Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Institute for Quality and Efficiency in Health Care

Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1

Entwurf vom 18.04.2013

Kontakt:

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Im Mediapark 8
50670 Köln

Tel: +49 (0)221 – 35685-0
Fax: +49 (0)221 – 35685-1
E-Mail: methoden@iqwig.de

Informationen zu diesem Dokument

Das Institut hat seit 2011 umfangreiche neue Aufgaben erhalten. Daran müssen auch die Allgemeinen Methoden angepasst werden, d. h. einige Abschnitte werden kurzfristig überarbeitet oder ergänzt. Statt wie bisher das Methodenpapier jeweils komplett zu aktualisieren, geht das Institut zu einer kontinuierlichen und schrittweisen Überarbeitung über.

Im ersten Teilschritt stellt das Institut Entwürfe einiger aktualisierter Abschnitte der Allgemeinen Methoden Version 4.0. vom 23.09.2011 und Ergänzungen öffentlich zur Diskussion. Im Anschluss an das Stellungnahmeverfahren werden die betreffenden Abschnitte ggf. überarbeitet und in die Allgemeinen Methoden integriert. Hieraus gehen dann die Allgemeinen Methoden Version 4.1 hervor.

Die vorliegenden Entwürfe für Aktualisierungen und Ergänzungen des Methodenpapiers betreffen folgende Abschnitte:

- 2.1.1 Bericht
- 2.2.3 Review der Produkte des Instituts
- Neuer Abschnitt 3.1.4 Endpunktbezogene Bewertung
- 3.1.5 Zusammenfassende Bewertung (vorher Abschnitt 3.1.4)
- 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V
- 7.3.8 Meta-Analysen
- Neuer Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens

Die vorgesehenen Änderungen betreffen im Wesentlichen die folgenden Aspekte:

- Abschnitte 2.1.1 und 2.2.3:
  Darstellung des externen Reviews für Vorberichte als optionalen Schritt
- Abschnitte 3.1.4 und 3.1.5:
  Konkretisierung der Anforderungen an die Beleglage zur Formulierung von Nutzenaussagen mit unterschiedlichen Aussagesicherheiten
- Abschnitt 3.3.3 und Anhang:
  Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens sowie dessen Rationale
- Abschnitt 7.3.8:
  Verwendung von Prädiktionsintervallen für Meta-Analysen mit zufälligen Effekten

2.1.1 Bericht

A) Ablauf der Berichterstellung

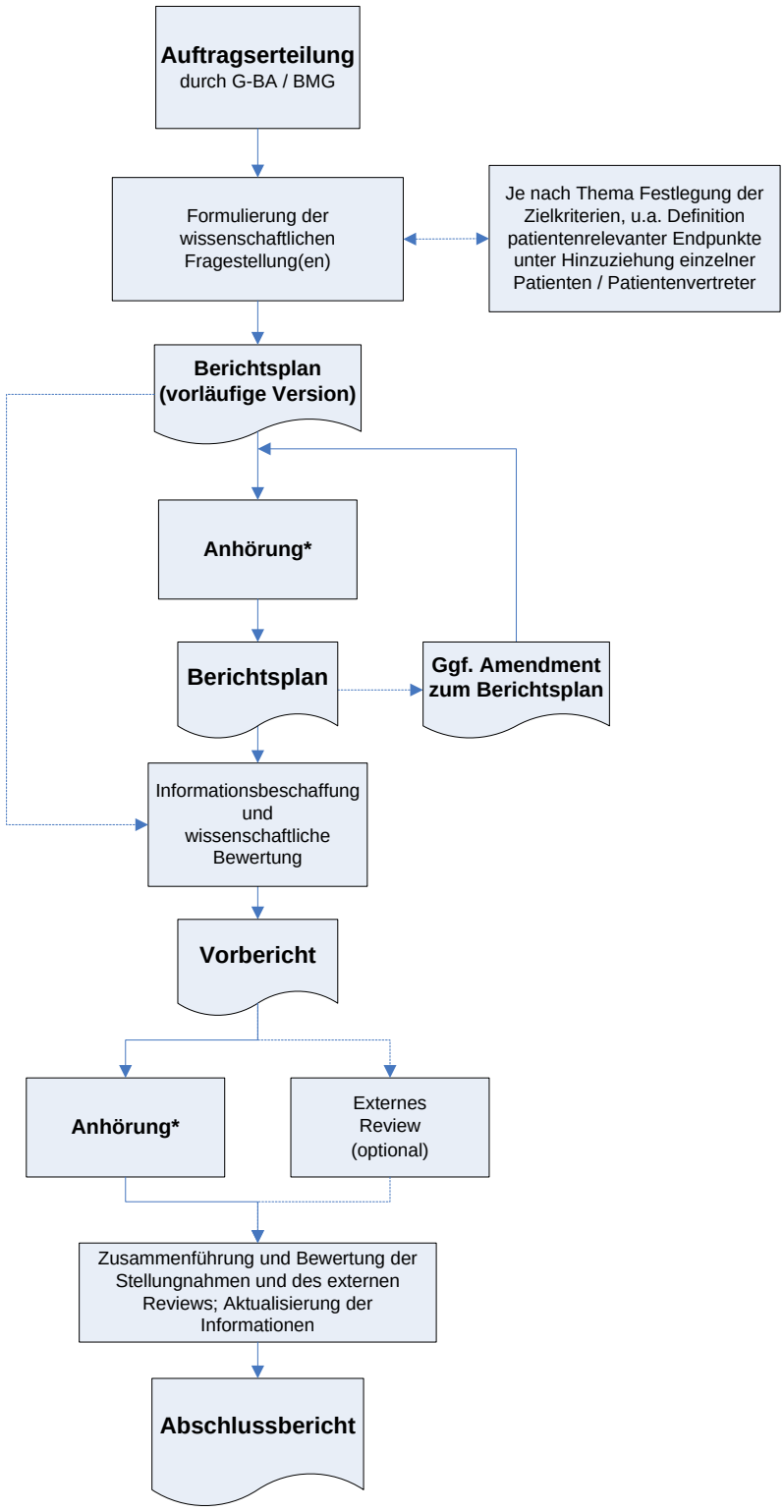
Der Ablauf der Berichterstellung ist in Abbildung 1 schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt. Dabei werden regelhaft externe Sachverständige beteiligt (siehe Abschnitt 2.2.1). Bei Bedarf wird auch der wissenschaftliche Beirat des Instituts einbezogen. Der interne Qualitätssicherungsprozess ist in diesem Flussdiagramm nicht dargestellt.

Nach der **Auftragserteilung** durch den G-BA oder das BMG erfolgt die Formulierung der wissenschaftlichen Fragestellung. Je nach Thema ist dabei auch die Festlegung von Zielkriterien erforderlich, z. B. bei Nutzenbewertungen. Regelhaft werden dabei, insbesondere zur Definition patientenrelevanter Endpunkte, entsprechende Patienten-organisationen beteiligt, gegebenenfalls wird auch die Meinung einzelner Betroffener eingeholt. Anschließend wird der Berichtsplan erstellt.

Der **Berichtsplan** beinhaltet die genaue wissenschaftliche Fragestellung einschließlich der Zielkriterien (z. B. patientenrelevante Endpunkte), die Ein- und Ausschlusskriterien der für die Bewertung zu verwendenden Informationen sowie die Darlegung der projektspezifischen Methodik der Beschaffung und Bewertung dieser Informationen. Die vorläufige Version des Berichtsplans wird zunächst dem Auftraggeber, dem Vorstand der Stiftung, dem Stiftungsrat und dem Kuratorium zugeleitet. Die Veröffentlichung auf der Website des Instituts erfolgt i. d. R. fünf Arbeitstage später.

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.

2.1.1 Bericht



* Die Anhörung erfolgt mittels Einholung schriftlicher Stellungnahmen. Optional wird eine mündliche wissenschaftliche Erörterung zur Diskussion unklarer Aspekte in den schriftlichen Stellungnahmen durchgeführt.

Abbildung 1: Ablauf der Berichtserstellung

2.1.1 Bericht

Nach Auswertung der Stellungnahmen und ggf. Durchführung der Erörterung wird der Berichtsplan in einer überarbeiteten Version zusammen mit der Dokumentation der Anhörung zum Berichtsplan zunächst dem Auftraggeber, dem Vorstand der Stiftung, dem Stiftungsrat und dem Kuratorium zugeleitet und i. d. R. fünf Arbeitstage später auf der Website des Instituts veröffentlicht. Der überarbeitete Berichtsplan ist Grundlage für die Erstellung des Vorberichts. Sofern weitere wesentliche methodische Änderungen im Verlauf der Vorberichtserstellung erforderlich sind, erfolgen diese i. d. R. in Form eines oder mehrerer Amendments zum Berichtsplan. Nach Veröffentlichung eines Amendments besteht i. d. R. ebenfalls Gelegenheit zur Stellungnahme zu diesem Amendment unter den o. g. Bedingungen.

Im **Vorbericht** werden die Ergebnisse der Informationsbeschaffung und der wissenschaftlichen Bewertung dargestellt. Um die Arbeit des Instituts nicht unangemessen zu verzögern, beginnt die Informationsbeschaffung und -bewertung bereits vor Abschluss der Anhörung zum Berichtsplan auf Basis der im vorläufigen Berichtsplan formulierten Kriterien. Das Ergebnis der Anhörung wird damit jedoch explizit nicht vorweggenommen, da diese Kriterien sich durch die Anhörung zum Berichtsplan in seiner vorläufigen Version ändern können. Dies kann auch zu einer Ergänzung und / oder Änderung der Informationsbeschaffung und -bewertung führen.

Der Vorbericht enthält die vorläufige Empfehlung an den G-BA. Er wird nach Fertigstellung zunächst dem Auftraggeber, dem Vorstand der Stiftung, dem Stiftungsrat und dem Kuratorium zugeleitet. Die Veröffentlichung auf der Website des Instituts erfolgt i. d. R. fünf Arbeitstage nach Versendung an den Auftraggeber.

Für eine Frist von mindestens vier Wochen wird dann der Öffentlichkeit **Gelegenheit zur Stellungnahme (Anhörung)** gegeben (schriftliche Stellungnahmen). Gegenstand des Stellungnahmeverfahrens sind insbesondere die im Vorbericht dargestellten Ergebnisse der Informationsbeschaffung und -bewertung. 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 Abschlussberichts.

