Bewertung des Verzerrungspotenzials, S. 118-119

„...In nicht verblindbaren Studien ist es zentral wichtig, dass eine adäquat verdeckte Zuteilung (Allocation Concealment) der Patienten zu den zu vergleichenden Gruppen gewährleistet ist. Weiterhin ist es erforderlich, dass die Zielvariable unabhängig vom (unverblindeten) Behandler ist bzw. unabhängig vom Behandler verblindet erhoben wird (verblindete Zielgrößen-erhebung)...."

⇨ Das Allocation Concealment ist in jedem RCT das wichtigste Qualitätskriterium, also nicht nur in RCTs mit unverblindbaren Interventionen (wie z.B. in der Chirurgie)

⇨ Die unabhängige oder verblindete Outcome-Messung ist ebenfalls immer von Relevanz

„... Für randomisierte Studien wird anhand dieser Aspekte das Verzerrungspotenzial zusammenfassend als „niedrig" oder „hoch" eingestuft. Ein niedriges Verzerrungspotenzial liegt dann vor, wenn mit großer Wahrscheinlichkeit ausgeschlossen werden kann, dass die Ergebnisse relevant verzerrt sind. Unter einer relevanten Verzerrung ist zu verstehen, dass sich die Ergebnisse bei Behebung der verzerrenden Aspekte in ihrer Grundaussage verändern würden...."

⇨ Es sollte darauf hingewiesen werden, dass es empirisch nachgewiesen ist, dass die verdeckte Zuteilung (Allocation Concealment) das wichtigste Beurteilungskriterium für die Qualität und interne Validität von RCTs ist (siehe Cochrane Handbook)

⇨ Warum wird nur eine zweistufige und nicht mehrstufige Bewertung verwendet wie im Cochrane Handbook beschrieben?

Literatur:

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

„...Eine Einstufung des Verzerrungspotenzials des Ergebnisses für einen Endpunkt als „hoch" führt nicht zum Ausschluss aus der Nutzenbewertung. Die Klassifizierung dient vielmehr der Diskussion heterogener Studienergebnisse und beeinflusst die Sicherheit der Aussage...."

⇨ Die Einstufung sollte auch Sensitivitätsprüfungen dienen. So kann in Meta-Analysen der Ausschluss von Studien mit hohem Verzerrungspotential klären helfen, wie stabil der Effekt bleibt.

Literatur:

- Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Kapitel 7.1.5 Interpretation von kombinierten Endpunkten, S. 119

⇨ Bedeutet die Kombination von Endpunkten beispielsweise die Berechnung des Gesamteffekts aus Tod und Krankenhaus-Aufnahme? Dieses würde Schwierigkeiten bei der Interpretation der Ergebnisse aufwerfen. Es können nach unserem Kenntnisstand nur gleiche Endpunkte zusammengefasst werden, d.h. Gesamteffekte z.B. für Tod, KH-Aufnahme etc. Die Aussagen im Methodenpapier zu dieser Thematik bedürfen weiterer Erklärung.

Literatur:

- Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Kapitel 7.1.6 Interpretation von Subgruppenanalysen, S. 120

„...Kein Beweischarakter: Subgruppenanalysen werden selten a priori geplant und sind selten Bestandteil des Studienprotokolls (ggf. auch von Amendments). Wenn Subgruppenanalysen bezüglich mehr oder weniger arbiträrer subgruppenbildender Merkmale post hoc stattfinden, können ihre Ergebnisse nicht als methodisch korrekte Prüfung einer Hypothese betrachtet werden...."

⇨ Falls aber Subgruppen-Analysen a-priori geplant wurden, sind auch diese aussagefähig. Dieses sollte explizit ergänzt werden.

Kapitel 7.1.7 Bewertung der Konsistenz publizierter Daten, S. 123

⇨ Es wird nicht explizit erwähnt, dass Rückfragen bei Autoren sehr hilfreich sein können. Dies gilt für das gesamte Kapitel 7.1 (siehe auch Cochrane Handbook)

Kapitel 7.2.1 *Einordnung systematischer Übersichten*, S. 124

„... Damit das Institut eine systematische Übersicht über Behandlungseffekte verwenden kann, muss diese gewisse Mindestvoraussetzungen erfüllen, d. h. sie darf nach dem Oxman-Guyatt-Index [266,379,381] oder anhand des AMSTAR-Instruments [445-447] nur geringe methodische Mängel aufweisen...."

⇨ Wir empfehlen, in diesem Kontext auch das QUOROM-Statement explizit zu nennen (Quality of Reporting of Meta-Analyses/ Cochrane Collaboration).

Kapitel 7.2.2 Nutzenbewertung auf Basis systematischer Übersichten, S. 125, 126

„...Alle qualitativ ausreichenden und thematisch relevanten systematischen Übersichten werden berücksichtigt. In der Regel sollten mindestens zwei Arbeiten von hoher Qualität, die unabhängig voneinander durchgeführt wurden, als Grundlage für die Erstellung eines Berichts auf Basis von Sekundärliteratur vorhanden sein, um die Konsistenz der Ergebnisse überprüfen zu können...."

⇨ Es ist nicht ohne weiteres einzusehen, warum mindestens zwei systematische Reviews (SR) vorliegen sollen. Ein hochwertiger SR bedeutet bereits das höchste Evidenzniveau. Wird die Forderung nach mehr als einem SR

aufrechterhalten, so müssen Literaturstellen angeführt werden, die diese sehr hohe Forderung empirisch begründen. Es sei daran erinnert, dass es nicht um einzelne RCT, sondern um einen SR geht, in den u. U. zahlreiche RCTs eingeflossen sind. Deshalb ist diese Hürde schwer nachzuvollziehen.

⇨ Zudem sind SR keine Sekundärliteratur, sie werden in allen hochrangigen Journals als Originalarbeiten anerkannt, da tatsächlich neues Wissen auf Basis der Originalstudien generiert wird. Der Begriff der Sekundärliteratur ist daher missverständlich.

„...C) Qualitätsbewertung der Publikationen einschließlich Mindestanforderungen..."

⇨ In diesem Abschnitt müssen die qualitativ wichtigsten Kernelemente des SR in Abgrenzung zum narrativen Review angeführt werden: 1.) die prospektiv formulierte, klare Fragestellung mit PICO-Schema (Population, Intervention, Kontrollen und Outcome), 2.) die umfassende, transparente und reliable Suche ,3.) die Qualitätsbewertung der eingeschlossenen Studien und 4.) die quantitative Datensynthese (Meta-Analyse), wenn diese möglich ist, oder alternativ die qualitative Zusammenfassung

Kapitel 7.2.3 Berücksichtigung publizierter Meta-Analysen, S. 127:

„...Von einer erhöhten Validität von Meta-Analysen basierend auf individuellen Patientendaten kann man allerdings nur dann ausgehen, wenn solche Analysen auch tatsächlich auf die Fragestellung der Bewertung des Instituts ausgerichtet sind und zudem eine hohe Ergebnissicherheit aufweisen. Für die Beurteilung der Ergebnissicherheit solcher Analysen ist eine maximale Transparenz Grundvoraussetzung. Dies bezieht sich sowohl auf die Planung als auch auf die Durchführung der Analysen. Diesbezüglich wesentliche, für die Durchführung von Meta-Analysen allgemeingültige Aspekte sind z. B. in einem Dokument der Europäischen Zulassungsbehörde EMA beschrieben [146]. ...."

⇨ Diese Individualdaten sind allerdings oft schwer zu bekommen, denn sie müssen dann für alle eingeschlossenen Studien verfügbar sein; so wird es schwierig sein, über z.B. 10-20-Jahreszeiträume alle Originaldaten sämtlicher Studien zu bekommen. Es handelt sich bei ihren Ausführungen eher um eine ausführliche Darstellung wenig realistischer Szenarien.

⇨ Die EMA kann Meta-Analysen mit individuellen Patientendaten einfacher veranlassen, da sich dieses Prozedere nur auf die Zulassungsstudien bezieht, d.h. es handelt sich um eine überschaubare Zahl von Zulassungsstudien (die eingereicht werden müssen) in einem ebenso überschaubaren Zeitfenster.

Kapitel 7.3.1 Darstellung von Effekten und Risiken, S. 127-128

„Darstellung von Effekten und Risiken
...Bei stetigen Variablen können Effekte in der Regel mithilfe von Mittelwerten sowie Differenzen von Mittelwerten – unter Umständen nach geeigneter Gewichtung – dargestellt werden...."

⇨ Die Art der Gewichtung sollte erläutert werden. An dieser Stelle könnte auch einleitend kurz gesagt werden, dass es sich bei der Berechnung des Gesamteffekts in Meta-Analysen letztlich um komplexe Mittelwertvergleiche handelt.

⇨ Es sollten v.a. die Kernelemente genannt werden: Für dichotome Outcomes werden RR oder OR verwendet, für kontinuierliche Outcomes die „Weighted

Mean Difference" (WMD), soweit gleiche Messinstrumente in den Originalstudien eingesetzt wurden, oder die „Standardized Mean Difference" (SMD), soweit unterschiedliche Messinstrumente genutzt wurden. Hier ist eine Präzisierung notwendig.

„... Zwingend erforderlich ist es, bei jeder Effektschätzung den Grad der statistischen Unsicherheit zu beschreiben. Häufige Methoden hierfür stellen die Berechnung des Standardfehlers sowie die Angabe eines Konfidenzintervalls dar. Wann immer möglich, gibt das Institut adäquate Konfidenzintervalle für Effektschätzungen mit der Information, ob es sich um ein- oder zweiseitige Konfidenzgrenzen handelt, sowie das gewählte Konfidenzniveau an."

⇨ Es ist aus unserer Sicht richtig, dass Konfidenzintervalle (KI) bevorzugt werden. Eine kurze Begründung dafür wäre jedoch hilfreich. Diese könnte beinhalten, dass das KI nicht nur die statistische Signifikanz, sondern im Gegensatz zum p-Wert darüber hinaus gleichzeitig auch die Effektstärke mit ihrer Präzision anzeigt.

„... Beide Aspekte – ob eine Hypothese ein- oder zweiseitig zu formulieren ist und ob für multiples Testen adjustiert werden muss – werden in der wissenschaftlichen Literatur immer wieder kontrovers diskutiert...."

⇨ Wie sieht dieser Diskurs aus? Eine Konkretisierung wäre hilfreich.

„...Bezüglich der Hypothesenformulierung ist traditionell von einem zweiseitigen Testproblem auszugehen. Ausnahmen hiervon bilden zum Beispiel Nichtunterlegenheitsstudien...."

⇨ Eine weitere Ausnahme sind sicher tödliche Krankheitsverläufe, sodass jeder Unterschied, den eine Intervention auslöst, nur in eine Richtung münden kann, nämlich die der Verbesserung. Damit wäre die einseitige Testung beispielsweise in einem solchen Fall angemessen.

Kapitel 7.3.3 Beurteilung klinischer Relevanz, S. 129-131

⇨ Zur Diskussion von klinischen Irrelevanzschwellen/Cut-Offs von SMD zur Bestimmung des patientenrelevanten Nutzens möchten wir Folgendes ausführen: An dieser Stelle werden im Methodenpapier für die Nutzenbewertung sehr hohe methodische Hürden eingezogen, die international nicht allgemeinverbindlich empfohlen werden. Hierauf wird auf S. 130 auch hingewiesen, sodass der Eindruck entsteht, hier sollten durch das IQWiG neue internationale Maßstäbe gesetzt werden. Dieses halten wir für problematisch, zumal gewichtige Gegenargumente einer ausführlicheren Diskussion im Methodenpapier bedürfen:

Die Standardized Mean Difference (SMD) in Meta-Analysen stellt einen Gesamtschätzer über verschiedene Messinstrumente und Skalen dar, wobei unterschiedliche Messinstrumente unterschiedlich starke patientenrelevante Effekte messen können. Die Einteilung von Stärken der SMD von 0,2 als kleinen, 0,5 als mittleren und 0,8 als starken Effekt ist lediglich eine grobe Orientierung nach Cohen, aber kein Absolutum. Im aktuellen Cochrane Handbook wird die Einteilung als „rule of thumb" beschrieben – und dies auch nur für die Effektstärke, nicht als allgemeingültige Regel zur Festlegung des patientenrelevanten Nutzens. Wie sich den Unterlagen der kürzlich stattgefundenen IQWIG-Veranstaltung „IQWIG im Dialog 2010" (18.06.2010; Internet: https://www.iqwig.de/index.1055.html) entnehmen lässt, werden in

den aktuellen Methodenhandbüchern einschlägiger nationaler und internationaler Institutionen der Evidenzbasierten Medizin wie der Cochrane Collaboration, dem NICE oder der Vorgängerversion des IQWIG-Methodenpapiers keine Handlungsempfehlungen zur generellen Anwendung eines „Cut-Offs" für den patientenrelevanten Nutzen abgegeben. Vielmehr lässt sich den Unterlagen zum erwähnten Symposium als auch dem Cochrane Handbook entnehmen, dass die kritische Diskussion zum Einsatz solcher Cut-Offs noch längst nicht abgeschlossen ist. Effektstärken sind darüber hinaus kontextabhängig zu betrachten, wobei sich die klinische Relevanz nicht nur über statistische Differenzen definiert, sondern der inhaltliche Bezug zur jeweiligen Fragestellung und zur jeweiligen Vergleichsgruppe zu berücksichtigen ist. Zudem sind Validierungen wie z.B. in Form der „Minimal Important Difference" wichtig, um zu bestimmen, ab welchen Skalenveränderungen von Patientenrelevanz auszugehen ist.

Verschärft wird der Cut-Off, wenn auch die Konfidenzintervalle dieses Kriterium allgemeingültig erfüllen müssen. Konfidenzintervalle sind ein Maß der Präzision, die u.a. von der Studiengröße abhängen, sodass sie auch unabhängig von der Effektstärke über die Größe beeinflussbar sind. Darüber hinaus stellt sich die Frage, ob die oben genannte Einteilung der Effektstärken nach Cohen auf medizinische Kontexte mit viel kleineren, spezifischeren Kollektiven unmittelbar übertragbar ist, da sie ursprünglich aus der Soziologie mit sehr großen Bevölkerungskollektiven stammt. Im Übrigen stellt sich die Frage, wie mit qualitativ hochwertigen systematischen Reviews umzugehen ist, wenn Meta-Analysen nur bedingt verfügbar sind, sodass anhand statistischer Kenngrößen Festlegungen ohnehin nicht getroffen werden können.

Der patientenrelevante Nutzen ist erst aus der Beantwortung einer Vielzahl weiterführender Fragen zu klären. Dieser Prozess kann erst am Ende der zu erarbeitenden Evidenzsynthese erfolgen, d.h. entsprechend den GRADE-Kriterien erst dann, wenn sämtliche Literatur gesichtet und evaluiert ist, um eine abschließende leistungsrechtliche Bewertung vornehmen zu können. Eine NICE-Guideline beispielsweise (NICE Guideline 26 (2005): Post-traumatic stress disorder; Internet: http://www.nice.org.uk/nicemedia/live/10966/29772/29772.pdf) nennt zur Bewertung des patientenrelevanten Nutzens von Interventionen u.a. in Anlehnung an die GRADE-Kriterien folgende Punkte, die zu klären und zu berücksichtigen sind:

- Grad der Evidenz
- Stärke der Evidenz
- Anwendbarkeit für und Übertragbarkeit auf die betroffenen Patientengruppen
- Ökonomische Bewertung
- Wertigkeit für die Gruppen, die maßgeblich zur Entwicklung beigetragen haben, und Wertigkeit für die Gesellschaft
- Praktische Erwägungen hinsichtlich der Umsetzbarkeit

Kapitel 7.3.4 Bewertung subjektiver Endpunkte bei offenen Studiendesigns, S. 132:

„...In verschiedenen empirischen Arbeiten wurde gezeigt, dass in randomisierten kontrollierten Studien mit fehlender Verblindung bei subjektiven Endpunkten im Mittel eine Verzerrung der Effekte zugunsten der untersuchten Intervention vorliegt. Zu den subjektiven Endpunkten gehören beispielsweise patientenberichtete Endpunkte sowie Endpunkte, deren Erhebung und Einschätzung stark von den behandelnden bzw. endpunkterhebenden Personen abhängen. Eine Zusammenfassung dieser Arbeiten liefern Wood et al. [521]. Demnach sind solche Ergebnisse potenziell hochgradig verzerrt. Ein allgemein akzeptierter Umgang mit diesem Problem im Rahmen von systematischen Übersichten existiert nicht. I. d. R. wird das Institut in dieser Situation keinen Beleg für einen Nutzen oder Schaden aus statistisch signifikanten Ergebnissen ableiten....“

⇨ Diese Passage ist schwer nachvollziehbar und bedarf weiterer Erläuterungen. Bestimmte Parameter sind begriffsimmanent subjektiv und hängen von den individuellen Präferenzen ab (z.B. Lebensqualitätsaspekte), was ja auch erwünscht ist. Das Methodenpapier betont mehrfach die aktive Einbeziehung von Patientenbedürfnissen. Bedeutet dieser Absatz hingegen, dass nun doch nur objektive Parameter einfließen sollen?

„...Eine Möglichkeit, dem hohen Verzerrungspotenzial bei subjektiven Endpunkten in offenen Studien Rechnung zu tragen, besteht in der Formulierung einer adjustierten Entscheidungsgrenze. Nur dann, wenn das Konfidenzintervall des interessierenden Gruppenunterschieds einen bestimmten Abstand zum Nulleffekt aufweist, wird der Interventionseffekt als so groß angesehen, dass er nicht allein durch Verzerrung zu erklären ist. Das Verfahren der Anwendung einer adjustierten Entscheidungsgrenze an sich stellt das Testen einer verschobenen Nullhypothese dar, wie es seit Jahrzehnten in der Anwendung ist, u. a. notwendigerweise bei der Prüfung von Äquivalenz- und Nichtunterlegenheitshypothesen [147]. Notwendig für die Anwendung adjustierter Entscheidungsgrenzen ist die prospektive Festlegung des konkreten Grenzwerts. Das Institut wird die Wahl des Grenzwerts bei Anwendung projektspezifisch durch empirische Daten, wie sie z. B. Wood et al. [521] liefern, begründen....“

⇨ Die Ausführungen suggerieren, dass versucht wird, ein qualitatives Problem quantitativ zu lösen. Es sollte überdacht werden, ob sich diese Problematik überhaupt stellt.

Kapitel 7.3.6 Nachweis der Verschiedenheit, S. 133-134

⇨ Forderung nach Äquivalenzstudien zum Nachweis der Gleichheit: Dieses Konzept fordert höchstes methodisches Niveau, berücksichtigt dabei allerdings nicht, dass diese Form von Studien international selten in der Literatur zu finden ist, sodass viele Studienergebnisse aus RCT nicht zu verwerten wären. Außerdem ist für kleine Unterschiede zur Erzielung einer großen Power eine große Teilnehmerzahl erforderlich. Dies wirft wiederum Machbarkeits- und Finanzierbarkeitsprobleme auf. Darüber hinaus ist die Validierung von Irrelevanzgrenzen erforderlich, die oft nicht verfügbar sind. Zusammenfassend entsteht dadurch ein „Knock-Out“-Kriterium für eine Vielzahl von RCT auf höchstem Evidenzniveau und infolgedessen die Aberkennung des Nutzens vieler medizinischer Verfahren.

„...Im Vergleich zu Überlegenheitsstudien besitzen Äquivalenzstudien spezielle methodische Probleme. Zum einen ist es häufig schwierig, Äquivalenzbereiche sinnvoll zu definieren [306]. Zum anderen schützen zum Beispiel die üblichen Designkriterien Randomisierung und Verblindung nicht mehr hinreichend sicher vor Verzerrungen [442]. Auch ohne Kenntnis der Therapiegruppe ist es zum Beispiel möglich, die Differenz der Behandlungsunterschiede zur Null und damit zur gewünschten Alternativhypothese hin zu verschieben. Des Weiteren ist mit dem Intention-to-Treat-Prinzip sehr

vorsichtig umzugehen, da dessen inadäquate Anwendung eine falsche Äquivalenz vortäuschen kann [271]. Somit ist bei der Bewertung von Äquivalenzstudien besondere Vorsicht geboten...."

⇨ Dieser Abschnitt bedarf dringend weiterer Erläuterung. Mit der jetzigen Formulierung wird nicht klar, welche Zielsetzung besteht. Es wäre auch sinnvoll, weitere Literatur als die zitierte heranzuziehen.

Kapitel 7.3.7 Adjustierung und multifaktorielle Verfahren, S. 134-135

⇨ Der Begriff „Multifaktorielle Verfahren" sollten durch „multivariate Verfahren" und „unifaktoriell" durch „univariat" ersetzt werden. Dies sind die üblicheren Bezeichnungen. Man könnte sonst an Faktorenanalyse denken, die hier nicht hingehört.

„...Um die Qualität einer solchen Analyse bewerten zu können, ist eine Darstellung der wesentlichen Aspekte im Rahmen der statistischen Modellbildung notwendig [216,412] sowie Angaben zur Güte des gewählten Modells (Goodness-of-Fit) [241]. Die wichtigsten Informationen hierbei sind in der Regel:
• eine eindeutige Beschreibung und A-priori-Festlegung der Zielvariablen und aller potenziell erklärenden Variablen
• das Messniveau und die Kodierung aller Variablen
• Angaben zur Selektion von Variablen und Wechselwirkungen
• eine Beschreibung, wie die Modellannahmen überprüft wurden
• Angaben zur Modellgüte
• eine Tabelle mit den wesentlichen Ergebnissen (Parameterschätzung, Standardfehler, Konfidenzintervall) für alle erklärenden Variablen..."

⇨ In diesem Abschnitt sollten die Begriffe der Wechselwirkung, Effektmodifikation und des Confoundings klar abgegrenzt werden.

Kapitel 7.3.8 Meta-Analysen, S. 136-140

⇨ Manchmal sind Meta-Analysen in SR nicht möglich bzw. sinnvoll; in diesen Fällen kann und muss qualitativ ohne Signifikanzprüfung zusammengefasst werden; ggf. kann auf Signifikanzprüfungen der einzelnen Studien zurückgegriffen werden. Zu dieser Situation sollte im Methodenpapier Stellung genommen werden (siehe Cochrane Handbook).

„... Liegen Informationen darüber vor, dass die Effekte der einzelnen Studien homogen sind, ist eine Meta-Analyse unter der Annahme fester Effekte ausreichend. Solche Informationen werden jedoch häufig nicht vorliegen, sodass bei der Evaluierung der Studien in ihrer Gesamtheit die Annahme zufälliger Effekte hilfreich ist [444]...."

⇨ Dies kann als Sensitivitätsprüfung erfolgen und sollte auch so benannt werden (Cochrane Handbook).

„...Des Weiteren ist zu beachten, dass die aus einem Modell mit festen Effekten berechneten Konfidenzintervalle für den erwarteten Gesamteffekt selbst bei Vorhandensein einer geringen Heterogenität im Vergleich zu Konfidenzintervallen aus einem Modell mit zufälligen Effekten eine substanziell kleinere Überdeckungswahrscheinlichkeit aufweisen können [64]. Das Institut verwendet daher vorrangig Modelle mit zufälligen Effekten und weicht nur in begründeten Ausnahmefällen auf Modelle mit festen Effekten aus. Dabei ist zu beachten, dass sich die meta-analytischen Ergebnisse von Modellen mit zufälligen und festen Effekten bei homogener Datenlage allenfalls marginal unterscheiden...."

⇨ Die primäre Verwendung von Random-Effects-Modellen (Modelle mit zufälligen Effekten) macht die Konfidenzintervalle größer, sodass Schwellenwerte schwerer überwunden werden können. Dies ist ein eher konservatives Vorgehen und konservativer als das der Cochrane Collaboration (Cochrane Handbook). Zu berücksichtigen ist, dass sich häufig auch homogene Datensätze in Reviews finden, sodass dann in jedem Fall Fixed-Effects-Modelle herangezogen werden sollte.

„...B) Heterogenität..."

⇨ An dieser Stelle sollte nicht so sehr die statistische, sonder vielmehr die inhaltliche Untersuchung der Heterogenität im Vordergrund stehen. In diesem Abschnitt des Methodenpapiers werden keinerlei inhaltliche Kriterien angeführt (wie z.B. unterschiedliche Populationen, unterschiedliche Interventionen etc.). Insgesamt handelt es sich bei den Ausführungen somit um eine Überbewertung der statistischen Möglichkeiten.

Für einen differenzierten Umgang mit Heterogenität ist u.a. auf das Kapitel 9.5.4 des aktuellen Cochrane Handbooks (Internet: http://www.cochrane.org/training/cochrane-handbook) hinzuweisen. Darin wird betont, dass der Umgang mit heterogenen Datensätzen nicht nur eine statistische, sondern auch eine inhaltliche Frage ist. Zunächst ist es wichtig, nach den Gründen zu suchen, die sowohl methodischer als auch klinischer Natur sein können. Dies kann u.a durch Subgruppen-Analysen erfolgen, um dann z.B. durch Ausschluss von Ausreißern mit homogenen Datensätzen arbeiten zu können (Fixed-Effects-Modelle). Eine andere Überlegung bezieht sich auf die Proportion der Gesamtvariation der Studienschätzer (NICE Guideline 26 (2005): Post-traumatic stress disorder; Internet: http://www.nice.org.uk/nicemedia/live/10966/29772/29772.pdf). Bei milder Ausprägung unter 30% wären z.B. Fixed-, bei starker Heterogenität über 50% Random-Effects-Modelle und bei moderater Ausprägung zwischen 30 und 50% wären je nach Forrest Plot und Fragestellung beide statistischen Modelle einsetzbar. Darüber hinaus kann es aus inhaltlichen Gründen auch möglich sein, die Heterogenität zu ignorieren bzw. zu akzeptieren und nur mit Fixed-Effects-Modellen zu arbeiten. Wenn allerdings Bedenken zur Heterogenität bestehen, auf Meta-Analysen aber nicht verzichtet werden soll, sind Random-Effects-Modelle angezeigt. Bestehen allerdings so erhebliche Bedenken zur Heterogenität, die eine quantitative Zusammenfassung mittels Meta-Analyse nicht erlauben (Äpfel-Birnen-Vergleich), wäre es sogar sinnvoll, ganz auf die Meta-Analyse zu verzichten. So sind viele Varianten des Umgangs mit Heterogenität möglich und begründbar. Dabei gehen Fixed-Effects-Modelle der Abbildung des besten Effekts nach, während Random-Effects-Modelle nach dem Durchschnittseffekt fragen. Dementsprechend finden sich bei Anwendung von Random-Effects-Modellen breitere Konfidenzintervalle, weil kleineren Studien darin ein höheres Gewicht als in Fixed-Effects-Modellen zukommt. Letzteres kann sich nachteilig auswirken, da kleinere Studien bias-anfälliger sind.

Grundsätzlich gilt es zu bedenken, dass die verschiedenen statistischen Modelle das Problem der Heterogenität nicht lösen. Vielmehr ist kontextbezogen zu argumentieren, da die Heterogenität auch ein Problem der a priori erarbeiteten Zielgerichtetheit und Genauigkeit der Fragestellung darstellt. In manchen Studien ist z.B. zu berücksichtigen (beispielsweise bei

Abbass et al. 2006), dass aufgrund sehr weiter Einschlusskriterien, die u.a. auch Ko-Morbiditäten erlauben, der Vorteil einer hohen externen Validität, also hohen Übertragbarkeit auf reale Versorgungssituationen, gegeben ist, andererseits aber der Nachteil von Heterogenität damit einhergeht, womit sich jeweils kontextbezogen ein Kontinuum von Abwägungsprozessen ergibt.

Die Ausführungen zu dieser Thematik im Methodenpapier bedürfen deshalb einer differenzierteren Betrachtung.

Literatur:

- Allan A. Abbass, Jeffrey T. Hancock, Julie Henderson, Steve R. Kisely Short-term psychodynamic psychotherapies for common mental disorders, 2006

"... C) Subgruppenanalysen im Rahmen von Meta-Analysen..."

⇨ Im Rahmen von SR und Meta-Analysen sind a-priori definierte Subgruppen-Analysen zunehmend anzutreffen, dann sollte die Bewertung auch eine höhere sein. Dies sollte explizit angeführt werden.

„...Für den denkbaren Fall, dass das Institut mit der regelmäßigen Aktualisierung einer systematischen Übersicht beauftragt wird, die so lange aktualisiert wird, bis eine Entscheidung auf der Basis eines statistisch signifikanten Resultats vorgenommen werden kann, wird das Institut jedoch die Anwendung von Methoden für kumulative Meta-Analysen mit Korrektur für multiples Testen in Erwägung ziehen...."

⇨ Was ist damit gemeint? Dieser Abschnitt bedarf einer näheren Erläuterung

Kapitel 7.3.9 Indirekte Vergleiche

S. 141

„...Um eine Kosten-Nutzen-Bewertung multipler Interventionen zu ermöglichen, kann das Institut unter Inkaufnahme einer – im Vergleich zum Ansatz der reinen Nutzenbewertung – geringeren Ergebnissicherheit auch indirekte Vergleiche zur Bewertung von Kosten-Nutzen-Verhältnissen heranziehen [252]...."

⇨ Warum wird nicht auch an anderen Stellen die Möglichkeit indirekter Vergleiche eingeräumt, sondern nur in dieser Situation? Dies bedarf einer näheren Erläuterung.

S. 140-141

⇨ Zur Einbeziehung indirekter Vergleiche in Bezug auf die Bewertung von Arzneimitteln ist zudem folgendes anzumerken:

- Im Rahmen der frühen Nutzenbewertung nach § 35a SGB V ist durch § 5 Absatz 5 der AM-NutzenV ausdrücklich vorgegeben, dass die Bewertung des Zusatznutzens auch auf der Basis von indirekten Vergleichen zwischen dem zu bewertenden Arzneimittel und der zweckmäßigen Vergleichstherapie erfolgen kann. Insofern ist es folgerichtig, dass das IQWiG diese neue Rechtslage durch den neuen Abschnitt „Indirekte Vergleiche" auch im vorliegenden Entwurf des Methodenpapiers behandelt.

- Es ist allerdings fachlich nicht nachvollziehbar, warum die Anwendbarkeit von indirekten Vergleichen auf das Verfahren zur frühen Nutzenbewertung nach § 35a SGB V sowie die Kosten-Nutzenbewertung von Arzneimitteln nach § 35b SGB V beschränkt bleiben soll.
- Da die Kosten-Nutzenbewertung des IQWiG entsprechend der gesetzlichen Vorgaben bzw. der Verfahrensordnung des G-BA stets ausschließlich auf Grundlage der Ergebnisse einer vorgelagerten (frühen) Nutzenbewertung erfolgt, wären – bei Beibehaltung der Nicht-Berücksichtigung indirekter Vergleiche in der Nutzenbewertung nach § 139a SGB V – inkonsistente und damit auch rechtsunsichere Bewertungsergebnisse zwischen Nutzen- und Kosten-Nutzenbewertung zwangsläufig: Da das IQWiG in der „reinen" Nutzenbewertung von Arzneimitteln indirekte Vergleichsstudien „i.d.R." von der Bewertung von vornherein aus methodischen Gründen ausschließt, die gleichen Studien dann aber anschließend im Rahmen der Kosten-Nutzenbewertung (aus guten Gründen) einschließt, können die betreffenden Bewertungsergebnisse des IQWiG nicht widerspruchsfrei sein.
- Im vorliegenden Methodenpapier muss daher klargestellt werden, dass (ggf. unter Berücksichtigung einer geringeren Ergebnissicherheit) indirekte Vergleiche auch in der „reinen" Nutzenbewertung von Arzneimitteln nach § 139a SGB V zu berücksichtigen sind und nicht, wie vorgesehen, von vornherein ausgeschlossen werden.

Kapitel 7.3.11 Darstellung von Verzerrungsarten, S. 141

⇨ Randomisierungsfehler als empirisch nachgewiesene, wichtigste Verzerrungsquelle im RCT werden auch hier nicht explizit angeführt.

„...Ein „Selection Bias" entsteht durch eine Verletzung der Zufallsprinzipien bei Stichprobenziehungen, d. h. bei der Zuteilung der Patienten zu den Interventionsgruppen..."

⇨ Die Unterscheidung von Auswahlverzerrungen ist hier unzureichend beschrieben: Es gibt eine wichtige Differenzierung des Selection Bias: 1.) Selection Bias bis zum Zeitpunkt der Randomisierung und 2.) Selection Bias bei Randomisierung mit unterschiedlichen Folgen: Der erste Fall ist später statistisch nicht korrigierbar, der zweite durch Adjustierung sehr wohl, weil er sich dann als Confounding niederschlagen kann, dem durch Adjustierung beizukommen ist (siehe auch Gordis 2009).

Literatur:
- Gordis L (2009): Epidemiology. Saunders Elsevier, Philadelphia

„... Ein „Selection Bias" entsteht durch eine Verletzung der Zufallsprinzipien bei Stichprobenziehungen, d. h. bei der Zuteilung der Patienten zu den Interventionsgruppen. Speziell bei Gruppenvergleichen kann ein Selection Bias zu systematischen Unterschieden zwischen den Gruppen führen. Sind dadurch wichtige Confounder in den Gruppen ungleich verteilt, so sind die Ergebnisse eines Vergleichs in aller Regel nicht mehr interpretierbar...."

⇨ Diese Schlussfolgerung ist nicht ganz nachzuvollziehen, da man in dieser Situation auch multivariat mit Adjustierung rechnen kann.

„... Neben der Vergleichbarkeit der Gruppen bezüglich potenzieller prognostischer Faktoren spielen die Behandlungsgleichheit und die Beobachtungsgleichheit aller Probanden eine entscheidende Rolle. Eine Verzerrung durch unterschiedliche Behandlungen (mit Ausnahme der zu untersuchenden Intervention) wird als „Performance Bias" bezeichnet. Eine Verletzung der Beobachtungsgleichheit kann zu einem „Detection Bias" führen. Die Verblindung ist ein wirksamer Schutz vor beiden Biasarten [275], die in der Epidemiologie als „Information Bias" zusammengefasst werden...."

⇨ Diese Passage ist schwer verständlich. Die Verblindung kann nicht vor beiden Biasarten schützen, sondern nur vor Detection Bias als Klassifikationsfehler schützen. Für den Performance Bias ist die Intention-To-Treat(ITT)-Analyse ein approbater Korrekturversuch. Darin wird so ausgewertet wie randomisiert wurde, unabhängig davon, welche Intervention die Teilnehmer bekommen haben und wie diese ausgeführt wurde. Die ITT-Analyse ist eine konservative Schätzung der Stabilität des Gesamteffektes. (siehe auch Gordis 2009 oder Cochrane Handbook 2011)

„... Protokollverletzungen und Studienaustritte können bei Nichtberücksichtigung in der Auswertung das Studienergebnis systematisch verzerren, was als „Attrition Bias" bezeichnet wird. Zur Verminderung von Attrition Bias kann in Studien, die eine Überlegenheit zeigen wollen, das Intention-to-Treat-Prinzip eingesetzt werden, das besagt, dass alle randomisierten Probanden in der Analyse berücksichtigt werden, und zwar in der durch die Randomisierung zugeordneten Gruppe, unabhängig von Protokollverletzungen [275,300]...."

⇨ ITT-Analysen beziehen sich v.a. auf den Performance Bias, nicht nur auf den Attrition Bias.

„... Beispielsweise empfiehlt die EMA in Sensitivitätsanalysen unterschiedliche Verfahren zum Umgang mit fehlenden Werten gegenüberzustellen [151]...."

⇨ Welche Verfahren sind gemeint? Eine weitere Erläuterung wäre hilfreich.

A 1.6 EBM Review Centers der Medizinischen Universität Graz (EBM RC Graz)

Autoren:
Horvath, Karl

Beleg - Hinweis - Anhaltspunkt	Die nun erweiterte Graduierung bei der Feststellung des (Zusatz-) Nutzens bzw. Schadens ist allein aus der Beschreibung im Text nur schwer zu erfassen. Hier wäre eine zusätzliche tabellarische oder grafische Darstellung zur Einschätzung der Beleglage hilfreich, die eine Zusammenschau der relevanten Komponenten wie Ergebnissicherheit (quantitativ, qualitativ), Effektgröße und Konsistenz der Effekte ermöglicht.
2.2.3 Begutachtung der Produkte des Instituts ..." Alle Produkte einschließlich der jeweiligen Zwischenprodukte unterliegen einem umfangreichen mehrstufigen internen Qualitätssicherungsverfahren."	Die hier angesprochenen, mehrstufigen Qualitätssicherungsverfahren sind im Entwurf nicht näher ausgeführt und von da her nicht transparent. Im Sinne der Transparenz des Verfahrens wären Angaben zu den vorgesehenen Qualitätssicherungsschritten - idealerweise auch zu den geplanten zeitlichen Rahmen - wünschenswert.
7.1.1 Kriterien zum Einschluss von Studien: " ... wenn bei mindestens 80 % der in der Studie eingeschlossenen Patienten dieses Kriterium erfüllt ... ... zu mindestens 80 % das Einschlusskriterium bezüglich Prüfintervention ..."	In diesem Zusammenhang ist unklar, wie vorgegangen wird, wenn die Einschlusskriterien gleichzeitig sowohl für die Studienpopulation als auch für die Intervention nur teilweise erfüllt sind. Aus den Angaben in den Publikationen wird vermutlich sehr oft nicht sicher bestimmbar sein, ob insgesamt mindestens 80% der berichtsrelevanten Population auch die gewünschte Intervention erhalten oder weniger. Hier müssen dezidiert auch die Auswirkungen angesprochen werden, die sich daraus für die Recherche und Studienselektion ergeben. Wenn Einschlusskriterien bei bis zu 20% der Population nicht erfüllt sein müssen, dann führt das unter Umständen dazu, dass z. B. Einschränkungen auf Krankheitsbilder (z. B. Hypertonie) in der Recherche, aber auch beim Screenen der Titeln und Abstracts nicht mehr gemacht werden können. Unklar ist außerdem, inwieweit diese Kriterien auch für berichtete Subgruppen innerhalb einer Studie gelten. Wenn z. B. nach Patienten mit Typ 2 Diabetes gesucht wird und in einer großen Studie auch eine diabetische Subgruppe (Typ 1 und Typ 2) berichtet wird, ist es dann ausreichend, wenn 80% der Teilnehmer in dieser Subgruppe Typ 2 Diabetes haben?
7.2.2 Nutzenbewertung auf Basis systematischer Übersichten "Die Anwendbarkeit einer Nutzenbewertung auf Basis systematischer Übersichten hängt von der Verfügbarkeit systematischer Übersichten ab, die ... ... qualitativ hochwertige Primärstudien enthalten ..."	Warum dürfen systematische Übersichten, die transparent und nachvollziehbar darstellen, dass es keine qualitativ hochwertigen Primärstudien gibt (also z. B. auch IQWiG-Berichte), nicht für eine Nutzenbewertung auf Basis von Sekundärliteratur herangezogen werden? Dies würde bedeuten, dass Fragestellungen, zu denen es keine hochwertigen Primärstudien gibt, nicht auf Basis von Sekundärliteratur bearbeitet werden können. Hier gibt es eine Diskrepanz zur Nutzenbewertung auf Basis von Primärliteratur, die bei derselben Studienlage sehr wohl eine Aussage treffen kann.

A 1.7 Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (gmds)

Autoren:
Haerting, Johannes.

[redacted]

Institut für Qualität und Wirtschaftlichkeit
im Gesundheitswesen (IQWiG)
Dillenburger Straße 27
51105 Köln

7. April 2011

Stellungnahme der GMDS zum Papier Allgemeine Methoden – Entwurf für Version 4.0

Das Methodenpapier 4.0 stellt im vom Vorgängerpapier bekannten Duktus eine aktualisierte und erweiterte Übersicht über das für die Berichtstätigkeit des IQWiG erforderliche Methodenspektrum dar. Dabei stammen die meisten Methoden aus dem Bereich der medizinischen Biometrie und klinischen Epidemiologie, einige zusätzlich auch aus dem Bereich der empirischen Sozialforschung.

Anlass der Aktualisierung des Methodenpapiers ist primär, dass durch das Arzneimittel-Neuordnungsgesetz dem G-BA und dem IQWiG neue Aufgaben hinsichtlich der Bewertung des Zusatznutzens von Arzneimitteln zugewiesen worden sind und folglich eine Erweiterung des Methodenspektrums erforderlich wurde. Inwieweit die in der Verordnung des BMG und in der Verfahrensordnung des G-BA festgelegten Dossiers methodisch akzeptable Bewertungen eines Zusatznutzens ermöglichen, wird die praktische Umsetzung zeigen. Das Methodenpapier fasst hier im Wesentlichen die für das Dossier erforderlichen Informationen zusammen.

Außerdem ist der Abschnitt Methoden der Versorgungsqualität neu gefasst worden. Im Methodenpapier 3.0 waren nur die Abschnitte Methoden zu Leitlinienbewertung, Qualitätsprüfung und Disease-Management-Programmen enthalten. Der neu enthaltene Abschnitt Versorgungsanalyse beschreibt auf 11 Seiten Ziele, Fragestellungen und Strukturen von Berichten zur Versorgungsanalyse. Hier ist festzustellen, dass dieser Abschnitt dem Anspruch eines Methodenpapiers nicht gerecht wird. Es ergeht die Empfehlung, in der aus dem Methodenpapier 3.0 bewährten Form sich auf die konkreten Arbeitsaufgaben des Instituts zu beschränken und die darauf bezogenen Methoden so konkret und verständlich wie möglich darzustellen.

Die Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. hat im Jahr 2010 eine Präsidiumskommission gebildet, die aus Sicht unserer Fachgesellschaft die Arbeiten des IQWiG kritisch begleiten wird. Zur Thematik des erweiterten Metho-

.../2

gmds
Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V.

Geschäftsstelle

[redacted]

Geschäftsführung

Beatrix Behrendt

Präsidium

Prof. Dr. Johannes Haerting (Halle/Saale), Präsident

Prof. Dr. Heike Bickeböller (Göttingen), 1. Vizepräsidentin

Prof. Dr. Klaus A. Kuhn (München), 2. Vizepräsident

Prof. Dr. Harald G. Schweim (Köln), Schatzmeister

PD Dr. Andreas Wienke (Halle/Saale), Schriftführer

Prof. Dr. Armin Koch (Hannover), Beisitzer

Prof. Dr. Andreas Ziegler (Lübeck), Beisitzer

Prof. Dr. Petra Knaup-Gregori (Heidelberg), Fachbereichsleiterin

Prof. Dr. Dieter Hauschke (Freiburg), Fachbereichsleiter

PD Dr. Antje Timmer (Bremen), Fachbereichsleiterin

Susanne Stolpe (Bochum), Sektionsleiterin

Sabine Kapsammer (Mannheim), Sektionsbeisitzerin

[redacted]

denpapiers wird sich die Kommission im Mai 2011 treffen und nach Diskussion mit Vertretern des IQWiG eine ausführlichere Stellungnahme abgeben. Es wird jedoch jetzt schon darauf hingewiesen, dass der vom IQWiG vorgegebene Zeitraum zur Kommentierung sehr knapp bemessen war. Besteht seitens des IQWIG die Erwartung, dass sich die Fachöffentlichkeit wissenschaftlich mit den vertretenen Positionen zur Nutzenbewertung auseinandersetzt, dann müssen andere Zeitlinien gefunden werden. Dies sollte im Hinblick auf eine notwendige Abstimmung mit den entsprechenden Fachgesellschaften in Zukunft von Seiten des IQWiG berücksichtigt werden.

Prof. Dr. J. Haerting
(Präsident der GMDS)

A 1.8 Lundbeck

Autoren:

Friede, Michael
Janetzky, Wolfgang
Kessel-Steffen, Markus

Gutachter:

Röhmel, Joachim.
Brieden, Andreas

Lundbeck GmbH

[redacted]

Institut für Qualität und Wirtschaftlichkeit
im Gesundheitswesen (IQWiG)
Herrn Prof. Dr. med. Windeler
Dillenburger Str. 27

51105 Köln

05.04.11
SciU-Dr.Fr/nb ✧ ☎ [redacted]
E-mail: [redacted]

Entwurf der Allgemeinen Methoden 4.0

Sehr geehrter Herr Professor Windeler,

das IQWiG hat den Entwurf der Allgemeinen Methoden 4.0 vom 09.03.2011 veröffentlicht und die Möglichkeit eröffnet, zu diesem Entwurf eine Stellungnahme abzugeben.

Die Lundbeck GmbH möchte diese Möglichkeit nutzen, um das IQWiG mit einer Stellungnahme zu unterstützen.

Sehr gerne stehen wir Ihnen für Ihre Fragen und einen weiteren Dialog, auch im Rahmen einer wissenschaftlichen Anhörung, zur Verfügung.

Mit freundlichen Grüßen

Lundbeck GmbH

i. V. Dr. Michael Friede
(Leitung Scientific Unit)

i. V. Dr. Markus Kessel-Steffen
(Health Care Affairs)

i.V. Dr. Wolfgang Janetzky
(Scientific Unit)

[redacted]

Stellungnahme der Lundbeck GmbH zum Entwurf der Methoden, Version 4.0 (Entwurf vom 09.03.2011)

[redacted]

A. Beschreibungen der Anregungen

1. Produkte des IQWiG

Das IQWiG erstellt unterschiedliche Produkte, die in Kapitel 2 ‚Produkte des Instituts' behandelt werden. Neu hinzugekommen sind die in Abschnitt 2.1 ‚Produktspezifische Verfahrensabläufe' genannten

- Dossierbewertungen und
- Stellungnahmen des IQWiG.

Das IQWiG kann vom Gemeinsamen Bundesausschuss (G-BA) gemäß §35a SGB V mit einer Dossierbewertung beauftragt werden. Dabei bewertet das IQWiG den Nutzen auf der Grundlage eines Dossiers, das vom Hersteller eines neuen Wirkstoffs eingereicht wurde.

Eine Stellungnahmemöglichkeit des Herstellers zur Bewertung seines Dossiers nach Erstellung der Bewertung durch das IQWiG ist nicht vorgesehen, vielmehr wird auf das weitere Verfahren des G-BA verwiesen.

Stellungnahmen des IQWiG können u. a. vom G-BA beauftragt werden, wenn sich nach Erstellung eines Produkts des IQWiG ein zusätzlicher Bearbeitungsbedarf ergibt. Auch bei Stellungnahmen ist keine Anhörung vorgesehen.

Vorschläge der Lundbeck GmbH zu den genannten Produkten

Die Lundbeck GmbH hat bereits zu den Entwürfen der Allgemeinen Methoden 2.0 und 3.0 Stellungnahmen abgegeben. Die dort genannten Punkte sollen an dieser Stelle nicht wiederholt werden, vielmehr wird auf die im Anhang befindlichen Stellungnahmen verwiesen.

Bezüglich der Dossierbewertungen, Rapid Reports und Stellungnahmen des IQWiG müssen u. a. Arzneimittelhersteller, Sachverständige der Fachkreise und Betroffene die Möglichkeit haben, sich an jedem Schritt des Verfahrens mit Stellungnahmen zu beteiligen. Nur so wird eine notwendige Transparenz des Verfahrens geschaffen. Daher sollte grundsätzlich das Recht zu Stellungnahmen nach Veröffentlichung eines Produkts des IQWiG eingeräumt werden. Das endgültige, dem Auftraggeber zur Verfügung zu stellende Produkt sollte sowohl die Bewertungen durch das IQWiG als auch die der Stellungnehmenden widerspiegeln.

2. Evidenzbasierte Medizin (EBM)

Das IQWiG gründet die Erstellung von Produkten auf die evidenzbasierte Medizin. Dabei bezieht sich das IQWiG auf die Definition von Sackett et al. (1996), wonach

> *„EbM […] der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen, wissenschaftlichen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten [ist]. Die Praxis der EbM bedeutet die Integration individueller klinischer Expertise mit der bestverfügbaren externen Evidenz aus systematischer Forschung."* (vgl. IQWiG 2011)

Entgegen der Aussage, dass im Rahmen der EbM auf bestverfügbare externe Evidenz Bezug genommen werden soll, kommt das IQWiG zu der Aussage, dass nur randomisierte, kontrollierte, verblindete klinische Studien (RCTs) bei der Bewertung des Nutzens eines Wirkstoffes grundlegend sind. Die Verwendung der bestverfügbaren Evidenz zur Beurteilung eines Wirkstoffes wird auch vom G-BA gefordert (BMG 2010). Die ausschließliche Beschränkung der Verwendung von RCT ist somit im Rahmen der Dossierbewertung nicht möglich und steht dem Willen des Gesetzgebers entgegen.

Vorschläge der Lundbeck GmbH zur Evidenz-basierten Medizin

Nach den internationalen Standards der Evidenz basierten Medizin sind in Abhängigkeit von der Fragestellung und der verfügbaren Evidenz alle Studien in eine Bewertung einzubeziehen. Hier kann eine Graduierung – wie im Rahmen der EbM vorgesehen – erfolgen. Ein Ausschluss von Daten aus nicht-randomisierten Studien ist nach internationalen Standards nicht statthaft.

3. Beurteilung klinischer Relevanz / Irrelevanz

Bei der Bewertung der klinischen Relevanz zieht das IQWiG zwei prinzipielle Möglichkeiten in Betracht. Eine Bewertung anhand von

- Mittelwertdifferenzen oder
- Responderanalysen.

Bei der Interpretation der Mittelwertdifferenzen nennt das IQWiG mehrere Möglichkeiten:

- eine skalenspezifische, validierte bzw. etablierte Irrelevanzschwelle für den Gruppenunterschied bei nicht Vorliegen dieser Schwelle:
- standardisierte Mittelwertdifferenzen (Hedges' g) mit einer Irrelevanzschwelle von 0,2

In der weiteren Diskussion der Bewertungsmöglichkeiten geht das IQWiG explizit auf die Bedeutung von Irrelevanzgrenze, Konfidenzintervall eines Effektschätzers und die Lage des Konfidenzintervalls zur Irrelevanzgrenze ein.

> *„Falls skalenspezifisch validierte bzw. etablierte Kriterien zur Relevanzbewertung nicht vorliegen, muss dafür auf ein allgemein statistisches Maß zurückgegriffen werden. In diesem Fall werden standardisierte Mittelwertdifferenzen (SMD in Form von Hedges' g) betrachtet. Als Irrelevanzschwelle wird dann 0,2 SMD verwendet: Liegt das zum beobachteten Effekt korrespondierende Konfidenzintervall vollständig oberhalb dieser Irrelevanzschwelle, wird davon ausgegangen, dass die Effektstärke nicht in einem sicher irrelevanten Bereich liegt, Dies soll gewährleisten, dass der beobachtete Effekt hinreichend sicher mindestens als klein angesehnen werden kann.“* (IQWiG 2011)

Die Anwendung der Grenze von 0.2 für SMD wird in Abschnitt 7.3.3 des Entwurfes für den Fall angekündigt, dass ‚skalenspezifische validierte bzw. etablierte Kriterien zur Relevanzbewertung nicht vorliegen'. Dies vermittelt den Eindruck, dass diese Regelung nur selten und im Ausnahmefall angewendet wird. Dieser Eindruck trügt

jedoch. Bei mehreren Nutzenbewertungsverfahren zu Memantine, SSRIs und SNRIs verwendet das IQWiG nachträglich selektierte ‚patientenrelevante' Endpunkte zur Beurteilung der Relevanz, für die sich meist keine Relevanz- und Irrelevanzschwellen in den Guidelines finden lassen. Es ist daher zu befürchten, dass das IQWiG in schematischer Anwendung seiner Entscheidungshierarchie zwangsläufig diese Default-Option in größerem Umfang verwenden wird, was das Risiko von Fehlbewertungen potentiell erhöhen kann.

Die Nutzenbewertung über das Kriterium smd ist nur auf Basis international akzeptierter Regeln möglich. Für dieses Kriterium ist der Umfang des Zusatznutzens zu definieren. Dies gilt umso mehr als dass Indikationsgebieten existieren, bei denen Fortschritte in kleinen Schritten erzielt werden können (z.B. im Bereich ZNS), längerfristig der Status quo zementiert wird, unabhängig davon, ob diese Fortschritte kosteneffizient sind oder nicht. Dies würde im gegenwärtigen Verfahrensablauf gar nicht mehr geprüft werden, da Kosten-Nutzenbewertungen nur dann durchgeführt werden, wenn der Beleg eines Zusatznutzens erbracht wurde. Die Gesundheitssysteme laufen dabei Gefahr, dass die Entwicklung nützlicher und eventuell sogar kostengünstigerer innovativer Therapien unterbleibt, die zwar möglicherweise nur wenig Zusatznutzen aufweisen, aber zu keiner oder einer akzeptablen Verteuerung der Therapie und damit zu einer kosteneffizienten Versorgung beitragen könnten. Solche Therapien können jedoch für bestimmte Patientengruppen wertvolle Therapiealternativen darstellen, wenn z.B. Alternativen versagt haben oder nicht verträglich waren.

Vorschläge zur Bewertung der klinischen Relevanz

Bei der Bewertung des Nutzens eines Wirkstoffes bezieht sich das IQWiG – wie bereits dargestellt - besonders auf

- Mittelwertdifferenzen und
- Responderanalysen.

Responderanalysen sind primärer zur Bewertung der klinischen Relevanz heranzuziehen, da die den Responderanalysen zugrunde liegenden Definitionen indikationsspezifisch sind und unmittelbar den Nutzen eines Wirkstoffs widerspiegeln. Daher besteht allgemeiner Konsens über die Bedeutung von Responderanalysen zur Bewertung der klinischen Relevanz (Broich et al., 2007, EMA 2008, FDA 2006, Kieser et al, 2004).

Auch das IQWiG hat die Relevanz von Responderanalysen in der Vergangenheit mehrfach bestätigt, u. a. im Berichtsplan zum Ergänzungsauftrag der Bewertung von Acetylcholinesterase-Hemmern (A09-05).

> *„Bei der Betrachtung von Responderanalysen ist es erforderlich, dass bei diesen Analysen ein validiertes bzw. etabliertes Responsekriterium angewendet wurde (im Sinne einer individuellen Minimal Important Difference [MID]). Liegt bei einer solchen Auswertung ein statistisch signifikanter Unterschied der Anteile der Responder zwischen den Gruppen vor, wird dies als relevanter Effekt angesehen, da die Responsedefinition bereits eine Schwelle der Relevanz (nämlich die MID) beinhaltet."* (IQWiG, 2010)

Der direkte Bezug der Responderdefinitionen zum Indikationsgebiet stellt sicher, dass signifikante Unterschiede zwischen den Behandlungsgruppen ein Beleg für den Nutzen sind (Winblad et al., 2001).

Im Gegensatz zu Reponderanalysen kann der Nutzen eines Wirkstoffes mit Hilfe von Effektgrößen nur indirekt abgeleitet werden, da Effektgrößen dimensionslose Zahlen darstellen, die vor dem Hintergrund des Krankheitsgeschehens mit medizinischem Sachverstand interpretiert werden (Cohen 1988, Brieden 2009, s. Anhang).

Im Entwurf der Allgemeinen Methoden rekurriert das IQWiG auf die Einteilung von Effektsstärken (Hedges' g) in Nutzengraduierungen, ohne die damit verbundene Problematik bewusst zu thematisieren bzw. darauf einzugehen, dass diese ‚Grenzwerte' (Hedges' g, Cohens' d) lediglich als Hilfskonstrukte anzusehen sind (IQWiG, 2009).

Die Verwendung von Effektstärken geht auf Cohen (1988) zurück, der dieses biometrische Maß in sozialwissenschaftlichen Studien zu Fragen der Lebensqualität entwickelt hat und vorschlug, Effektstärken mit Werten von 0,2 als „klein", von 0,5 als „mittel" und ab 0,8 als „groß" zu operationalisieren. Cohen hat jedoch ausdrücklich darauf hingewiesen, dass die Zahlenwerte rein subjektiv festgelegt sowie die verwendeten Begrifflichkeiten willkürlich gewählt sind und daher missverstanden werden können:

> *„The values chosen had no more reliable a basis than my own intuition."*
>
> *„... the author proposes, as a convention, effect size values to serve as operational definitions of the qualitative adjectives "small", "medium" and "large". This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as "large" are sometimes understood as absolute, sometimes as relative; and thus run a risk of being misunderstood."*
> (Cohen, 1988)

Des weiteren macht Cohen selbst darauf aufmerksam, dass die standardisierte und dimensionslose Effektgröße oder –stärke d (entsprechend auch die leicht modifizierte Größe Hedge's g) eine rein rechnerische technische Größe ist, die mit fachlicher Substanz gefüllt werden muss, d.h. die (minimale) relevante Differenz (minimal important difference, MID), die im Zähler von d bzw. g steht, muss sachlich, d.h. in diesem Zusammenhang medizinisch gut begründet werden und Akzeptanz in der medizinischen Fachwelt besitzen.

Die Nennergröße von d und g, die die Gesamtvariabilität in der untersuchten Population beschreibt, wird dabei durch viele Faktoren beeinflusst. Je heterogener eine Population zusammengesetzt ist, desto größer wird erwartungsgemäß die Variabilität ausfallen. Die großen Multicenterstudien, die in der Phase 3 der klinischen Entwicklung durchgeführt werden, sind bekanntermaßen hinsichtlich der Zielvariablen sehr heterogen, so dass das Verhältnis aus MID und Gesamtvariabilität zwangsläufig kleiner wird, obwohl sich an der durch die MID ausgedrückten Relevanz des Effektes selbst nichts geändert hat.

Die Überprüfung von verschiedenen Studienresultaten aus klinischen Studien aus der Literatur und in der ClinicalTrials.gov Studien-Datenbank zeigt, dass auch rein formal sehr kleine d- bzw. g-Werte mit relevanten und akzeptierten Effekten von Arzneimitteln im Einklang stehen können. Eine schematische Anwendung bestimmter Grenzwerte für

d und g, auch wenn dies nur auf den Fall fehlender Informationen hinsichtlich der MID beschränkt bleiben sollte, geht daher an der Realität vorbei und vergrößert das Risiko einer falschen Nutzenbewertung. Dieses Risiko wird noch dadurch gesteigert, dass diese Kriterien nicht als Argumente in der Diskussion und Abwägung von Nutzenbewertungen, sondern zum Entscheidungskriterium erhoben werden, ob bezogen auf den jeweiligen Endpunk überhaupt Nutzen zu attestieren sei oder nicht. Zudem erweckt das IQWiG damit auch den Eindruck, dass sich primär medizinische Bewertungsprobleme, die ganz klar als Werturteil zu klassifizieren sind, durch schematische Anwendung standardisierter Effektgrößen auflösen ließen.

Das IQWiG selbst hat mittlerweile konzediert, dass Cohen's d (Hedge's g) nicht als Relevanzmaß anzusehen ist, sondern nur als Ersatzgröße, die in letzter Instanz herangezogen wird, wenn nichts anderes vorhanden ist. Trotz dieser pragmatischen Einschätzung ändert sich aber nichts an der Tatsache, dass es als verteilungsbasiertes Maß für diesen Zweck prinzipiell ungeeignet ist und insbesondere in der bisher praktizierten Operationalisierung als Entscheidungskriterium nicht verwendet werden sollte.

Die Frage der klinischen Relevanz ist letztlich ein Werturteil, das nicht allein aus statistischer Sicht beantwortet werden kann. Dieses Urteil hängt unter anderem von der Indikation und den vorhandenen Alternativen ab.

Des Weiteren sieht der Entwurf der Allgemeinen Methoden neben der Einführung einer wissenschaftlich nicht begründbaren Irrelevanzgrenze vor, die untere Grenze des Konfidenzintervalls des Effektschätzers mit der Irrelevanzgrenze in Relation zu setzen. Das Erreichen des Effektschätzers dieser Grenze genügt demnach nicht aus, um klinische Relevanz zu zeigen. Die Grenze 0,2 wird auf das untere Konfidenzintervall ausgedehnt, was einer Verschiebung der Nullhypothese entspricht. Auch dieses Vorgehen ist wissenschaftlich nicht begründbar.

Die genannte Kritik findet sich auch in einem Diskussionspapier von Professor Röhmel, das dem G-BA mit der Stellungnahme der Lundbeck GmbH im laufenden Verfahren zu Memantin vorgelegt wurde. Demnach ist die Vorgehensweise des IQWiG bei der Nutzenbewertung rein biostatistisch getrieben (Röhmel, 2010; siehe Anhang). Explizit wird auf die starke Limitation der biometrischen Ableitung der Irrelevanzgrenze verwiesen. Professor Röhmel folgert:

> *„Aus meiner Sicht ist mit den vorgebrachten biostatistischen Ableitungen eine Irrelevanzgrenze von Cohen's d=0.2 (im normierten Fall für σ=1) auch nicht begründbar. Für mittlere relevante Effekte δ_r= 0.5σ liegt die empfohlene klinische Irrelevanzgrenze δ_i= 0.2σ aus biometrischer Sicht zu hoch.*"
> (Röhmel, 2010; siehe Anhang)

Sind diese Faktoren erreicht, d.h. liegt das untere Konfidenzintervall vollständig oberhalb der Irrelevanzgrenze, handelt es sich laut Methodenpapier allerdings noch nicht um einen relevanten Effekt, sondern um einen nicht sicher irrelevanten Effekt. Wann ein Studienergebnis als relevant angesehen werden kann geht aus dieser Methodik nicht hervor. Es entsteht eine nicht definierte „Grauzone".

Das IQWiG würde sich entgegen den anerkannten Standards zur Bewertung der klinischen Relevanz setzen (Brieden, 2009; siehe Anhang), da bei einer Bewertung mittels dimensionsloser Effetktschätzer medizinische Sachkenntnis einbezogen werden muss.

Die im Entwurf zu den Allgemeinen Methoden beschriebene Vorgehensweise wurde u. a. bei der Bewertung von Memantin gewählt. Dass eine rein biometrisch Bewertung ohne medizinische Sachkenntnis die Bewertung ad absurdum führt, zeigt auch die deutliche Kritik an der Vorgehensweise des IQWiG der S3-Leitlinie „Demenz":

> *„Die Absprechung des Nutzens von Memantin bei der mittelschweren bis schweren Demenz durch das IQWiG ist somit durch ein formal statistisches Kriterium begründet. Dieses formalisierte Vorgehen der Festlegung einer Effektgröße von d=0.2 als nicht relevant ist empirisch für die mittelschwere bis schwere Demenz nicht begründet."* (DGN/DGPPN, 2009)

Molnar et al. (2009) weisen in einem systematischen Review zur Bewertung der klinischen Signifikanz unter Antidementiva darauf hin, dass keine evidenz-basierten Schwellenwerte existieren, mit denen die klinische Relevanz auf Basis von Effektstärken bewertet werden kann.

> *„Effect size does not tell us that a result actually is clinically important and does not set evidence-based thresholds at which outcome measures are felt to be clinically important."* (Molnar et al., 2009)

Insgesamt gesehen handelt es sich bei diesem Vorgehen um eine willkürliche, methodisch nicht validierte und wissenschaftlich nicht fundierte Methodik. Die Entscheidung über Nutzen bzw. Zusatznutzen einer Therapie und damit im späteren Verlauf über die Erstattungsfähigkeit darf nicht auf Grundlage dieser Methode basieren.

4. Heterogenität in Meta-Analysen

Es wird ausgeführt, dass das Institut bei „zu hoher" Heterogenität keine Meta-Analyse durchführen wird, wenn die einzelnen Studienergebnisse nicht einen deutlichen und gleichgerichteten Effekt zeigen. Eine Aussage über die Definition von „zu hoch" wird hierbei nicht gemacht.

Vorschlag der Lundbeck GmbH zur Heterogenität in Meta-Analysen:

Die Grenzen, die zu einer veränderten Vorgehensweise oder gar veränderten Beurteilung führen, wie zu hohe Heterogenität, sollten genannt werden. Es sollte bei vorliegen „zu hoher" Heterogenität die verantwortlichen Faktoren untersucht werden. Eine Nutzenbewertung darf aufgrund „zu hoher" Heterogenität einer Meta-Analyse nicht ausbleiben.

5. Indirekte Vergleiche

Die im Punkt 7.3.9 beschriebene Methodik der Indirekten Vergleiche ist sehr allgemein gefasst und beschreibt auch die Komplexizität und die Unausgereiftheit dieses Verfahrens:

> *„…Allerdings gibt es noch zahlreiche ungelöste methodische Probleme, sodass gegenwärtig von einer routinemäßigen Anwendung dieser Methoden im Rahmen der Nutzenbewertung abzuraten ist" (IQWiG 2011)*

Jedoch im gleichen Abschnitt kommt das IQWiG zu dem Schluss, dass von Indirekte Vergleiche in der Nutzenbewertung abzuraten ist, sehr wohl aber in der Kosten-Nutzenbewertung Anwendung finden kann:

> *In bestimmten Situationen wie z. B. bei Bewertungen des Nutzens von Arzneimitteln mit neuen Wirkstoffen [116] sowie bei Kosten-Nutzen-Bewertung (siehe unten) kann es jedoch erforderlich sein, indirekte Vergleiche einzubeziehen und daraus Aussagen für die Nutzenbewertung unter Berücksichtigung einer geringeren Ergebnissicherheit abzuleiten. (IQWiG 2011)*

Obwohl es sich bei den Indirekten Vergleichen um ein sehr komplexes Verfahren handelt, bleibt das Institut in seiner Beschreibung sehr unkonkret und legt sich auf keine Vorgehensweise fest. Damit sind Tür und Tor für eine Anwendung dieses Tools in allen Situationen geöffnet.

Vorschlag der Lundbeck GmbH zu Indirekten Vergleichen :

Um die Nachvollziehbarkeit der Entscheidung für Indirekte Vergleiche transparent zu gestallten, sollten eindeutige Kriterien vom IQWiG beschrieben, und auf schwammige Formulierungen verzichtet werden. Es wäre auch zu überlegen, ob die Möglichkeit eines vereinfachten Vorgehens im Bereich der Indirekten Vergleiche mögliche wäre.

B - Fazit:

1. Bei allen Produkten des IQWiG ist im Sinne eines transparenten Verfahrens eine Stellungnahmemöglichkeit einzuräumen.
2. Bei der Bewertung des Nutzens eines Wirkstoffs ist die verfügbare Evidenz zu berücksichtigen. Eine Beschränkung auf RCTs kann dazu führen, dass einzelne Fragestellungen nicht beantwortet werden und ein Bias entsteht.
3. Zur Bewertung der klinischen Relevanz sind Responderanalysen die bisher best verfügbare Methode, da sie unmittelbare Aussagen zum Nutzen einer Therapie zulassen. Sie sind daher als primär anzusehen.
4. Die Effektstärke (z.B. Cohen's d) kann berechnet werden, um statistische Unterschiede zwischen zwei Gruppen zu zeigen. Die mathematische Ableitung des Nutzens auf der Basis einer berechneten Effektstärke und willkürlich gesetzter Irrelevanzgrenzen entspricht jedoch nicht den internationalen Standards und ist nicht statthaft.
5. Wenn die berechnete Effektstärke aussagefähig sein soll, muss sie unter Einbezug medizinischer Sachkenntnis interpretiert und hinsichtlich der klinischen Relevanz bewertet werden. Dies wird im Entwurf der Allgemeinen Methoden des IQWiG nicht deutlich.

Literatur

Nr.	Feldbe-zeichnung	Text
1	Au:	Broich, K.
	TI:	Outcome measures in clinical trials on medical products fort he treatment of dementia: A European regulatory perspective
	SO:	Int Psychogeriatr 2007, 19(3): 509-524
2	Au:	Bundesministerium für Gesundheit (BMG)
	TI:	Gesetz zur Neuordnung des Arzneimittelmarktes in der gesetzlichen Krankenversicherung (Arzneimittelmarktneuordnungsgesetz – AMNOG)-
	SO:	http://www.bmg.bund.de/krankenversicherung/arzneimittelversorgung/amnog/amnog.html
3	Au:	Cohen. J.
	TI:	Statistical Power Analysis for the Behavioral Science
	SO:	2nd Edition, 1988
4	Au:	DGN / DGPPN (Hrsg.)
	TI:	S3-Leitlinien "Demenzen"
	SO:	http://media.dgppn.de/mediadb/media/dgppn/pdf/leitlinien/s3-leitlinie-demenz-kf.pdf, 2009
5	Au:	EMA
	TI:	Guideline on Medical products for the treatment of Alzheimer's disease and other dementias
	SO:	CPMP/EWP/553/95 Rev. 1 http://www.EMA.europa.eu/pdfs/human/ewp/055395en.pdf, 2008
6	Au:	FDA
	TI:	Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Label Claims. 32 Seiten
	SO:	http://www.fda.gov/ScienceResearch/SpecialTopics/RunningClinicalTrials/ProposedRegulationsandDraftGuidances/default.htm, 2006
7	Au:	IQWiG
	TI:	Allgemeine Methoden (Entwurf). Version 4.0 v. 09.03.2011
	SO:	https://www.iqwig.de/download/IQWiG_Entwurf_Methoden_Version_4-0.pdf, 2011
8	Au:	IQWiG
	TI:	Bupropion, Mirtazapin und Reboxitin bei der Behandlung der Depression – Dokumentation und Würdigung der Anhörung zum Vorbericht. Auftrag A05-20C, Version 1.0 v. 05.11.2009
	SO:	http://www.iqwig.de/download/A05-20C_DWA_VB_Bupropion_Mirtazapin_und_Reboxetin_bei_Depressionen.pdf, 2009
9	Au:	IQWIG
	TI:	Cholinesterasehemmer bei Alzheimer Demenz – Ergänzungsauftrag: Rivastigmin-Pflaster und Galantamin, 2010
	SO:	http://www.iqwig.de/download/A09-05_Berichtsplan_Cholinesterasehemmer%E2%80%93Ergaenzungsauftrag_Rivastigmin-Pflaster_Galantamin.pdf
10	Au:	Kieser, M. et al.
	TI:	Power and sample size determination when assessing the clinical relevance of trial results by ‚responder analyses'
	SO:	Statistics Med 2004, 23: 3287-3305
11	Au:	Molnar, F.J. et al.
	Ti:	Systematic Review of Measures of Clinical Significance Employed in Randomized Controlled Trials of Drugs in Dementia
	SO:	J Am Geriatr Soc 2009, 57: 536-546

12	Au:	Winblad, B. et al.
	Ti:	Pharmacotherapy of Alzheimer's Disease: is there a Need to Redifine Treatment Success?
	SO:	Int J. Geriatr Psychiatry 2001, 16: 653-666

Schwellenwert Bestimmung zur Beurteilung klinischer Relevanz und Responsekriterien bei psychometrischer Skalen.