Der **Abschlussbericht**, der auf dem Vorbericht aufbauend die Bewertung der wissenschaftlichen Erkenntnisse unter Berücksichtigung der Ergebnisse der Anhörung zum Vorbericht enthält, stellt das abschließende Produkt der Berichtserstellung dar. Der Abschlussbericht und die Dokumentation der Anhörung zum Vorbericht werden zunächst dem Auftraggeber, dem Vorstand der Stiftung sowie dem Stiftungsrat und anschließend (i. d. R. vier Wochen später) dem Kuratorium der Stiftung zugeleitet. I. d. R. weitere vier Wochen später erfolgt die Veröffentlichung dieser Dokumente auf der Website des Instituts. Sollten Stellungnahmen zu Abschlussberichten 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 Berichts) aus Sicht des Instituts erforderlich erscheint.

Der Auftraggeber entscheidet über die Beauftragung des Instituts. Eine solche Aktualisierung unterliegt den allgemeinen methodischen und verfahrenstechnischen Anforderungen an Produkte des Instituts.

B) Allgemeine Anmerkungen zum Stellungnahmeverfahren (Anhörung)

Stellungnahmeberechtigte Organisationen

Das Institut hat nach § 139a Abs. 5 SGB V zu gewährleisten, dass in allen wichtigen Abschnitten des Bewertungsverfahrens den Sachverständigen der medizinischen, pharmazeutischen und gesundheitsökonomischen Wissenschaft und Praxis, den Arzneimittelherstellern, den für die Wahrnehmung der Interessen der Patientinnen und Patienten und die Selbsthilfe chronisch kranker und behinderter Menschen maßgeblichen Organisationen sowie dem oder der Beauftragten der Bundesregierung für die Belange der Patientinnen und Patienten Gelegenheit zur Stellungnahme gegeben wird. Die Stellungnahmen sind in die Entscheidung einzubeziehen. Diesen Anforderungen wird dadurch Rechnung getragen, dass Anhörungen zum Berichtsplan und zum Vorbericht durchgeführt werden und der Personenkreis der Stellungnahmeberechtigten nicht begrenzt wird. Darüber hinaus werden alle Institutsprodukte nach § 139a vor der Publikation dem Kuratorium des Instituts zugeleitet. Im Kuratorium sind Patientenorganisationen, der oder die Patientenbeauftragte der Bundesregierung, Organisationen der Leistungserbringer und der Sozialpartner und die Selbstverwaltungsorgane des Gemeinsamen Bundesausschusses vertreten.

Formale Anforderungen

Um die Arbeit des Instituts nicht unangemessen zu verzögern, müssen die Stellungnahmen bestimmten formalen Anforderungen genügen. Weiterführende Angaben zum Stellungnahmeverfahren einschließlich der Bedingungen für die Teilnahme an einer wissenschaftlichen Erörterung finden sich in einem Leitfaden, der auf der Website des Instituts abgerufen werden kann.

Veröffentlichung der Stellungnahmen

Stellungnahmen, die den formalen Anforderungen genügen, werden auf der Website des Instituts in einem gesonderten Dokument (Dokumentation und Würdigung der Anhörung) veröffentlicht. Zur Gewährleistung der Transparenz werden mit den Stellungnahmen eingereichte Unterlagen, die nicht öffentlich zugänglich sind (z. B. Manuskripte), ebenfalls veröffentlicht.

Vorlage von Unterlagen im Rahmen der Anhörung

Im Rahmen der Anhörung sowohl zum Berichtsplan als auch zum Vorbericht besteht die Möglichkeit, qualitativ angemessene Unterlagen jeglicher Art, die aus Sicht des jeweiligen Stellungnehmenden für die Beantwortung der Fragestellung des Berichts geeignet sind, vorzulegen. Falls die in dem Berichtsplan definierte Suchstrategie z. B. auf randomisierte kontrollierte Studien beschränkt ist, können im Rahmen des Stellungnahmeverfahrens

trotzdem nicht randomisierte Studien eingereicht werden. In solchen Fällen ist aber zusätzlich eine adäquate Begründung für die Validität der kausalen Interpretation der in solchen Studien beschriebenen Effekte erforderlich.

2.2.3 Review der Produkte des Instituts

Das Review der Produkte des Instituts hat insbesondere zum Ziel, eine hohe wissenschaftliche Qualität der Produkte zu gewährleisten. Darüber hinaus können für einzelne Produkte auch andere Ziele wesentlich sein, z. B. die Allgemeinverständlichkeit.

Alle Produkte einschließlich der jeweiligen Zwischenprodukte unterliegen einem umfangreichen mehrstufigen internen Qualitätssicherungsverfahren. Darüber hinaus kann im Verlauf der Erstellung von Berichten und z. T. auch von Gesundheitsinformationen (siehe 2.1.5) ein externes Reviewverfahren als optionaler weiterer Schritt der Qualitätssicherung durchgeführt werden. Die Auswahl der internen und externen Reviewer erfolgt primär auf Basis ihrer methodischen und/oder fachlichen Expertise.

Die Identifikation externer Reviewer kann 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. Auch für die externen Reviewer ist die Darlegung potenzieller Interessenkonflikte erforderlich.

Die Auswahl der externen Reviewer erfolgt durch das Institut. Eine Höchstgrenze von Reviewern gibt es nicht. Die externen Reviews werden hinsichtlich ihrer Relevanz für das jeweilige Produkt geprüft. Eine Veröffentlichung der externen Reviews erfolgt nicht. Die Namen der externen Reviewer 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.

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

3.1.4 Endpunktbezogene Bewertung

Die Nutzenbewertung und die Einschätzung der Stärke der Ergebnis(un)sicherheit orientieren sich an internationalen Standards der evidenzbasierten Medizin, wie sie z. B. von der GRADE-Gruppe erarbeitet werden [25].

Medizinische Interventionen werden im Vergleich zu einer anderen Intervention oder Scheinintervention (z. B. Placebo) oder keiner Intervention bezüglich ihrer Auswirkungen auf definierte patientenrelevante Endpunkte in ihrem (Zusatz-)Nutzen und Schaden zusammenfassend beschrieben. Dafür wird zunächst für jeden vorher definierten patientenrelevanten Endpunkt einzeln aufgrund der Analyse vorhandener wissenschaftlicher Daten eine Aussage zur Belegbarkeit des (Zusatz-)Nutzens und Schadens in 4 Abstufungen getroffen: Es liegt entweder ein Beleg, ein Hinweis, ein Anhaltspunkt oder keine dieser drei Situationen vor. Der letzte Fall tritt ein, wenn keine Daten vorliegen oder die vorliegenden Daten keine der drei übrigen Aussagen zulassen.

Je nach Fragestellung beziehen sich die Aussagen auf das Vorhandensein oder das Fehlen eines (Zusatz-)Nutzens und Schadens. Die Voraussetzung für Aussagen über das Fehlen eines (Zusatz-)Nutzens bzw. Schadens sind gut begründete Definitionen von Irrelevanzbereichen (siehe Abschnitt 7.3.6).

Ein wichtiges Kriterium zur Ableitung von Aussagen zur Beleglage ist die Ergebnissicherheit. Grundsätzlich ist jedes Ergebnis einer empirischen Studie oder einer systematischen Übersicht über empirische Studien mit Unsicherheit behaftet und daher auf seine Ergebnissicherheit zu prüfen. Hierbei ist zu unterscheiden zwischen qualitativer und quantitativer Ergebnissicherheit. Die qualitative Ergebnissicherheit wird bestimmt durch das Studiendesign, aus dem sich Evidenzgrade ableiten lassen (siehe Abschnitt 7.1.3), sowie durch (endpunktbezogene) Maßnahmen zur weiteren Vermeidung oder Minimierung möglicher Verzerrungen (z. B. verblindete Zielgrößenerhebung, Auswertung auf Basis aller eingeschlossenen Patientinnen und Patienten, ggf. mithilfe des Einsatzes adäquater Ersetzungsmethoden für fehlende Werte, ggf. Einsatz adäquater, valider Messinstrumente), die in Abhängigkeit vom Studiendesign bewertet werden müssen (siehe Abschnitt 7.1.4). Neben der qualitativen Ergebnissicherheit gibt es quantitativ messbare Unsicherheiten aufgrund statistischer Gesetzmäßigkeiten, die wiederum in unmittelbarem Zusammenhang mit dem Stichprobenumfang, d. h. der Anzahl der in einer Studie untersuchten Patientinnen und Patienten bzw. der Anzahl der in einer systematischen Übersicht enthaltenen (Primär-) Studien, sowie mit der in bzw. zwischen den Studien beobachteten Variabilität stehen. Falls die zugrunde liegenden Daten dies zulassen, lässt sich die statistische Unsicherheit als Standardfehler bzw. Konfidenzintervall von Parameterschätzungen quantifizieren und beurteilen (Präzision der Schätzung).

3.1.4 Endpunktbezogene Bewertung

Das Institut verwendet die folgenden drei Kategorien zur Graduierung des Ausmaßes der qualitativen Ergebnissicherheit auf Einzelstudien- und Endpunktebene:

- **hohe qualitative Ergebnissicherheit**: Ergebnis einer randomisierten Studie mit niedrigem Verzerrungspotenzial.
- **mäßige qualitative Ergebnissicherheit**: Ergebnis einer randomisierten Studie mit hohem Verzerrungspotenzial.
- **geringe qualitative Ergebnissicherheit**: Ergebnis einer nicht randomisiert vergleichenden Studie.

Bei der Ableitung der Beleglage für einen Endpunkt sind die Anzahl der vorhandenen Studien, deren qualitative Ergebnissicherheiten sowie die in den Studien gefundenen Effekte von zentraler Bedeutung. Liegen mindestens 2 Studien vor, wird zunächst unterschieden, ob sich aufgrund der vorhandenen Heterogenität im Rahmen einer Meta-Analyse (siehe Abschnitt 7.3.8) sinnvoll ein gemeinsamer Effektschätzer bilden lässt oder nicht. Im Fall homogener Ergebnisse, die sich sinnvoll poolen lassen, muss der gemeinsame Effektschätzer statistisch signifikant sein. Sind die geschätzten Effekte zu heterogen, um sinnvoll einen gepoolten gemeinsamen Effektschätzer zu bilden, wird unterschieden zwischen "nicht gleichgerichteten", "mäßig gleichgerichteten" und "deutlich gleichgerichteten" Effekten, die wie folgt definiert sind.

Falls das Prädiktionsintervall zur Darstellung der Heterogenität in einer Meta-Analyse mit zufälligen Effekten (siehe Abschnitt 7.3.8) dargestellt wird und den Nulleffekt nicht überdeckt, liegen gleichgerichtete Effekte vor. Anderenfalls (keine Darstellung des Prädiktionsintervalls oder dieses überdeckt den Nulleffekt) liegen gleichgerichtete Effekte in folgender Situation vor:

Die Effektschätzer von zwei oder mehr Studien zeigen in eine Richtung. Für diese „gerichteten“ Studien gelten alle folgenden Bedingungen:

- Das Gesamtgewicht dieser Studien ist $\geq 80\%$.
- Mindestens 2 dieser Studien zeigen statistisch signifikante Ergebnisse.
- Mindestens 50 % des Gewichts dieser Studien basiert auf statistisch signifikanten Ergebnissen.