Diskussionspapier
Joachim Röhmel

Hintergrund

Bei der Beurteilung von Medikamenteneffekten in Bezug auf „Beeinträchtigung der Aktivitäten des täglichen Lebens“ in der Indikation „moderate bis schwere Erkrankung bei der Alzheimerschen Demenz (AD)“ spielen zwei psychometrische Skalen eine bedeutende Rolle: ADCS-ADL-sev[1] bei Patienten mit schwerer AD und ADCS-ADL_{23}[2] bei Patienten mit moderater AD. Die ADCS-ADL-sev besteht aus 19 Items, die einen ganzzahligen Summenscore zwischen 0 und 54 ergeben können. ADCS-ADL_{23} hat 23 Items, und die Ergebnisse in den 23 Items werden zu einem ganzzahligen Summenscore zwischen 0 und 78 addiert.

Im Abschussbericht A05-19C des IQWiG zu Memantin bei AD werden für diese Skalen und die darauf beruhende Nutzenbewertung verbindliche rein biostatistische Definitionen zur klinischen Relevanz getroffen allein basierend auf Streuungsmaße, Mittelwertsunterschiede und zugehörige Konfidenzintervalle. In der Veranstaltung „IQWiG im Dialog“ vom 18. Juni 2010 sind diese Definitionen noch einmal durch biostatistische Ableitungen erläutert worden. In dem vorliegenden Diskussionspapier will ich **zum Einen** kritisch auf diese biostatistischen Ableitungen eingehen. **Zum Weiteren** will ich darauf hinweisen, dass es sich im biostatistischen Sinne bei den verwendeten Skalen um diskrete und daher keineswegs um kontinuierlich verteilte Variable (Endpunkte klinischer Studien) mit unterliegender Normalverteilung handelt. Deswegen sind Begriffe wie Mittelwertsdifferenz und Streuung nur eingeschränkt deutbar wie unter Normalverteilung, insbesondere bei Nichtvorliegen von Normalverteilung oder wenigstens bei unsymmetrischen Verteilungen. **Abschließend** will ich diskutieren, wie der Vorschlag, für eine Responder – Nonresponder –Analyse als Abstandskriterium etwa die Hälfte der Streuung zu benutzen[3], sich bei diskret verteilten Daten auswirkt.

---

[1] **Aus IQWiG Arbeitsbericht A05-19C:“**Das bei den entsprechenden Studien als ADCS-ADL-19 beziehungsweise modifizierte ADCS-ADL bezeichnete Instrument ist ein Inventar von 19 Aktivitäten. Der Grad an Unabhängigkeit bei der Durchführung dieser Aktivitäten wird vom Kliniker mithilfe der Angaben vom Pflegenden beziehungsweise Angehörigen geschätzt. Die einzelnen Wertungen der Items werden zu einem Score addiert, wobei dieser mit abnehmenden Fähigkeiten der Patienten sinkt. Der Score liegt zwischen 0 und 54.“

[2] **Aus IQWiG Arbeitsbericht A05-19C:“**Das in den Publikationen der entsprechenden Studien als ADCS-ADL_{23} bezeichnete Instrument ist ein Inventar von 23 Aktivitäten. Der Grad an Unabhängigkeit bei der Durchführung dieser Aktivitäten wird vom Kliniker mithilfe der Angaben vom Pflegenden beziehungsweise Angehörigen geschätzt. Die einzelnen Wertungen der Items werden zu einem Score addiert, wobei dieser mit abnehmenden Fähigkeiten der Patienten sinkt. Der Score liegt zwischen 0 und 78.“

[3] **Aus IQWIG Berichte – Jahr: 2010 Nr. 74 version 1.0 vom 1.Juli 2010:** „In Ermanglung einer validierten MID könnte hierfür z.B. eine Differenz von0,5 Standardabweichungen gewählt werden (Median aus den vorliegenden Studien), da es für die Eignung eines solchen Kriteriums, zumindest im Bereich Lebensqualität, empirische Evidenz gibt[8].“

Zur biostatistischen Ableitung der Irrelevanzgrenze Cohen's d=0.2

Die Diskussion um Schwellenwerte zur klinischen Relevanz von beobachteten Unterschieden zwischen Therapiegruppen in Endpunkten klinischer Studien ist in jüngster Zeit durch IQWiG Entscheidungen wiederbelebt worden. Das bedeutet z.B. für das IQWiG, dass statistisch signifikante Effekte auch in Bezug auf ihre klinische Relevanz hin untersucht werden, und klinische Relevanz nicht konstatiert wird, wenn der Effektschätzer samt dem zugehörigen Konfidenzintervall für den wahren Effekt nicht vollständig über einem Schwellenwert liegt, der unter Berufung auf Cohen[4] bei 20% der Streuung angesetzt wurde. Einschränkend ist hinzuzufügen, dass eine solche Regel nur dann angewendet werden soll, wenn klinische Relevanz nicht anders bzw. auf klinischem Konsens beruhend durch einen anderen Schwellenwert festgelegt ist. Da aber solche anderweitigen Festlegungen Mangelware sind, kommt dem Vorgehen des IQWiG eine generelle Bedeutung zu, insbesondere bei den neurologischen und psychiatrischen Erkrankungen, in denen Skalen (wie die oben genannten) eingesetzt werden und eine Diskussion über die klinische Relevanz von gefundenen Unterschieden zwischen medikamentösen Therapien bisher nur ansatzweise, aber nicht abschließend stattgefunden hat. In der Veranstaltung „IQWiG im Dialog 2010 am 18. Juni 2010 hat PD Dr. Stefan Lange in einem (gemeinsam mit Thomas Kaiser, Yvonne Beatrice-Schüler, Guido Skipka, Volker Vervölgyi, Beate Wieseler erarbeiteten) Vortrag (downloadbar von der web-Seite des IQWiG) erläutert, wie sich die Gruppe[5] vorstellt, dass aus einer bekannten Relevanzschranke δ_r sich eine Irrelevanzschranke δ_i ableiten ließe, die dann in dem Sinne benutzt werden könne, dass klinische Relevanz nicht zu konstatieren ist, wenn der Effektschätzer samt dem zugehörigen Konfidenzintervall für den wahren Effekt nicht vollständig über δ_i liegt. Zunächst ist positiv festzustellen, dass in den Vorstellungen der Gruppe ein deutlicher Unterschied gemacht wird zwischen Schwellenwerten zur Relevanz und Schwellenwerten zur Irrelevanz. Dabei wird mit dem oft vorhandenen Missverständnis aufgeräumt, dass eine Schwellenwert Δ für klinische Relevanz eine vollständige Aufteilung in relevante und irrelevante Effekte verursacht: alle Werte unterhalb Δ bedeuten Irrelevanz und alle Werte über Δ bedeuten Relevanz. Zwischen anerkannt relevanten Effekten und anerkannt irrelevanten Effekten liegt vielmehr ein mehr oder weniger großer Graubereich.

Zum Thema „relevante klinische Effekte" gibt es aus meiner Sicht in vielen Indikationen und für viele klinische Endpunkte eine große Menge an Informationen, wenn man zustimmen kann, dass Effekte, für die in der Vergangenheit (und zur Zeit) klinische Studien geplant wurden (werden), als klinisch relevant bezeichnet werden dürfen. Andernfalls müsste man unterstellen, dass Planungen zu Studien, die oft mit Zulassungsbehörden international und wissenschaftlich beraten wurden, sich auf nicht relevante Effekte bezogen hätten. Deswegen kann man der Annahme der Gruppe zustimmen, dass es Sinn macht, von einer bekannten Relevanzschranke δ_r auszugehen. Daraus leitet die Gruppe rein biostatistisch eine Irrelevanzschranke δ_i ab nach der Formel

$$\delta_i = \delta_r - z_{0.975}\sqrt{2\frac{\sigma^2}{n}}$$

Dabei ist $z_{0.975}$ das 97.5% Quantil der Normalverteilung (i.e. gleich 1.96) entsprechend einem Signifikanztest auf dem einseitigen 2.5% Niveau, σ die angenommene (und als im

[4] z.B. Cohen J. Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, 1988, 2nd edition

[5] ich nenne die Autorengruppe im Folgenden der Einfachheit halber „Gruppe"

Wesentlichen bekannt vorausgesetzte) Streuung und n die Stichprobengröße pro Gruppe. Dieses n wiederum ist nach der bekannten Stichprobenumfangsformel zu berechnen, wenn (1-β) die Power betragen soll:

$$n = 2(z_{0.975}+z_{1-\beta})^2 \frac{\sigma^2}{\delta_r^2}$$

Nimmt man die Power als 90% (bzw. β=10%) an, so ergibt sich annähernd $\delta_i = 0.4\ \delta_r$, und somit für einen mittleren relevanten Effekt von $\delta_r = 0.5\sigma$ eine Irrelevanzschranke von $\delta_i=0.2\sigma$. Dies ist nun gerade die Irrelevanzschranke, die das IQWiG in der Vergangenheit verschiedentlich benutzt hat, und so scheint sich eine biometrische Begründung dieser zunächst willkürlichen Festsetzung zu ergeben.

Dies könnte auch suggerieren, dass bei vernünftiger Studienplanung (z.B. α=2.5% einseitig, β=10%) und bei einer mittleren relevanten Effektgröße von $\delta_r = 0.5\sigma$ irrelevante Effekte unter $\delta_i=0.2\sigma$ ausgeschlossen werden können. Dies ist aber selbst unter der Annahme der Normalverteilung nicht richtig. Im genannten Fall beträgt die Power der Studie, irrelevante Effekte unter $\delta_i=0.2\sigma$ auszuschließen gerade 50%. Mit dieser Power würde kein vernünftiger Sponsor eine Studie planen noch ein Ethik Komitee eine solche Studie befürworten. Um im genannten Fall wieder eine Power von 90% zu erreichen, müsste die Studie ursprünglich mit einer Power von 99.97% versehen werden, was zu einer Fallzahlerhöhung von 274% führte.

Fazit: aus meiner Sicht ist mit den vorgebrachten biostatistischen Ableitungen eine Irrelevanzgrenze von Cohen's d=0.2 (im normierten Fall für σ=1) auch nicht begründbar. Für mittlere relevante Effekte $\delta_r = 0.5\sigma$ liegt die empfohlene klinische Irrelevanzgrenze $\delta_i = 0.2\sigma$ aus biostatistischer Sicht zu hoch. Irrelevanzgrenzen führen in jedem Fall zu erhöhtem Aufwand für klinische Studien.

Eigenheiten psychometrischer Skalen am Beispiel von ADCS-ADLsev und ADCS-ADL$_{23}$

ADCS-ADL-sev und ADCS-ADL$_{23}$ sind diskrete Skalen mit einem ganzzahligen Wertevorrat von 0 bis 58 bzw. 0 bis 78. Dieser Wertevorrat ist jedoch nur theoretisch und wird bei weitem nicht ausgeschöpft. Bei diskreten Skalen und darüber hinaus bei nicht normal verteilten kontinuierlich verteilten Endpunkten ist die Rolle des Mittelwerts und der Streuung kritischer zu sehen. Dies beeinflusst auch Berechnungen von „effect size", da dieser genau aus Mittelwertsdifferenz und Streuung gebildet wird. Es ist daher möglich, dass das Effektmaß „effect size" hier aus biostatistischer Sicht den psychometrischen Skalen nicht gerecht wird. Laut Abschlussbericht A05-19C zentrieren sich die beobachteten Werte für Patienten mit moderater bis schwerer AD bei Studienbeginn (baseline) und für ADCS-ADL-sev um die Werte 30 bis 36 (abhängig von der Studienpopulation) und bei ADCS-ADL$_{23}$ um die Werte 50-53. Bei dem progredienten Verlauf der Krankheit ist bei Wiederholung der Messung nach 24 Wochen eher damit zu rechnen, dass der Scorewert weiter absinkt. Zieht man den baseline Wert vom Wert nach 24 Wochen ab (post-prä), so wird sich in der Mehrzahl ein negativer Wert ergeben. Im IQWiG Arbeitsbericht A05-19C bei ADCS-ADL-sev wurde eine Zentrierung der (ganzzahligen) post-prä Änderungen zwischen 0 und -5 Punkten beobachtet; bei Stichproben-Streuungen gab es Werte, die zwischen 6.5 und 8 lagen. Damit sind auch eine große Zahl positiver Änderungen wahrscheinlich oder aber die Häufigkeitsverteilung der Änderungen ist schief, was ebenfalls zu großen Streuungen führen kann. Es ist allgemein anerkannte Tatsache, dass die Stichprobenstreuung sehr sensitiv gegenüber Abweichungen von der Normalverteilung reagiert. Diese Möglichkeit, dass die post-prä Änderungen in den

Skalen (neben der Diskretheit) auch aus einer nicht normal verteilten Grundgesamtheit stammen könnten, wird vom IQWiG (IQWIG Berichte – Jahr: 2010 Nr. 74 version 1.0 vom 1.Juli 2010**)** durchaus eingeräumt, ohne allerdings daraus auch Konsequenzen für das „effect size“ zu ziehen. Diese könnten z.B. darin bestehen bei der Anwendung von „effect size“ eine Überprüfung der zugrunde liegenden Verteilungen durchzuführen auf Probleme wie Schiefe, Ausreißer oder andere Erscheinungen, welche die Stichproben Mittelwerte und Streuung verzerren können.

Fazit: Für die dem „effect size“ zugrunde liegenden Größen Mittelwertdifferenz und Streuung muss sicher gestellt werden, dass sie im verwendeten Zusammenhang sinnvoll sind und nicht zu Verzerrungen führen. Auch Alternativen müssten diskutabel sein bei ordinal kategoriell skalierten Endpunkten.

Dichotomisierung (Responder-Nonresponder) bei Skalen

Wegen der Unklarheit bei der Beurteilung von Mittelwertsdifferenzen basierend auf psychometrischen Skalen wird in Europäischen Guidelines zur Durchführung klinischer Studien im Bereich Neurologie/Psychiatrie immer öfter eine sogenannte Responder Analyse verlangt. Diese kann auf der Einschätzung einer einzelnen Skala beruhen, wie aber auch auf der Kombination der Einschätzungen mehrere Skalen[6]. Üblicherweise geschieht dies durch Festsetzung eines sogenannten „cut-off“. Bei der ADAS-cog (ebenfalls eine Skala mit Werten zwischen 0 und 70) z.B. ist ein Responsekriterium implizit etabliert, nämlich eine Verschlechterung um 4 Punkte oder mehr innerhalb eines halben Jahres als deutliche Verschlechterung (Non-responder) im kognitiven Bereich anzusehen. Warum hier 4 Punkte gewählt wurden, ist fachsubstantiell nicht belegt worden. Vermutlich ist diese Zahl als Konsens aus Diskussionen mit Zulassungsbehörden entstanden. Zwar gibt es methodische Vorschläge, wie auf der individuellen Patientenebene „cut-offs“ begründet werden könnten, doch sind diese häufig sehr aufwendig, da sie auf Auswertungen von Skalen in anderen Ebenen (z.B. Patientenzufriedenheit) angewiesen sind, und dann eben auch oft nur approximativ „cut-offs“ begründen können.

Zwar ist es auf der einen Seite wünschenswert, „cut-offs“ inhaltlich begründen zu können, auf der anderen Seite hat ein so gewählter „cut-off“ in vielen Fällen nur marginal veränderte Response Raten zur Folge gegenüber „cut-offs“ ,welche in der Nähe des begründeten „cut-off“ liegen und weniger gut begründet erscheinen. Bei diskreten Zahlen mit post-prä Änderungen überwiegend im einstelligen ganzzahligen Bereich sind zudem gar nicht zu viele „cut-offs“ möglich bzw. führen zu schwer bearbeitbaren Proportionen in der Nähe von 0 oder 100%.

Beispiel:
Im bereits genannten Arbeitspapier zu Memantin bei AD (IQWIG Berichte – Jahr: 2010 Nr. 74 version 1.0 vom 1.Juli 2010) wird für ADCS-ADL-sev bzw. ADCS-ADL_{23} ein „cut-off“ C in der Nähe von 50% der Streuung vorgeschlagen, da für diese wenigstens im Bereich der Lebensqualität empirische Evidenz gibt, und auch der 4 Punkte „cut-off“ bei der ADAS-cog in diesem Bereich liegt. Nimmt man weiter an, dass der Effekt unter Plazebo im Mittel -3.5 Punkte (Abfall) betrug, der unter Verum -2.2 (was einer Mittelwertsdifferenz von 1.3 Punkten entspricht) und variiert die Streuung S zwischen 7.0, 6.5 und 6.0 so ergeben sich bei Annahme einer Normalverteilung und „cut-offs“ C=-4, -3, -2, und -1 folgende theoretischen Non-Responder Anteile:

[6] Kieser M, Röhmel J, Friede T. Power and sample size determination when assessing the clinical relevance of trial results by “responder analyses”. Stat in Med 23, 2004, 3287-3306

	S=7.0			S=6.5			S=6.0		
C	%Verum	%Pla	%Diff	%Verum	%Pla	%Diff	%Verum	%Pla	%Diff
-4	52.8	60.1	-7.3	53.1	60.9	-7.8	53.3	61.8	-8.5
-3	47.2	54.5	-7.3	46.9	54.9	-8.0	46.7	55.3	-8.6
-2	41.5	48.9	-7.4	40.9	48.8	-7.9	40.1	48.7	-8.6
-1	36.0	43.2	-7.2	35.0	42.7	-7.7	33.8	42.1	-8.3

Die Annahme der Normalverteilung ist hier nur zur Illustration gemacht worden. Im konkreten Fall und bei ausreichend großem Datenumfang sind natürlich die empirischen diskreten Wahrscheinlichkeitsverteilungen als Basis zu nehmen.

Fazit: bei gegebener Streuung hat der „cut-off“ nur einen marginalen Anteil an der %Differenz zwischen den Gruppen. Wenn die zugrunde liegende Verteilungen nicht approximativ normal verteilt sind, so könnten sich hieraus natürlich größere bzw. kleinere Abweichungen ergeben.

Cohen's d:
Idee, Bedeutung, Interpretation, Anwendung und mögliche Fehlerquellen

Prof. Dr. Andreas Brieden

1. Überblick

Das sogenannte Cohen's *d* ist ein wichtiges und viel angewendetes Maß für die Quantifizierung der Effektgröße. Im Hinblick auf eine korrekte Interpretation der jeweiligen Größe ist ein genaues Verständnis der theoretischen Hintergründe und Ideen von zentraler Bedeutung. Insbesondere die Verwendung der in diesem Kontext häufig benutzten qualitativen Adjektive „small", „medium" und „large" in Verbindung mit den zugeordneten Werten *d=0,2* (kleiner Effekt), *d=0,5* (mittlerer Effekt) und *d=0,8* (großer Effekt) birgt Risiken, auf die bereits der Namensgeber selbst hinweist:

> „This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as ‚small', ‚medium', and ‚large' are sometimes understood as absolute, sometimes as relative; and thus they run a risk of being misunderstood." (Cohen, 1988, S. 12)

Insbesondere ist bei der Interpretation der spezifischen Umstände der jeweils konkreten Anwendung Rechnung zu tragen:

> „Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined … Large effects are frequently at issue in such fields as sociology, economics, and experimental and physiological psychology…" (Cohen, 1988, S. 13)

Wird obiger Ratschlag befolgt und die spezifische Anwendungssituation berücksichtigt, dann können „verblüffende" Feststellungen getroffen werden:

1. Ein "kleiner" Effekt bei einer Anwendung kann mitunter für diese viel bedeutendere Konsequenzen haben, als ein "großer" Effekt bei einer anderen Anwendung.

Doch es ist noch nicht einmal notwendig, verschiedene Anwendungen zu betrachten:

2. In Abhängigkeit von der Ausgangssituation kann bei derselben Anwendung ein „kleiner Effekt" größere Konsequenzen nach sich ziehen als ein „großer" Effekt.

In den nachfolgenden Kapiteln werden diese Erkenntnisse im Rahmen sehr allgemeiner Modellannahmen hergeleitet, wobei detailliertere mathematische und statistische Ausführungen bewusst in einen Anhang ausgegliedert sind. Auch ohne fundierte methodische Kenntnisse sollen die Ideen hinter dem wichtigen Konzept der Effektgrößen verständlich und die nachfolgende Empfehlung bei ihrer Verwendung nachvollziehbar sein.

> ***Die Anwendung von Cohen's d ist von zentraler Bedeutung, die Interpretation ist aber immer vor dem spezifischen Anwendungshintergrund durchzuführen.***

2. Ursache, Effekt und Konsequenz

Die prinzipielle Idee hinter einem statistischen Test und die grundlegenden Anforderungen an ihn lassen sich an einem einfachen Gedankenbeispiel erklären.

Angenommen bei einem Würfelspiel hängt der Erfolg vom Werfen der „6“ ab und die Spielteilnehmer möchten sicherstellen, dass der Würfel hinreichend fair ist, dass also in etwa einem Sechstel aller Fälle die „6“ gewürfelt wird. Gesucht ist ein Verfahren, dass im Wesentlichen zwei Eigenschaften aufweist. Erstens, gelangt der Test zu dem Ergebnis, der Würfel sei „unfair“ also gezinkt, so muss diese Aussage mit hoher Wahrscheinlichkeit stimmen. Zweitens sollte das Verfahren mit hoher Wahrscheinlichkeit garantieren, dass eine Verletzung der Hypothese, der Würfel ist fair, auch tatsächlich entdeckt wird. Dabei ist es naheliegend, eine gewisse Toleranzgrenze einzuräumen, d.h. eine geringe Abweichung von der Quote 1/6 zuzulassen.

Die Testentscheidung basiert auf der Auswertung von Testdaten, die etwa im Rahmen einer Stichprobe, hier etwa das wiederholte Werfen des Würfels erhoben werden. Die Entscheidung zu Gunsten oder zu Ungunsten der „Fairnesshypothese“ ist letztendlich eine Festlegung der Ursachen für ein beobachtetes Stichprobenergebnis. Liegt etwa ein erhöhter Anteil „6er“ vor und wird die Fairnesshypothese nicht verworfen, so wird als Ursache für das gehäufte Auftreten der Zufall verantwortlich gemacht. Mit anderen Worten, der Spieler hatte einfach Glück bzw. einfach Pech, die beobachteten Würfe sind im Rahmen der natürlichen, zufälligen Schwankungen zu erklären. Wird die Fairnesshypothese dagegen verworfen, wird als Ursache ein systematischer Grund unterstellt. Dabei wird der Würfel aber erst dann als „unfair“ angesehen, wenn der Effekt, sprich die offensichtliche Häufung der „6er“, hinreichend groß ist. Die Frage, welche Abweichung, also welche Effektgröße, noch akzeptabel ist, hängt unmittelbar von den aus dem Spielergebnis resultierenden Konsequenzen, beispielsweise dem monetären Gewinn oder Verlust ab.

Obiges Gedankenspiel ist übertragbar auf eine Vielzahl von unterschiedlichsten Anwendungsgebieten wie etwa Qualitätsmanagement oder auch Ernährungswissenschaften.

Im Rahmen des Qualitätsmanagements besteht die Aufgabe das von Konsumenten geforderte Qualitätsniveau durch entsprechende Qualitätskontrollen sicherzustellen. Insbesondere bedeutet dieses, dass Chargen einer Produktion, die dieses Niveau deutlich unterschreiten, auch tatsächlich als mangelhaft erkannt werden, dass so genannte Konsumentenrisiko also möglichst gering ist, siehe etwa Schwarze, 1986, S. 198. Die Festlegung, was unter „deutlicher“ Unterschreitung des Niveaus zu verstehen ist, also die Festlegung der Effektgröße, hängt von der zu erwartenden Reaktion der Kunden ab. Erst bei

Abwägung dieser Konsequenzen kann die Effektgröße, hier also das Ausmaß der Unterschreitung des Qualitätsniveaus, seriös angegeben werden.

In den Ernährungswissenschaften ist folgendes Szenario denkbar. Angenommen die derzeitige Ernährung einer Population ist im Hinblick auf die Versorgung mit Nährstoffen nicht ausreichend, um langfristig eine gewisse Entwicklung oder sogar das Überleben der Population zu gewährleisten. Dann muss die Ernährung nachweislich der Art umgestellt werden, dass nicht nur die aktuelle Ernährung verbessert wird, sondern etwa nach der Umstellung die Spezies einen vorgegebenen Mindestprozentsatz an Nährstoffen im Blut aufweist.

Für alle drei Beispiele ist ersichtlich, dass die als relevant festzulegende Effektgröße - synonym wird der Begriff Effektstärke verwendet - jeweils von den aus ihr resultierenden Konsequenzen abhängt, die also fallspezifisch zu berücksichtigen sind. Ziel der nachfolgenden Kapitel ist es, diese Beobachtung zu formalisieren und in den allgemeinen statistischen Kontext zu stellen.

3. Signifikanz und Effekt

Die Tatsache, dass schlussletztendlich die Konsequenzen die entscheidende Bedeutung haben, mindert nicht die Bedeutung der Effekte und ihrer Festlegung im Kontext der statistischen Tests. Im Rahmen eines solchen Tests wird eine Nullhypothese gegen die konträre Alternativhypothese getestet. Verschiedenste Typen von Hypothesen sind dabei denkbar, so können etwa statistische Variablen auf Unabhängigkeit oder auf einen speziellen Verteilungstyp getestet werden. Von besonderer Bedeutung sind die sogenannten parametrischen Tests, bei denen Hypothesen beispielsweise über Mittelwert und Streuung der zu Grunde liegenden Verteilung von Interesse sind. Bei der Testentscheidung gibt es zwei mögliche Fehlerquellen. Entweder die Nullhypothese wird fälschlicherweise abgelehnt, dann liegt ein Fehler 1. Art oder auch α-Fehler vor. Wird die Nullhypothese fälschlicherweise angenommen, also die korrekte Alternativhypothese fälschlicherweise abgelehnt, so liegt ein Fehler 2. Art oder auch β-Fehler vor. Die Wahrscheinlichkeit für einen Fehler 1. Art ist durch das **Signifikanzniveau α** nach oben beschränkt, das bei der Durchführung des Tests frei wählbar ist. Gängige Wahlen für das Signifikanzniveau sind fünf, ein oder auch ein Zehntel Prozent. Eigentliches Ziel des Tests ist es, die Nullhypothese zu verwerfen, d.h. es wird die Korrektheit der Alternativhypothese vermutet. Folglich ist mit möglichst großer Wahrscheinlichkeit eine korrekte Alternativhypothese anzunehmen, somit sollte die mit β bezeichnete Wahrscheinlichkeit für einen Fehler 2. Art, die korrekte Alternative zu Gunsten der falschen Nullhypothese zu verwerfen, möglichst gering sein. Für die Größe **1-β**, die sogenannte **Teststärke**, ist entsprechend ein möglichst hoher Wert erstrebenswert. Da letztlich nicht sicher ist – ansonsten müsste nicht getestet werden – welche der Hypothesen korrekt ist, sollten sowohl α als auch β klein gewählt werden. Die voneinander unabhängige Wahl ist aber im Allgemeinen nicht möglich, vielmehr kann der β-Fehler nicht besser als durch $1-\alpha$ nach oben beschränkt werden. Mit anderen Worten, die beiden Ziele der Fehlerreduzierung stehen sich konträr gegenüber (vgl. etwa Schira, 2005, S. 473 ff.). Die Lösung des Problems besteht in der Verwendung einer spezifischen Alternativhypothese, also in der Festlegung der gewünschten **Effektstärke *ES***. Dann kann mit Hilfe des vierten zu wählenden Parameters, dem **Stichprobenumfang *n***, der Test so konzipiert werden, dass beide Fehlerwahrscheinlichkeiten gering sind. Ein weiterer Nebeneffekt der Festlegung der Effektstärke ist ebenfalls bedeutend. Egal zu welchem Grad die Nullhypothese verletzt ist, mit Hilfe einer hinreichend großen Stichprobe kann diese Verletzung signifikant nachgewiesen werden. Folglich besteht die Gefahr, Effekte als statistisch signifikant auszuweisen, die für die Anwendung keinerlei Relevanz besitzen. Ein Zahlenbeispiel für diese Problematik findet sich beispielsweise in Bortz & Lienert, 2008, S. 49. Die hier zu Grunde liegenden Überlegungen werden mit Hilfe der Definition der Konsequenz in Kapitel 5 zunächst formalisiert und danach ausführlich analysiert.

Aufgrund der funktionalen Abhängigkeit zwischen Signifikanzniveau α, Teststärke 1-β, Effektstärke *ES* und Stichprobenumfang *n* sind theoretisch vier verschiedene Analysen möglich, bei denen jeweils ein Parameter als abhängig und die verbleibenden drei als variabel betrachtet werden. Das Buch *Statistical Power Analysis for the Behavioral Sciences* (Cohen, 1988) kann als Standardwerk für diese Thematik angesehen werden. Bei den Analysen konzentriert sich Cohen auf die Untersuchung von Teststärke beziehungsweise Stichprobenumfang als abhängige Parameter. Insbesondere die Effektstärke ist also jeweils vorzugeben. Für verschiedene Typen statistischer Tests werden daher zunächst geeignete, normierte Effektstärken eingeführt, darunter das sogenannte Cohen's *d* als Maß für die Effektstärke beim Vergleichstest für zwei Mittelwerte, das in den folgenden Kapiteln näher analysiert wird.

4. Effektstärke Cohen's d und ihre Interpretation

Die Vorgabe einer Effektstärke ist von fundamentaler Bedeutung für die Durchführung eines statistischen Tests. In Abhängigkeit von ihr können die spezifische Alternativhypothese formuliert und die weiteren freien Parameter so fixiert werden, dass etwa eine gewünschte Teststärke erzielt wird. Es liegt nahe, zu verlangen, dass Effektgrößen dimensionslos und somit insbesondere unabhängig von der Maßeinheit der Ursprungsdaten sein sollen. Ferner ist es vor dem Hintergrund der oben beschriebenen Analysen sinnvoll, Effektstärken unabhängig vom Stichprobenumfang n zu definieren, weil dieser die Rolle eines separaten Parameters übernimmt.

Ein für eine Vielzahl von Anwendungen wichtiger Test ist der Test auf Gleichheit der Mittelwerte μ_1 und μ_2 zweier Verteilungen, die annahmegemäß die gleiche Standardabweichung σ aufweisen. Als Effektstärke schlägt Cohen die durch σ normierte Differenz $d = \frac{\mu_2 - \mu_1}{\sigma}$ der beiden Mittelwerte vor, die offensichtlich beide formalen Anforderungen an eine Effektstärke erfüllt. Mit Hilfe dieses Ausdrucks, mittlerweile etabliert unter dem Namen Cohen's *d*, können Nullhypothese und insbesondere eine spezifische Alternativhypothese formuliert werden. Dazu bedarf es selbstverständlich der Vorgabe eines konkreten, üblicher Weise positiven Wertes für *d*, der laut Cohen jeweils anwendungsspezifisch gewählt werden muss.

Eine naheliegende Frage ist die nach der Interpretation des normierten *d*-Wertes. $d = \frac{\mu_2 - \mu_1}{\sigma}$ ist gleichbedeutend mit $\mu_2 = \mu_1 + \sigma \cdot d$, d.h. die Differenz zwischen den Mittelwerten ist $\sigma \cdot d$. Unter der Annahme, dass beide Verteilungen symmetrisch sind, sind jeweils 50 Prozent der Werte kleiner oder gleich dem jeweiligen Mittelwert der Verteilung. Somit kann zu jedem Wert für *d* derjenige Prozentsatz der Werte der ersten Verteilung angegeben werden, die kleiner sind als der Mittelwert $\mu_2 = \mu_1 + d \cdot \sigma$. Da im Falle von $d = 0$ und folglich $\mu_2 = \mu_1$ dieses 50 Prozent sind, wird der jeweilige Wert um jeweils 0,5 reduziert. Sind beide Verteilungen beispielsweise normalverteilt, so ergeben sich, wie in Abbildung 7-1 dargestellt, die Werte

$7{,}9\% = 57{,}9\% - 50{,}0\%$ für $d = 0{,}2$ sowie

$19{,}1\% = 69{,}1\% - 50{,}0\%$ für $d = 0{,}5$ und

$28{,}8\% = 78{,}8\% - 50{,}0\%$ für $d = 0{,}8$.

Eine mögliche Interpretation ist etwa, dass durch den eingetretenen Effekt in Größe von *d=0,8* insgesamt 28,8% der Ausprägungen der ersten Zufallsvariablen im Gegensatz zu vorher nicht mehr größer sind als 50% der Ausprägungen der zweiten Zufallsvariablen. Cohen schlägt insgesamt drei stark verwandte Interpretationen vor (Cohen, 1988, S. 21), die jeweils mit Wahrscheinlichkeiten hinterlegt sind. Nachfolgend wird der mit *d* verbundene Effekt verstanden als derjenige Prozentsatz der Beobachtungen der ersten Verteilung, der zwischen den beiden Mittelwerten liegt; eine formale Definition befindet sich im Anhang.

Cohen selbst führt zur Benennung obiger Werte für Cohen's *d* die Begriffe „klein", „mittel" und „groß" ein, nicht jedoch ohne, wie bereits im ersten Kapitel erwähnt, gleichzeitig vor fehlerhafter Interpretation zu warnen. Eine erste potentielle Fehlerquelle ist die Unterstellung einer Normalverteilung, die nicht immer den Daten zu Grunde liegt. Zur Verdeutlichung listet Abbildung 4-1 für verschiedene Standardverteilungen die jeweiligen Effekte auf. Bereits bei diesen einfachen Verteilungen treten erhebliche Größenunterschiede bei den Effekten auf. So ist beispielsweise ein im Sinne von Cohen mittlerer Effekt bei der logistischen Verteilung fast genauso groß wie ein großer Effekt bei der Gleichverteilung.

Effekt *E(d)*	Cohen's *d*		
Verteilung	0,2	0,5	0,8
Gleichverteilung	5,8	14,4	23,1
Allgemeine β -Verteilung	6,7	16,5	25,7
Dreiecksverteilung	7,8	18,3	27,3
Normalverteilung	7,9	19,1	28,8
Logistische Verteilung	9,0	21,2	31,0

Abbildung 4-1 Effekte für unterschiedliche Verteilungen

Mittels zielgerichteter Konstruktion spezieller Verteilungen gelingt es, obige Differenzen noch signifikant zu vergrößern. Jedoch bedarf es zum Nachweis der Bedeutung der anwendungsspezifischen Interpretation des Cohen's *d* nicht unter Umständen künstlich und sehr theoretisch wirkender Verteilungen. Einen wesentlich größeren Einfluss hat das nachfolgend eingeführte Konzept der Konsequenz, mit Hilfe dessen die Relevanz des Effektes in gewissem Sinne quantifiziert wird.

5. Effekt versus Konsequenz

Zur weiteren Verdeutlichung, dass Cohen's *d* nicht losgelöst von der jeweiligen Anwendung und der jeweiligen Konstellation innerhalb dieser interpretiert werden kann, reicht die Betrachtung einer sehr einfachen Funktion. Diese wertet das beobachtete Merkmal hinsichtlich der resultierenden Konsequenzen K, die mit der jeweils angenommenen Effektgröße *d* einhergeht, aus. Dabei ist ein erwünschtes Resultat erzielt, falls ein Wert *x* oberhalb einer kritischen Grenze *k* eintritt, ansonsten liegt kein erwünschtes Resultat vor. Beispielsweise wird zum Bestehen einer Klausur eine gewisse Punktzahl benötigt oder zum Überleben benötigt eine Spezies einen Mindestprozentsatz an Nährstoffen im Blut. Dieses kann durch die Funktion $g(x)=\begin{cases}0 & \text{falls } x<k,\\ 1 & \text{falls } x\geq k,\end{cases}$ dargestellt werden.

Die Konsequenzen K lassen sich dann quantifizieren durch die Differenz der Prozentsätze der Beobachtungen der beiden Zufallsvariablen, die jeweils die kritische Grenze erreichen oder überschreiten. Die im Anhang aufgeführten Berechnungen zeigen allgemein, dass die Konsequenzen nicht nur von der Effektgröße sondern auch vom gewählten kritischen Wert abhängen, d.h. $K=K(d,k)$. Speziell folgt, dass in der Regel nur im Fall $k=\mu_2$ Konsequenz und Effekt gleich groß sind, d.h. $E(d)=K(d,k)$ gilt. Neben diesen rot markierten Werten listet Tabelle 5-1 für weitere Kombinationen kritischer Grenzen und Effektgrößen die jeweils resultierende Konsequenz auf, wobei Normalverteilungen unterstellt werden. Es zeigt sich beispielsweise, dass ein „kleiner" Effekt bei einer kritischen Grenze von 0,2 eine größere Konsequenz nach sich zieht, als ein „großer" Effekt bei einer kritischen Grenze von 2,3. Bezogen etwa auf das Nährstoffbeispiel überleben also bei einem kleinen Effekt 7,9% der Population, bei einem großen Effekt lediglich 5,6%.

K(k,d)	Cohen's *d*		
kritische Grenze *k*	0,2	0,5	0,8
0,0	7,9%	19,1%	28,8%
0,2	**7,9%**	19,7%	30,5%
0,5	7,4%	**19,1%**	30,9%
0,8	6,2%	17,0%	**28,8%**
1,1	4,8%	13,9%	24,6%
1,4	3,4%	10,3%	19,3%
1,7	2,2%	7,1%	13,9%
2,0	1,3%	4,4%	9,2%
2,3	0,7%	2,5%	5,6%

Tabelle 5-1 Standardisierte Konsequenzen der Normalverteilung

Im Anhang sind für weitere Standardverteilungen entsprechende Tabellen aufgeführt, ein Blick auf die Zahlen belegt unmittelbar die zu Beginn getroffenen Ausführungen:

1. Ein “kleiner” Effekt bei einer Anwendung kann mitunter für diese viel bedeutendere Konsequenzen haben, als ein “großer“ Effekt bei einer anderen Anwendung.

2. In Abhängigkeit von der Ausgangssituation kann bei derselben Anwendung ein „kleiner Effekt“ größere Konsequenzen nach sich ziehen als ein „großer“ Effekt.

Diese im Folgenden theoretisch untermauerten Fakten seien noch einmal eher umgangssprachlich formuliert:

> „Schon ein kleiner Tropfen kann das Fass zum Überlaufen bringen.“ Die dadurch verursachten Konsequenzen können mitunter enorm sein.

6. Fazit

Obige Ausführungen zu Effekt und Konsequenz unterstreichen mehr als deutlich die Notwendigkeit, anwendungsspezifische Informationen in die Interpretation der Ergebnisse einfließen zu lassen. Entscheidend ist dabei jeweils die Beurteilung der durch den Effekt ausgelösten Konsequenzen, wobei obige Beurteilungsfunktion lediglich ein einfaches Beispiel darstellt. Wie Cohen selbst schreibt, sind daher die Begriffe kleiner, mittlerer und großer Effekt mit Augenmaß zu verwenden.

> „The terms ‚small', ‚medium', and ‚large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation […]. In the face of this relativity, there is a certain risk inherent in offering operational definitions for these terms for use in power analysis in as diverse a field of inquiry as behavioral science. This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the effect size index is available." (Cohen, 1988, S. 25)

Liegt ein fundiertes Verständnis der Effektgrößen jedoch nicht vor und wird die Methode als Standardschema begriffen, so können als Ergebnis schwere Fehlinterpretation auftreten.

München, im Dezember 2009

7. Anhang

7.1. Allgemeine Ausführungen

7.1.1. Effekt *E(d)*

Cohen schlägt verschiedene, eng miteinander verwandte Interpretationen für die Ausprägung der Effektgröße vor. Dabei wird vorausgesetzt, dass beide betrachteten Zufallsvariablen normalverteilt sind und identische Streuung aufweisen, die entsprechenden Dichtefunktionen also durch eine Verschiebung ineinander überführt werden können. Eine genauere Analyse dieser Interpretation zeigt, dass die Normalverteilungsannahme hierfür nicht notwendig ist, entscheidend sind die Symmetrie der Verteilungen und die Verschiebungseigenschaft. Dann kann in Abhängigkeit von der Effektgröße die Frage beantwortet werden, wie viel Prozent der Beobachtungen der ersten Zufallsvariablen nur bedingt durch den eingetretenen Effekt in der Größe *d* kleiner sind, als der annahmegemäß größere Mittelwert der zweiten Zufallsvariable. Diese Idee kann wie folgt formalisiert werden:

Sei F_1 die Verteilungsfunktion einer Zufallsvariablen X_1 mit Erwartungswert μ_1 und Varianz σ. Eine zweite Zufallsvariable habe identische Varianz und Erwartungswert $\mu_2 \geq \mu_1$. Dann berechnet sich die Wahrscheinlichkeit, dass die erste Zufallsvariable einen Wert zwischen μ_1 und μ_2 annimmt mittels $F_1(\mu_2) - F_1(\mu_1)$. Ohne die Verschiebung, also im Falle eines Effektes der Größe 0, wäre dieser Wert 0. Deshalb kann $F_1(\mu_2) - F_1(\mu_1)$ interpretiert werden, als der mit der Effektstärke *d* verbundene Effekt.

Definition 1 Effekt E(d)

Der mit einer Effektstärke in der Größe d verbundene Effekt ist definiert als

$$E(d) := F_1(\mu_2) - F_1(\mu_1) = F_1(\mu_1 + \sigma \cdot d) - F_1(\mu_1).$$

Um bei der nachfolgenden, beispielhaften Berechnung von Effekten für verschiedene Verteilungen größtmögliche Vergleichbarkeit zu erreichenden, werden die Verteilungen geeignet standardisiert. Dazu werden die jeweils freien Parameter so gewählt, dass die resultierende Ausgangsverteilung Erwartungswert 0 und Varianz 1, die zweite Verteilung nach eingetretenem Effekt gleiche Varianz, aber Erwartungswert *d* besitzt. Bezeichne dazu F_1^s die Verteilungsfunktion der standardisierten Zufallsvariablen $\tilde{X} := \frac{X - \mu_1}{\sigma}$. Dann folgt

$$E(d) = F_1(\mu_2) - F_1(\mu_1) \overset{\tilde{X}:=\frac{X-\mu_1}{\sigma}}{=} F_1^s(d) - F_1^s(0).$$

Folglich kann der Effekt $E(d)$ direkt mittels zugehöriger standardisierter Verteilungsfunktion berechnet werden; die entsprechende Formel wird jeweils angegeben. Die Standardisierung bedeutet keine Einschränkung der Allgemeingültigkeit der erzielten Aussagen. Zur exemplarischen Verdeutlichung wird nachfolgend bei der Normalverteilung und bei der Gleichverteilung die Standardisierung explizit in den Berechnungen ausgeführt.

Letztendlich ist für die hier angegeben Interpretation lediglich die Symmetrie der Verteilungen und die Verschiebung des Mittelwertes notwendig. Die auch in den nachfolgenden Graphiken dargestellte Parallelverschiebung der jeweiligen Dichtefunktion vereinfacht jedoch das Verständnis. Darüber hinaus erweist sich diese Eigenschaft jedoch als hilfreich für die einfache Berechnung der aus einem Effekt tatsächlich resultierenden Konsequenzen.

7.1.2. Konsequenz *K(k,d)*

Zum Beleg, dass Effektgrößen nie losgelöst von der konkreten Anwendungssituation interpretiert werden dürfen, wird ein einfaches Szenario unterstellt. Ab der kritischen Grenze *k* tritt für die Ausprägungen der Zufallsvariablen eine gewünschte Konsequenz ein. Folglich ist zu bewerten, wie viele Ausprägungen der zweiten Zufallsvariablen im Vergleich zur ersten zusätzlich, also lediglich bedingt durch den beobachteten Effekt, die kritische Grenze erreichen beziehungsweise überschreiten.

Definition 2 Konsequenz K(k,d)

Die durch den Effekt in Größe d bei einer kritischen Grenze von k verursachte Konsequenz ist definiert als

$$K(k,d) = \big(1 - F_2(k)\big) - \big(1 - F_1(k)\big) = F_1(k) - F_2(k),$$

wobei F_2 die Verteilungsfunktion der zweiten Zufallsvariablen bezeichnet.

Zur möglichst einfachen Berechnung der durch einen Effekt in Größe *d* verursachten Konsequenzen *K* wird die Standardannahme getroffen, dass die Verteilung der zweiten Zufallsvariablen durch Parallelverschiebung der Verteilung der ersten Zufallsvariablen entsteht, d.h. es gilt $F_2(x) = F_1(x + \mu_1 - \mu_2)$. Dahinter steht die Annahme beziehungsweise das Verständnis, dass der Effekt der Veränderung der Mittelwerte dadurch eintritt, dass alle beobachteten Werte um den gleichen, konstanten Betrag verschoben werden.

Für die Differenz der Prozentsätze der Beobachtungen der beiden Zufallsvariablen, die jeweils die kritische Grenze *k* erreichen oder überschreiten, also die Konsequenz $K = K(k,d)$, gilt dann

$$K(k,d) = \big(1 - F_2(k)\big) - \big(1 - F_1(k)\big) = F_1(k) - F_2(k) = F_1(k) - F_1\big(k - (\mu_2 - \mu_1)\big).$$

Die Transformation $\tilde{X} = \dfrac{X - \mu_1}{\sigma}$ führt die kritische Grenze k über in die standardisierte kritische Grenze $\tilde{k} = \dfrac{k - \mu_1}{\sigma}$ und es folgt

$$K(k,d) = F_1(k) - F_1\big(k - (\mu_2 - \mu_1)\big) = F_1^s\left(\frac{k - \mu_1}{\sigma}\right) - F_1^s\left(\frac{k - (\mu_2 - \mu_1) - \mu_1}{\sigma}\right) = F_1^s\left(\tilde{k}\right) - F_1^s\left(\tilde{k} - d\right).$$

Insbesondere folgt die Übereinstimmung von Effekt und Konsequenz in einem wichtigen Spezialfall.

Theorem 1

Im Falle $k = \mu_2$ stimmen Effekt und Konsequenz überein, d.h.

$$K(k,d) = E(d).$$

Beweis

Aus $k = \mu_2$ folgt $\tilde{k} = d$ und

$$K(k,d) = K(\mu_2,d) = F_1^s\left(d\right) - F_1^s\left(0\right) = E(d).$$

7.1.3. Verallgemeinerung

Obige Definition der Konsequenz leitet sich ab aus der speziell gewählten Funktion

$$g(x) = \begin{cases} 0 & \text{falls } x < k, \\ 1 & \text{falls } x \geq k. \end{cases}$$

Mit Hilfe der Definition des Erwartungswertes einer Funktion kann das Konzept der Konsequenz verallgemeinert werden. Der Erwartungswert μ einer Funktion *g*, die auf eine Zufallsvariable *X* angewendet wird, ist definiert als

$$\mu\big[g(X)\big] := \int g(x)\, f(x) dx,$$

wobei *f* die zugehörige Dichtefunktion bezeichnet. Wird mit dem Effekt in der Größe *d* die Dichtefunktion f_2^d assoziiert, so verändert sich der Erwartungswert, der als eine Bewertung der jeweiligen Realisationen der Zufallsvariable interpretiert werden kann, von

$$\mu_1[g(X)] = \int g(x)\, f_1(x)dx \quad \text{zu} \quad \mu_2^d[g(X)] = \int g(x)\, f_2^d(x)dx.$$

Die von der Bewertungsfunktion *g* und vom Effekt in Höhe *d* abhängige Konsequenz ist dann

$$K(g,d) := \mu_2^d[g(X)] - \mu_1[g(X)].$$

Die Bewertungsfunktion *g* ist dabei anwendungsspezifisch zu definieren und „vererbt" die verwendete Maßeinheit auf die Konsequenz. Ist *g* beispielsweise eine monetäre Bewertung, so ist die Konsequenz ebenfalls monetär zu interpretieren.

7.2. Spezielle Verteilungen

7.2.1. Normalverteilung

7.2.1.1. *Formeln*

Eine normalverteilte Zufallsvariable *X* hat die Dichtefunktion

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\quad -\infty\le x\le+\infty$$

mit den Parametern μ und $\sigma>0$.

Es gilt $E(X)=\mu$ und $Var(X)=\sigma^2$ sowie allgemein für $\mu_1\le\mu_2$

$$E(d)=F_1(\mu_2)-F_1(\mu_1)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mu_1}^{\mu_2}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2}dx\quad\overset{\tilde{X}:=\frac{X-\mu_1}{\sigma}}{=}\quad\frac{1}{\sqrt{2\pi}}\int_0^d e^{-\frac{1}{2}\tilde{x}^2}d\tilde{x}=F_1^s(d)-F_1^s(0).$$

Bei Wahl von $\mu=0$ und $\sigma=1$ gilt $Var(X)=1,0=\mu_1$ und $d=\mu_2$.

Der mit d verbundene Effekt $E(d)$ ist im standardisierten Fall

$$E(d)=\frac{1}{\sqrt{2\pi}}\int_0^d e^{-\frac{1}{2}x^2}dx.$$

In Abhängigkeit von d und der standardisierten kritischen Grenze $\tilde{k}$ folgt für die Konsequenz

$$K(\tilde{k},d)=F_1^s(\tilde{k})-F_1^s(\tilde{k}-d)=\frac{1}{\sqrt{2\pi}}\int_{\tilde{k}-d}^{d}e^{-\frac{1}{2}\tilde{x}^2}d\tilde{x}.$$

7.2.1.2. *Effekte*

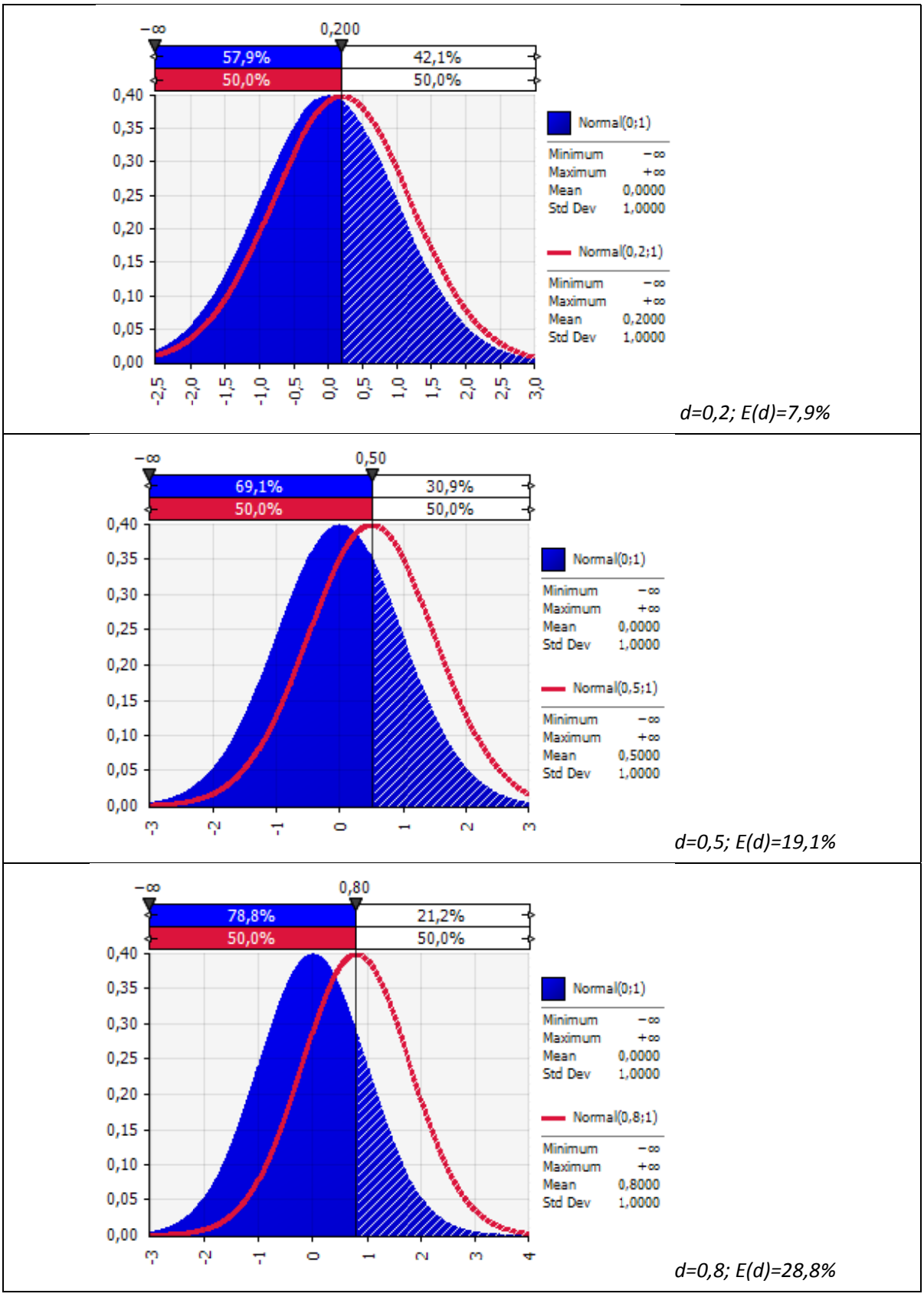
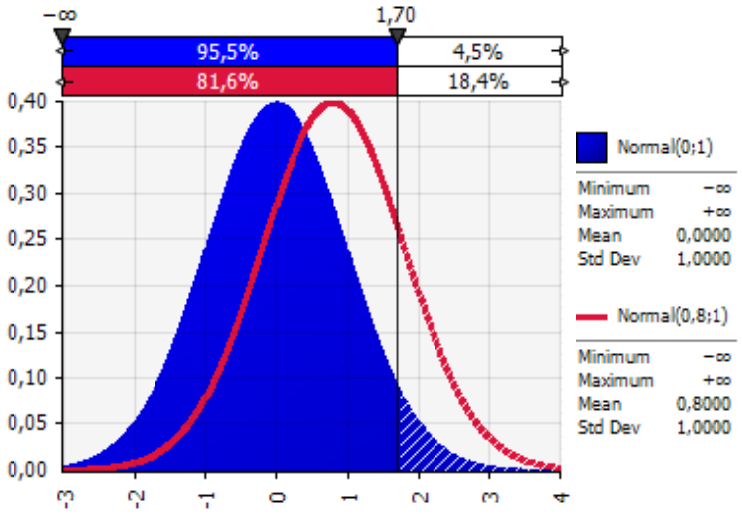


Abbildung 7-1 Effekte der Normalverteilung bei Cohen's *d = 0,2 / 0,5 / 0,8*

7.2.1.3. *Konsequenzen*



d=0,8 & k=1,7

$$K(k,d) = \underbrace{\left(1-F_2(k)\right)}_{=18,4\%=1-81,6\%} - \underbrace{\left(1-F_1(k)\right)}_{=4,5\%=1-95,5\%} = \underbrace{F_1(k)}_{=95,5\%} - \underbrace{F_2(k)}_{=81,6\%} = 13,9\%$$

Abbildung 7-2 Konsequenz der Normalverteilung bei Cohen's *d=0,8* & kritischer Grenze *k=1,7*

K(k,d)	Cohen's *d*		
kritische Grenze *k*	0,2	0,5	0,8
0,0	7,9%	19,1%	28,8%
0,2	7,9%	19,7%	30,5%
0,5	7,4%	19,1%	30,9%
0,8	6,2%	17,0%	28,8%
1,1	4,8%	13,9%	24,6%
1,4	3,4%	10,3%	19,3%
1,7	2,2%	7,1%	13,9%
2,0	1,3%	4,4%	9,2%
2,3	0,7%	2,5%	5,6%

Tabelle 7-1 Standardisierte Konsequenzen der Normalverteilung

7.2.2. Gleichverteilung

7.2.2.1. *Formeln*

Eine gleichförmig oder auch rechteckverteilte Zufallsvariable *X* hat die Verteilungsfunktion

$$F(x)=\begin{cases} 0 & \text{für} \quad x<a \\ \dfrac{x-a}{b-a} & \text{für} \quad a\le x\le b \\ 1 & \text{für} \quad x>b \end{cases}, \text{ mit den Parametern } a<b.$$

Es gilt $E(X)=\dfrac{a+b}{2}$ und $Var(X)=\dfrac{(b-a)^2}{12}$ und allgemein für $a\le\mu_1\le\mu_2\le b$

$$E(d)=\int_{\mu_1}^{\mu_2}(b-a)^{-1}dx \overset{\tilde{X}=(X-\mu_1)/\sigma}{=} \int_0^d \sigma(b-a)^{-1}d\tilde{x}=\int_0^d\frac{(b-a)}{\sqrt{12}}(b-a)^{-1}d\tilde{x}=\frac{d}{\sqrt{12}}.$$

Bei Wahl von $a=-\sqrt{12}/2$ und $b=\sqrt{12}/2$ gilt $Var(X)=1,0=\mu_1$ und $d=\mu_2$.

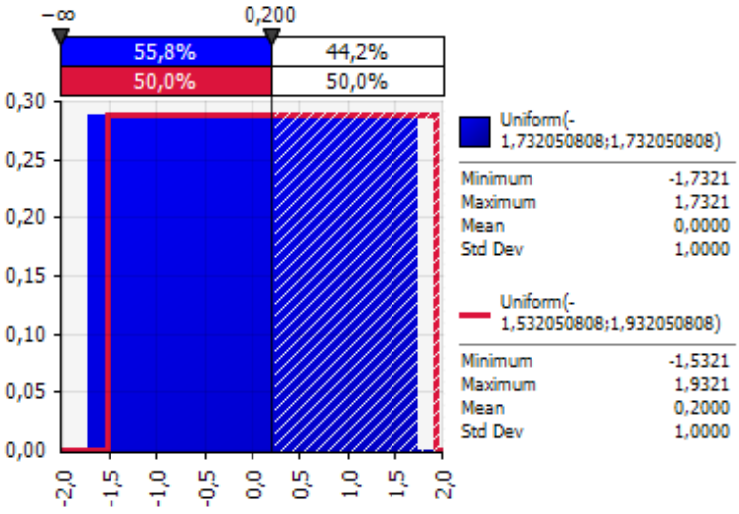
Der mit *d* verbundene Effekt $E(d)$ ist im standardisierten Fall

$$E(d)=F(d)-F(0)=\frac{d}{\sqrt{12}}.$$

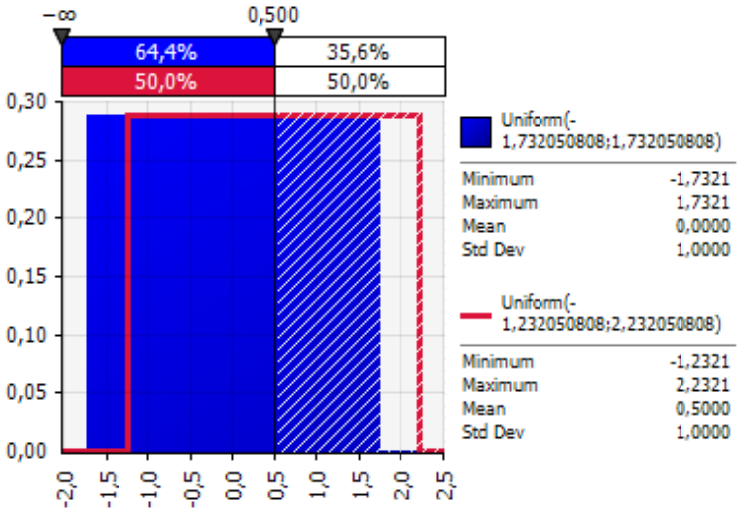
In Abhängigkeit von d und der standardisierten kritischen Grenze $\tilde{k}$ folgt beispielsweise im Fall $a\le\tilde{k}-d\le\tilde{k}\le b$ für die Konsequenz

$$K(\tilde{k},d)=F_1^s\left(\tilde{k}\right)-F_1^s\left(\tilde{k}-d\right)=\frac{\tilde{k}-a}{\sqrt{12}}-\frac{\tilde{k}-d-a}{\sqrt{12}}=\frac{d}{\sqrt{12}}.$$

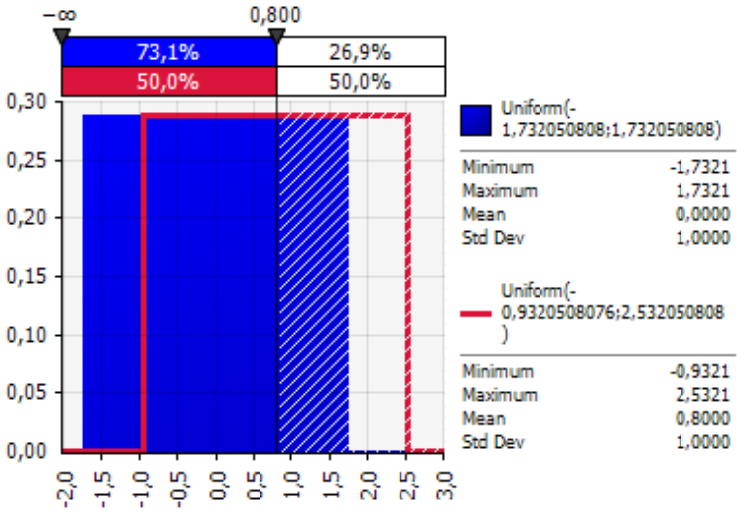
7.2.2.2. *Effekte*



d=0,2; E(d)=5,8%



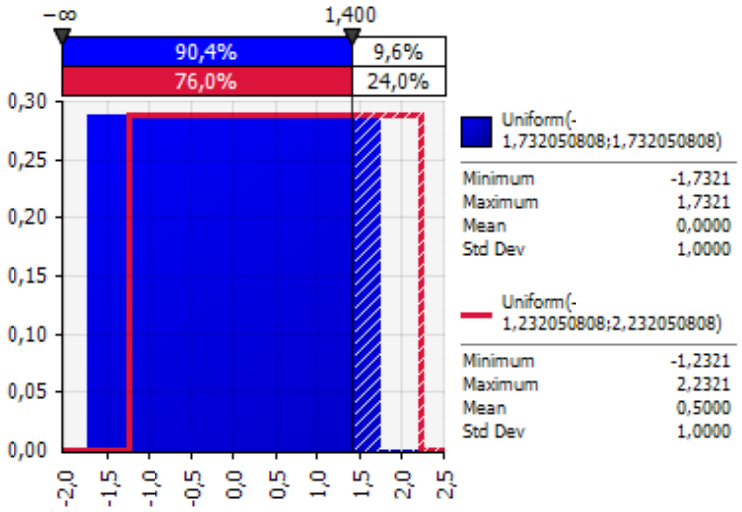
d=0,5; E(d)=14,4%



d=0,8; E(d)=23,1%

Abbildung 7-3 Effekte der Gleichverteilung bei Cohen's *d = 0,2 / 0,5 / 0,8*

7.2.2.3. *Konsequenzen*



d=0,5 & k=1,4

$$K(k,d) = \underbrace{\left(1-F_2(k)\right)}_{=24,0\%=1-76,0\%} - \underbrace{\left(1-F_1(k)\right)}_{=9,6\%=1-90,4\%} = \underbrace{F_1(k)}_{=90,4\%} - \underbrace{F_2(k)}_{=76,0\%} = 14,4\%$$

Abbildung 7-4 Konsequenz der Gleichverteilung bei Cohen's *d=0,5* & kritischer Grenze *k=1,4*

K(k,d)	Cohen's *d*		
kritische Grenze *k*	0,2	0,5	0,8
0,0	5,8%	14,4%	23,1%
0,2	5,8%	14,4%	23,1%
0,5	5,8%	14,4%	23,1%
0,8	5,8%	14,4%	23,1%
1,1	5,8%	14,4%	23,1%
1,4	5,8%	14,4%	23,1%
1,7	5,8%	14,4%	23,1%
2,0	0,0%	6,7%	15,4%
2,3	0,0%	0,0%	6,7%

Tabelle 7-2 Standardisierte Konsequenzen der Gleichverteilung

7.2.3. Dreiecksverteilung

7.2.3.1. *Formeln*

Eine dreiecksverteilte Zufallsvariable *X* hat die Verteilungsfunktion

$$F(x)=\begin{cases}\dfrac{(x-a)^2}{(b-a)(c-a)} & \text{für a} \le x \le \text{c} \\ 1-\dfrac{(b-x)^2}{(b-a)(b-c)} & \text{für c} \le x \le \text{b}\end{cases}$$

, mit den Parametern $a<c<b$.

Es gilt $E(X)=\dfrac{a+b+c}{3}$ und $Var(X)=\dfrac{(a-b)^2+(a-c)^2+(c-b)^2}{36}$.

Bei Wahl von $a=-\sqrt{6}$, $b=\sqrt{6}$ und $c=0$ gilt $Var(X)=1,0=\mu_1$ und $d=\mu_2$.

Der mit *d* verbundene Effekt $E(d)$ ist im standardisierten Fall

$$E(d)=F(d)-F(0)=\frac{1}{2}-\frac{\left(\sqrt{6}-d\right)^2}{12}.$$

In Abhängigkeit von d und der standardisierten kritischen Grenze $\tilde{k}$ folgt beispielsweise im Fall $0\le\tilde{k}-d\le\tilde{k}\le\sqrt{6}$ für die Konsequenz

$$K(\tilde{k},d)=F_1^s\left(\tilde{k}\right)-F_1^s\left(\tilde{k}-d\right)=\frac{\left(\sqrt{6}-\left(\tilde{k}-d\right)\right)^2-\left(\sqrt{6}-\tilde{k}\right)^2}{2\sqrt{6}\sqrt{6}}=\frac{d\left(2\left(\sqrt{6}-\tilde{k}\right)+d\right)}{12}.$$

7.2.3.2. *Effekte*

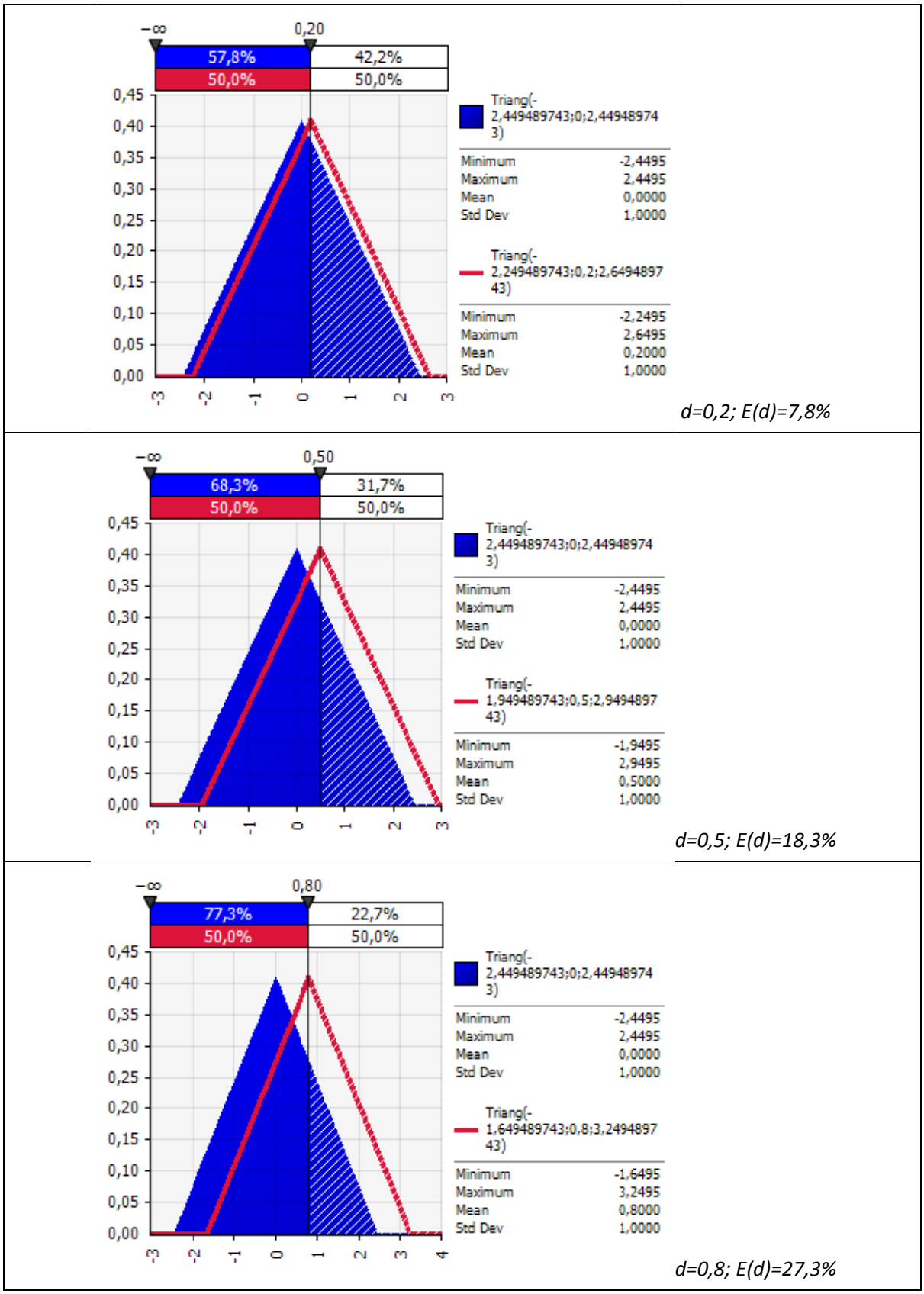
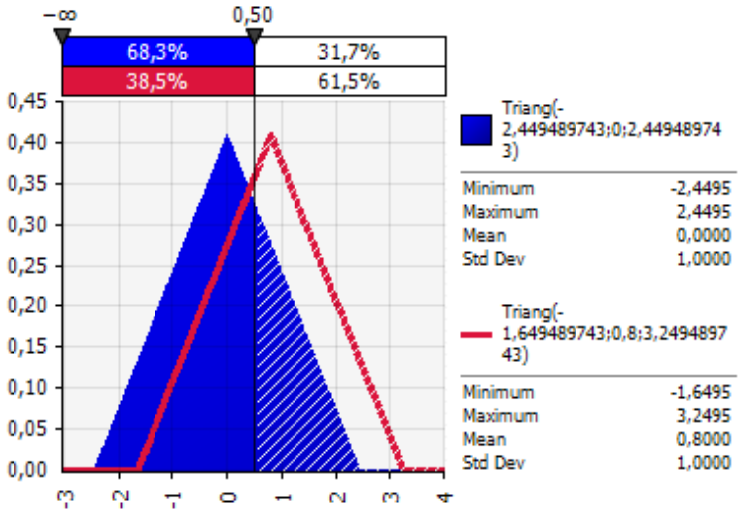


Abbildung 7-5 Effekte der Dreiecksverteilung bei Cohen's *d = 0,2 / 0,5 / 0,8*

7.2.3.3. *Konsequenzen*



d=0,8 & k=0,5

$$K(k,d) = \underbrace{\left(1-F_2(k)\right)}_{=61,5\%=1-38,5\%} - \underbrace{\left(1-F_1(k)\right)}_{=31,7\%=1-68,3\%} = \underbrace{F_1(k)}_{=68,3\%} - \underbrace{F_2(k)}_{=38,5\%} = 29,8\%$$

Abbildung 7-6 Konsequenz der Normalverteilung bei Cohen's *d=0,8* & kritischer Grenze *k=0,5*

K(k,d)	Cohen's *d*		
kritische Grenze *k*	0,2	0,5	0,8
0,0	7,8%	18,3%	27,3%
0,2	7,8%	19,3%	29,3%
0,5	6,8%	18,3%	29,8%
0,8	5,8%	15,8%	27,3%
1,1	4,8%	13,3%	23,3%
1,4	3,8%	10,8%	19,3'%
1,7	2,8%	8,3%	15,3%
2,0	1,8%	5,8%	11,3%
2,3	0,8%	3,3%	7,3%

Tabelle 7-3 Standardisierte Konsequenzen der Dreiecksverteilung

7.2.4. Logistische Verteilung

7.2.4.1. *Formeln*

Eine logistisch verteilte Zufallsvariable *X* hat die Verteilungsfunktion

$$F(x) = \frac{1}{1+e^{-(x-\alpha)/\beta}}, \quad -\infty \le \mathrm{x} \le +\infty,$$

mit den Parametern α und $\beta > 0$.