Die Gewichte der Studien kommen hierbei in der Regel aus einer Meta-Analyse mit zufälligen Effekten (siehe Abschnitt 7.3.8). Falls keine Meta-Analyse sinnvoll ist, entspricht die relative Fallzahl dem Gewicht.

Wann gleichgerichtete Effekte mäßig oder deutlich gleichgerichtet sind, wird wenn möglich anhand der Lage des Prädiktionsintervalls entschieden. Da das Prädiktionsintervall in der Regel jedoch nur dargestellt wird, falls mindestens 4 Studien vorliegen (siehe Abschnitt

7.3.8), hängt die Einstufung in mäßige gleichgerichtete und deutlich gleichgerichtete Effekte von der Anzahl der Studien ab.

- 2 Studien: Gleichgerichtete Effekte sind immer deutlich gleichgerichtet.
- 3 Studien:
  - Alle drei Studien weisen statistisch signifikante Ergebnisse auf: Die gleichgerichteten Effekte sind deutlich gleichgerichtet.
  - Nicht alle drei Studien weisen statistisch signifikante Ergebnisse auf: Die gleichgerichteten Effekte sind mäßig gleichgerichtet.
- 4 oder mehr Studien:
  - Das Prädiktionsintervall überdeckt nicht den Nulleffekt: Die gleichgerichteten Effekte sind deutlich gleichgerichtet.
  - Das Prädiktionsintervall überdeckt den Nulleffekt: Die gleichgerichteten Effekte sind mäßig gleichgerichtet.

Für den Fall, dass die vorhandenen Studien dieselbe qualitative Ergebnissicherheit aufweisen oder nur eine Studie vorliegt, lassen sich mit diesen Definitionen die regelhaften Anforderungen an die Beleglage zur Ableitung von Aussagen mit unterschiedlichen Aussagesicherheiten definieren. Das Institut unterscheidet hierbei die 3 verschiedenen Aussagesicherheiten „Beleg“, „Hinweis“ und „Anhaltspunkt“.

In der Regel wird an die Aussage eines Belegs die Anforderung zu stellen sein, dass eine Meta-Analyse von Studien mit hoher qualitativer Ergebnissicherheit einen entsprechenden statistisch signifikanten Effekt zeigt. Falls eine Meta-Analyse nicht durchführbar ist, sollten mindestens 2 voneinander unabhängig durchgeführte Studien mit hoher qualitativer Ergebnissicherheit und einem statistisch signifikantem Effekt vorliegen, deren Ergebnis nicht durch weitere vergleichbare ergebnissichere Studien infrage gestellt wird (Konsistenz der Ergebnisse). Bei den 2 voneinander unabhängig durchgeführten Studien muss es sich nicht um solche mit exakt identischem Design handeln. Welche Abweichungen im Design zwischen Studien noch akzeptabel sind, hängt von der Fragestellung ab. Eine Meta-Analyse von Studien mit mäßiger qualitativer Ergebnissicherheit oder eine einzelne Studie mit hoher qualitativer Ergebnissicherheit kann trotz statistisch signifikantem Effekt demnach in der Regel nur Hinweis liefern. Soll aus lediglich einer Studie im Ausnahmefall ein Nutzenbeleg abgeleitet werden, so sind an eine solche Studie und ihre Ergebnisse besondere Anforderungen zu stellen [157]. Eine Meta-Analyse von Studien mit geringer qualitativer Ergebnissicherheit oder eine einzelne Studie mit mäßiger qualitativer Ergebnissicherheit liefert bei statistisch signifikantem Effekt in der Regel nur einen Anhaltspunkt. Die regelhafte Operationalisierung ist übersichtlich in Tabelle 2 zu finden. In begründeten Fällen beeinflussen weitere Faktoren diese Einschätzungen. Die Betrachtung von Surrogatendpunkten (siehe Abschnitt 3.1.2), das Vorliegen schwerwiegender Designmängel bei einer Studie, oder auch begründete Zweifel an der Übertragbarkeit auf die Behandlungssituation in

Deutschland können z. B. zu einer Verringerung der Aussagesicherheit führen. Auf der anderen Seite können z. B. große Effekte oder eine eindeutige Richtung eines vorhandenen Verzerrungspotenzials eine Erhöhung der Sicherheit begründen.

Tabelle 2: Anforderungen an die Beleglage für die unterschiedlichen Aussagesicherheiten beim Vorliegen von Studien derselben qualitativen Ergebnissicherheit

	Anforderung		
Aussage	**Anzahl der Studien**	**qualitative Ergebnissicherheit**	**Effekt(e)[a]**
Beleg	≥ 2	hoch	homogen, Meta-Analyse statistisch signifikant
	≥ 2	hoch	heterogen deutlich gleichgerichtet
Hinweis	≥ 2	mäßig	homogen, Meta-Analyse statistisch signifikant
	≥ 2	mäßig	heterogen deutlich gleichgerichtet
	≥ 2	hoch	heterogen mäßig gleichgerichtet
	1	hoch	statistisch signifikant
Anhalts-punkt	≥ 2	gering	homogen, Meta-Analyse statistisch signifikant
	≥ 2	gering	heterogen deutlich gleichgerichtet
	≥ 2	mäßig	heterogen mäßig gleichgerichtet
	1	mäßig	statistisch signifikant
a: Zur Erläuterung des Begriffs: siehe Text.			

Liegen mehrere Studien mit unterschiedlicher qualitativer Ergebnissicherheit vor, so werden zunächst nur die Studien mit der höherwertigen Ergebnissicherheit betrachtet und auf dieser Grundlage Aussagen zur Beleglage gemäß Tabelle 2 abgeleitet. Bei der Ableitung von Aussagen zur Beleglage für den gesamten Studienpool gelten dann folgende Grundsätze.

- Die Aussagen zur Beleglage bei Beschränkung auf die höherwertigeren Studien werden durch Hinzunahme der übrigen Studien nicht abgeschwächt, sondern allenfalls aufgewertet.

- Die für einen Beleg notwendige Bestätigung (Replikation) eines statistisch signifikanten Ergebnisses einer Studie hoher qualitativer Ergebnissicherheit kann durch eine oder mehrere Ergebnisse mäßiger (jedoch nicht geringer) qualitativer Ergebnissicherheit erbracht werden. Dabei sollte das Gewicht der Studie hoher qualitativer Ergebnissicherheit eine angemessene Größe haben (zwischen 25 und 75%).
- Ist das meta-analytische Ergebnis für die höherwertigeren Studien nicht statistisch signifikant bzw. liegen für diese Studien keine gleichgerichteten Effekte vor, sind die Aussagen zur Beleglage auf der Grundlage der Ergebnisse des gesamten Studienpools abzuleiten, wobei die Aussagesicherheit durch die minimale qualitative Ergebnissicherheit aller einbezogenen Studien bestimmt wird.

Nach diesen Definitionen und Grundsätzen wird für jeden Endpunkt einzeln eine entsprechende Nutzenaussage abgeleitet. Überlegungen zur endpunktübergreifenden Bewertung finden sich im nachfolgenden Abschnitt (siehe Abschnitt 3.1.5).

3.1.5 Zusammenfassende Bewertung

Die im Rahmen der Ableitung von Aussagen zur Beleglage für jeden patientenrelevanten Endpunkt einzeln getroffenen Aussagen werden anschließend – soweit möglich – in einem bewertenden Fazit in Form einer Nutzen-Schaden-Abwägung zusammengefasst. Beim Vorhandensein von Belegen eines (Zusatz-)Nutzens und/oder eines Schadens bzgl. der Zielgrößen 1 bis 3 aus Abschnitt 3.1.1 stellt das Institut

1. den Nutzen,
2. den Schaden und
3. ggf. eine Nutzen-Schaden-Abwägung dar,

soweit dies aufgrund der vorliegenden Daten möglich ist. Hierbei werden alters-, geschlechts- und lebenslagenspezifische Besonderheiten berücksichtigt.

Eine Möglichkeit der gleichzeitigen 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 gleichzeitigen 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 gleichzeitige Würdigung von Nutzen und Schaden wird themenspezifisch konkretisiert und sollte – wenn dies prospektiv möglich ist – im Berichtsplan oder andernfalls im Vorbericht beschrieben werden. Eine quantitative Gewichtung unter Verwendung von Summenscores oder Indizes sollte prospektiv zum Zeitpunkt der Auswahl der zu untersuchenden Endpunkte erfolgen.

Häufig werden sogenannte Nutzwerte für Gesundheitszustände erhoben, die die von den Befragten positiv wie negativ empfundenen Aspekte in einer Indexzahl ausdrücken sollen. Unter Integration der Dauer der entsprechenden Gesundheitszustände können diese Nutzwerte bspw. in sogenannte qualitätsadjustierte Lebensjahre (QALYs = Quality-Adjusted Life Years) überführt werden. Aufgrund der ethischen und methodischen Probleme gerade der häufig verwendeten QALYs [122,137,138,520] sollten alternative Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzerhebung angewendet werden. Dazu zählen u. a. der Analytic Hierarchy Process (AHP) und die Conjoint-Analyse (CA).

Für das AHP [135,136] wird ein Entscheidungsproblem in sog. Kriterien zerlegt. Diese werden dann in eine Hierarchie gebracht. So kann z. B. ein Arzneimittel nach den Kriterien Wirkung, Nebenwirkung und Lebensqualität beurteilt werden. Das Kriterium Wirkung kann

in weitere Subkriterien zerlegt werden, die Endpunkten entsprechen können [253]. Die am AHP Teilnehmenden werden dann jeweils binär zu den Kriterien befragt, d. h., sie müssen auf einer vorgegebenen Skala wählen, wie viel mehr ihnen ein Kriterium als ein anderes Kriterium bedeutet. Mittels eines Verfahrens der Matrizenmultiplikation [427-429] können über den sog. rechten Eigenvektor die Gewichte für die Kriterien bzw. Subkriterien ermittelt werden, die sich zu 1 aufsummieren müssen.

Die CA gehört zur Gruppe der sogenannten Stated-Preference-Techniken [61]. Eine Entscheidung wird ebenfalls in sogenannte Attribute zerlegt. Die Befragten werden anschließend mit einem Set von (theoretischen) Szenarien konfrontiert, die jeweils für alle Attribute eine Effektgröße gegenüberstellen. Aus der Wahl der Szenarien wird dann über ein Regressionsmodell ein Gewichtungsfaktor für jedes Attribut errechnet. Diese Gewichte können wiederum auf 1 normiert werden.