Es gilt $E(X) = \alpha$ und $Var(X) = \frac{\beta^2 \pi^2}{3}$.

Bei Wahl von $\alpha = 0$ und $\beta = \frac{\sqrt{3}}{\pi}$ gilt $Var(X) = 1,0 = \mu_1$ und $d = \mu_2$.

Der mit *d* verbundene Effekt ist im standardisierten Fall

$$E(d) = F(d) - F(0) = \frac{1}{1+e^{-\pi d/\sqrt{3}}} - \frac{1}{2}.$$

In Abhängigkeit von d und der standardisierten kritischen Grenze $\tilde{k}$ folgt für die Konsequenz

$$K(\tilde{k}, d) = F_1^s\left(\tilde{k}\right) - F_1^s\left(\tilde{k} - d\right) = \frac{1}{1+e^{-\pi \tilde{k}/\sqrt{3}}} - \frac{1}{1+e^{-\pi\left(\tilde{k}-d\right)/\sqrt{3}}}.$$

7.2.4.2. *Effekte*

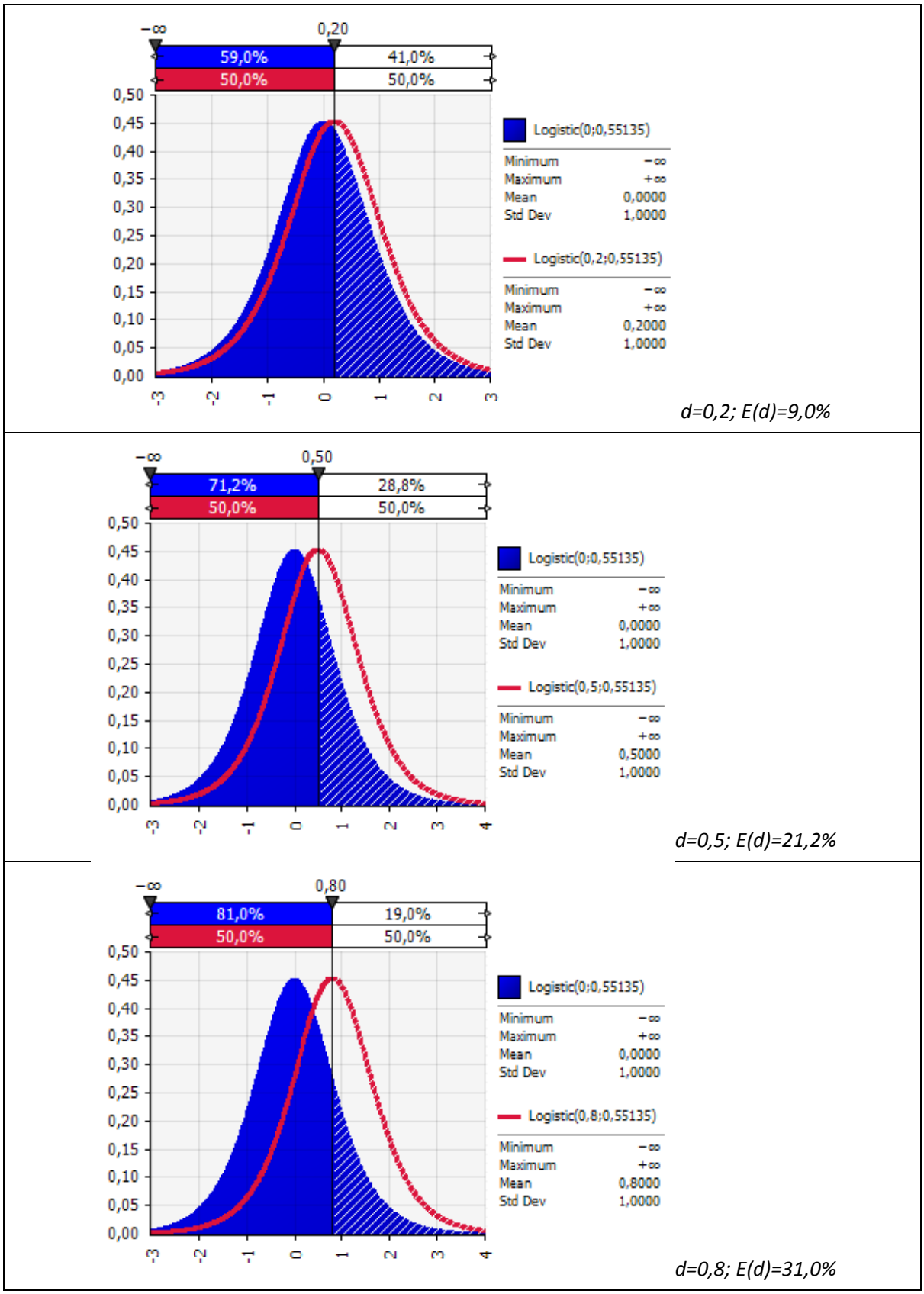
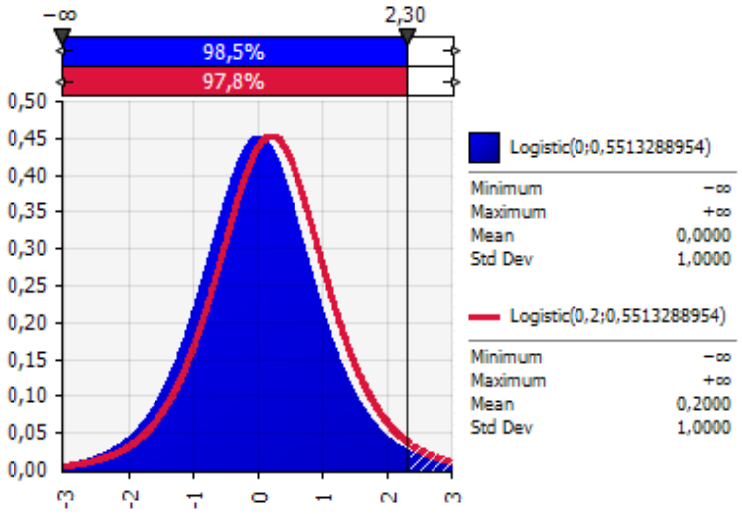


Abbildung 7-7 Effekte der logistischen Verteilung bei Cohen's *d = 0,2 / 0,5 / 0,8*

7.2.4.3. *Konsequenzen*



d=0,2 & k=2,3

$$K(k,d) = \underbrace{\left(1-F_2(k)\right)}_{=2,2\%=1-97,8\%} - \underbrace{\left(1-F_1(k)\right)}_{=1,5\%=1-98,5\%} = \underbrace{F_1(k)}_{=98,5\%} - \underbrace{F_2(k)}_{=97,8\%} = 0,7\%$$

Abbildung 7-8 Konsequenz der logistischen Verteilung bei Cohen's *d=0,2* & kritischer Grenze *k=2,3*

K(k,d)	Cohen's *d*		
kritische Grenze *k*	0,2	0,5	0,8
0,0	9,0%	21,2%	31,0%
0,2	9,0%	22,3%	33,8%
0,5	7,9%	21,2%	34,5%
0,8	6,2%	17,7%	31,0%
1,1	4,3%	13,2%	24,7%
1,4	2,9%	9,0%	17,9
1,7	1,8%	5,8%	11,9%
2,0	1,1%	3,6%	7,6%
2,3	0,7%	2,2%	4,7%

Tabelle 7-4 Standardisierte Konsequenzen der logistischen Verteilung

7.2.5. Allgemeine Beta-Verteilung

7.2.5.1. *Formeln*

Eine allgemein betaverteilte Zufallsvariable *X* hat die Dichtefunktion

$$f(x)=\begin{cases}\dfrac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta+1}} & \text{wenn } a\le x\le b,\\ 0 & \text{sonst,}\end{cases}$$

mit den Parametern $\alpha,\beta>0$, $a,b\in\mathbb{R}$ und der Betafunktion $B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt.$

Es gilt $E(X)=\dfrac{b\alpha+a\beta}{\alpha+\beta}$ und $Var(X)=\dfrac{(b-a)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Beispielsweise bei Wahl von $\alpha=\beta=2$ und $-a=b=\sqrt{5}$ gilt $Var(X)=1,0=\mu_1$ und $d=\mu_2$.

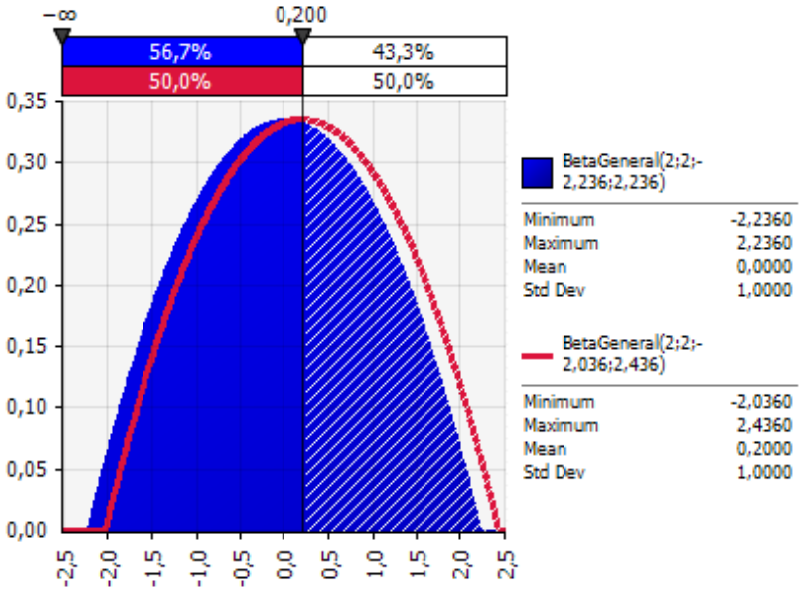
Der mit *d* verbundene Effekt ist in diesem standardisierten Fall

$$E(d)=\int_0^{\min\{b,d\}}\frac{(s-a)^{\alpha-1}(s-t)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta+1}}ds\,.$$

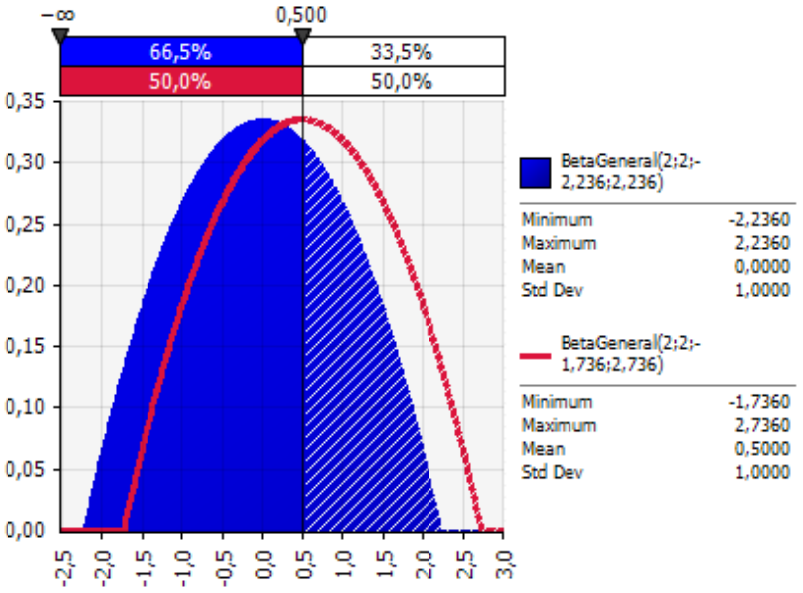
In Abhängigkeit von d und der standardisierten kritischen Grenze $\tilde{k}$ folgt für die Konsequenz etwa im Fall $-\sqrt{5}\le\tilde{k}-d\le\tilde{k}\le\sqrt{5}$

$$K(\tilde{k},d)=F_1^s(\tilde{k})-F_1^s(\tilde{k}-d)=\int_{\tilde{k}-d}^{\tilde{k}}\frac{(s-a)^{\alpha-1}(s-t)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta+1}}ds.$$

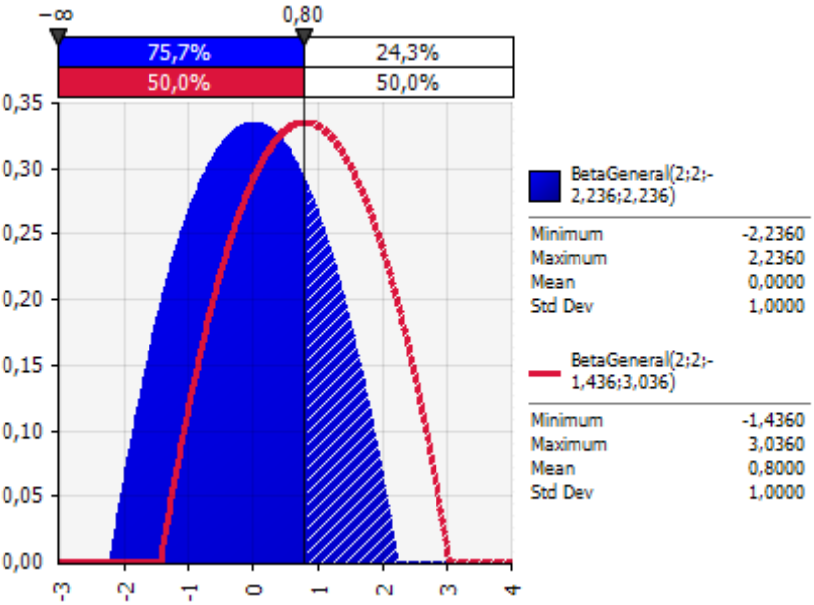
7.2.5.2. *Effekte*



d=0,2; E(d)=6,7%



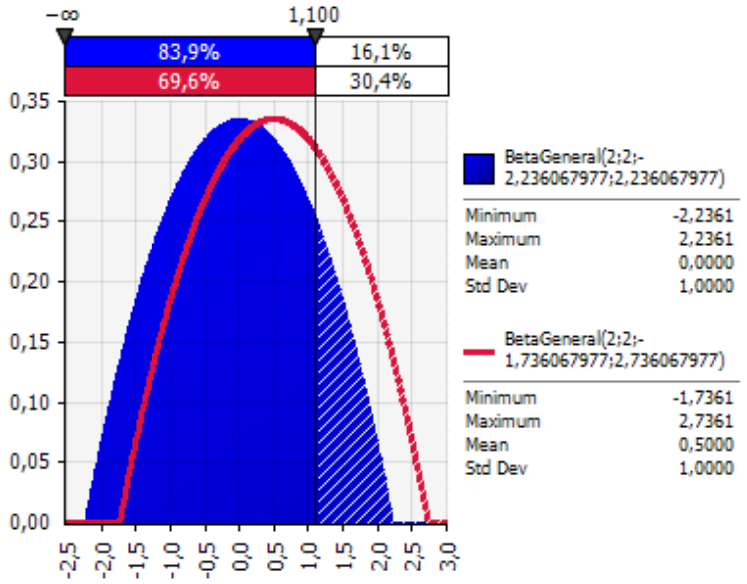
d=0,2; E(d)=16,5%



d=0,2; E(d)=25,7%

Abbildung 7-9 Effekte der allgemeinen Beta-Verteilung bei Cohen's *d = 0,2 / 0,5 / 0,8*

7.2.5.3. *Konsequenzen*



d=0,5 & k=1,1

$$K(k,d) = \underbrace{\left(1-F_2(k)\right)}_{=30,4\%=1-69,6\%} - \underbrace{\left(1-F_1(k)\right)}_{=16,1\%=1-83,9\%} = \underbrace{F_1(k)}_{=83,9\%} - \underbrace{F_2(k)}_{=69,6\%} = 14,3\%$$

Abbildung 7-10 Konsequenz der allgemeinen Beta-Verteilung bei Cohen's *d=0,5* & kritischer Grenze *k=1,1*

K(k,d)	Cohen's *d*		
kritische Grenze *k*	0,2	0,5	0,8
0,0	6,7%	16,5%	25,7%
0,2	6,7%	16,7%	26,3%
0,5	6,5%	16,5%	26,5%
0,8	6,1%	15,7%	25,7%
1,1	5,3%	14,3%	23,9%
1,4	4,4%	12,2%	21,2%
1,7	3,2%	9,6%	17,4%
2,0	1,8%	6,4%	12,8%
2,3	0,3%	2,7%	7,2%

Tabelle 7-5 Standardisierte Konsequenzen der allgemeinen Beta-Verteilung

8. Literaturverzeichnis

Bortz, J., & Lienert, G. (2008). *Kurzgefasste Statistik für die klinische Forschung* (3. Auflage Ausg.). Berlin: Springer.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Edition Ausg.). Lawrence Erlbaum Associates.

Schira, J. (2005). *Statistsiche Methoden der VWL und BWL* (2. Auflage Ausg.). München: Pearson Studium.

Schwarze, J. (1986). *Grundlagen der Statistik II* (7. Auflage Ausg.). Herne/Berlin: Verlag Neue Wirtschafts-Briefe.

Cohen's d:
Sinn und Zweck einer Schätzung

Prof. Dr. Andreas Brieden

1. Überblick

Effektstärken wie beispielsweise Cohen's d spielen eine wichtige Rolle bei der Durchführung statistischer Signifikanztests. Insbesondere ermöglicht die Vorgabe einer Effektstärke nicht nur die Kontrolle des sogenannten Fehlers erster Art, sondern gleichzeitig auch die Kontrolle des Fehlers zweiter Art, siehe etwa Cohen J., 1988. Sorgfältig zu unterscheiden von der Vorgabe einer Effektstärke ist das Schätzen der Effektstärke, wobei das Schätzen nochmals unterteilt werden kann in die sogenannte Punkt- und in die sogenannte Intervallschätzung, die etwa Cohen selbst für Effektstärken vorgeschlagen hat (Cohen J. , 1994).

Prinzipiell stehen Testen und Schätzen in einem sehr engen Zusammenhang, wobei die Entscheidung zur Verwendung eines Tests oder einer Schätzung durchaus situations- beziehungsweise anwendungsbezogen erfolgen sollte. Unabdingbar ist aber in beiden Fällen eine korrekte Interpretation der Effektstärken, egal ob vorgegeben oder geschätzt.

Die Frage, welche Methode zu präferieren ist, wird durchaus kontrovers diskutiert, s. etwa Brandstätter, 2009, wobei Argumente wie etwa leichtere Verständlichkeit im Rahmen dieser Diskussion aus wissenschaftlicher Sicht nicht nachvollziehbar sind. Im zweiten Kapitel wird diese Fragestellung aus dem Blickpunkt der jeweiligen Anwendung beleuchtet, woraus ersichtlich wird, dass eine allgemein gültige Antwort auf obige Frage nicht gegeben werden kann. Vielmehr ist der jeweilige Entscheidungskontext zu berücksichtigen.

Die Bedeutung der Effektstärken im Rahmen von Tests im Allgemeinen und von Cohen's d im Speziellen wurde ausführlich behandelt in Brieden, 2009. Die Effektstärke Cohen's *d* gibt dabei die durch die Standardabweichung normierte Differenz zwischen Mittelwerten an. Ihre Darstellung im Kontext von Schätzungen erfolgt in dieser Ausarbeitung, wobei insbesondere im vierten und fünften Kapitel die Beispiele aus obiger Arbeit ohne expliziten Verweis weiter Verwendung finden. Dazu bedarf es im dritten Kapitel zunächst einiger allgemeiner Ausführungen zu Schätzern. Von besonderem Interesse ist hier die prinzipielle Interpretation einer Intervallschätzung, da diese aus wissenschaftlicher Sicht nicht eindeutig ist. Dazu bedarf es einer kurzen Diskussion zweier unterschiedlicher Konzepte über Wahrscheinlichkeiten. Zum einem wird die klassische Sichtweise des Wahrscheinlichkeitsbegriffs und zum anderen der Begriff der subjektiven Wahrscheinlichkeit beleuchtet.

Das vierte Kapitel widmet sich dann der Schätzung der Effektstärke Cohen's *d*. Neben theoretischen Aussagen zu konkreten Verteilungen steht zunächst die Frage im Mittelpunkt, welcher Effekt im Sinne von Cohen mit einer Punkt- bzw. mit einer Intervallschätzung assoziiert werden kann. In diesem Kontext wird auch beantwortet, wie Relevanzgrenzen zu

interpretieren sind. Dabei spielt jeweils die den ursprünglichen Daten zu Grunde liegende Verteilung eine Rolle.

Bereits in Brieden, 2009, wurde ausführlich dargestellt, dass Effekte im Sinne von Cohen nicht der alleinige Maßstab sein können, sondern vielmehr auch die durch einen eingetretenen Effekt ausgelöste Konsequenz beachtet werden muss. Im fünften Kapitel wird die Interpretation der Schätzer um das Konzept der Konsequenz erweitert.

Die Ausführungen schließen im sechsten Kapitel mit einem Fazit, das hier wie folgt zusammengefasst werden kann.

> ***Schätzer für Cohen's d sind von zentraler Bedeutung, ihre Interpretation ist aber immer vor dem spezifischen Anwendungshintergrund durchzuführen.***

2. Testen oder Schätzen

Die Entscheidung über Testen oder Schätzen ist abhängig vom jeweiligen Szenario zu treffen. Dieses sei an zwei exemplarischen Fällen verdeutlicht. Der zentrale Unterschied liegt dabei in den Konsequenzen einer Ablehnung der Nullhypothese, obwohl die Relevanzgrenze nicht erreicht wurde.

Für das erste Szenario wird als Beispiel das Qualitätsmanagement herangezogen. Bei der Qualitätskontrolle ist sicherzustellen, dass etwa minderwertige Lieferungen nicht akzeptiert werden. Dabei ist bekannt, dass die Kunden geringfügige Qualitätsminderungen akzeptieren, ab einer gewissen „Schmerzgrenze“ jedoch nachhaltiger Schaden für das Unternehmen entsteht. Diese Relevanzgrenze ist somit die im Rahmen des Tests vorzugebene Effektstärke. Hiernach wird die Teststärke festgelegt, also die Wahrscheinlichkeit, dass Lieferungen, die die Relevanzgrenze unterschreiten, also auch tatsächlich nicht akzeptiert werden. Durch entsprechende Wahl des Stichprobenumfangs kann diese Teststärke dann auch realisiert werden. In Abbildung 1 ist die Situation eines solchen Tests fiktiv dargestellt. Hierbei gibt der Wert auf der *x*-Achse die Abweichung nach unten vom Mindestniveau an. Als Relevanzgrenze wurde eine Abweichung um 2,68 gewählt, die mit einer Wahrscheinlichkeit von 85 Prozent, also der Teststärke, auch tatsächlich entdeckt wird. Der Fehler 2. Art ist somit 15 Prozent, das Signifikanzniveau, also der Fehler 1. Art, wurde auf 5 Prozent festgesetzt. Für den Test auf die Hypothese „H_0: Es gibt keine Abweichung nach unten.“ ergibt sich die kritische Grenze von 1,64. Liefert eine Stichprobe ein Ergebnis kleiner oder gleich 1,64 wird die Nullhypothese angenommen, ansonsten, wird sie verworfen, also die Alternativhypothese „H_1: Es gibt eine Abweichung nach unten.“ akzeptiert. In blau beziehungsweise rot dargestellt ist die vom Stichprobenumfang abhängige Verteilung der Schätzfunktion unter der Annahme, dass die Abweichung Null beziehungsweise 2,68 beträgt. Hierbei wird der Einfachheit halber unterstellt, dass die Abweichung normalverteilt ist[1]. Nach Durchführung des Tests liefert der Vergleich der Prüfgröße, also das Ergebnis der Stichprobe, mit dem kritischen Wert von 1,64 die Testentscheidung. Wird die Nullhypothese angenommen, so besteht lediglich ein Restrisiko von 15 Prozent, dass das Niveau tatsächlich jenseits der Relevanzgrenze liegt und auf Seiten der Kunden tatsächlich relevante Reaktionen zu erwarten sind. Da dieser Prozentsatz, die Wahrscheinlichkeit für den Fehler 2. Art, aber Dank der Festlegung der Relevanzgrenze frei gewählt wurde, besteht kein Problem.

[1] Daraus folgt selbst bei kleinen Stichproben und bekannter Varianz die exakte Normalverteilung für das Stichprobenmittel. Ansonsten sind die Aussagen approximativ gültig.

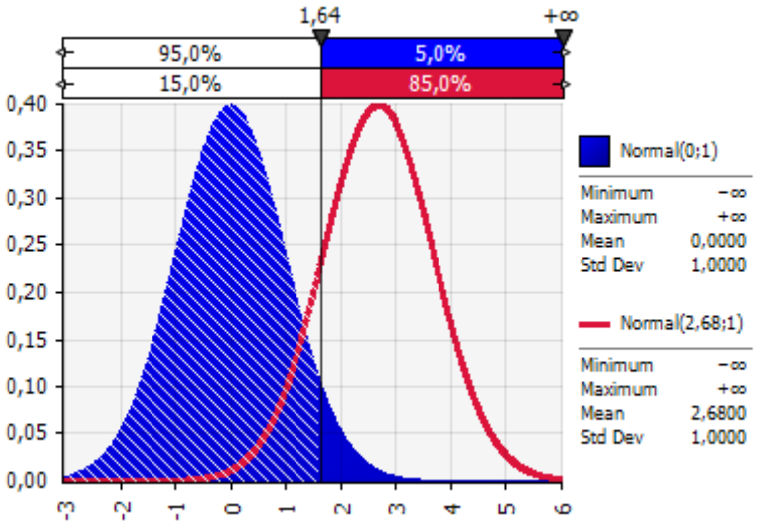


Abbildung 1: Signifikanzniveau und Teststärke

Wird die Nullhypothese verworfen, ist die Wahrscheinlichkeit, dieses fälschlicherweise getan zu haben, definitionsgemäß gleich dem Fehler erster Art, also ebenfalls durch Vorgabe des Signifikanzniveaus auf einem akzeptablen Level. Angenommen, die Ablehnung der Nullhypothese ist korrekt, dann kann dieses zum einen korrekt sein, weil die wahre Abweichung jenseits der Relevanzgrenze von 2,68 liegt. Dann hat das Qualitätsmanagement perfekt funktioniert, da genau für das Aufspüren dieser Abweichung der Test konzipiert wurde. Die Entscheidung ist aber auch korrekt, falls der wahre Wert zwischen 0 und 2,68 liegt. Es liegt also ebenfalls verminderte Qualität der Lieferung vor, allerdings in einem Ausmaß, das bei den Kunden keine relevanten Reaktionen hervorrufen dürfte. Die Testentscheidung, die in einer Ablehnung der Lieferung resultiert, stellt aber auch hier kein Problem dar. Denn auch wenn die Relevanzgrenze nicht überschritten wurde, das vorgegebene Mindestniveau (in diesem Fall Abweichung Null) wurde in jedem Fall nicht erreicht.

Allgemeiner formuliert liegt also die Situation vor, dass Veränderungen jenseits der Relevanzgrenze mit entsprechender Teststärke entdeckt werden sollen, aber die Entscheidung gegen die Nullhypothese auch bei geringeren Veränderungen nicht problematisch ist. Veränderungen können dabei sowohl negativer Natur, wie bei der Qualitätskontrolle, als auch positiver Natur sein.

Im Gegensatz zu obigem Szenario, in dem die Ablehnung der Nullhypothese keine negativen Auswirkungen hat – und der Test deshalb auch geeignet ist – ist im zweiten Szenario die Annahme der Alternativhypothese problematisch, falls die gewünschte Relevanzgrenze nicht tatsächlich erreicht wird.

Als Beispiel dient hier etwa ein Test auf die Wirksamkeit eines neuen Medikaments. Bei entsprechenden Studien ist es von zentraler Bedeutung, nicht nur statistische Signifikanz zu betrachten, sondern vielmehr auch die sogenannte klinische Relevanz in Betracht zu ziehen, siehe etwa Bortz & Lienert, 2008. Welche Verbesserung gegenüber dem Status Quo dabei als notwendige Mindestverbesserung bestimmt wird, also welche Relevanzgrenze gefordert ist, hängt dabei von der jeweiligen Anwendung ab und ist von den Fachleuten vorzugeben. Der Einfachheit halber seien die gleichen Zahlen wie in obigem Szenario gewählt, die Argumentation ist jedoch entsprechend anzupassen. Wird also als klinische Relevanzgrenze 2,68 festgelegt, dann sollte bei einem Test, im Falle des tatsächlichen Erreichens oder sogar Überschreitens dieser Grenze, die Ablehnung der Nullhypothese „keine Wirksamkeit“ auch mit entsprechend hoher Wahrscheinlichkeit gewährleistet sein. Als Teststärke sei wiederum 85 Prozent gewählt, der Test wird durchgeführt und die Nullhypothese annahmegemäß tatsächlich abgelehnt. Dann gibt es wieder zwei beziehungsweise drei Möglichkeiten. Im ersten Fall kommt die Ablehnung dadurch zu Stande, dass die tatsächliche Veränderung gleich oder größer der Relevanzgrenze 2,68 ist. Dann liefert der Test exakt das gewünschte Resultat. Im zweiten beziehungsweise zweiten und dritten Fall liegt die tatsächliche Veränderung unterhalb der Relevanzgrenze. Entweder ist die Nullhypothese tatsächlich richtig – der Fall dessen Wahrscheinlichkeit durch die Vorgabe des Signifikanzniveaus kontrolliert wird. Oder aber es liegt eine Veränderung vor, die jedoch nicht klinisch relevant ist (im Beispiel zwischen 0 und 2,68). In diesem Fall liefert der Test zwar ebenfalls eine Ablehnung der Nullhypothese, die Interpretation, dass die Relevanzgrenze mit Wahrscheinlichkeit 85 Prozent auch tatsächlich erreicht ist, ist aber offenkundig falsch. Sind die Kosten für die Einführung des Medikamentes über das Erreichen der Relevanzgrenze gerechtfertigt, entsteht ein entscheidender Unterschied zum ersten Beispiel hinsichtlich der Beurteilung der Sinnhaftigkeit des Tests. Beim ersten Beispiel ist auch das Ablehnen der Nullhypothese bei Nichtereichen der Relevanzgrenze nicht weiter tragisch, da ein etwa vertraglich vereinbartes Mindestniveau nicht erreicht wurde. Aber warum sollten Kosten akzeptiert werden, wenn eine von Experten geforderte Mindestverbesserung nicht erreicht wird?

Erweist sich im Rahmen der jeweils spezifischen Anwendung die Annahme der Alternativhypothese als problematisch, falls die gewünschte Relevanzgrenze nicht tatsächlich erreicht wird, so rückt die folgende Frage in den Mittelpunkt:

Welche Veränderung liegt tatsächlich vor? Die Beantwortung dieser Frage wird mit Hilfe verschiedener Schätzmethoden versucht, die im nachfolgenden Kapitel beschrieben werden.

3. Punkt- und Intervallschätzer sowie Konfidenzintervalle

Neben dem Testen ist das Schätzen eine Hauptaufgabe der induktiven Statistik. Ziel ist dabei, mit Hilfe einer Stichprobe ein möglichst genaues Bild über die Eigenschaften der zu Grunde liegenden Verteilung zu erhalten. Theoretischer Hintergrund hierfür ist der Hauptsatz der Statistik, der besagt, dass die mit Hilfe von Stichproben gewonnene empirische Verteilungsfunktion mit wachsendem Stichprobenumfang gegen die tatsächliche Verteilungsfunktion konvergiert. In vielen Anwendungen ist es jedoch nicht notwendig, die volle Information zu ermitteln, es reicht mitunter bereits die Kenntnis einiger Parameter wie etwa der Mittelwert oder die Standardabweichung einer Verteilung aus. In diesen Fällen wird zunächst ein Punktschätzer betrachtet.

Ein Punktschätzer ist allgemein gesprochen eine Rechenvorschrift, wie aus dem Ergebnis einer zukünftig durchzuführenden Stichprobe der gesuchte Parameter ermittelt oder besser geschätzt wird. Ein Punktschätzer ist somit eine Zufallsvariable, da das konkrete Ergebnis von der konkreten Stichprobe abhängig ist. Für einen Schätzer gibt es mehrere Möglichkeiten, im Regelfall wird jedoch ein Schätzer ausgewählt, der gewissen Schätzprinzipen „gehorcht". Eine erste Forderung ist die sogenannte Erwartungstreue. Ein Schätzer heißt erwartungstreu, wenn im Mittel der richtige Wert exakt getroffen, oder besser erwartet werden kann. Ein weiteres Prinzip ist das Maximum-Likelihood-Prinzip; es wird derjenige Wert ausgewählt, der das beobachtete Stichprobenergebnis gegenüber allen alternativen Wahlen mit der größten Wahrscheinlichkeit hervorbringt. Ein Beispiel für einen Schätzer, der beide Prinzipien erfüllt ist das Stichprobenmittel, also die Verwendung des arithmetischen Mittels der Stichprobe als Schätzer für den Erwartungswert der Verteilung. Ein Problem ist offenkundig, es wird lediglich eine Aussage über den zu erwartenden Durchschnitt getroffen. Was ist aber bezüglich der Abweichung einer einzelnen, konkret durchgeführten Schätzung vom wahren Wert zu erwarten? Eine zusätzliche, wünschenswerte Eigenschaft ist daher die sogenannte Effizienz, da ein effizienter Schätzer unter allen erwartungstreuen Schätzern, definitionsgemäß minimale Varianz (also auch Streuung) um den wahren Wert aufweist. Das Stichprobenmittel erfüllt ebenfalls dieses Prinzip und hat darüber hinaus eine weitere fundamentale Eigenschaft, mit deren Hilfe Intervallschätzer konstruiert werden können.

Der zentrale Grenzwertsatz besagt, dass die Zufallsvariable $Z := \left(\bar{X}_n - \mu\right) \Big/ \left(\sigma_X / \sqrt{n}\right)$ approximativ standardnormalverteilt ist. Dabei bezeichnet $\bar{X}_n$ das Stichprobenmittel bei einer Stichprobe vom Umfang *n*, μ den wahren Erwartungswert und σ_X die Standardabweichung der Zufallsvariablen *X*. Diese Aussage gilt unabhängig von der Verteilung von *X* und gilt im Falle einer Normalverteilung exakt für beliebigen Stichprobenumfang. Muss die Standardabweichung geschätzt werden, gelten analoge

Aussagen unter Verwendung der sogenannten t-Verteilung. Zur Vereinfachung der Ausführungen wird im Folgenden davon ausgegangen, dass die Standardabweichung bekannt ist und daher nicht geschätzt werden muss. Für eine standardnormalverteilte Zufallsvariable kann für eine beliebige Wahrscheinlichkeit *p* derjenige Wert $z_{p/2}$ ermittelt werden, so dass $P\left(-z_{p/2} \le Z \le z_{p/2}\right) = p$ gilt, also die Wahrscheinlichkeit, einen Wert zwischen $-z_{p/2}$ und $z_{p/2}$ zu erhalten, gleich *p* ist. Abbildung 2 illustriert die Situation für $p = 90\%$. Im Intervall von -1,645 bis +1,645 wird die Realisierung der Zufallsvariable mit Wahrscheinlichkeit 90% erfolgen; dieses Intervall heißt deshalb Konfidenzintervall für *Z* mit Konfidenzniveau 90%.

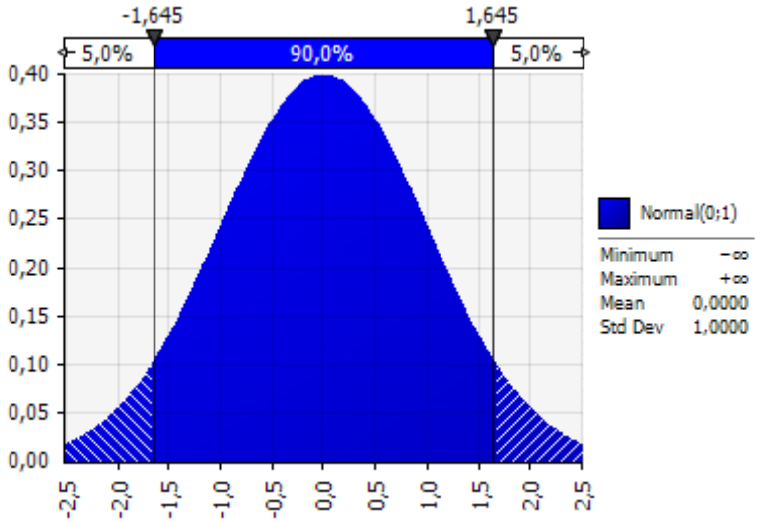


Abbildung 2: 90% Konfidenzintervall

Welchen Einfluss hat im Kontext der Konfidenzintervalle der Stichprobenumfang? Der Einfachheit halber sei annahmegemäß etwa $\mu = 0$ und $\sigma_X = 1$. Abbildung 3 zeigt die resultierenden Verteilungen für Stichproben vom Umfang *n=25* (blau) und *n=100* (rot). Aus der Grafik wird ersichtlich, dass die Konfidenzintervalle offensichtlich mit zunehmendem Stichprobenumfang kleiner werden, die Vorhersage für $\bar{X}$ bei gleichen Konfidenzniveau also präziser wird. Grund hierfür ist wegen $Z = \left(\bar{X}_n - \mu\right) \Big/ \left(\sigma_X / \sqrt{n}\right)$ die Äquivalenz von $P\left(-z_{p/2} \le Z = \left(\bar{X}_n - \mu\right) \Big/ \left(\sigma_X / \sqrt{n}\right) \le z_{p/2}\right) = p$ und $P\left(\mu - \frac{\sigma_X}{\sqrt{n}} z_{p/2} \le \bar{X}_n \le \mu + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right) = p.$

Somit liefert der Intervallschätzer $\left[\mu - \frac{\sigma_X}{\sqrt{n}} z_{p/2}; \mu + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right]$ ein Intervall der Länge

$2\frac{\sigma_X}{\sqrt{n}} z_{p/2}$, das bei bekannter Varianz und vorgegebenen Konfidenzniveau nur noch vom Stichprobenumfang abhängt.

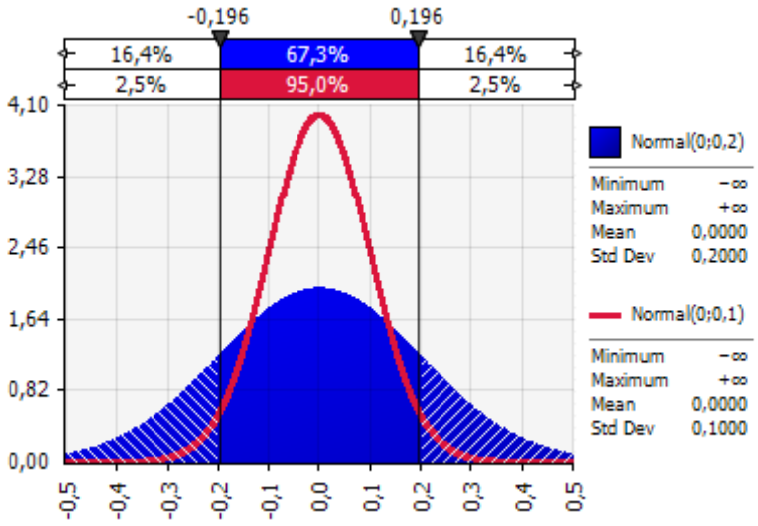


Abbildung 3: Stichprobenverteilungen für *n=25* und *n=100*

Nach Durchführung der Stichprobe ist die Frage von Interesse, welcher Wert tatsächlich vorliegt. Für ihre Beantwortung ist obiges Intervall aber nicht geeignet, da eine Aussage über das als unbekannt vorausgesetzte μ getroffen werden soll. Ebenfalls äquivalent zu obigen Gleichungen ist die Gleichung $P\left(\bar{X}_n - \frac{\sigma_X}{\sqrt{n}} z_{p/2} \le \mu \le \bar{X}_n + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right) = p$, die bedeutet, dass der zu schätzende Wert μ mit der Wahrscheinlichkeit p im Intervall $\left[\bar{X}_n - \frac{\sigma_X}{\sqrt{n}} z_{p/2}, \bar{X}_n + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right]$ liegen wird. Zu beachten ist hierbei die a priori-Aussage, da die Wahrscheinlichkeitsaussage vor Realisierung der Stichprobe getroffen wird. Obiges Intervall ist somit ein Intervallschätzer, also eine Rechenvorschrift, die ein zufallsabhängiges Intervall liefert. Wiederum handelt es sich um ein Konfidenzintervall zum Konfidenzniveau *p*, dieses Mal wird jedoch der Parameter μ mit der Wahrscheinlichkeit *p* im Intervall enthalten sein. Unumstritten ist die Interpretation, dass bei 100 Durchführungen der Stichprobe jeweils vom Umfang *n* der Parameter μ im Schnitt 100p mal im jeweiligen Intervall enthalten ist. Welche Interpretation ist jedoch korrekt, wenn die Stichprobe durchgeführt wurde und mit Hilfe des tatsächlichen realisierten Stichprobenmittels $\bar{x}_n$ der Intervallschätzer das konkrete Intervall $\left[\bar{x}_n - \frac{\sigma_X}{\sqrt{n}} z_{p/2}, \bar{x}_n + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right]$ geliefert hat? Während einige Autoren einzig obige Interpretation, also die Situation vor der Realisierung akzeptieren, s. etwa von Auer, 2006, wird in anderen Lehrbüchern sogar eine „Dichtefunktion“ für den wahren Parameter μ in

Abhängigkeit von der Realisierung $\left[\bar{x}_n - \frac{\sigma_X}{\sqrt{n}} z_{p/2}, \bar{x}_n + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right]$ des Punktschätzers angegeben, siehe etwa Schira, 2005. Hätte beispielsweise in obigem Beispiel bei einem Stichprobenumfang von *n=25* der Punktschätzer $\bar{X}_n$ den Wert $\bar{x}_n = 0,3$ geliefert, so zeigt Abbildung 4 die resultierende „Dichtefunktion“ für μ.

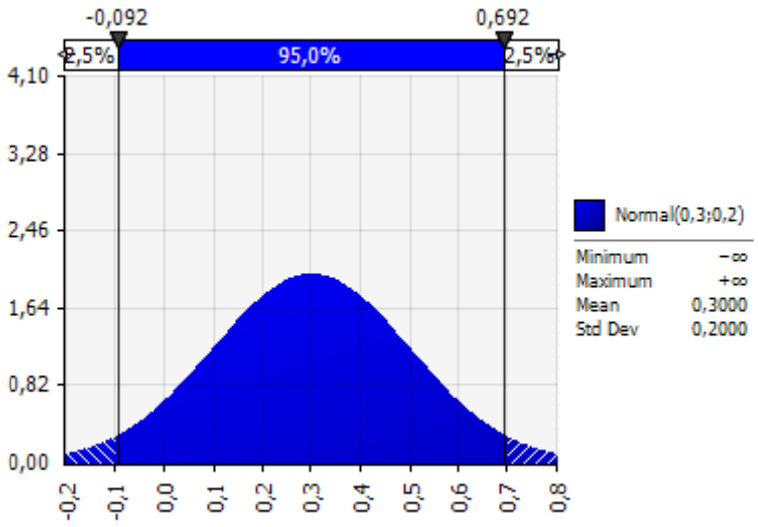


Abbildung 4: „Dichtefunktion“ für µ bei Punktschätzung 0,3

Die unterschiedlichen Sichtweisen sind begründet in einer unterschiedlichen Interpretation von Wahrscheinlichkeiten. Sobald die Stichprobe realisiert ist, der Wert $\bar{x}_n$ also vorliegt, ist der wahre Parameter μ entweder im Intervall $\left[\bar{x}_n - \frac{\sigma_X}{\sqrt{n}} z_{p/2}, \bar{x}_n + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right]$ enthalten oder er ist nicht enthalten. Im Sinne des klassischen Wahrscheinlichkeitsbegriffs ist die Wahrscheinlichkeit somit 1 oder 0. Zur Verdeutlichung werde die Zufallsvariable „Ergebnis eines einmaligen Wurfes eines „fairen“ Würfels“ betrachtet. Die Wahrscheinlichkeit eine 1 zu würfeln ist vor dem Wurf 1/6. Sobald gewürfelt wurde ist das Ereignis entweder eingetreten, seine Wahrscheinlichkeit ist 1, oder das Ereignis ist nicht eingetreten, dann ist die Wahrscheinlichkeit 0. Zufall ist nach dem Wurf nicht mehr im Spiel. Genau dasselbe gilt nach der Berechnung des Stichprobenmittels, denn der wahre Wert μ ist ein zwar immer noch unbekannter, aber trotzdem ein konkreter Wert, der ebenfalls nicht vom Zufall abhängt. Während bei Verwendung des klassischen Wahrscheinlichkeitsbegriffs die Diskussion hier endet, kann bei Verwendung der sogenannten subjektiven Wahrscheinlichkeiten weiter argumentiert werden. Denn es besteht ein entscheidender Unterschied zwischen den beiden Beispielen. Beim Würfelbeispiel sind alle Informationen für die Entscheidung über das Eintreten des Ereignisses „1 wird gewürfelt“ bekannt. Beim Stichprobenmittel fehlt jedoch nach wie vor die Kenntnis des gesuchten Parameters μ. Eine

Parallele ergibt sich dann, wenn der Wurf zwar ausgeführt, aber der Würfel noch unter einem Würfelbecher versteckt ist. Dann hat die realisierte Zufallsvariable weiterhin den Charakter einer Zufallsvariable und mögliche Ereignisse können subjektiv mit Wahrscheinlichkeiten belegt werden. In der Situation, dass die Zufallsvariable Stichprobenmittel $\overline{X}_n$ realisiert wurde, also ein konkreter Wert $\overline{x}_n$ berechnet wurde, ist μ nach wie vor unbekannt. Somit kann die Frage gestellt werden, mit welcher Wahrscheinlichkeit der wahre Parameter μ in einem gegebenen Intervall liegt. Wird der subjektive Wahrscheinlichkeitsbegriff verwendet - und er ist wie gesagt nicht unumstritten - dann und nur dann kann auch das Intervall $\left[\overline{x}_n - \frac{\sigma_X}{\sqrt{n}} z_{p/2}, \overline{x}_n + \frac{\sigma_X}{\sqrt{n}} z_{p/2}\right]$ als ein Konfidenzintervall für μ interpretiert werden. Bei den nachfolgenden Ausführungen wird somit die Akzeptanz des subjektiven Wahrscheinlichkeitsbegriffs implizit vorausgesetzt.

4. Schätzer und Konfidenzintervalle für Cohen's d und ihr Effekt

Die Effektstärke Cohen's *d* ist definiert als durch die gemeinsame Standardabweichung σ normierte Differenz der Mittelwerte μ_1 und μ_2 zweier Verteilungen, d.h. $d = \frac{\mu_2 - \mu_1}{\sigma}$. Bei einer Schätzung für *d*, unabhängig ob bei einer Punkt- oder Intervallschätzung, ist zunächst zu klären, welche Werte unbekannt sind. Im Regelfall sind dieses die beiden Mittelwerte μ_1 und μ_2 und mitunter ebenfalls die Standardabweichung σ. Die nachfolgenden Aussagen und Interpretationen sind davon abhängig, ob die Standardabweichung σ bekannt oder unbekannt ist. Im zweiten Fall werden die Aussagen technisch anspruchsvoller, ohne jedoch inhaltlich stark abzuweichen. Deswegen wird nachfolgend die Kenntnis von σ vorausgesetzt. In diesem Fall folgt für normalverteilte Zufallsvariablen X^1 und X^2, dass der Schätzer $\hat{d}^n = \frac{\bar{X}_n^2 - \bar{X}_n^1}{\sigma}$ für Cohen's *d* exakt normalverteilt ist; Verteilungsaussagen für den Fall unbekannter Varianz finden sich etwa in Hedges, L. 1981. Ferner ist $\hat{d}^n$ erwartungstreu und hat Varianz $\sqrt{2/n}$. Abbildung 5 zeigt exemplarisch die (unbekannte) Dichtefunktion des Schätzers $\hat{d}^n$ für den Fall *n=50* und *d=0,6*. Ein 90%-Konfidenzintervall für den Schätzer ist der Bereich von 0,271 bis 0,929.

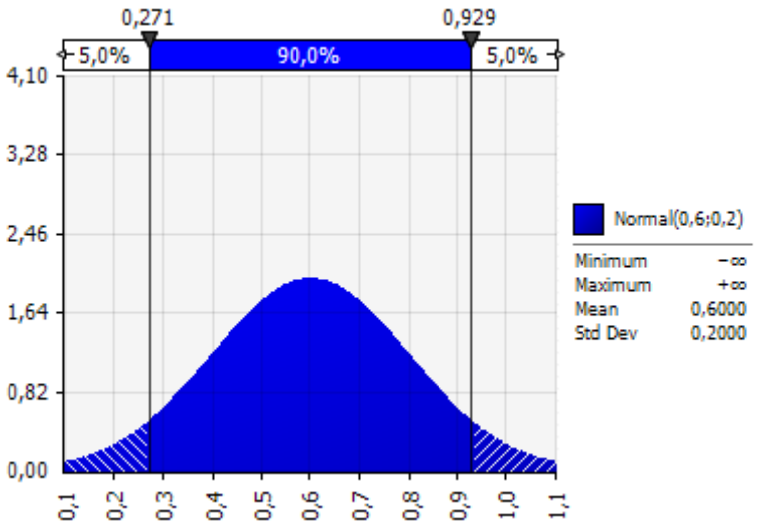


Abbildung 5: Konfidenzintervall für den Effektschätzer

Die Frage der korrekten Interpretation eines Stichprobenergebnisses sei anhand eines fiktiven Beispiels erläutert. Nach Realisierung der Stichprobe ergibt sich beispielsweise ein konkreter Schätzwert $\hat{d}^n$ für d in Höhe von *d=0,5*. Die Akzeptanz subjektiver Wahrscheinlichkeiten vorausgesetzt, ergibt sich die „Dichtefunktion" und ein 90%-Konfidenzintervall für den wahren, aber natürlich unbekannten Wert für Cohen's *d* gemäß Abbildung 6.

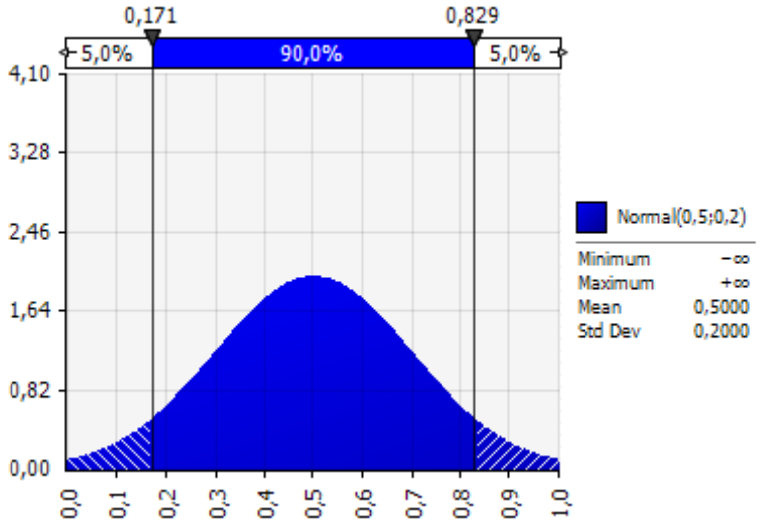


Abbildung 5: „Konfidenzintervall“ für die Effektstärke

Die gezeigte „Dichtefunktion“ spiegelt im Sinne der subjektiven Wahrscheinlichkeiten das Maximum-Likelihood-Prinzip wieder. Denn liefert die Stichprobe den Schätzwert *d=0,5*, so ist die dargestellte „Dichtefunktion“ diejenige, die dieses Ergebnis mit der größten Wahrscheinlichkeit produziert. In Konsequenz kann die Punktschätzung *d=0,5* auch als „wahrscheinlichster“ Wert für den wahren Wert interpretiert werden. Die Interpretation des Wertes im Sinne des resultierenden Effekts erfolgt hiernach genauso wie in Brieden, 2009, beschrieben. Ist im Vorfeld eine Relevanzgrenze festgelegt worden, so kann der geschätzte Wert mit dieser verglichen werden. Eine Entscheidung, ob die Relevanzgrenze nun tatsächlich über oder unterschritten wurde, ist aufgrund des einfachen Größer-Kleiner-Vergleichs prinzipiell möglich. Fundierter ist selbstverständlich eine Betrachtung und Analyse der gesamten Verteilung. Mitunter wird hier die untere Intervallgrenze eines Konfidenzintervalls zu einem vorab gewählten Niveau herangezogen. Die Sinnhaftigkeit dieses Vorgehens kann durch die folgende Überlegung widerlegt werden.

Angenommen die Relevanzgrenze beträgt 0,5 und entspricht der tatsächlich vorliegenden Effektgröße. Dann liefert der Schätzer $\hat{d}^n$ aufgrund seiner Verteilung in 50% der Fälle einen Wert *x*, der kleiner ist als 0,5. Die linke Grenze jedes Konfidenzintervalls um x, egal welcher Länge, also egal bei welchem Stichprobenumfang, ist in diesen Falle somit immer kleiner als 0,5. Dann wird nach obigem Vorgehen das Vorliegen des Effekts aber „verneint“. Die Chance, einen vorliegenden, relevanten Effekt nachzuweisen, ist also nicht größer als 50%.

Zur weiteren Verdeutlichung dieser Interpretationsweise empfiehlt es sich an dieser Stelle, den Zusammenhang zwischen einem Signifikanztest und einem subjektiven Konfidenzintervall aufzuzeigen. Aus Abbildung 6 kann der Annahme- und der

Ablehnungsbereich eines Tests mit der Nullhypothese $H_0 : d \leq 0$ zum Signifikanzniveau 95% abgeleitet werden. Ab einem Testergebnis von 0,329 wird die Nullhypothese verworfen. Ferner ist ein einseitiges, subjektives Konfidenzintervall zum Niveau 95 Prozent erkennbar, das zu einer realisierten Schätzung von 0,329 gehört und bei 0,0 beginnt. Bei jedem größeren Wert als 0,329 enthält das resultierende Konfidenzintervall den Wert 0,0 nicht mehr.

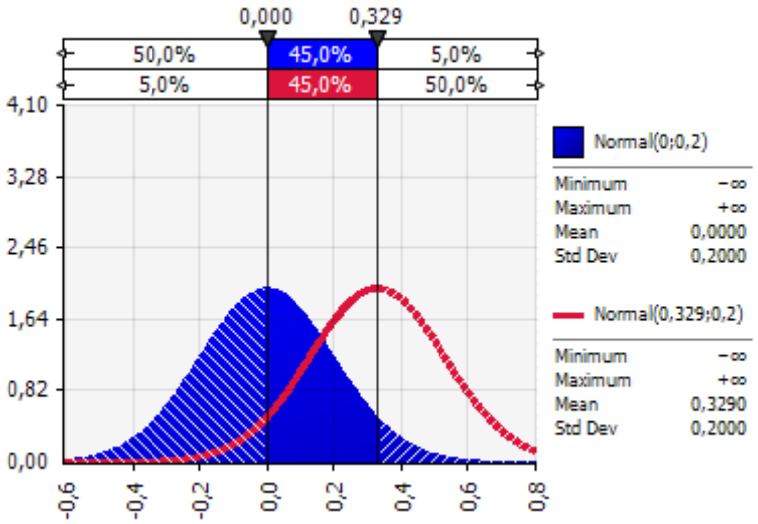


Abbildung 6: Testen versus Schätzen

Dieses bedeutet bezüglich der obigen Interpretationsweise letztendlich das Folgende. Angenommen die linke Intervallgrenze des Konfidenzintervalls ist 0,1. Dann hat die Schätzung einen Wert ergeben, der zur Nichtablehnung der Nullhypothese $H_0 : d \leq 0{,}1$ führt. Die Nichtablehnung einer Nullhypothese ist aber ohne Kenntnis der Wahrscheinlichkeit für einen β-Fehler lediglich ein schwaches Indiz für ihre Korrektheit.

Um zu einer geeigneten Interpretation der subjektiven Dichtefunktion im Sinne der Deutung von Cohen's *d* zu gelangen, sei folgendes Beispiel betrachtet. Im Rahmen der Produktionsplanung sind Reparaturkosten einer erforderlichen Maschine zu berücksichtigen. Es wird geschätzt, dass mit einer Wahrscheinlichkeit von jeweils 0,25 eine beziehungsweise drei Reparaturen notwendig sind. Zwei Reparaturen treten mit einer Wahrscheinlichkeit von 0,5 auf, so dass insgesamt im Schnitt zwei Reparaturen anfallen. Die Kosten für eine Reparatur betragen 2.000 Euro, 3.000 Euro für zwei und 3.500 Euro für drei Reparaturen. Für die Kostenplanung sind jetzt aber nicht die Kosten für die erwartete Reparaturanzahl in Höhe von 3.000 Euro relevant. Vielmehr sind die erwarteten Kosten zu berechnen. Mit einer Wahrscheinlichkeit von jeweils 0,25 treten Kosten in Höhe von 2.000 Euro beziehungsweise 3.500 Euro und mit einer Wahrscheinlichkeit von 0,5 Kosten in Höhe von 3.000 Euro auf.

Somit sind im Schnitt $0,25 \cdot 2.000€ + 0,5 \cdot 3.000€ + 0,25 \cdot 3.000€ = 2.750€$ Kosten zu erwarten. Sind die erwarteten Kosten akzeptabel, kann mit der Planung fortgefahren werden. Übertragen auf Cohen's *d* bedeutet dieses das Folgende. Mit Hilfe der aus dem Ergebnis der Stichprobe resultierenden (subjektiven) Dichtefunktion können die Wahrscheinlichkeiten ermittelt werden, mit denen jeweils die mit der Effektstärke *d* identifizierten Effekte *E(d)* auftreten. Zu berechnen ist folglich der erwartete Effekt $\mu(E(d))$. Ist im Vorfeld eine Relevanzgrenze *r* festgelegt worden, so ist der erwartete Effekt mit dem Effekt der Relevanzgrenze *E(r)* zu vergleichen.

Zur Verdeutlichung des obigen Konzepts dienen die folgenden Zahlenbeispiele für geeignet standardisierte Verteilungen. Dabei wird aus Vereinfachungsgründen diskretisiert. Während $\mu(E(d))$ sich unter Verwendung der normalverteilten Dichtfunktion f(d) als das Integral $\mu(E(d)) := \int_{-\infty}^{\infty} E(d) f(d) dd$ berechnet, werden hier vereinfachend die folgenden Wahrscheinlichkeiten fixiert. Die Stichprobe hat den Wert *d=0,5* ergeben, der dann subjektiv mit einer Wahrscheinlichkeit von *p=0,5* auch tatsächlich korrekt ist. Die Werte d=0,2 und d=0,8 sind mit der subjektiven Wahrscheinlichkeit p=0,25 korrekt. Tabelle 1 weist in der fünften Spalte den resultierenden, erwarteten Effekt aus. In der sechsten Spalte steht derjenige Wert, der den gleichen Effekt nach sich zieht. Im Sinne der Interpretation bedeutet dieses, dass die vorab gewählte Relevanzgrenze mit diesem Wert verglichen werden muss.

Effekt *E(d)*	Cohen's *d*				
Verteilung	0,2 (*p*=0,25)	0,5 (*p*=0,5)	0,8 (p=0,25)	$\mu(E(d))$	„effektgleiche" Relevanzgrenze
Gleichverteilung	5,8%	14,4%	23,1%	14,4%	0,5
Allgemeine β - Verteilung	6,7%	16,5%	25,7%	16,6%	0,5
Dreiecksverteilung	7,8%	18,3%	27,3%	17,9%	0,49
Normalverteilung	7,9%	19,1%	28,8%	18,7%	0,49
Logistische Verteilung	9,0%	21,2%	31,0%	20,6%	0,48

Tabelle 1: erwartete Effekte für unterschiedliche Verteilungen

Im Falle der konkreten Anwendung ist es selbstverständlich unabdingbar $\mu(E(d)) = \int_{-\infty}^{\infty} E(d) f(d) dd$ exakt zu berechnen, obige Berechnungen dienen nur zur konzeptionellen Darstellung der Idee der Interpretation der Schätzungen für Cohen's *d*.

Dabei ist es aber auf jeden Fall entscheidend, die Relevanzgrenze anwendungsspezifisch vorzugeben.

5. Schätzer für Cohen's d und ihre Konsequenz

Die Bedeutung der anwendungsspezifischen Festlegung der Relevanzgrenze wird belegt durch das Konzept der Konsequenz. Letztlich muss nämlich nicht der Effekt, sondern die aus ihr resultierende Konsequenz betrachtet werden. In Fortführung des Beispiels aus Brieden, 2009, wird in Tabelle 2 in Abhängigkeit von der jeweiligen kritischen Grenze die erwartete Konsequenz berechnet, wobei die Punktschätzung annahmegemäß den Wert d=0,5 hervorgebracht hat. Eine kritische Grenze von beispielsweise 0,5 bedeutet dabei, dass erst ab Erreichen dieses Wertes ein Effekt Relevanz erhält. Und ist in diesem Fall Cohen's *d*=0,2, dann erreichen 7,4% mehr als vorher die kritische Grenze.

K(k,d)	Cohen's *d*			
kritische Grenze *k*	0,2 (p=0,25)	0,5 (p=0,5)	0,8 (p=0,25)	$\mu(E(d))$
0,0	7,9%	19,1%	28,8%	18,7%
0,2	7,9%	19,7%	30,5%	19,5%
0,5	7,4%	19,1%	30,9%	19,1%
0,8	6,2%	17,0%	28,8%	17,3%
1,1	4,8%	13,9%	24,6%	14,3%
1,4	3,4%	10,3%	19,3%	10,8%
1,7	2,2%	7,1%	13,9%	7,6%
2,0	1,3%	4,4%	9,2%	4,8%
2,3	0,7%	2,5%	5,6%	2,8%

Tabelle 2: Standardisierte, erwartete Konsequenzen der Normalverteilung

Im Wesentlichen ist zu erkennen, dass die erwartete Konsequenz jeweils in der gleichen Größenordnung liegt, wie die Konsequenz für *d=0,5*. Ein ähnliches Bild ergibt sich, falls die Punktschätzung etwa den Wert *d=0,8* liefert und für die Werte *d=0,5* und *d=1,1* dann die Wahrscheinlichkeiten 0,25 angesetzt werden. Es gelten somit für geschätzte Werte für Cohen's *d* die analogen Aussagen wie im „ungeschätzten" Fall.

1. Ein geschätzt "kleiner" Effekt bei einer Anwendung kann mitunter für diese viel bedeutendere Konsequenzen haben, als ein geschätzt "großer" Effekt bei einer anderen Anwendung.

2. In Abhängigkeit von der Ausgangssituation kann bei derselben Anwendung ein geschätzt „kleiner Effekt" größere Konsequenzen nach sich ziehen als ein geschätzt „großer" Effekt.

6. Fazit

Obige Ausführungen zu Effekt und Konsequenz unterstreichen auch für den Fall der Schätzung von Cohen's *d* mehr als deutlich die Notwendigkeit, anwendungsspezifische Informationen in die Interpretation der Ergebnisse einfließen zu lassen. Entscheidend ist wiederum dabei jeweils die Beurteilung der durch den Effekt ausgelösten Konsequenzen. Bei der Interpretation der Konfidenzintervalle ist Vorsicht geboten und ein Verständnis des subjektiven Wahrscheinlichkeitsbegriff vorausgesetzt.

München, im Dezember 2009

7. Literaturverzeichnis

Bortz, J., & Lienert, G. (2008). *Kurzgefasste Statistik für die klinische Forschung* (3. Auflage Ausg.). Berlin: Springer.

Brandstätter, E. (1999). Konfidenzintervalle als Alternative zu Signifikanztests. *Methods of Psychological Research Online , Vol. 4* (No. 2), 1-17.

Brieden, A. (2009). Cohen'sd: *Idee, Bedeutung, Interpretation, Anwendung und mögliche Fehlerquellen.*

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Edition Ausg.). Lawrence Erlbaum Associates.

Cohen, J. (1994). The Earth Is Round (p<.05). *American Psychologist* , 997-1003.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational and Behavioral Statistics , Volume 6* (Number 2), 107-128.

Schira, J. (2005). *Statistsiche Methoden der VWL und BWL* (2. Auflage Ausg.). München: Pearson Studium.

von Auer, L. (2006). *Ökonometrie: Eine Einführung* (3. Auflage Ausg.). Berlin, Heidelberg, New York: Springer.

A 1.9 Merz Pharmaceuticals

Autoren:
Zügel, Martin.
Göbel, Claus

MERZ PHARMACEUTICALS

[redacted]

Herrn Professor
Dr. med. J. Windeler
Leiter des Instituts für Qualität und Wirtschaftlichkeit
Im Gesundheitswesen
Dillenburger Str. 27
51105 Köln

[redacted]

7. April 2011

Stellungnahme zum Entwurf der allgemeinen Methoden, Version 4.0 vom 09.03.2011

Sehr geehrter Herr Professor Windeler,

im März hat das IQWiG den Entwurf der allgemeinen Methoden, Version 4.0, fertig gestellt und gibt die Gelegenheit zur Stellungnahme. Unser Haus hat sich in der Vergangenheit an der Diskussion um die Nutzenbewertung beteiligt und hält einen breiten Konsens zwischen den verschiedenen Beteiligten für sehr wichtig. Daher möchten wir unsere Sichtweise auch jetzt wieder in die Weiterentwicklung des Nutzenbewertungsverfahrens einbringen.

Wir möchten im Folgenden auf das Kapitel 7.3.3. Beurteilung klinischer Relevanz (S. 129ff) eingehen, zu dem wir im Rahmen des laufenden Verfahrens zur Nutzenbewertung von Antidementiva (Memantin) bereits mehrfach Stellung genommen haben.

Im zweiten Satz sollte der Begriff *Systemebene* erläutert werden. Beim dritten Satz *„Dabei können auch ökonomische Überlegungen eine Rolle spielen"* bleibt unklar, in welchem Zusammenhang ökonomische Überlegungen mit der klinischen Relevanz stehen sollen. Bedeutet dies, dass je nach dem, welche Ressourcen für eine Therapie zur Verfügung stehen, die klinische Relevanz unterschiedlich bewertet werden soll? Aus unserer Sicht sollte die klinische Relevanz unabhängig von ökonomischen Betrachtungen beurteilt werden.

Im dritten Absatz wird ausgeführt, dass bei Fehlen von validierten bzw. etablierten Relevanzkriterien für eine Skala auf ein allgemeines statistisches Maß, die standardisierte Mittelwertsdifferenz (SMD) in Form von Hedges' g (vergleichbar mit Cohen's d) zurückgegriffen werden muss. Die Festsetzung einer fixen Irrelevanzschwelle von 0,2 SMD erscheint hier willkürlich und wird nicht ausreichend begründet.