3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V

Eine Nutzenbewertung eines Arzneimittels gemäß § 35a SGB V basiert auf einem Dossier des pharmazeutischen Unternehmers. In diesem Dossier legt der pharmazeutische Unternehmer folgende Angaben vor:

1. Zugelassene Anwendungsgebiete
2. Medizinischer Nutzen
3. Medizinischer Zusatznutzen im Verhältnis zur zweckmäßigen Vergleichstherapie
4. Anzahl der Patientinnen und Patienten und Patientengruppen, für die ein therapeutisch bedeutsamer Zusatznutzen besteht
5. Kosten der Therapie für die gesetzliche Krankenversicherung
6. Anforderung an eine qualitätsgesicherte Anwendung

Die Anforderungen an die Form und den Inhalt des Dossiers sind in Dossiervorlagen beschrieben, die Bestandteil der Verfahrensordnung des G-BA sind [193]. Im Dossier ist vom pharmazeutischen Unternehmer unter Angabe der Aussagekraft der Nachweise darzulegen, mit welcher Wahrscheinlichkeit und in welchem Ausmaß ein Zusatznutzen des zu bewertenden Arzneimittels im Vergleich zur zweckmäßigen Vergleichstherapie vorliegt. Diese Angaben sollen sowohl bezogen auf die Anzahl der Patientinnen und Patienten als auch bezogen auf die Größe des Zusatznutzens gemacht werden. Die Kosten sind für das zu bewertende Arzneimittel und die zweckmäßige Vergleichstherapie anzugeben (gemessen am Apothekenabgabepreis und unter Berücksichtigung der Fach- und Gebrauchsinformation).

Die Wahrscheinlichkeit des Zusatznutzens beschreibt die Aussagesicherheit zum Zusatznutzen. Das Ausmaß des Zusatznutzens soll im Dossier gemäß den Kategorien der Arzneimittel-Nutzenbewertungsverordnung beschrieben werden (erheblicher, beträchtlicher, geringer, nicht quantifizierbarer Zusatznutzen, kein Zusatznutzen belegt, Nutzen des zu bewertenden Arzneimittels geringer als Nutzen der zweckmäßigen Vergleichstherapie) [72].

Mit der Nutzenbewertung werden die Validität und die Vollständigkeit der Angaben im Dossier geprüft. Dabei wird auch geprüft, ob die vom pharmazeutischen Unternehmer gewählte Vergleichstherapie als zweckmäßig im Sinne des § 35a SGB V und der Arzneimittel-Nutzenbewertungsverordnung gelten kann. Darüber hinaus bewertet das Institut die in den vorgelegten Unterlagen beschriebenen Effekte unter Berücksichtigung ihrer Ergebnissicherheit. In dieser Bewertung werden die qualitative und die quantitative Ergebnissicherheit der vorgelegten Nachweise sowie die Größe der beobachteten Effekte und deren Konsistenz gewürdigt. Die Nutzenbewertung erfolgt auf Basis der im vorliegenden Methodenpapier beschriebenen Standards der evidenzbasierten Medizin, die Bewertung der Kosten auf Basis der Standards der Gesundheitsökonomie. Als Ergebnis der Bewertung legt

das Institut eigene Schlussfolgerungen vor, die die Schlussfolgerungen des pharmazeutischen Unternehmers bestätigen oder begründet von diesen abweichen können.

Die Operationalisierung der Feststellung des Ausmaßes des Zusatznutzens umfasst 3 Schritte:

1. Im 1. Schritt gilt es, für jeden Endpunkt separat die Wahrscheinlichkeit für das Vorliegen eines Effekts zu prüfen (qualitative Aussage). Dazu werden die Kriterien zur Ableitung von Aussagen zur Beleglage (siehe Abschnitt 3.1.4) angewendet. Je nach Güte der Evidenz wird die Wahrscheinlichkeit demnach als Anhaltspunkt, Hinweis oder Beleg eingestuft.
2. Im 2. Schritt ist für die Endpunkte, für die im ersten Schritt zumindest ein Anhaltspunkt für das Vorliegen eines Effekts attestiert wurde, jeweils separat das Ausmaß der Effektstärke festzustellen (quantitative Aussage). Folgende quantitative Aussagen sind möglich: erheblich, beträchtlich, gering, nicht quantifizierbar.
3. Im 3. und letzten Schritt gilt es, anhand aller Endpunkte unter Würdigung der Wahrscheinlichkeit und des Ausmaßes auf Endpunktebene im Rahmen einer Gesamtschau die Gesamtaussage zum Zusatznutzen festzustellen.

Zur Feststellung des Ausmaßes auf Endpunktebene im 2. Schritt sind die Qualität der Zielgröße sowie die Effektstärke maßgeblich. Die Rationale für diese Operationalisierung findet sich im Anhang „Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens“. Das grundsätzliche Konzept sieht vor, für relative Effektmaße Schwellenwerte für Konfidenzintervalle in Abhängigkeit von anzustrebenden Effekten abzuleiten, die wiederum von der Qualität der Zielgrößen und den Ausmaßkategorien abhängen.

Das Ausmaß auf Endpunktebene wird nicht in jedem Fall zu quantifizieren sein. Falls z. B. ein statistisch signifikanter Effekt für ein ausreichend valides Surrogat vorliegt, eine verlässliche Schätzung für den jeweiligen Effekt der patientenrelevanten Zielgröße jedoch nicht möglich ist, lässt sich der (patientenrelevante) Effekt nicht quantifizieren. In solchen und ähnlichen Situationen wird mit entsprechender Begründung ein Effekt nicht quantifizierbarem Ausmaßes attestiert.

Vom Fall eines quantifizierbaren Effekts ausgehend, richtet sich das weitere Vorgehen nach der Skala der Zielgröße. Es werden folgende Skalen unterschieden:

- binär (Analysen von Vierfeldertafeln)
- Zeit bis Ereignis (Überlebenszeitenanalysen)
- stetig oder quasi-stetig mit jeweils vorliegenden Responderanalysen (Analysen von Mittelwerten und Standardabweichungen)
- sonstige (z. B. Analysen von nominalen Daten).

Im Folgenden wird zunächst das Verfahren für binäre Zielgrößen beschrieben. Die übrigen Skalen werden nachfolgend auf dieses Verfahren zurückgeführt.

Vom Effektmaß relatives Risiko ausgehend werden Zähler und Nenner immer so gewählt, dass sich der Effekt (sofern vorhanden) als Wert < 1 realisiert. D. h. ein Effekt ist umso stärker, je niedriger der Wert ist.

A) Binäre Zielgrößen

Zur Feststellung des Ausmaßes des Effekts bei binären Zielgrößen wird das zweiseitige 95%-Konfidenzintervall für das relative Risiko – ggf. selbst berechnet – herangezogen. Falls mehrere Studien quantitativ zusammengefasst wurden, findet das meta-analytische Ergebnis für das relative Risiko Anwendung.

Je nach Qualität der Zielgröße muss das Konfidenzintervall vollständig unterhalb eines bestimmten Schwellenwertes liegen, um das Ausmaß als gering, beträchtlich oder erheblich anzusehen. Entscheidend ist also, dass die obere Grenze des Konfidenzintervalls kleiner als der jeweilige Schwellenwert ist.

Es werden folgende drei Kategorien für die Qualität der Zielgröße gebildet:

- Gesamtmortalität
- Schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen sowie gesundheitsbezogene Lebensqualität
- Nicht schwerwiegende (bzw. nicht schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen

Die Schwellenwerte sind für jede Kategorie separat festgelegt und umso größer (im Sinne näher an 1), je schwerwiegender das Ereignis ist. Die Schwellenwerte sind umso kleiner (im Sinne weiter entfernt von 1), je höher das Ausmaß ist. Die folgende Tabelle NT1 gibt die für die 3 Ausmaßkategorien (gering, beträchtlich, erheblich) zu unterschreitenden Schwellenwerte für jede der 3 Kategorien der Qualität der Zielgrößen wieder.

Tabelle NT1: Schwellenwerte zur Feststellung des Ausmaßes eines Effekts

		Zielgrößenkategorie		
		Gesamt-mortalität	Schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen und gesundheitsbezogene Lebensqualität[a]	Nicht schwerwiegende (bzw. nicht schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen
Ausmaßkategorie	Erheblich	0,85	0,75 und Risiko ≥ 5 %[b]	Nicht besetzt
	Beträchtlich	0,95	0,90	0,80
	Gering	1,00	1,00	0,90

a: Voraussetzung ist die Verwendung eines validierten bzw. etablierten Instruments sowie eines validierten bzw. etablierten Responsekriteriums.

b: Risiko muss für mindestens eine der beiden zu vergleichenden Gruppen mindestens 5 % betragen.

Das relative Risiko kann generell auf zwei Arten berechnet werden, je nachdem ob sich das Risiko auf Ereignisse oder Gegenereignisse bezieht (z. B. Tod vs. Überleben, Response vs. Non-Response). Für die vorgelagerte Signifikanzaussage ist dies unerheblich, da der p-Wert diesbezüglich bei einer Einzelstudie invariant ist und bei einer Meta-Analyse eine untergeordnete Rolle spielt. Das gilt jedoch nicht für den Abstand der Konfidenzintervall-grenzen zum Nulleffekt. Daher muss zur Festlegung des Ausmaßes des Effekts für jede binäre Zielgröße anhand inhaltlicher Kriterien unter Berücksichtigung der Art des Endpunkts und der zugrunde liegenden Erkrankung entschieden werden, welches Risiko betrachtet wird – das für das Ereignis oder das für das Gegenereignis.

B) Zeit bis Ereignis

Zur Feststellung des Ausmaßes des Effekts bei Zielgrößen „Zeit bis zu einem Ereignis“ wird das zweiseitige 95%-Konfidenzintervall für das Hazard Ratio benötigt. Falls mehrere Studien quantitativ zusammengefasst wurden, wird das meta-analytische Ergebnis für das Hazard Ratio herangezogen. Liegt das Konfidenzintervall für das Hazard Ratio nicht vor, wird es anhand der zur Verfügung stehenden Angaben approximiert, sofern möglich [N11]. Für die Ausmaßfeststellung werden dieselben Grenzen wie für das relative Risiko angelegt (Tabelle NT1).

Liegt kein Hazard Ratio vor und ist dies auch nicht berechenbar oder das vorliegende Hazard Ratio ist nicht sinnvoll interpretierbar (z. B. wegen wesentlicher Verletzung der Proportional Hazard Annahme), ist zu eruieren, ob sich aus den Angaben ein relatives Risiko (bezogen auf einen sinnvollen Zeitpunkt) berechnen lässt. Auch bei transienten (vorübergehenden) Ereignissen, für die als Zielgröße „Zeit bis zum Ereignis“ gewählt wurde, ist zu eruieren, ob

diese Operationalisierung adäquat ist. Ggf. ist auch hier die Berechnung eines relativen Risikos zu einem Zeitpunkt angezeigt.

C) Stetige oder quasi-stetige Zielgrößen mit jeweils vorliegenden Responderanalysen

Zur Feststellung des Ausmaßes des Zusatznutzens bei stetigen oder quasi stetigen Zielgrößen werden Responderanalysen herangezogen. Dazu bedarf es eines validierten bzw. etablierten Responsekriteriums bzw. Cut-off-Werts. Anhand der Responderauswertungen (Vierfeldertafeln) werden die relativen Risiken daraus direkt berechnet. Entsprechend Tabelle NT1 wird anschließend das Ausmaß des Effekts festgestellt.

D) Sonstige Zielgrößen

Für den Fall sonstiger Zielgrößen, für die auch keine Responderauswertungen mit daraus ableitbaren relativen Risiken vorliegen, ist im Einzelfall zu eruieren, ob relative Risiken approximiert werden können [N1], um die entsprechenden Schwellenwerte zur Ausmaßfeststellung anzulegen. Anderenfalls ist das Ausmaß als nicht quantifizierbar festzustellen.

Für den dritten Schritt der Operationalisierung, der Gesamtaussage zum Ausmaß des Zusatznutzens bei gemeinsamer Betrachtung aller Endpunkte, ist eine strenge Formalisierung nicht möglich, da für die hierzu zu treffenden Werturteile gegenwärtig keine ausreichende Abstraktion bekannt ist. Das Institut wird im Rahmen seiner Nutzenbewertung die Aussagen zur Wahrscheinlichkeit und zum Ausmaß der Effekte vergleichend gegenüberstellen und einen begründeten Vorschlag für eine Gesamtaussage unterbreiten.

7.3.8 Meta-Analysen

A) Allgemeines

In der Literatur verwendete Begriffe wie „Literaturübersicht“, „systematische Übersicht“, „Meta-Analyse“, „gepoolte Analyse“ oder „Forschungssynthese“ sind häufig unterschiedlich definiert und nicht klar voneinander abgegrenzt [150]. Das Institut verwendet folgende Begriffe und Definitionen: Bei einer „nicht systematischen Übersicht“ handelt es sich um eine Beschreibung und Bewertung von Studienergebnissen zu einer definierten Thematik ohne eine ausreichend systematische und reproduzierbare Identifikationsmethode der relevanten Forschungsergebnisse zu dieser Thematik. Eine quantitative Zusammenfassung von Daten mehrerer Studien wird als „gepoolte Analyse“ bezeichnet. Wegen der fehlenden Systematik und der inhärenten subjektiven Komponente sind Übersichten und Auswertungen, die nicht auf einer systematischen Literaturrecherche basieren, sehr anfällig für Verzerrungen. Eine „systematische Übersicht“ beruht auf einer umfassenden systematischen Vorgehensweise und Studienbewertung, um mögliche Biasquellen zu minimieren. Sie kann – muss aber nicht – eine quantitative Zusammenfassung der Studienergebnisse beinhalten. Eine „Meta-Analyse“ wird verstanden als eine statistische Zusammenfassung von Ergebnissen mehrerer Studien im Rahmen einer systematischen Übersicht. Sie basiert in den meisten Fällen auf aggregierten Studiendaten aus Publikationen. Dabei wird aus den in einzelnen Studien gemessenen Effektstärken, unter Berücksichtigung der Fallzahlen und der Varianzen, ein Gesamteffekt berechnet. Effizientere Auswertungsverfahren sind möglich, wenn individuelle Patientendaten aus den Studien verfügbar sind. Unter einer „Meta-Analyse mit individuellen Patientendaten“ (IPD = Individual Patient Data) wird die Auswertung von Daten auf Patientenebene im Rahmen eines allgemeinen statistischen Modells mit festen oder zufälligen Effekten verstanden, in das die Studie als Effekt und nicht als Beobachtungseinheit eingeht. Unter einer „prospektiven Meta-Analyse“ versteht das Institut die a priori geplante statistische Zusammenfassung der Ergebnisse mehrerer prospektiv gemeinsam geplanter Studien. Sollte es zur jeweiligen Fragestellung auch noch andere Studien geben, so müssen diese jedoch auch in der Auswertung berücksichtigt werden, um den Charakter einer systematischen Übersicht zu bewahren.

Die übliche Darstellung der Ergebnisse einer Meta-Analyse erfolgt mittels Forest Plots, in denen die Effektschätzer der einzelnen Studien und des Gesamteffekts inklusive der Konfidenzintervalle grafisch aufgetragen werden [326]. Es kommen zum einen Modelle mit festem Effekt zum Einsatz, die (z. B. durch die Inverse der Varianz) gewichtete Mittelwerte der Effektstärken liefern. Es werden zum anderen aber auch häufig Modelle mit zufälligen Effekten gewählt, bei denen eine Schätzung der Varianz zwischen den einzelnen Studien (Heterogenität) berücksichtigt wird. Die Frage, in welchen Situationen welches Modell eingesetzt werden soll, wird seit Langem kontrovers diskutiert [154,457,515]. Liegen Informationen darüber vor, dass die Effekte der einzelnen Studien homogen sind, ist eine Meta-Analyse unter der Annahme eines festen Effekts ausreichend. Solche Informationen

werden jedoch häufig nicht vorliegen, sodass bei der Evaluierung der Studien in ihrer Gesamtheit die Annahme zufälliger Effekte hilfreich ist [458]. 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 festem Effekt aus. Dabei ist zu beachten, dass sich die meta-analytischen Ergebnisse von Modellen mit zufälligen und festen Effekten bei homogener Datenlage allenfalls marginal unterscheiden. Des Weiteren wird das Institut, wie im Folgenden beschrieben, nur dann stark heterogene Studienergebnisse meta-analytisch zusammenfassen, wenn plausible Gründe für die Heterogenität ersichtlich sind, die eine Zusammenfassung trotzdem rechtfertigen.

B) Heterogenität

Bevor eine Meta-Analyse durchgeführt wird, muss zunächst überlegt werden, ob die Zusammenfassung der betrachteten Studien überhaupt sinnvoll ist, da die Studien bezüglich der Fragestellung vergleichbar sein müssen. Darüber hinaus werden sich in den zusammenzufassenden Studien trotz Vergleichbarkeit häufig heterogene Effekte zeigen [243]. In dieser Situation ist es erforderlich, die Heterogenität der Studien bezüglich der Ergebnisse zu untersuchen [198]. Das Vorliegen von Heterogenität kann statistisch getestet werden, wobei diese Verfahren in der Regel eine sehr niedrige Power haben. Daher wird für diese Tests empfohlen, ein Signifikanzniveau von 0,1 bis 0,2 zu wählen [274,300]. Daneben gilt es auch, das Ausmaß der Heterogenität zu quantifizieren. Zu diesem Zweck gibt es spezielle statistische Methoden wie z. B. das I^2-Maß [242]. Für dieses Maß existieren Untersuchungen, die eine grobe Einschätzung der Heterogenität zulassen (z. B. die Kategorien wahrscheinlich unbedeutend (0 – 40 %), mittelmäßig (30 – 60 %), substanziell (50 – 90 %) und erheblich (75 – 100 %) [110]. Ist die Heterogenität der Studien zu groß, so ist eine statistische Zusammenfassung der Studienergebnisse unter Umständen nicht sinnvoll [110]. Die Spezifizierung, wann eine „zu große“ Heterogenität vorliegt, ist kontextabhängig. In der Regel wird von einer statistischen Zusammenfassung abgesehen, falls der Heterogenitätstest einen *p*-Wert unter 0,2 liefert. Es spielt auch die Lage der Effekte eine Rolle. Zeigen die einzelnen Studien einen deutlichen und gleichgerichteten Effekt, dann kann auch die Zusammenfassung heterogener Ergebnisse mittels eines Modells mit zufälligen Effekten zu einer Nutzenaussage führen. In dieser Situation kann ggf. aber auch ohne quantitative Zusammenfassung eine positive Nutzenaussage getroffen werden (siehe Abschnitt 3.1.4). In den übrigen Situationen führt das Institut keine Meta-Analyse durch. In diese Entscheidung sollten jedoch neben statistischen Maßzahlen auch inhaltliche Gründe einfließen, die nachvollziehbar darzustellen sind. In diesem Zusammenhang spielt auch die Wahl des Effektmaßes eine Rolle. Es kann sein, dass die Wahl eines bestimmten Maßes zu großer Studienheterogenität führt, ein anderes Maß jedoch nicht. Bei binären Daten sind relative

Effektmaße häufig stabiler als absolute, da sie nicht so stark vom Basisrisiko abhängen [188]. In solchen Fällen sollte die Datenanalyse über ein relatives Effektmaß erfolgen; für die deskriptive Darstellung können dann unter Umständen absolute Maße für spezifische Basisrisiken hieraus abgeleitet werden.

Bei einer großen Heterogenität der Studien ist es notwendig, mögliche Ursachen hierfür zu untersuchen. Unter Umständen lassen sich mittels Meta-Regressionen Faktoren finden, die die Heterogenität der Effektstärken erklären können [492,509]. In einer Meta-Regression wird die statistische Assoziation zwischen den Effektstärken der einzelnen Studien und den Studiencharakteristika untersucht, sodass möglicherweise Studiencharakteristika gefunden werden können, die einen Erklärungswert für die unterschiedlichen Effektstärken, also die Heterogenität, haben. Wichtig ist jedoch, dass man bei der Interpretation der Ergebnisse die Einschränkungen solcher Analysen berücksichtigt. Selbst wenn eine Meta-Regression auf randomisierten Studien basiert, kann aus ihr nur die Evidenz einer Beobachtungsassoziation abgeleitet werden, nicht jedoch ein kausaler Zusammenhang [492]. Besonders schwierig zu interpretieren sind Meta-Regressionen, die versuchen, eine Beziehung zwischen den unterschiedlichen Effektstärken und den durchschnittlichen Patientencharakteristika der einzelnen Studien aufzuzeigen. Solche Analysen unterliegen den gleichen Beschränkungen wie die Ergebnisse ökologischer Studien der Epidemiologie [206]. Aufgrund der hohen Anfälligkeit für Verzerrungen, die bei auf aggregierten Daten basierenden Analysen auch nicht durch Adjustierung ausgeglichen werden können, sind hier gesicherte Schlussfolgerungen nur auf der Basis individueller Patientendaten möglich [466,492] (siehe Abschnitt 7.2.3).