Merz Pharmaceuticals GmbH

[redacted]

Zwar hatte Cohen für dieses biometrische Maß die Werte 0,2 als kleine, 0,5 als mittlere und ≥ 0,8 als große Effektsstärken vorgeschlagen, aber gleichzeitig darauf hingewiesen, dass diese Grenzen relativ sind, vom jeweiligen Forschungsgebiet abhängen und dass die zugrunde gelegten Werte ausschließlich intuitiv gewählt wurden [Cohen 1988]:

> *„The terms ‚small', ‚medium' and ‚large' are relative, not only to each other, but to the area of behavioural science or even more particularly to the specific content and research method being employed in any given investigation."*
>
> *(...)*
>
> *„The values chosen had no more a reliable basis than my own intuition."*

Auch andere Autoren weisen auf die Nachteile der vom IQWiG vorgeschlagenen Methode hin:

> *"Effect size does not tell us that a result actually is clinically important and does not set evidence-based thresholds at which outcome measures are felt to be clinically important. Furthermore, effect size is not stable."* [Molnar 2009, S. 537].

In derselben Arbeit nennt der Autor die Vor- und Nachteile von Cohen's d:

> Vorteile:
> *"Cheap; Fast; May be a* ***useful starting point*** *when there are no empirically derived or opinion-based measures of clinical importance"*
>
> Nachteile:
> *"May be* ***unstable*** *because may vary from study to study (because based on standard deviation and sample size of study). This feature is undesirable: 'the determination of clinically significant change should not depend on the vagaries of a particular client sample.'* [Molnar 2009, S. 539]

Zusätzlich zur Einführung einer Irrelevanzschwelle wird vom IQWiG gefordert, dass für einen Nutzenbeleg die untere Grenze des Konfidenzintervalls oberhalb der Grenze von 0,2 SMD liegen muss.

Dies bedeutet, dass für eine Substanz mit einem kleinen aber nachweisbaren Effekt von 0,2 SMD niemals ein Nutzenbeleg erbracht werden kann, da die untere Grenze des Konfidenzintervalls selbst bei sehr großen Patientenzahlen immer unter 0,2 liegen wird.

Aus unserer Sicht entspricht die rein mathematische Ableitung des Nutzens auf der Basis einer berechneten Effektstärke und willkürlich gesetzter Relevanzgrenzen nicht den internationalen Standards, dieser Abschnitt sollte daher korrigiert werden.

Wir sind Ihnen dankbar, wenn Sie unsere Anliegen bei der Überarbeitung der Methoden in der Version 4.0 berücksichtigen. Sollte es eine mündliche Anhörung zum Methodenpapier geben, so würden wir uns gerne daran beteiligen.

Mit freundlichen Grüßen

Dr. med. Martin Zügel
CEO Merz Pharmaceuticals GmbH

Dr. med. Claus Göbel
Head Global Clinical Development Operations

Literatur

Cohen J: Statistical power analysis for the behavioural sciences. Hillsadale: NJ: Erlbaum. 1988

Molnar FJ; Man-Son-Hing M; Fergusson D: Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia. J Am Geriatr Soc / 57(3)/ 536-546 /2009

A 1.10 MSD SHARP & DOHME GMBH

Autoren:
Meinhardt, Erik
Krobot, Karl J.

MSD SHARP & DOHME GMBH
[redacted]

MSD

[redacted]
Institut für Qualität und Wirtschaftlichkeit
im Gesundheitswesen
Prof. Dr. med. Jürgen Windeler
Dillenburger Straße 27
51105 Köln

08. April 2011

Entwurf der Version 4.0 der Allgemeinen Methoden

Sehr geehrter Herr Professor Windeler,

wir begrüßen die Möglichkeit der Kommentierung des Entwurfs und übersenden Ihnen hiermit unsere Stellungnahme. Wir bitten, die genannten Punkte zu berücksichtigen.

Für Rückfrage stehen wir Ihnen gerne zur Verfügung.

Mit freundlichen Grüßen
MSD SHARP & DOHME GMBH

Erik Meinhardt
Direktor Market Access

Dr. Dr. Karl J. Krobot, MPH (USA)
Direktor Outcomes Research / HTA

Anlage

Stellungnahme der MSD SHARP & DOHME GMBH vom 08.04.2011

zum Entwurf der

Version 4.0 der Allgemeinen Methoden vom 09.03.2011

Kontakt:

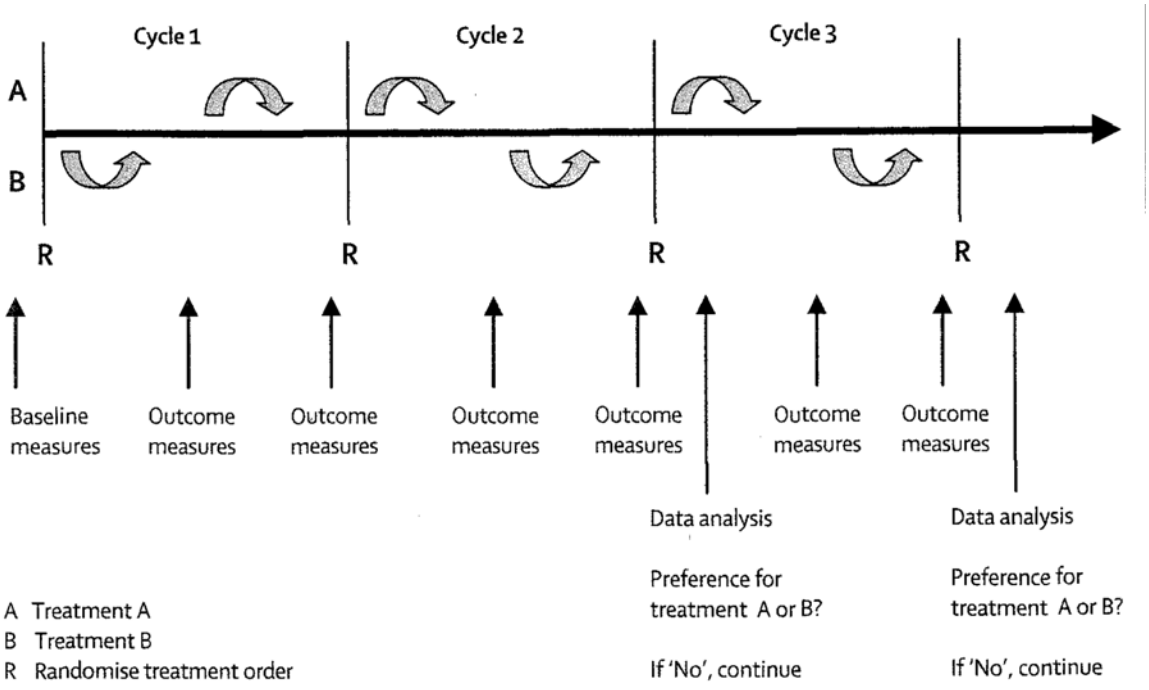
Dr. Dr. Karl J. Krobot, MPH (UNC)
Director Outcomes Research / HTA
MSD SHARP & DOHME GMBH
[redacted]

Wir danken für die Gelegenheit zur Kommentierung des Entwurfs der neuen Methoden 4.0 und unterbreiten fünf Empfehlungen.

1. "Nutzen im Einzelfall" und "Randomisierte n-of-1-Studien" (S. 10)

Wir stimmen zu, dass Aussagen über den Nutzen im Sinne von Erfolgsaussagen für den Einzelfall grundsätzlich nicht möglich sind bzw. umgekehrt auf Einzelfällen basierende Erfahrungen – abgesehen von Ausnahmen (z. B. dramatischer Effekt) – für eine Nutzenbewertung ungeeignet sind.

Wir stimmen jedoch nicht zu, dass dies auch für die typische randomisierte „n-of-1-Studie“ zutrifft, in welcher nicht nur ein einzelner, sondern so viele Patienten verblindet mehrere Versuchs- und Kontrollstudienperioden in zufälliger Reihenfolge durchlaufen, bis hinreichend ergebnissicher beurteilt werden kann, wie viele Patienten nur von Therapie A, nur von Therapie B bzw. wie viele Patienten von beiden Therapien gleichermaßen profitieren.



Hankey GJ. Are n-of-1 trials of any practical value to clinicians and researchers. In Rothwell PM, ed.: Treating individuals. From randomized trials to personalized medicine. pp231-243. 2007

Insofern liefern diese Studien z. B. in Schmerzindikationen head-to-head Informationen, die im Parallelgruppenvergleich verborgen bleiben, aber für Patienten, behandelnde Ärzte bzw. HTA-Verfahren höchst relevant sein können. Die Durchführungsvoraussetzungen hatten Sie in den Methoden 3.0 genannt. Wir empfehlen daher, getrennte Kapitel anzulegen:

- "Nutzen im Einzelfall"
- "Randomisierte n-of-1-Studien".

2. Verhältnis von Zulassung und Nutzenbewertung (S. 34)

Wir zitieren von S. 34:

Die Notwendigkeit für die Betrachtung von Surrogatendpunkten kann im Rahmen der frühen Nutzenbewertung von Arzneimitteln (siehe Abschnitt 1.1) eine besondere Bedeutung haben, da in den Zulassungsverfahren primär die Wirksamkeit, aber nicht immer der patientenrelevante Nutzen untersucht wird.

Hier bitten wir zu konkretisieren:

Die Notwendigkeit für die Betrachtung von Surrogatendpunkten kann im Rahmen der frühen Nutzenbewertung von Arzneimitteln (siehe Abschnitt 1.1) eine besondere Bedeutung haben, da in den Zulassungsverfahren primär die Wirksamkeit, aber nicht immer der patientenrelevante Zusatznutzen untersucht wird.

Wir belegen dies mit dem Schreiben des Bundesministeriums für Gesundheit vom 25. Oktober 2010[1], welches wir auszugsweise einkopieren.

Erläuterung:

Ein Arzneimittel ist nur verkehrsfähig, wenn es arzneimittelrechtlich zugelassen ist, d.h. wenn Qualität, Sicherheit und Wirksamkeit des Arzneimittels nachgewiesen wurden. Die arzneimittelrechtlichen Zulassung wird nur erteilt, wenn das Nutzen-Risikoverhältnis des Arzneimittels positiv ist, was eine Bewertung der positiven therapeutischen Wirkungen des Arzneimittels im Verhältnis zu seinem Risiko umfasst. Es ist daher nicht möglich, dass für ein zugelassenes Arzneimittel keine hinreichenden Nachweise für den Nutzen nach anerkanntem Stand

der medizinischen Erkenntnisse vorliegen, sonst hätte die Zulassung nicht erteilt werden dürfen. Behauptungen, es bestehe kein Zusammenhang zwischen arzneimittelrechtlicher Zulassung und patientenrelevantem Nutzen sind somit falsch. Aufgrund der Zulassung sind Arzneimittel grundsätzlich für die Behandlung der zugelassenen Indikationen geeignet.

Die Zulassungsentscheidung ist ein Verwaltungsakt, an den auch der G-BA insoweit gebunden ist, als er die bei der Zulassung geprüften Kriterien unter dem Aspekt des medizinischen Nutzens eines Arzneimittels nicht abweichend von der Beurteilung der für die Zulassung nach dem AMG zuständigen Behörde bewerten darf. Bereits nach geltendem Recht ist der G-BA daher an den Inhalt der Zulassungsentscheidung gebunden und nicht befugt, hiervon abweichende Regelungen zu treffen (z.B. BSG Urteil vom 31.5.2006, B 6 KA 13/05 R).

Bei der Zulassung wird das Nutzen-Risiko Verhältnis nach dem jeweils gesicherten Stand der wissenschaftlichen Erkenntnis allerdings nur für das Arzneimittel geprüft, für das ein Antrag auf Zulassurg gestellt wird. Die arzneimittelrechtliche Zulassung sagt weder etwas über die Wirtschaftlichkeit eines Arzneimittels, noch trifft sie Aussagen zur Zweckmäßigkeit des Arzneimittels, die sich aus dem Vergleich eines Arzneimittels mit anderen Arzneimitteln und Behandlungsformen im Therapiegebiet ergibt. Der G-BA ist somit befugt, auf Grundlage der Feststellungen der Zulassungsbehörden über Verordnungseinschränkungen und –ausschlüsse wegen Unzweckmäßigkeit oder Unwirtschaftlichkeit von Arzneimitteln in seinen Richtlinien nach § 92 Absatz 1 Satz 2 Nr. 6 SGB V zu entscheiden.

3. Einzelstudienbelege (S. 38)

Wir zitieren aus S. 38:

Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und deren Ergebnisse besondere Anforderungen zu stellen [146]. Meta-Analysen und qualitative Zusammenfassungen von Studien mit endpunktbezogen mäßiger Ergebnissicherheit oder Einzelstudienergebnisse mit hoher Ergebnissicherheit können trotz statistisch signifikanter Effekte demnach in der Regel allenfalls einen Hinweis liefern.

Hier sehen wir zwingenden Ergänzungsbedarf:

Einzelstudien, aufgrund derer eine Zulassung oder Zulassungserweiterung erteilt wurde, liefern regelhaft einen Nutzenbeleg. Andernfalls sind an eine solche Studie und deren Ergebnisse besondere Anforderungen zu stellen [146]. Meta-Analysen und qualitative Zusammenfassungen von Studien mit endpunktbezogen mäßiger Ergebnissicherheit oder sonstige Einzelstudienergebnisse mit hoher Ergebnissicherheit können trotz statistisch signifikanter Effekte in der Regel allenfalls einen Hinweis liefern.

Einzelstudienbeispiele aus eigener Forschung mit bis zu mehr als 20.000 Patienten bzw. 900 Zentren exemplifizieren dies[2-10]:

- Zur "Prävention der symptomatischen Herzinsuffizienz"
- Zur "Senkung der kardiovaskulären Mortalität und Morbidität"
- Zur "Reduktion des Schlaganfallrisikos"
- Zur "Therapie der Osteoporose, um das Risiko für Wirbel- und Hüftfrakturen zu vermindern".

Design, Laufzeit und Anzahl der eingeschlossenen Patienten führen zu einer Ergebnissicherheit, die nicht unter "Hinweis" subsummiert werden kann und darf. Diese Studien zu wiederholen, wäre darüber hinaus auch ethisch unvertretbar.

Die nachfolgende Tabelle und die angefügten Fachinformationen[11-14] belegen dies im Detail.

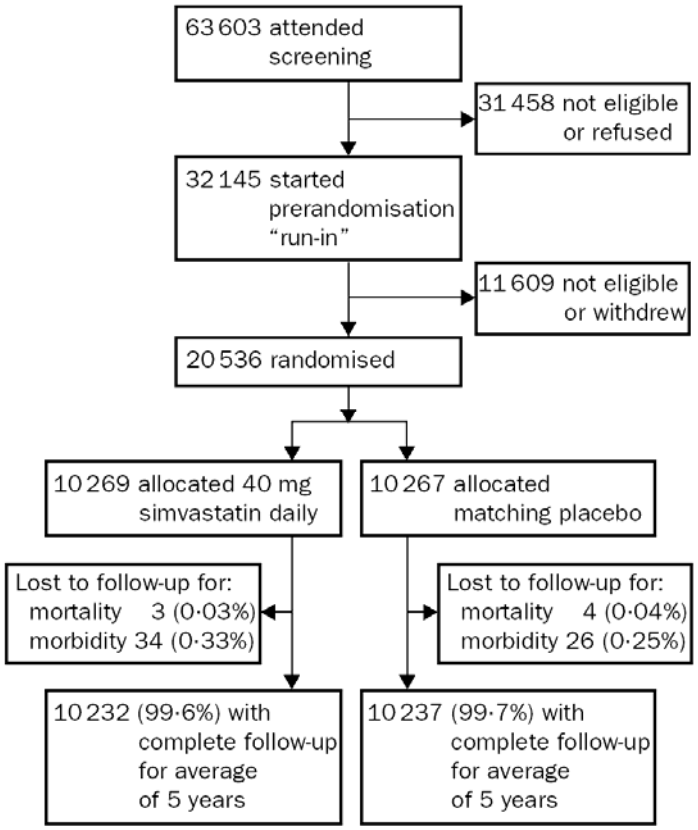
Studie	Intervention	Patienten	Zentren	4.1. Anwendungsgebiet (Auszug aus den beigefügten Fachinformationen)
SOLVD-Behandlungs-studie	Enalapril (XANEF®)	2.569	83	– Behandlung der symptomatischen Herz-insuffizienz
SOLVD-Präventions-studie	Enalapril (XANEF®)	4.228	83	– Prävention der symptomatischen Herz-insuffizienz bei Patienten mit asymptoma-tischer linksventrikulärer Dysfunktion
4S	Simvastatin (ZOCOR®)	4.444	94	Zur Senkung kardiovaskulärer Mortalität und Morbidität bei Patienten mit manifester athe-rosklerotischer Herzerkrankung oder Diabe-tes mellitus, deren Cholesterinwerte normal oder erhöht sind. Begleitend zur Korrektur anderer Risikofaktoren und kardioprotektiver Therapie (siehe 5.1).
Heart Protection Study (HPS)	Simvastatin (ZOCOR®)	20.536	69	Zur Senkung kardiovaskulärer Mortalität und Morbidität bei Patienten mit manifester athe-rosklerotischer Herzerkrankung oder Diabe-tes mellitus, deren Cholesterinwerte normal oder erhöht sind. Begleitend zur Korrektur anderer Risikofaktoren und kardioprotektiver Therapie (siehe 5.1).
LIFE	Losartan (LORZAAR®)	9.193	945	• Reduktion des Schlaganfallrisikos bei erwachsenen hypertonen Patienten mit EKG-dokumentierter linksventrikulärer Hypertrophie (siehe Abschnitt 5.1: LIFE-Studie, ethnische Zugehörigkeit).
RENAAL	Losartan (LORZAAR ®)	1.513	250	• Behandlung einer Nierenerkrankung bei erwachsenen Patienten mit Hypertonie und Typ-2-Diabetes mellitus mit einer Proteinurie ≥0,5 g/Tag als Teil einer anti-hypertensiven Behandlung.
FIT	Alendronat (FOSAMAX®)	6.459	34	FOSAMAX®[1] ist indiziert zur Therapie der Osteoporose bei postmenopausalen Frauen, um das Risiko für Wirbel- und Hüft-frakturen zu vermindern.

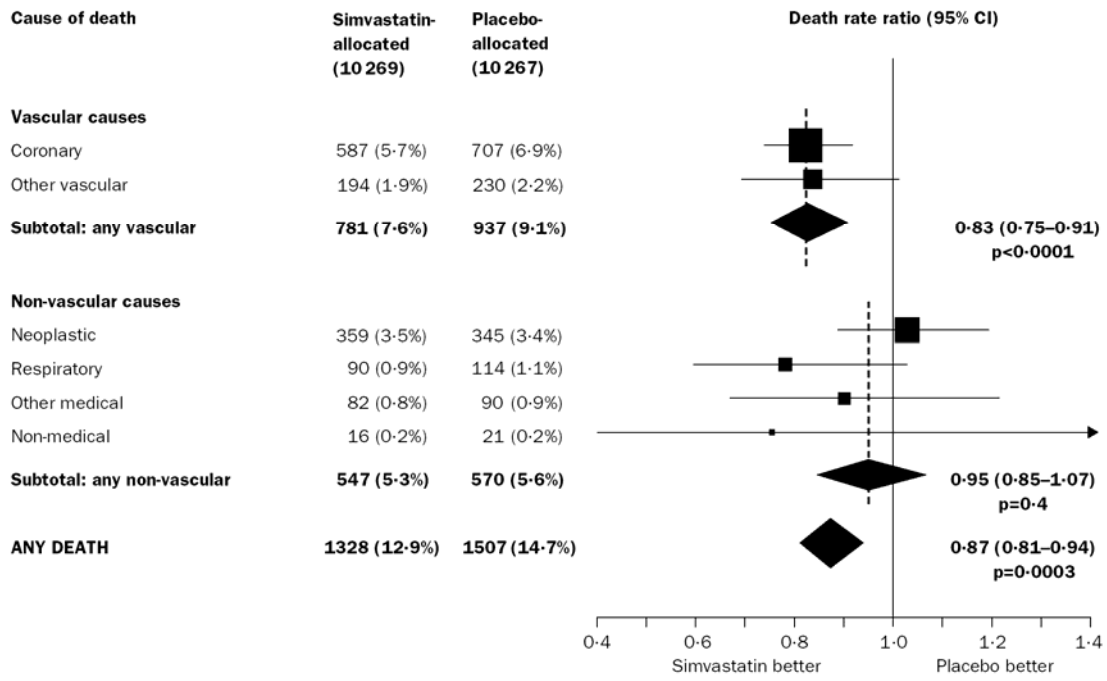
4. Ergebnissicherheit (S. 39)

Wir zitieren aus S. 38 (Unterstreichung durch uns):

Grundsätzlich ist jedes Ergebnis einer empirischen Studie oder einer systematischen Übersicht über empirische Studien unsicher. Hierbei ist zu unterscheiden zwischen qualitativer und quantitativer Ergebnisunsicherheit.

Wir vermuten, intendiert ist hervorzuheben, dass wir nie den direkten Blick auf den wahren Wert haben, sondern uns einer Studie bedienen müssen. Dies ist richtig. Andererseits liegt jedermann fern, z. B. das Ergebnis der Heart Protection Study[10] als "unsicher" zu bezeichnen. Im Gegenteil, die Studie ist qualitativ (doppelblind) wie quantitativ (p=0.0003) höchst ergebnissicher:





Eine Möglichkeit, solchen Studien Rechnung zu tragen, wäre:

Grundsätzlich ist jedes Ergebnis einer empirischen Studie oder einer systematischen Übersicht über empirische Studien auf Ergebnisunsicherheit zu prüfen. Hierbei ist zu unterscheiden zwischen qualitativer und quantitativer Ergebnisunsicherheit.

5. Kriterien zum Einschluss von Studien (S. 114)

Auf S. 114 findet sich folgender Absatz:

Für das Einschlusskriterium bezüglich Population reicht es aus, wenn bei mindestens 80 % der in der Studie eingeschlossenen Patienten dieses Kriterium erfüllt ist. Liegen für solche Studien entsprechende Subgruppenanalysen vor, wird auf diese Analysen zurückgegriffen. Studien, bei denen das Einschlusskriterium bezüglich Population bei weniger als 80 % der in der Studie eingeschlossenen Patienten erfüllt ist, werden nur dann eingeschlossen, wenn entsprechende Subgruppenanalysen vorliegen. Ist beides in einer Studie nicht der Fall, wird die Studie aus der Nutzenbewertung ausgeschlossen

Nicht berücksichtigt ist hier die Möglichkeit, dass Indikationen auf das Vorliegen bestimmter Unverträglichkeiten bzw. Kontraindikationen eingeschränkt erteilt werden. In diesen Fällen entsteht das einschränkende Merkmal mitunter erst im Laufe des Zulassungsverfahrens; war also mitunter kein Einschlusskriterium der (für die Frühbewertung einzig zugrundeliegenden) Zulassungsstudie - mitunter für nachträgliche Subgruppenanalysen noch nicht einmal erhoben. Dennoch muss sichergestellt sein, dass die Studie in die Nutzen(früh)bewertung eingeschlossen wird, denn auf ihr basiert die Zulassung bzw. Zulassungserweiterung.

In anderen Fällen setzt die Anwendung eines Arzneimittels voraus, dass bestimmte (Vor)therapien nicht ausreichen. In solchen Fällen wird eine Studie mitunter bei "gesünderen" Patienten durchgeführt, als später als Anwendungsgebiet spezifiziert, damit die Intervention ethisch vertretbar überhaupt über einen ausreichend langen Zeitraum gegen Placebo geprüft werden kann. Auch in diesen Fällen muss sichergestellt sein, dass die Studie in die Nutzen(früh)bewertung eingeschlossen wird, denn auf ihr basiert die Zulassung bzw. Zulassungserweiterung.

Wir regen daher an, auf S. 114 voranzustellen:

Studien, auf welchen die jeweilige Zulassung bzw. Zulassungserweiterung basiert, werden regelhaft eingeschlossen.

Ebenfalls einzubringen ist die im Methodenpapier an anderer Stelle (S. 45) beschriebene Situation, dass die Ergebnisse von Studien außerhalb des Zulassungsstatus auch dann als "anwendbar" anzusehen sind, wenn *"hinreichend sicher plausibel ist, dass die Effektschätzer patientenrelevanter Endpunkte nicht wesentlich durch das betreffende Merkmal (z. B. geforderte Vorbehandlung) beeinflusst werden"*.

Aus Gründen der Konsistenz empfehlen wir daher sinngemäß zu ergänzen:

Für das Einschlusskriterium bezüglich Population reicht es aus, wenn bei mindestens 80 % der in der Studie eingeschlossenen Patienten dieses Kriterium erfüllt ist. Liegen für solche Studien entsprechende Subgruppenanalysen vor, wird auf diese Analysen zurückgegriffen. Studien, bei denen das Einschlusskriterium bezüglich Population bei weniger als 80 % der in der Studie eingeschlossenen Patienten erfüllt ist, werden nur dann eingeschlossen, wenn entsprechende Subgruppenanalysen vorliegen bzw. wenn hinreichend sicher plausibel ist, dass die Effektschätzer patientenrelevanter Endpunkte nicht wesentlich durch das betreffende Merkmal (z. B. geforderte Vorbehandlung) beeinflusst werden. Ist beides in einer Studie nicht der Fall, wird die Studie aus der Nutzenbewertung ausgeschlossen.

Referenzen

1. Schreiben des Bundesministeriums für Gesundheit vom 25.10.2010 (in Sachen Verordnungseinschränkungen und -ausschlüsse: Glinide zur Behandlung des Diabetes mellitus). URL: http://www.g-ba.de/downloads/40-268-1410/2010-06-17_AM-RL3_Glinide_BMG_2.pdf (abgerufen am 09.05.2011).
2. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. The SOLVD Investigators. N Engl J Med 1991;325(5):293-302.
3. Effect of enalapril on mortality and the development of heart failure in asymptomatic patients with reduced left ventricular ejection fractions. The SOLVD Investigattors. N Engl J Med 1992;327(10):685-91.
4. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). Lancet 1994;344(8934):1383-9.
5. Black DM, Thompson DE, Bauer DC, et al. Fracture risk reduction with alendronate in women with osteoporosis: the Fracture Intervention Trial. FIT Research Group. The Journal of clinical endocrinology and metabolism 2000;85(11):4118-24.
6. Brenner BM, Cooper ME, de Zeeuw D, et al. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. N Engl J Med 2001;345(12):861-9.
7. Collins R, Armitage J, Parish S, Sleigh P, Peto R. MRC/BHF Heart Protection Study of cholesterol-lowering with simvastatin in 5963 people with diabetes: a randomised placebo-controlled trial. Lancet 2003;361(9374):2005-16.
8. Collins R, Armitage J, Parish S, Sleight P, Peto R. Effects of cholesterol-lowering with simvastatin on stroke and other major vascular events in 20536 people with cerebrovascular disease or other high-risk conditions. Lancet 2004;363(9411):757-67.
9. Dahlof B, Devereux RB, Kjeldsen SE, et al. Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. Lancet 2002;359(9311):995-1003.
10. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360(9326):7-22.
11. Fachinformation FOSAMAX einmal wö. 70mg Tabletten. MSD 2010 Dez.
12. Fachinformation XANEF 5mg, 10mg, 20mg, Cor 2,5mg. MSD 2010 Nov.
13. Fachinformation ZOCOR-ZOCOR FORTE. MSD 2010 Oct.
14. Fachinformation LORZAAR 12,5-50-100mg. MSD 2010 Dec.

A 1.11 Verband forschender Arzneimittelhersteller e.V. (vfa)

Autoren:
Throm, Siegfried

vfa. Die forschenden Pharma-Unternehmen

Stellungnahme

zu

Allgemeine Methoden des IQWiG

Entwurf für Version 4.0

vom 09.03.2011

Der vfa ist der Verband der forschenden Pharma-Unternehmen in Deutschland. Er vertritt die Interessen von 44 weltweit führenden Herstellern und ihren über 100 Tochter- und Schwesterfirmen in der Gesundheits-, Forschungs- und Wirtschaftspolitik. Die Mitglieder des vfa repräsentieren rund zwei Drittel des gesamten deutschen Arzneimittelmarktes und beschäftigen in Deutschland rund 90.000 Mitarbeiter.

Nachfolgend nimmt der vfa Stellung zum Entwurf für die Version 4.0 der „Allgemeinen Methoden des IQWiG":

Kapitel 2.1.4 Stellungnahme des IQWiG, S. 20/21

Hier wird ein neues IQWiG-Produkt, und zwar eine Stellungnahme, beschrieben. Darin ist je nach Fragestellung vorgesehen, externe Sachverständige zu beteiligen. Sollte die Stellungnahme ein oder mehrere Arzneimittel betreffen, regen wir an, auch den/die entsprechenden pharmazeutischen Unternehmer in die Erarbeitung mit einzubeziehen.

Kapitel 3.1.1 Definition des patientenrelevanten medizinischen Nutzens, S. 32 (Kapitel 3.1.4 Zusammenfassende Bewertung, S. 38f. entsprechend)

Hier heißt es mit Verweis auf höchstrichterliche Rechtsprechung, wonach bestimmte (Nutzen)Aspekte erst dann notwendigerweise zu bewerten sind, wenn die therapeutische Wirksamkeit hinreichend belegt ist: „Bei besonders schwerwiegenden oder gar lebensbedrohlichen Erkrankungen ist es beispielsweise in der Regel nicht allein ausreichend, eine Verbesserung der Lebensqualität durch die Anwendung einer zu bewertenden Intervention nachzuweisen, wenn nicht gleichzeitig mit ausreichender Sicherheit ausgeschlossen werden kann, dass schwerwiegende Morbidität oder gar Mortalität in einem nicht mehr akzeptablen Ausmaß ungünstig beeinflusst werden." Weiter heißt es auf S. 33, dass es sinnvoll sein kann, eine Hierarchisierung von Endpunkten vorzunehmen. Allgemeine Nutzen- und Schadenaussagen würden sich dann in erster Linie auf Belege bezüglich der höher gewichteten Zielgrößen stützen.

Im Rahmen der vergleichenden Nutzenbewertung hat das IQWiG Interventionen bezogen auf patientenrelevante Outcomes miteinander zu vergleichen. Eine Hierarchisierung der Outcomes in dem Sinne, dass etwa der Nachweis eines patientenrelevanten (Zusatz-)Nutzens bzw. die Abwesenheit eines (Zusatz-)Schadens für einen Outcome eine notwendige Bedingung für die Beurteilung weiterer Outcomes ist, erscheint ohne die Angabe von Kriterien, in welchen Fällen und wie eine Hierarchisierung vorgenommen werden soll, willkürlich und intransparent. Zudem obliegt eine solche Hierarchisierung nicht dem IQWiG als wissenschaftlicher Instituti-

on, sondern dem G-BA als bewertender Institution. In Absenz entsprechender Hierarchisierungsstudien kann diese Entscheidung nicht basierend auf wissenschaftlicher Evidenz entschieden werden, sondern bedarf einer gesellschaftlichen Abwägung. Im Hinblick auf den Verweis auf höchstrichterliche Rechtsprechung ist anzumerken, dass die Wirksamkeit durch die Zulassung belegt ist.

Kapitel 3.1.2 Surrogate des patientenrelevanten medizinischen Nutzens, S. 33ff.

Das Kapitel zur Berücksichtigung von Surrogatparametern schlägt primär die Verwendung von Metaanalysen von mehreren randomisierten Studien zur Validierung von Surrogatparametern vor. Dies verkennt, dass Surrogatparameter insbesondere dann zur Anwendung kommen, wenn harte Endpunktstudien in der Regel im Rahmen von RCT nicht sinnvoll erhoben werden können, beispielsweise auf Grund des zeitlichen Bezugs bis zum Auftreten des harten Endpunkts und der damit verbundenen Beobachtungsdauer, der notwendigen einzuschließenden Patientenzahl, die meist ethisch kaum vertretbar sein dürfte.

Ferner heißt es, dass die Validierung in möglichst identischen Patientenkollektiven stattfinden muss. Dies ist jedoch aufgrund der sich stetig ändernden Versorgungsstruktur und bei entsprechend kleinen Patientenpopulationen nahezu ausgeschlossen.

Die an die Validierung von Surrogatparameter angelegten Hürden sind hier derart hoch gewählt, dass eine adäquate Validierung im Sinne des Methodenpapiers kaum möglich erscheint bzw. gleich eine Endpunktstudie statt einer Studie mit Surrogatparametern durchgeführt werden sollte.

Insbesondere wenn es sich um etablierte Surrogatparameter handelt, die zudem von den Zulassungsbehörden gefordert werden, sollten diese auch als Beleg für einen Zusatznutzen dienen. Hierzu verweisen wir auch auf die ausführliche Stellungnahme des vfa zum Rapid Report „Aussagekraft von Surrogatendpunkten in der Onkologie" vom 23.03.2011 (Anlage).

Kapitel 3.3.3 Nutzenbewertung von Arzneimitteln gem. § 35a SGB V, S. 48

Im Absatz 2 heißt es im letzten Satz: „Die Kosten sind für das zu bewertende Arzneimittel und die zweckmäßige Vergleichstherapie als direkte Kosten anzugeben (gemessen am Apothekenabgabepreis und unter Berücksichtigung der Fach- und Gebrauchsinformation)." Hier möchten wir anmerken, dass der diesem Satz zugrunde liegende Text der Arzneimittel-Nutzenbewertungsverordnung (AM-NutzenV) („Bestehen bei Anwendung der Arzneimittel entsprechend der Fach- und Gebrauchsinformation regelhaft Unterschiede

bei der notwendigen Inanspruchnahme ärztlicher Behandlung oder bei der Verordnung sonstiger Leistungen zwischen dem zu bewertenden Arzneimittel und der zweckmäßigen Vergleichstherapie, sind die damit verbundenen Kostenunterschiede für die Feststellung der den Krankenkassen tatsächlich entstehenden Kosten zu berücksichtigen.") u.E. auf die zulassungsgemäße Anwendung des Arzneimittels abhebt, und nicht darauf, dass sich die Unterschiede bei der Inanspruchnahme von Leistungen aus der Fach- oder Gebrauchsinformation ergeben.

Der vfa regt an, dass das IQWiG bei der Nutzenbewertung gem. § 35a SGB V ggf. die Möglichkeit zur Kommunikation mit dem betroffenen pharmazeutischen Unternehmen nutzt, z. B. für die rasche Abklärung von Unklarheiten. Dies sollte trotz des engen Zeitplans für die Nutzenbewertung gem. § 35a SGB V möglich sein und hat sich bei Zulassungsverfahren als nützlich für beide Seiten erwiesen.

Gemäß dem letzten Absatz will das IQWiG u. a. prüfen, „ob die vom pharmazeutischen Unternehmen gewählte Vergleichstherapie als zweckmäßig im Sinne der AM-NutzenV gelten kann." Dies kann aber nur dann der Fall sein, wenn der G-BA diese Vergleichstherapie nicht bereits im Rahmen einer Beratung zum Dossier bestimmt hat und das IQWiG ausdrücklich mit dieser Festlegung betraut. Denn gem. § 6 der AM-NutzenV legt der G-BA die zweckmäßige Vergleichstherapie fest. Bei einer frühen Beratung (vor Phase III-Studien) darf von Ergebnissen bezüglich der zweckmäßigen Vergleichstherapie nur dann abgewichen werden, wenn es zwischenzeitlich bedeutsame Änderungen beim Stand der medizinischen Wissenschaft gegeben hat. Kapitel 5 Evidenzbasierte Gesundheitsinformation für Bürger und Patienten.

Das Kapitel 5 zu den Gesundheitsinformationen wurde in der Version 4.0 gegenüber der Vorgängerversion stark erweitert. Der vfa begrüßt, dass auf die Patientensichtweise und ihre Erfahrungen im Umgang mit ihren Erkrankungen bei der Erstellung der Informationen wesentlich stärker eingegangen werden soll. Damit werden die Patientenbedürfnisse berücksichtigt. Die Einbeziehung der Patienten als Nutzer in die Qualitätskontrolle der Gesundheitsinformationen wird begrüßt.

Gleichwohl ist nicht transparent, nach welchen Kriterien die Patientenvertreter ausgewählt werden und in welchem Umfang sie in die Gestaltung der „Gesundheitsinformationen" eingebunden sind.

Bei den Quellen, die das Institut berücksichtigt, um die Informationswünsche der Bürger und Patienten zu erschließen, fehlen die forschenden Pharma-Unternehmen.

Kapitel 6 Informationsbeschaffung

Hier heißt es im 3. Absatz: „Werden dem Institut Daten übermittelt, die nicht publiziert werden dürfen, so können diese nicht inhaltlich in Bewertungen des Instituts einfließen, da dies dem Transparenzgebot widerspricht."

Hier wird ein Fehlverständnis des IQWiG deutlich. Der vom IQWiG unterstellte Normenkonflikt zwischen Transparenzgebot (§ 139a Abs. 4 S. 2 SGB V) und Bewertungsentscheidung existiert nicht. Die deutsche Rechtsordnung kennt kein allgemeines „Symmetrieprinzip", wonach nur solche Informationen und Daten einer hoheitlichen Entscheidung zugrunde gelegt werden dürften, die der Allgemeinheit uneingeschränkt öffentlich zugänglich gemacht werden können. Vielmehr sind auch Informationen und Daten, in deren Veröffentlichung bspw. ein Unternehmen im Fall von Betriebs- und Geschäftsgeheimnissen nicht eingewilligt hat, zwingend zu berücksichtigen, was sich bereits aus dem Rechtsgedanken des § 20 Abs. 2 SGB X als Ausdruck des allgemeinen Rechtsstaatsprinzips ergibt und im konkreten Fall auch für das IQWiG und seine Bewertungen gilt. Auch das IQWiG unterliegt der Pflicht zum Treffen der richtigen Sachentscheidung auf der Grundlage vollständig ermittelter Tatsachen. Die Berücksichtigung aller dem IQWiG vorliegenden Informationen und Daten unabhängig von der Frage nach deren Veröffentlichbarkeit muss in der Bewertungspraxis des IQWiG gewährleistet sein; daher muss dieser Satz gestrichen werden.

Kapitel 6.1.3 Eigene Informationsbeschaffung durch das Institut, S. 108

In Absatz 2 heißt es: „Basiert die Nutzenbewertung auf systematischen Übersichten, wird Vollständigkeit im Sinne einer Berücksichtigung aller verfügbaren Primärstudien nicht angestrebt."

Der vfa geht davon aus, dass dies nicht für Nutzenbewertungen von Arzneimitteln gilt. Da das IQWiG die vollständige Einreichung von Studien von den Firmen verlangt, sollten auch bei eigener Informationsbeschaffung alle Studien berücksichtigt werden.

Kapitel 7.1.4 Aspekte der Bewertung des Verzerrungspotenzials, S. 119

Der Satz „Für nicht randomisierte vergleichende Studien wird in der Regel keine zusammenfassende Bewertung der Verzerrungsaspekte durchgeführt, da die Ergebnisse dieser Studien aufgrund der fehlenden Randomisierung generell ein hohes Verzerrungspotenzial besitzen." sollte gestrichen werden. Er widerspricht nämlich der Verfahrensordnung des G-BA, nach der die Möglichkeit besteht, von Nicht-RCTs eine Metaanalyse zu machen. In jedem Fall muss

die bestverfügbare Evidenz bewertet werden, auch wenn diese lediglich aus Nicht-RCTs besteht.

Hierzu wäre es hilfreich, wenn das Methodenpapier dahingehend ergänzt würde, wie und mit welchen Methoden das Verzerrungspotential von nicht randomisierten Studien minimiert werden kann.

Kapitel 7.1.5 Interpretation von kombinierten Endpunkten, S. 120

Entsprechend dem vorliegenden Entwurf dürfen kombinierte Endpunkte nur dann zusammengefasst werden, sofern die einzelnen Bestandteile eine ähnliche „Schwere" darstellen. Dies ist äußerst vage und gibt keinen Anhaltspunkt, wann eine ähnliche Schwere vorliegt. Bei einer isolierten Betrachtung bleibt zu berücksichtigen, dass die Studien in der Regel auf das Erreichen einer signifikanten Verbesserung des kombinierten Endpunktes „gepowered" wurden und nicht auf das Erreichen der Teilparameter.

Zudem ergeben sich kombinierte Endpunkte zumeist durch die Vorgaben der Zulassungsbehörden sowie wissenschaftlicher Standards in der entsprechenden Indikation. Des Weiteren ist unklar wie diese restriktive Definition mit der Vorgabe in § 7 der Arzneimittel-Nutzenbewertungsverordnung vereinbar ist, dass Bewertungen eines Arzneimittels nicht den Feststellungen der Zulassungsbehörde über Qualität, Wirksamkeit und Unbedenklichkeit eines Arzneimittels widersprechen dürfen.

Kapitel 7.3.3 Beurteilung klinischer Relevanz, S. 129

Hier betreffen unsere Einwände die Wahl einer standardisierten Irrelevanzschwelle von 0.2 für den Fall, dass skalenspezifische validierte bzw. etablierte Kriterien zur Relevanzbewertung nicht vorliegen. Dies soll gemäß IQWiG gewährleisten, dass der beobachtete Effekt hinreichend sicher mindestens als ‚klein' im Sinne der von Cohen eingeführten standardisierten Effektgröße angesehen werden kann. Die Einwände richten sich auf die Verwendung der standardisierten Effektgröße als Relevanzmaß per se wie auch auf eine unrealistische Festlegung der Irrelevanzschwelle bei der Durchführung von Shifthypothesen**.**

Cohen's Effektgröße
Cohen's klassischer Ansatz, eine standardisierte Effektgröße (standardisierte Mittelwertdifferenz, SMD) einzuführen, um die Abschätzung der notwendigen Stichprobengröße oder die Trennschärfe wissenschaftlicher Untersuchungen möglichst einfach (d. h. auch ohne Computerunterstützung) und vergleichend zwischen verschiedenen Studien und Ergebniskenngrößen (Endpunkten) vornehmen zu können, hat enorme Meriten und ist aus den Lehrbüchern der Statistik nicht mehr wegzudenken. Bereits Cohen selbst

hat aber darauf aufmerksam gemacht, dass die standardisierte und dimensionslose Effektgröße oder –stärke d (entsprechend auch die leicht modifizierte Größe Hedge's g) eine rein rechnerische technische Größe ist, die mit fachlicher Substanz gefüllt werden muss, d. h. die (minimale) relevante Differenz (minimal important difference, MID), die im Zähler von d bzw. g steht, muss sachlich, d. h. in diesem Zusammenhang medizinisch, gut begründet werden und Akzeptanz in der medizinischen Fachwelt besitzen.

Die Nennergröße von d und g, die die Gesamtvariabilität in der untersuchten Population beschreibt, wird dabei durch viele Faktoren beeinflusst. Je heterogener eine Population zusammengesetzt ist, desto größer wird erwartungsgemäß die Variabilität ausfallen. Die großen Multicenterstudien, die in der Phase 3 der klinischen Entwicklung durchgeführt werden, sind bekanntermaßen hinsichtlich der Zielvariablen sehr heterogen (im Vergleich etwa zu ‚kleinen' Phase 2 Studien mit stark selektierten Patientenpopulationen), so dass das Verhältnis aus MID und Gesamtvariabilität zwangsläufig kleiner wird, obwohl sich an der durch die MID ausgedrückten Relevanz des Effektes selbst nichts geändert hat.

Die Überprüfung von verschiedenen Studienresultaten aus klinischen Studien aus der Literatur und in der ClinicalTrials.gov Studien-Datenbank zeigt, dass auch rein formal sehr kleine d- bzw. g-Werte mit relevanten und akzeptierten Effekten von Arzneimitteln im Einklang stehen können. In der Praxis tritt dies nicht nur im Einzelfall, sondern in diversen Indikationsfeldern auf, z. B. in Studien zur Harninkontinenz, Veränderung des HbA1c bei Diabetes mellitus, bei Symptomveränderungen der benignen Prostata-Hyperplasie (BPH) oder beim Nachweis der Verringerung von Herz-Kreislauf-Risikofaktoren, um nur einige Beispiele zu nennen. Eine schematische Anwendung bestimmter Grenzwerte für d und g, auch wenn dies nur auf den Fall fehlender Informationen hinsichtlich der MID beschränkt bleiben sollte, geht daher an der Realität vorbei und vergrößert das Risiko einer falschen Nutzenbewertung. Dieses Risiko wird noch gesteigert dadurch, dass diese Kriterien nicht als Argumente in der Diskussion und Abwägung von Nutzenbewertungen, sondern zum Entscheidungskriterium erhoben werden, ob überhaupt Nutzen zu attestieren sei oder nicht (bezogen auf den jeweiligen Endpunkt). Zudem erweckt das IQWiG damit auch den Eindruck, dass sich primär medizinische Bewertungsprobleme, die ganz klar als Werturteil zu klassifizieren sind, durch schematische Anwendung standardisierter Effektgrößen auflösen ließen.

Das IQWiG selbst hat mittlerweile konzediert, dass Cohen's d (Hedge's g) nicht als Relevanzmaß anzusehen ist, sondern nur als Ersatzgröße, die in letzter Instanz herangezogen wird, wenn nichts anderes vorhanden ist. Trotz dieser pragmatischen Einschätzung ändert sich aber nichts an der Tatsache, dass es als verteilungsba-

siertes Maß für diesen Zweck prinzipiell ungeeignet ist und insbesondere in der bisher praktizierten Operationalisierung als Entscheidungskriterium nicht verwendet werden sollte.

Die Frage der klinischen Relevanz ist letztlich ein Werturteil, das nicht allein aus statistischer Sicht beantwortet werden kann. Dieses Urteil hängt unter anderem von der Indikation und den vorhandenen Alternativen ab.

Verschobene Nullhypothese (Shifthypothese)
Das IQWiG setzt im Falle fehlender Informationen zu den MIDs nachträglich eine Relevanzgrenze von d=0.2 fest und spricht dann den Studienergebnissen die Relevanz ab, bei denen in einer Meta-Analyse von mehreren Studien das Konfidenzintervall für die standardisierte Effektgröße nicht vollständig oberhalb dieser vorgegebenen Relevanzschranke liegt. Diese Vorgehensweise ist identisch mit dem Testen einer „verschobenen Nullhypothese", da nicht gegen 0, sondern gegen eine standardisierte Effektgröße von d=0.2 getestet wird.

„Verschobene Nullhypothesen" wurde bereits in den achtziger und neunziger Jahren des letzten Jahrhunderts speziell in deutschen Fachkreisen intensiv diskutiert. Ursprünglich wurde dieser Gedanke von Victor (On Clinically Relevant Differences and Shifted Nullhypotheses, Methods Inf Med 1987) in die Öffentlichkeit getragen, später dann von Röhmel und Trampisch für das Indikationsgebiet ‚Periphere Arterielle Verschlusskrankheit', speziell für Claudicatio intermittens-Patienten, weiter entwickelt. Fazit der damaligen Untersuchungen war, dass die Einführung der Relevanzschwelle Delta in die Nullhypothese, die von der unteren Grenze des Konfidenzintervalls nicht unterschritten werden sollte, zu unsinnigen Resultaten führen kann, da im Falle der tatsächlichen Existenz eines klinisch relevanten Unterschieds Delta die Wahrscheinlichkeit nur sehr gering (2.5 %) wäre, dass eine Studie überhaupt erfolgreich ist. Aus diesem Grund hat Röhmel darauf hingewiesen, dass das Delta nicht gleich der MID gesetzt werden darf und ähnlich wie bei den Nicht-Unterlegenheits-Margen in Non-Inferiority Studien deutlich kleiner gewählt werden muss. In der Tat haben Röhmel und Trampisch diese Schranke (im Sinne einer Irrelevanzschranke) für die Claudicatio intermittens-Studien deutlich niedriger festgelegt, um zu pragmatischen Lösungen für Studien in diesem Bereich zu kommen. Diese prinzipiellen methodischen Überlegungen wurden 1995/1996 publiziert (VASA 24/25), und haben auch heute an ihrer Gültigkeit nichts verloren.

Allgemeine Schlussfolgerung aus den damaligen Diskussionen war, dass der zunächst sehr naheliegende Gedanke, die Relevanzschranke mit in die Formulierung der Superioritätshypothese einzubeziehen, zu unsinnigen Ergebnissen, in jedem Fall aber zu überhöhten Fallzahlen in klinischer Studien führt, die nicht

machbar und auch ethisch nicht vertretbar wären. Die europäische Guideline zur peripheren arteriellen Verschlusskrankheit, die zunächst die Einbeziehung der Relevanzschranke implizit vorsah, wurde aufgrund dieser Überlegungen später revidiert und die Forderung nach Implementierung einer Shifthypothese daher gänzlich aufgegeben. Insgesamt lässt sich feststellen, dass die Einbeziehung einer Irrelevanzschranke analog zu dem Claudicatio intermittens-Ansatz bisher keine größere internationale Resonanz erfahren hat. Das IQWiG weist im Methodenpapier selbst darauf hin, dass Irrelevanzschwellen zwischen Gruppen nur in wenigen Leitlinien zu finden sind. Die Festlegung einer Irrelevanzschwelle im Indikationsfeld der Claudicatio intermittens ist daher eher als (deutscher) Ausnahmefall zu betrachten.

Festlegung von SMD=0.2 als Grenze hinsichtlich der Irrelevanzschwelle

Schaut man etwas genauer auf das Claudicatio intermittens-Beispiel, auf das sich Stefan Lange in seinem letztjährigen ‚IQWiG im Dialog'-Vortrag bezogen hat, und standardisiert die in diesem Indikationsfeld akzeptierte Irrelevanzschwelle, dann lässt sich feststellen, dass die IQWiG-Grenze von 0.2 dort deutlich unterschritten wird. Die gezogene Grenze liegt sogar deutlich unter 0.1 (im log-transformierten Ansatz für das Verhältnis der Quotienten der Gehstreckensteigerungen errechnet sich eine Grenze zwischen 0.06 und 0.08 abhängig von der Annahme über die Variabilität). Gemessen an diesem vielzitierten Beispiel ist die generelle Forderung des IQWiG nach einem Überschreiten der Grenze von 0.2 eindeutig als zu hoch zu bezeichnen.

Die Anwendung der Grenze von 0.2 für SMD wird in Abschnitt 7.3.3 des Entwurfes für den Fall angekündigt, dass ‚skalenspezifische validierte bzw. etablierte Kriterien zur Relevanzbewertung nicht vorliegen'. Dies vermittelt den Eindruck, dass diese Regelung nur selten und im Ausnahmefall angewendet wird. Dieser Eindruck trügt jedoch. Bei mehreren Nutzenbewertungsverfahren zu Memantine, SSRIs, SNRIs, aber auch bei dem geplanten Bewertungsverfahren für die Zweitlinientherapie in der Rheumatischen Arthritis verwendet das IQWiG nachträglich selektierte ‚patientenrelevante' Endpunkte zur Beurteilung der Relevanz, für die sich meist keine Relevanz- und Irrelevanzschwellen in den Guidelines finden lassen. Es ist daher zu befürchten, dass das IQWiG in schematischer Anwendung seiner Entscheidungshierarchie zwangsläufig diese Default-Option in größerem Umfang verwenden wird, was das Risiko von Fehlbewertungen potentiell erhöhen kann.

SMD=0.2 als zusätzliche Hürde

Für die Nutzenbewertung von Arzneimitteln hat die Einführung von zusätzlichen Hürden dieser Art den vordergründigen Effekt eines ‚Filters', der Arzneimittel mit ‚zu geringem Zusatznutzen' von vorneherein geringere Chancen gibt, (Zusatz-) Nutzen nachweisen zu

können. Längerfristig hätte dies allerdings zur Folge, dass in Indikationsgebieten, bei denen Fortschritte nur in vielen kleinen Schritten erzielt werden können (z. B. in der Onkologie), längerfristig der Status quo zementiert wird, unabhängig davon, ob diese Fortschritte kosteneffizient sind oder nicht. Dies würde ja im gegenwärtigen Verfahrensablauf gar nicht mehr geprüft werden, da Kosten-Nutzenbewertungen nur dann durchgeführt werden, wenn der Beleg eines Zusatznutzens erbracht wurde. Die Gesundheitssysteme laufen dabei Gefahr, dass die Entwicklung nützlicher und eventuell sogar kostengünstigerer innovativer Therapien unterbleibt, die zwar möglicherweise nur wenig Zusatznutzen aufweisen, aber zu keiner oder einer akzeptablen Verteuerung der Therapie und damit zu einer kosteneffizienten Versorgung beitragen könnten. Solche Therapien können jedoch für bestimmte Patientengruppen wertvolle Therapiealternativen darstellen, wenn z. B. Alternativen versagt haben oder nicht verträglich waren.

Kapitel 7.3.4 Bewertung subjektiver Endpunkte bei offenen Studiendesigns, S. 132

Die hier vorgeschlagene adjustierte Entscheidungsgrenze mit einem an Wood et al. orientierten konkreten Grenzwert wird aus folgenden Gründen kritisch gesehen. So warnen Wood et al. vor dem unkritischen Gebrauch der Zahl 0,75 als generellem Bias bei offenen Studien mit subjektiven Endpunkten. Multivariate Verfahren sind eine Voraussetzung für die Adjustierung von Regressionskoeffizienten, die Vermeidung von Cofounding und das Bereinigen von Effektschätzern. Die bisher vorhandenen Verfahren sind jedoch diesbezüglich noch nicht ausreichend getestet und auf methodische Robustheit zur allgemeinen Verwendung in der Praxis erprobt.

Die Verwendung einer harten Schranke hält der vfa für methodisch fraglich. Als Alternative werden Verfahren diskutiert, die das Ergebnis einer Metaanalyse hinsichtlich eines bestehenden Bias-Risikos korrigieren können. Hierzu werden Bayesianische Verfahren zur Effektschätzer-Modifikation (Shrinkage) vorgeschlagen. Problematisch ist bei diesem Ansatz die fehlende empirische Grundlage zur Bestimmung wesentlicher Parameter, die in das Shrinkage-Verfahren eingehen. Obwohl theoretisch damit die Grundlage für eine interessante Strategie geschaffen ist, fehlen zu deren Anwendung die Parameter, welche die Realität in der klinischen Forschung angemessen quantifizieren.

Instrumente wie der SF-36, Fragen zum global patient assessment und andere Instrumente der Outcome-Messung können jedoch einen hohen Grad an Objektivität aufweisen, so dass sich eine Adustierung erübrigt.

Kapitel 7.3.9 Indirekte Vergleiche, S. 140f.

In diesem Kapitel wird beschrieben, dass für bestimmte Fragestellungen indirekte Vergleiche vonnöten sind, um Interventionen miteinander in Beziehung zu setzen. Leider bleibt das Methodenpapier hier äußerst vage und gibt keine konkreten Vorschläge oder Beispiele, wie indirekte Vergleiche vorzunehmen sind. An dieser Stelle wären entsprechende Methodenbeispiele wünschenswert.

Da es sich hierbei um ein Gebiet handelt, in dem noch viel Fragen offen sind, bietet der vfa gerne einen wissenschaftlichen Dialog hierzu an.

Weiterhin sollte hier das einfache und robuste Verfahren von Bucher et al. (*Bucher 1997; Bucher HC, Guyatt GH, Griffith LE, Walter SD.; The results of direct and indirect treatment comparisons in meta- analysis of randomized controlled trials. Journal of Clinical Epidemiology 1997; 50: 683-691.)* ergänzt werden.

Kapitel 7.4.2 Konsultationstechniken, S. 146:

Hier werden Interviews mit Key Informants, z. B. Patientenvertretern, zur Ermittlung patientenrelevanter Endpunkte aufgeführt. Da patientenrelevanten Endpunkten eine erhebliche Bedeutung bei Nutzenbewertungen zukommt, sollte hier unbedingt ergänzt werden, wie solche Key Informants ausgewählt werden und wie sichergestellt wird, dass die Ergebnisse repräsentativ sind.

Stand: 08.04.2011

A 2 Stellungnahmen von Privatpersonen

A 2.1 Nothacker, Monika

WG Methodenpapier - Bewertung von Leitlinien.pdf

Sehr geehrte Damen und Herren,

im zur Kommentierung freigegebenen Entwurf der Version 4.0 Ihres Methodenpapiers (Abschnitt
4.2.2 Methodische Bewertung von Leitlinien) verweisen Sie darauf, dass das IQWiG aktiv an der
Überarbeitung des DELBI
mitarbeitet. In Deutschland wird zur Leitlinienbewertung hauptsächlich DELBI und nicht AGREE
angewendet.
Für die überarbeitete Version des DELBI wird eine Validierungsstudie durchgeführt werden.
Auf diesen Aspekt sollte unseres Erachtens im Methodenpapier hingewiesen werden.

Wir gehen davon aus, dass das validierte Instrument dann auch vom IQWiG angewendet werden wird.

Mit freundlichen Grüßen

i.A. Monika Nothacker

Dr. Monika Nothacker, MPH
wiss. Mitarbeiterin
Ärztliches Zentrum für Qualität in der Medizin
[redacted]

A 2.2 Seidel, Gabriele

Von: [redacted]
Gesendet: Dienstag, 5. April 2011 12:06
An: methoden@iqwig.de
Betreff: Anmerkungen Methoden

Sehr geehrtes IQWiG-Team,

mit Interesse haben wir Ihren "Entwurf der Allgemeinen Methoden 4.0" gelesen, insbesondere den
Abschnitt 5 "Evidenbasierte Gesundheitsinformation für Bürger und Patienten".
Wir freuen uns, dass Sie bei 5.5.3 "Evaluation" die Magisterarbeit von Frau Hirschberg (236) zitiert haben,
die im Rahmen der Nutzertestung durch die Patientenuniversität an der MHH entstanden ist. Zu der
Literaturangabe haben wir zwei Bitten bzw. Anregungen:
1. Die Literaturangabe -Masterarbeit von Frau Hirschberg" würden wir ersetzen mit der Veröffentlichung
der Masterarbeit im Grinverlag (ISBN): Da die Herausgeberinnen der Reihe, Marie-Luise Dierks und
Gabriele Seidel, zugleich Projektleiterinnen der Nutzertestung sind, wäre es schön, wenn Sie
die Veröfffentlichung aufnehmen könnten:
"Hirschberg I. (2010): Bewertung und Wirkung von evidenzbasierten Gesundheitsinformationen - die
Perspektive der Nutzer. In: Schriftenreihe Patientenorientierung und Gesundheitskompetenz; Band 1,
Dierks ML, Seidel G. (Ed). München: GRIN-Verlag; 2010."
2. Da die Magisterarbeit nur einen Teil der Untersuchung abdeckt, und unser offizieller Abschlussbericht
seit einiger Zeit dem Ressort vorliegt, wollten wir vorschlagen, den Bericht zusätzlich zu zitieren (so wie
bei Schmacke et al., Literaturangabe 433).
"Seidel G, Hirschberg I, Kreusel I, Dierks ML. Nutzertestung von Gesundheitsinformationen des Instituts
für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG).
Abschlussbericht Oktober 2010.
Patientenuniversität an der Medizinischen Hochschule Hannover, Institut für Epidemiologie, Sozialmedizin
und Gesundheitssystemforschung."
Für die Berücksichtigung wären wir Ihnen dankbar und stehen Ihnen für Rückfragen gerne zur Verfügung.

Mit freundlichen Grüßen
G. Seidel

Dr. Gabriele Seidel
Patientenuniversität an der
Medizinischen Hochschule Hannover
Abtl. Epidemiologie, Sozialmedizin und
Gesundheitssystemforschung
[redacted]

A 2.3 Vach, Werner

Stellungnahme zu ausgewählten Punkten des Entwurfs für Version 4.0 der “Allgemeinen Methoden” des IQWiG

Meine Stellungnahme gliedert sich in fünf Punkte. Die ersten vier beziehen sich auf die Evaluation diagnostischer Methoden, der letzte auf die Zusammenfassung von Nichtunterlegenheitstudien.

Direkter vs. indirekter Zusatznutzen

Die Bewertung des Nutzen diagnostischer Verfahren geschieht heutzutage in weiten Bereichen durch die Bewertung des Zusatznutzens im Vergleich zu etablierten diagnostischen Methoden. Dabei kann es sich um einen indirekten Nutzen handeln, wenn es in erster Linie darum geht, die diagnostische Genauigkeit zu verbessern, so dass der Patient von einer angemesseneren Behandlung profitieren kann, oder die Dauer bis zur diagnostischen Entscheidung zu verkürzen und damit schneller eine angemessene Behandlung zu erreichen. Aber es kann auch um einen direkten Nutzen gehen, wenn die bisherige diagnostische Methode z.B. mit Nebenwirkungen verbunden ist. Viele neue bildgebenden Verfahren oder molekulare Marker bieten die Möglichkeit, bisherige invasive diagnostische Methoden durch nicht-invasive zu ersetzen, und haben somit einen direkten Nutzen.

Viele der Aussagen in Abschnitt 3.5 (bzw. auch in 3.1.1) machen Sinn, wenn es um den indirekten Nutzen geht, sind aber teilweise irreführend, wenn es um den direkten Nutzen geht. So heißt es z.B. auf Seite 51: “Unter patientenrelevanten Endpunkten sind in diesem Zusammenhang dieselben Nutzenkategorien zu verstehen wie bei der Bewertung therapeutischer Interventionen, nämlich Mortalität, Morbidität und gesundheitsbezogene Lebensqualität. Deren Beeinflussung durch diagnostische Maßnahmen lässt sich dabei durch die Vermeidung risikobehafteter(er) bzw. komplikationsträchtiger(er) Interventionen oder durch den gezielt(er)en Einsatz von Interventionen erzielen.” Beim direkten Nutzen ist die diagnostische Maßnahme aber schon selber die Intervention, die vermieden wird.

Ich habe daher den Eindruck, dass die “Allgemeinen Methoden” von einer konsequenteren Trennung dieser beiden Situationen profitieren würden.

Bewertung diagnostischer Verfahren vs. Bewertung diagnostischer Kette

Das IQWiG weist auf Seite 50 ganz richtig auf den hohen Stellenwert randomisierter Studien hin. Allerdings geht das IWQiG in diesem Zusammenhang nicht auf das dahinterliegende, grundlegende Problem ein, nämlich das im Falle des indirekten Nutzens eines diagnostischen Verfahrens es eben nicht möglich ist, das Verfahren an sich zu bewerten, sondern nur die diagnostische Kette, d.h. das diagnostische Verfahren im Zusammenhang mit bestimmten, nachfolgenden Maßnahmen. Das zitierte

Beispiel von der FDA ist ja gerade von der Art, dass nicht eine einzelne diagnostische Maßnahme, sondern eben der Verbund aus Diagnose und Therapie bewertet werden soll.

Auch hier würde es der Verständlichkeit und Interpretierbarkeit des Abschnittes 3.5. meines Erachtens gut tun, wenn deutlicher gemacht wird, dass in vielen Fällen die Bewertung einer bestimmten diagnostischen Kette und nicht eines Verfahrens selber angestrebt wird. Es ist hier auch sehr unglücklich, dass das IQWiG auf Seite 51 im dritten Abschnitt die Bewertung der diagnostischen Kette als Gegensatz zu randomisierten Studien einführt. Ich denke, es wäre sachlich richtiger zwischen Ansätzen, die die direkte Bewertung der Kette als Gesamtheit oder die indirekte durch eine separate Bewertung der Einzelteile anstreben, zu unterscheiden.

Ich möchte mir hier auch den Hinweis erlauben, dass die zitierte randomisierte Studie von van Tinteren et al. nicht unbedingt ein geeignetes Beispiel ist, da es fragwürdig ist, ob sie einen patientenorientierten Outcome im Sinne des IQWiG (also Mortalität, Morbidität und gesundsheitsbezogene Lebensqualität) benutzt. Der Hauptendpunkt dieser Studie ist die “futile thoracotomy“, also eine als nutzlos eingeschätzte Managementmassnahme, was man wohl höchstens wohlwollend als einen (nicht validierten) Surrogatendpunkt (für was?) ansehen kann.

Darüberhinaus halte ich den Hinweis, dass randomiserte Studien zu diagnostischen Ketten mit moderatem Aufwand durchführbar sind, für irreführend. Wenn man sich an die Kriterien des IQWiG hält, erfordern derartige Studien in der Regel hohe Fallzahlen, da ja schon die bisherigen diagnostischen Methoden eine gewisse Güte haben, und es daher in der Natur der Sache liegt, dass es für maximal 30% aller Patienten zu einer verbesserten Entscheidung kommen kann, die dann wiederum nur in einem Teil durch eine verbesserte Behandlungsentscheidung langfristig zu einem besseren Outcome führt. Insofern spricht viel dafür, dass randomisierte Studien, die den sonstigen Anforderungen des IQWiGs genügen, eher in seltenen Fällen mit moderatem Aufwand durchführbar sind.

Nutzen von randomisierten Therapiestudien zur Nutzenbewertung diagnostischer Maßnahmen

Auf Seite 51 schreibt das IQWiG “... inwieweit belegt ist, dass die aus den Testergebnissen resultierenden Konsequenzen mit einem Nutzen verbunden sind. Für den (zumeist anzunehmenden) Fall therapeutischer Konsequenzen lassen sich solche Belege aus randomisierten Interventionsstudien (mit patientenrelevanten Endpunkten) ableiten, in denen ein bestimmtes (Test-)Ergebnis des zu prüfenden diagnostischen Verfahrens als Einschlusskriterium definiert wurde.” Diese Aussage erscheint mir unlogisch. Eine solche Studie kann nicht den Nutzen des Tests zeigen, es sei denn dass es naheliegend ist, dass der beobachtete Unterschied in Falle des gegenteiligen Testergebnisses nicht gegeben ist.

Vergleich von Testcharakteristika

Auf Seite 52 führt das IQWiG aus, dass in einigen Fällen ein direkter Vergleich der Testcharakteristika ausreichend ist. Dabei werden Studien “mit zufälliger Zuordnung der Reihenfolge der (voneinander unabhängigen und möglichst verblindeten) Testdurchführung bei denselben Patienten” und “mit zufälliger Zuordnung der Tests auf verschiedene Patienten” als gleichwertig gegeneinander gestellt. Dies ist meines Erachtens nicht richtig. Studien mit Testdurchführung bei denselben Patienten sind hier höherwertig, da sie tatsächlich den Schluss erlauben, dass zwei diagnostische Methoden zu den gleichen Ergebnissen kommen. Studien mit zufälliger Zuordnung der Tests auf verschiedene Patienten erlauben aber nur den Schluss, dass Testcharakteristika wie Sensitiviät und Spezifizität gleich sind. Dies kann aber dazu führen, dass zwei diagnostische Methoden als gleichwertig angesehen werden, obwohl der indirekte Nutzen verschieden ist. So kann z.B. ein neues diagnostische Verfahren besser Patienten in einem späten Stadium als das bisherige, aber schlechter in einem frühen Stadium erkennen (oder generell mit ungünstiger bzw. günstiger Prognose).

Systematische Zusammenfassung von Nichtunterlegenheitsstudien

Das IQWiG erwähnt zwar an verschiedenen Stellen Nichtunterlegenheitsstudien, geht aber nicht auf die besondere Problematik der Zusammenfassung solcher Studien ein.

Die besondere Problematik ergibt sich daraus, dass Nichtunterlegenheitstudien an bestimmte Vorbedingungen geknüpft sind, die direkt den Zusatznutzen der neuen Behandlung betreffen. Nichtunterlegenheitsstudien setzen voraus, dass die neue Behandlung einen impliziten Zusatznutzen besitzt. Dieser implizite Zusatznutzen macht es erst sinnvoll, nicht die Überlegenheit, sondern die Nichtunterlegenheit hinsichtlich des Hauptnutzens als Anforderung zu stellen. Insofern steckt hinter Nichtunterlegensheitstudien eigentlich immer die Frage der Abwägung zweier verschiedener Nutzen, häufig in der Form von geringerer Belastung, weniger Nebenwirkungen oder geringeren Kosten vs. einer möglicherweise leicht verringerten Wirksamkeit. Daher sind eigentlich Methoden, wie sie vom IQWiG auch zur Kosten-Nutzen-Abwägung vorgeschlagen werden, angebracht.

Auf dem Niveau der einzelnen Studie ist diese Abwägung jedoch häufig nicht möglich, da der implizite Nutzen mit den üblichen Fallzahlen nicht aufzeigbar ist (z.B. bei seltenen Nebenwirkungen). Diese Situation ändert sich jedoch in einer systematischen Zusammenfassung, weil eben durch die höheren Fallzahlen es möglich wird, den impliziten Nutzen zu quantifizieren und damit gegen einen möglichen Wirksamkeitsverlust abzuwägen. In vielen Nichtunterlegenheitsstudien wird auch auf eine Quantifizierung des impliziten Nutzens verzichtet, obwohl diese möglich wäre. Auch hier könnte in einer systematischen Zusammenfassung eine Lücke geschlossen werden, in dem z.B. Daten zu Kosten aus anderen Quellen einbezogen werden.

Zusammenfassend halte ich es für wünschenswert, wenn in den „Allgemeinen Methoden“ des IQWiGs eine klare Aussage erfolgt, dass es bei der Zusammenfassung von Nichtunterlegenheitsstudien nicht

das Ziel ist, die Studien durch eine metaanalytische Zusammenfassung als eine große Nichtunterlegenheitsstudie zu analyisieren, sondern eine Abwägung zwischen dem (impliziten) Zusatznutzen und einem möglichen Wirksamkeitsverlust vorzunehmen.

Freiburg, den 8.4.2011

Werner Vach
Professor für klinische Epidemiologie
Institut für Medizinische Biometrie und Medizinische Informatik
Universitätsklinikum Freiburg