Zur Darstellung der Heterogenität im Rahmen einer Meta-Analyse mit zufälligen Effekten verwendet das Institut Prädiktionsintervalle [N5,N6,N8]. Im Gegensatz zu einem Konfidenzintervall, das die Präzision eines geschätzten Effekts quantifiziert, überdeckt ein 95 %-Prädiktionsintervall den wahren Effekt einer einzelnen (neuen) Studie mit Wahrscheinlichkeit 95 %. Dabei ist es wichtig zu beachten, dass ein Prädiktionsintervall nicht zur Beurteilung der statistischen Signifikanz eines Effekts herangezogen wird. Das Institut folgt dem Vorschlag von Guddat et al. [N5], das Prädiktionsintervall deutlich unterscheidbar vom Konfidenzintervall in den Forest Plot einzufügen. In Ergänzung zur üblichen Darstellung der Punkt- und Intervallschätzung des Effekts in Form einer Raute wird eine weitere Zeile für das Prädiktionsintervall hinzugefügt, in der selbiges in Form eines Rechtecks präsentiert wird. Die Anwendung von Meta-Analysen mit zufälligen Effekten und zugehörigen Prädiktionsintervallen im Fall von sehr wenigen Studien (z. B. weniger als 5) wird in der Literatur kritisch gesehen, da eine mögliche Heterogenität nur sehr unpräzise geschätzt werden kann [N6]. Das Institut stellt Prädiktionsintervalle in Forest Plots von Meta-Analysen mit zufälligen Effekten in der Regel dar, wenn mindestens 4 Studien vorhanden sind.

Prädiktionsintervalle werden auch in Forest Plots verwendet, wenn aufgrund zu starker Heterogenität kein Gesamteffekt geschätzt und dargestellt wird. In diesen heterogenen Situationen ist das Prädiktionsintervall eine wertvolle Hilfe bei der Beurteilung, ob die

Studieneffekte gleichgerichtet sind oder nicht und ob es sich im ersten Fall um deutlich gleichgerichtete oder mäßig gleichgerichtete Effekte handelt (siehe Abschnitt 3.1.4).

C) Subgruppenanalysen im Rahmen von Meta-Analysen

Neben den allgemeinen Aspekten, die bei der Interpretation von Subgruppenanalysen beachtet werden müssen (siehe Abschnitt 7.1.6), gibt es besondere Aspekte, die bei Subgruppenanalysen im Rahmen von Meta-Analysen eine Rolle spielen. Während im Allgemeinen post hoc durchgeführte Subgruppenanalysen auf Studienebene kritisch zu interpretieren sind, ist man in einer systematischen Übersicht dennoch auf die Verwendung der Ergebnisse solcher Analysen auf Studienebene angewiesen, wenn im Rahmen der systematischen Übersicht genau diese Subgruppen untersucht werden sollen. Analog zum Vorgehen, Studien mit zu großer Heterogenität nicht mithilfe von Meta-Analysen zusammenzufassen, sollten auch Ergebnisse von Subgruppen nicht zu einem gemeinsamen Effektschätzer zusammengefasst werden, wenn sich die Subgruppen zu stark voneinander unterscheiden. Das Institut interpretiert im Rahmen von Meta-Analysen die Ergebnisse eines Heterogenitäts- oder Interaktionstests bezüglich wichtiger Subgruppen in der Regel wie folgt. Ein zum Niveau $\alpha = 0{,}05$ signifikantes Ergebnis wird als Beleg unterschiedlicher Effekte, ein zum Niveau $\alpha = 0{,}20$ signifikantes Ergebnis wird als Hinweis auf unterschiedliche Effekte in den Gruppen gewertet. Liegt mindestens ein Hinweis auf unterschiedliche Effekte in den Subgruppen vor, so werden neben dem Gesamteffekt auch die einzelnen Subgruppenergebnisse berichtet. Liegt ein Beleg für unterschiedliche Effekte in den Subgruppen vor, so werden die Ergebnisse aller Subgruppen nicht zu einem gemeinsamen Effektschätzer gepoolt. Bei mehr als zwei Subgruppen werden – wenn möglich – die paarweisen statistischen Tests auf das Vorliegen von Subgruppeneffekten durchgeführt und Paare, die zum Niveau $\alpha = 0{,}20$ nicht statistisch signifikant sind, zu einer Gruppe zusammengefasst. Die Ergebnisse der verbleibenden Gruppen werden getrennt berichtet und es werden getrennte Nutzenaussagen für diese Gruppen abgeleitet [468].

D) Geringe Zahl von Ereignissen

Ein häufiges Problem in Meta-Analysen bei binären Daten ist das Vorhandensein von sogenannten Nullzellen, also die Beobachtung von keinem einzigen Ereignis in einer Interventionsgruppe einer Studie. Das Institut folgt dem üblichen Vorgehen, beim Auftreten von Nullzellen den Korrekturwert von 0,5 zu jeder Zellhäufigkeit der entsprechenden Vierfeldertafel zu addieren [110]. Dieses Vorgehen ist adäquat, wenn nicht zu viele Nullzellen vorkommen. Im Fall einer insgesamt geringen Zahl von Ereignissen ist es unter Umständen notwendig, auf andere Methoden zurückzugreifen. Bei sehr seltenen Ereignissen kann die sogenannte Peto-Odds-Ratio-Methode verwendet werden, die keinen Korrekturterm beim Vorliegen von Nullzellen erfordert [58,110].

Kommen sogar Studien vor, in denen in beiden Studienarmen kein Ereignis beobachtet wird (sogenannte Doppelnullstudien), so werden diese Studien in der Praxis häufig aus der meta-analytischen Berechnung ausgeschlossen. Dieses Verfahren sollte vermieden werden, wenn zu viele Doppelnullstudien auftreten. Es gibt mehrere Methoden, um den Ausschluss von Doppelnullstudien zu vermeiden. Unter Umständen kann als Effektmaß die absolute Risikodifferenz verwendet werden, die gerade bei sehr seltenen Ereignissen häufig nicht zu den sonst üblichen Heterogenitäten führt. Ein in der Praxis bislang selten angewendetes Verfahren stellt die logistische Regression mit zufälligen Effekten dar [504]. Neuere Verfahren wie exakte Methoden [496] oder die Anwendung der Arcus-Sinus-Differenz [425] stellen interessante Alternativen dar, sind aber noch nicht ausreichend untersucht. Das Institut wird in Abhängigkeit der jeweiligen Datensituation ein geeignetes Verfahren auswählen und ggf. mithilfe von Sensitivitätsanalysen die Robustheit der Ergebnisse untersuchen.

E) Meta-Analysen diagnostischer Studien

Auch die Ergebnisse von Studien zur diagnostischen Güte können mithilfe meta-analytischer Techniken statistisch zusammengefasst werden [126,273]. Wie in Abschnitt 3.5 ausgeführt, sind Studien, die allein die diagnostische Güte untersuchen, jedoch meist von nachrangiger Bedeutung in der Bewertung diagnostischer Verfahren, sodass auch Meta-Analysen von Studien zur diagnostischen Güte einen in gleicher Weise eingeschränkten Stellenwert haben.

Für eine Meta-Analyse von Studien zur diagnostischen Güte gelten die gleichen grundlegenden Prinzipien wie für Meta-Analysen von Therapiestudien [126,409]. Dies beinhaltet insbesondere die Notwendigkeit einer systematischen Literaturübersicht, die Bewertung der methodischen Qualität der Primärstudien, die Durchführung von Sensitivitäts-analysen und die Untersuchung des möglichen Einflusses von Publikationsbias.

Bei Meta-Analysen diagnostischer Studien ist in der Praxis in den meisten Fällen mit Heterogenität zu rechnen, daher empfiehlt sich hier in der Regel die Verwendung von Modellen mit zufälligen Effekten [126]. Eine solche meta-analytische Zusammenfassung von Studien zur diagnostischen Güte kann durch getrennte Modelle für Sensitivität und Spezifität erfolgen. Bei Interesse an einer summarischen Receiver-Operating-Characteristic (ROC)-Kurve und / oder einem zweidimensionalen Schätzer für Sensitivität und Spezifität haben jedoch neuere bivariate Meta-Analysen mit zufälligen Effekten Vorteile [221,410]. Diese Verfahren ermöglichen auch die Berücksichtigung erklärender Variablen [220]. Die grafische Darstellung der Ergebnisse erfolgt entweder über die separate Darstellung der Sensitivitäten und Spezifitäten in Form modifizierter Forest Plots oder eine zweidimensionale Abbildung der Schätzer für Sensitivität und Spezifität. Analog zu den Konfidenz- und Prädiktions-intervallen in Meta-Analysen von Therapiestudien können bei bivariaten Meta-Analysen von diagnostischen Studien Konfidenz- und Prädiktionsregionen im ROC-Raum dargestellt werden.

F) Kumulative Meta-Analysen

Es wird seit einiger Zeit verstärkt diskutiert, ob man bei wiederholten Aktualisierungen systematischer Übersichten die darin enthaltenen Meta-Analysen als kumulative Meta-Analysen mit Korrektur für multiples Testen berechnen und darstellen sollte [51,65,66,384,493,528]. Das Institut verwendet standardmäßig die übliche Form von Meta-Analysen und greift in der Regel nicht auf Methoden für kumulative Meta-Analysen zurück.

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.

Anhang: Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens

In diesem Abschnitt wird die Rationale für das methodische Vorgehen bei der Feststellung des Ausmaßes des Zusatznutzens gemäß der Arzneimittel-Nutzenbewertungsverordnung (AM-NutzenV) beschrieben.

Gemäß § 5 Abs. 4 Satz 1 der Arzneimittel-Nutzenbewertungsverordnung (AM-NutzenV) ist im Dossier darzulegen und folgerichtig auch zu bewerten, „in welchem Ausmaß ein Zusatznutzen vorliegt“. In § 5 Abs. 7 der AM-NutzenV findet sich dazu eine Einteilung in 6 Kategorien: (1) erheblicher Zusatznutzen, (2) beträchtlicher Zusatznutzen, (3) geringer Zusatznutzen, (4) nicht quantifizierbarer Zusatznutzen, (5) kein Zusatznutzen belegt, (6) geringerer Nutzen. Weiterhin liefert § 5 Abs. 7 der AM-NutzenV für die Kategorien 1 bis 3 eine Definition sowie beispielhafte, besonders zu berücksichtigende Kriterien als Orientierung für die Darlegung und Bewertung. Die dort formulierten Kriterien beschreiben sowohl qualitative Momente (Art der Zielgrößen) als auch explizit quantitative Momente (z. B. „erhebliche Verlängerung der Überlebensdauer“ vs. „moderate Verlängerung der Überlebensdauer“). Darüber hinaus ist offensichtlich eine Hierarchisierung dieser Zielgrößen intendiert, da teilweise dieselben Attribute (z. B. das Attribut „relevant“) bei unterschiedlichen Zielgrößen zu einem unterschiedlichen Ausmaß führen. In Tabelle NT2 sind die diesbezüglichen Angaben für die primär relevanten Ausmaßkategorien erheblicher, beträchtlicher und geringer Zusatznutzen aufgeführt. Es ergibt sich für die Nutzenbewertung die Aufgabe, auf der Basis dieser Vorgaben das Ausmaß des Zusatznutzens zu operationalisieren.

Die in § 5 Abs. 7 der AM-NutzenV gelieferten Kriterien für das Ausmaß des Zusatznutzens benennen (Rechts-) Begriffe, die zum Teil eindeutig bestimmt (z. B. „Überlebensdauer“, „schwerwiegende Nebenwirkungen“), teilweise weniger eindeutig bestimmt sind (z. B. „Abschwächung schwerwiegender Symptome“). Darüber hinaus sind die Kategorien nicht für alle aufgeführten Kriterien erschöpfend besetzt, z. B. werden für die „Überlebensdauer“ nur Beispiele für die Kategorien „erheblicher“ und „beträchtlicher“ Zusatznutzen genannt.

Durch die Formulierung „insbesondere“ in § 5 Abs. 7 zu den Kategorien 1-3 macht der Verordnungsgeber deutlich, dass die den Kategorien zugeordneten Kriterien nicht abschließend zu verstehen sind. Es ist nicht davon auszugehen, dass der Verordnungsgeber einer weniger als „moderaten Verlängerung der Überlebensdauer“ nicht zumindest einen „geringen Zusatznutzen“ anerkennen wollte. Weiterhin erscheint die Zielgröße (gesundheitsbezogene) Lebensqualität, die in § 2 Abs. 3 der AM-NutzenV explizit als Nutzenkriterium formuliert wird, überhaupt nicht in der Kriterienliste für das Ausmaß des Zusatznutzens.

Tabelle NT2: Feststellung des Ausmaßes des Zusatznutzens – Kriterien gemäß AM-NutzenV

Ausmaßkategorie	**Erheblich** **nachhaltige** und gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **große Verbesserung** des therapierelevanten Nutzens	Heilung	Erhebliche Verlängerung der Überlebensdauer	Langfristige Freiheit von schwerwiegenden Symptomen	Weitgehende Vermeidung schwerwiegender Nebenwirkungen
	Beträchtlich gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **deutliche Verbesserung** des therapierelevanten Nutzens	Spürbare Linderung der Erkrankung	Moderate Verlängerung der Überlebensdauer	Abschwächung schwerwiegender Symptome	Relevante Vermeidung schwerwiegender Nebenwirkungen Bedeutsame Vermeidung anderer Nebenwirkungen
	Gering gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **moderate und nicht nur geringfügige Verbesserung** des therapierelevanten Nutzens			Verringerung von nicht schwerwiegenden Symptomen	Relevante Vermeidung von Nebenwirkungen

In einem ersten Schritt ist es also sinnvoll, die Kriterienliste anzupassen und durch qualitativ und quantitativ gleichwertige Kriterien zu ergänzen. In Tabelle NT3 sind die Ergänzungen der Vorgaben der AM-NutzenV aufgeführt. Dabei wurden die Aspekte „Heilung“ und „spürbare Linderung der Erkrankung“ nicht explizit berücksichtigt. Der Begriff der „Heilung“ bedarf grundsätzlich einer Operationalisierung, die sich regelhaft auf Kriterien stützen wird, die sich auch in den Endpunkten Mortalität und Morbidität abbilden lassen (z.B. Überleben über mindestens einen definierten Zeitraum in der Onkologie). Da der Begriff „Heilung“ in der AM-NutzenV ausschließlich mit einem erheblichen Zusatznutzen verknüpft wird, ist die jeweilige konkrete Operationalisierung anhand der verwendeten Endpunkte daraufhin zu prüfen, ob sie einer relevanten Verbesserung der Mortalität bzw. schwerwiegender Ereignisse gleich kommt. Die Verkürzung der Symptomdauer, z.B. bei banalen Infektionskrankheiten, ist in diesem Sinne nicht als Heilung anzusehen.

Ausgehend von diesen Ergänzungen ist eine Umstrukturierung der Zielgrößenkategorien angezeigt, um die in der AM-NutzenV intendierte Hierarchisierung der Zielgrößen abzubilden und gemäß § 5 Abs. 7 der AM-NutzenV den Schweregrad der Erkrankung zu berücksichtigen. Dazu werden die Zielgrößen gemäß ihrer Bedeutung wie folgt gruppiert (siehe Tabelle NT4):

1. Gesamtmortalität
2. • schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen)
   • schwerwiegende (bzw. schwere) Nebenwirkungen
   • Gesundheitsbezogene Lebensqualität
3. • nicht schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen)
   • nicht schwerwiegende (bzw. schwere) Nebenwirkungen.

Der gesundheitsbezogenen Lebensqualität wird die gleiche Bedeutung beigemessen wie schwerwiegenden (bzw. schweren) Symptomen, Folgekomplikationen und Nebenwirkungen. Die möglichen Ausmaßkategorien für die nicht schwerwiegenden Zielgrößen bleiben auf „beträchtlich“ und „gering“ beschränkt.

Die Vorgaben der AM-NutzenV machen deutlich, dass zur Feststellung des Ausmaßes des Zusatznutzens zunächst auf Endpunktebene eine Effektstärkenbeschreibung zu erfolgen hat. Für jede Zielgröße wird separat die Effektstärke – unabhängig von ihrer Richtung – in die 3 Ausmaßkategorien (gering, beträchtlich, erheblich) eingestuft. Im Rahmen einer Gesamtabwägung sind diese einzelnen Ausmaße anschließend zu einer globalen Aussage zum Ausmaß des Zusatznutzens zusammenzufassen. Das schrittweise Vorgehen ist in Abschnitt 3.3.3 beschrieben.

Tabelle NT3: Feststellung des Ausmaßes des Zusatznutzens – Kriterien gemäß AM-NutzenV mit Ergänzungen*

		Zielgrößenkategorie			
		Gesamtmortalität	*Symptome (Morbidität)*	*Gesundheitsbezogene Lebensqualität*	*Nebenwirkungen*
Ausmaßkategorie	**Erheblich** **nachhaltige** und gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **große Verbesserung** des therapierelevanten Nutzens	Erhebliche Verlängerung der Überlebensdauer	Langfristige Freiheit von schwerwiegenden *(bzw. schweren)* Symptomen *(bzw. Folgekomplikationen)*	*Erhebliche Verbesserung der Lebensqualität*	Weitgehende Vermeidung schwerwiegender *(bzw. schwerer)* Nebenwirkungen
	Beträchtlich gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **deutliche Verbesserung** des therapierelevanten Nutzens	Moderate Verlängerung der Überlebensdauer	Abschwächung schwerwiegender *(bzw. schwerer)* Symptome *(bzw. Folgekomplikationen)* *Bedeutsame Verringerung von nicht schwerwiegenden (bzw. schweren) Symptomen*	*Bedeutsame Verbesserung der Lebensqualität*	Relevante Vermeidung schwerwiegender *(bzw. schwerer)* Nebenwirkungen Bedeutsame Vermeidung anderer *(nicht schwerwiegender bzw. schwerer)* Nebenwirkungen
	Gering gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **moderate und nicht nur geringfügige Verbesserung** des therapierelevanten Nutzens	*Jegliche Verlängerung der Überlebensdauer*	*Jegliche Verringerung schwerwiegender (bzw. schwerer) Symptome (bzw. Folgekomplikationen)* Verringerung von nicht schwerwiegenden *(bzw. schweren)* Symptomen	*Relevante Verbesserung der Lebensqualität*	*Jegliche (statistisch signifikante) Verringerung schwerwiegender (bzw. schwerer) Nebenwirkungen* Relevante Vermeidung von *(anderen, nicht schwerwiegenden bzw. schweren)* Nebenwirkungen

*Ergänzungen gegenüber AM-NutzenV *kursiv* gesetzt

Tabelle NT4: Feststellung des Ausmaßes des Zusatznutzens – hierarchisierte Kriterien gemäß AM-NutzenV mit Ergänzungen*

		Zielgrößenkategorie			
		Gesamtmortalität	**Schwerwiegende *(bzw. schwere)*** Symptome *(bzw. Folgekomplikationen)* und Nebenwirkungen	***Gesundheits-bezogene Lebensqualität***	**Nicht schwerwiegende *(bzw. nicht schwere)*** Symptome *(bzw. Folgekomplikationen)* und Nebenwirkungen
Ausmaßkategorie	**Erheblich** **nachhaltige** und gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **große Verbesserung** des therapierelevanten Nutzens	Erhebliche Verlängerung der Überlebensdauer	Langfristige Freiheit bzw. weitgehende Vermeidung	*Erhebliche Verbesserung*	*Nicht besetzt*
	Beträchtlich gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **deutliche Verbesserung** des therapierelevanten Nutzens	Moderate Verlängerung der Überlebensdauer	Abschwächung bzw. relevante Vermeidung	*Bedeutsame Verbesserung*	Bedeutsame Vermeidung
	Gering gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte **moderate und nicht nur geringfügige Verbesserung** des therapierelevanten Nutzens	*Jegliche Verlängerung der Überlebensdauer*	*Jegliche Verringerung*	*Relevante Verbesserung*	Relevante Vermeidung

*Ergänzungen gegenüber AM-NutzenV *kursiv* gesetzt

Entsprechend § 2 Abs. 3 der AM-NutzenV ist der Begriff „Nutzen“ als Effekt definiert und in § 2 Abs. 4 der AM-NutzenV der Begriff „Zusatznutzen“ als ein solcher Effekt im Vergleich zu der zweckmäßigen Vergleichstherapie. Daraus kann abgeleitet werden, dass die Feststellung des Ausmaßes des Zusatznutzens neben der hierarchischen Betrachtung von Zielgrößen auch auf der Basis von Effektstärken zu erfolgen hat.

Zu den Fragen, welche Effektstärken für die einzelnen Zielgrößen zu welcher Ausmaßkategorie führen und welche Effektmaße für diese Bewertung zu wählen sind, finden sich in der AM-NutzenV keine Angaben. Diese Fragen können prinzipiell nur bedingt methodisch beantwortet werden. Dennoch besteht die Notwendigkeit, das in den Dossiers dargelegte Ausmaß des Zusatznutzens zu bewerten (§ 7 Abs. 2, AM-NutzenV) und selbst Aussagen zum Ausmaß zu machen. Um hierbei zunächst die im weiteren Abwägungsprozess notwendigerweise zu treffenden Werturteile möglichst gering zu halten und diese explizit zu machen, bedarf es einer

- expliziten Operationalisierung, um ein transparentes und nachvollziehbares Verfahren sicherzustellen, sowie einer
- abstrakten Operationalisierung, um größtmögliche Konsistenz zwischen den Nutzenbewertungen zu erzielen.

Vor diesem Hintergrund ist zunächst die Wahl eines geeigneten Effektmaßes zu treffen. Es sei zunächst die Situation binärer Daten (Analyse von Vierfeldertafeln) im Fokus. Relative Effektmaße – hierunter fallen im Wesentlichen das relative Risiko (RR) und das Odds Ratio (OR) – haben in diesem Zusammenhang gegenüber absoluten Maßen wie der Risikodifferenz (RD) folgende Vorteile:

- Die Risikodifferenz beschreibt nicht die Effektivität einer Therapie als solche, da sie stark vom Basisrisiko in der Kontrollgruppe abhängt. Dieses variiert jedoch zwischen Regionen, Populationen und im Zeitverlauf sowie insbesondere auch zwischen verschiedenen Vergleichstherapien. Eine Risikodifferenz muss daher als beschreibendes Maß einer konkreten Studie, nicht als fixe Maßzahl eines Therapieverfahrens aufgefasst werden, ein Problem auch und vor allem für Meta-Analysen [N10]. Diese hohe Sensitivität für Rahmenbedingungen stellt die Übertragbarkeit von absoluten Effektmaßen aus Studien in die Versorgung in Frage. Daher ist es übliche Praxis, Effekte in klinischen Studien vorzugsweise als relatives Risiko, Odds Ratio bzw. Hazard (oder auch Incidence) Ratio auszudrücken [N2].
- Die Höhe der Risikodifferenz wird von der Höhe des Basisrisikos (absolutes Risiko in der Kontrollgruppe) begrenzt. Liegt dieses bei 1 %, dann kann die Risikodifferenz niemals über 0,01 liegen, beträgt es 10 %, dann nicht über 0,1 usw. Die Risikodifferenz könnte nur dann ihr Optimum 1 erreichen, wenn das Basisrisiko bei 100 % läge. Würde nun beispielsweise eine mindestens 20 %ige absolute Risikoreduktion als wesentliche therapeutische Verbesserung definiert, so wäre (für diese beispielhafte Forderung) bei

Erkrankungen mit (langfristigen) Überlebensraten > 80 % grundsätzlich kein erheblicher Zusatznutzens (für den entsprechenden Endpunkt) mehr darstellbar.

- Ein weiterer Nachteil der Verwendung von absoluten Risikoreduktionen als Effektmaß zur Operationalisierung zur Feststellung des Ausmaßes des Zusatznutzens ist zudem, dass ein genauer Zeitpunkt definiert werden muss, bei dem diese absolute Risikoreduktion bestimmt wird (z.B. nach 1, 2, 5 oder 10 Jahren), sofern es dazu keine allgemein akzeptierten Festlegungen gibt (z.B. 30-Tage-Mortalität bei Myokardinfarkt).

Zusammenfassend mögen absolute Risikoreduktionen in einer individuellen Entscheidungssituation eher handlungsleitend sein, für allgemeine Aussagen im Sinne einer Bewertung des Zusatznutzens eines Arzneimittels sind dagegen relative Effektmaße besser geeignet.

Den relativen Maßen ist gemeinsam, dass der Nulleffekt (kein Gruppenunterschied) bei 1 liegt. Im Folgenden wird auf Effekte unterhalb von 1 eingegangen. Hierauf können Effekte oberhalb von 1 durch Kehrwertbildung zurückgeführt werden. Das Konzept sieht vor, dass ein 95 %-Konfidenzintervall im Sinne einer verschobenen Hypothesengrenze einen jeweiligen Schwellenwert unterschreitet, damit das Ergebnis als erheblicher, beträchtlicher oder geringer Zusatznutzen eingestuft wird. Ein solches inferenzstatistisches Vorgehen hat gegenüber der Betrachtung von Punktschätzern zwei wesentliche Vorteile: (i) die Präzision der Schätzung fließt in die Bewertung ein; (ii) Die statistischen Irrtumsmöglichkeiten lassen sich damit einhergehend auf übliche kleine Werte (z. B. 5 %) beschränken.

Die Schwellenwerte variieren bezüglich der in Tabelle NT4 abgebildeten 2 Dimensionen Zielgrößenkategorie und Ausmaßkategorie des Effekts. Die Schwellenwerte sollten umso näher an 1 liegen (unterhalb von 1), je mehr Bedeutung einer Zielgröße zugemessen wird. Dadurch wird der Anforderung der AM-NutzenV nach einer Berücksichtigung der Krankheitsschwere Rechnung getragen. Dem gegenüber sollten die Schwellenwerte umso weiter weg von 1 liegen (unterhalb von 1), je stärker das Ausmaß des Effekts attestiert wird.

Der oben beschriebenen expliziten und abstrakten Operationalisierung folgend ist eine Rasterung der Schwellenwerte von 0,05 vorgesehen [N7]. Im Folgenden wird kurz die weitere Entwicklung der Methodik erläutert, die zu diesen Schwellenwerten geführt hat. Die weiteren Ausführungen werden aufzeigen, dass diese Wahl von 0,05 in der Praxis anwendbar ist und zu vernünftigen Aussagen führt.

Den Ausgangspunkt bildete die Frage, welche Größenordnung die tatsächlichen Effekte haben sollten, um z. B. das Ausmaß „erheblich“ innezuhaben. Dazu wurde ursprünglich ein relatives Risiko von 0,50 – von Djulbegovic et al. [N3] als Anforderung für einen „Durchbruch“ postuliert – als Effekt erheblichen Ausmaßes für die Zielgröße Gesamtmortalität verankert [N7].

Es stellte sich für diesen tatsächlichen Effekt (0,5) die Frage, wie der Schwellenwert gewählt werden muss, um mit einer adäquaten Power die Ausmaßkategorie „erheblich“ auch erreichen zu können. Die entsprechenden Überlegungen dazu können im Detail der ersten durch das Institut durchgeführten Dossierbewertung entnommen werden [N7], werden aber auch am

Ende dieses Anhangs noch einmal aufgegriffen. Sie führten dazu, dass für einen Schwellenwert von 0,85 die gleichzeitige Anforderung nach Realisierbarkeit und Stringenz als erfüllt angesehen werden kann.

Im nächsten Schritt mussten dann für die Ausmaßmatrix die übrigen tatsächlichen Effekte festgelegt und die dazugehörigen Schwellenwerte ermittelt werden. Dabei war zu beachten, dass die Anforderungen von der Zielgrößenkategorie „Mortalität“ ausgehend für weniger schwerwiegende Zielgrößen zunehmen und von der Ausmaßkategorie „erheblich“ ausgehend für niedrigere Ausmaßkategorien abnehmen sollten. Eine Rasterung von 1/6 für die tatsächlichen Effekte erwies sich dabei als pragmatische Lösung. Nachfolgend werden die Schwellenwerte für die jeweiligen Ausmaßkategorien beschrieben.

1. Gesamtmortalität

Jegliche zum üblichen Irrtumsniveau 5 % statistisch signifikante Verlängerung der Überlebensdauer wird zumindest als „geringer Zusatznutzen“ eingestuft, da für die Gesamtmortalität die Anforderung „mehr als geringfügig“ bereits durch den Endpunkt selbst als erfüllt angesehen wird. Demnach beträgt der auf das 95 %-Konfidenzintervall bezogene Schwellenwert hier 1. Als „beträchtlicher“ Effekt wird eine Verlängerung der Überlebensdauer bezeichnet, wenn ein Schwellenwert von 0,95 unterschritten wird. Als „erheblich“ wird eine Verlängerung der Überlebensdauer bewertet, wenn der Schwellenwert von 0,85 durch die obere Grenze des 95 %-Konfidenzintervalls unterschritten wird.

2.
- **schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen)**
- **schwerwiegende (bzw. schwere) Nebenwirkungen**
- **gesundheitsbezogene Lebensqualität**

Auch für schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und schwerwiegende (bzw. schwere) Nebenwirkungen entspricht jegliche statistisch signifikante Verminderung zumindest einem „geringen“ Effekt, weil die Anforderung „mehr als geringfügig“ bereits durch die Qualität des Endpunkts selbst erfüllt ist. Ein „beträchtlicher“ Effekt erfordert – in Abgrenzung zu gewünschten Effekten für die Gesamtmortalität – die Unterschreitung eines Schwellenwerts von 0,90. Voraussetzung für die Einstufung eines Effekts für diese Zielgrößen als „erheblich“ ist ebenfalls in Abgrenzung zu Effekten bezüglich der Gesamtmortalität die Unterschreitung eines Schwellenwerts von 0,75. Um aus diesen Zielgrößen einen erheblichen Effekt ableiten zu können, soll außerdem das Risiko für das untersuchte Ereignis in mindestens einer der zu vergleichenden Gruppen 5 % oder höher sein. Dieses zusätzliche Kriterium stützt die Relevanz des Ereignisses auf Populationsebene und trägt den besonderen Anforderungen an diese Kategorie des Zusatznutzens Rechnung.

Voraussetzung zur Feststellung des Ausmaßes des Effekts für Endpunkte zur gesundheitsbezogenen Lebensqualität ist, dass sowohl die eingesetzten Instrumente als auch die Responsekriterien validiert oder zumindest unzweifelhaft etabliert sind. Liegen solche Ergebnisse dichotom im Sinne von Respondern / Non-Respondern vor, gelten dieselben im

vorherigen Absatz genannten Kriterien (Risiko für die Kategorie „erheblich“ soll mindestens 5 % betragen) wie für schwerwiegende Symptome.

3. • nicht schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen)
• nicht schwerwiegende (bzw. schwere) Nebenwirkungen

Die Festlegung der Schwellenwerte für die nicht schwerwiegenden (bzw. schweren) Symptome (bzw. Folgekomplikationen) und die nicht schwerwiegenden (bzw. schweren) Nebenwirkungen berücksichtigt den im Vergleich zu den Kategorien 1 und 2 verminderten Schweregrad. Die Einstufung eines Effekts für diese Zielgrößen als „erheblich“ ist grundsätzlich nicht angezeigt. Voraussetzung für die Einstufung eines Effekts als „beträchtlich“ ist die Unterschreitung eines Schwellenwerts von 0,80. Ein „geringer Zusatznutzen“ erfordert die Unterschreitung eines Schwellenwerts von 0,90. Dies ist in der in § 5 Abs. 7 der AM-NutzenV formulierten Anforderung an einen geringen Zusatznutzen, dass es sich um eine moderate und nicht nur geringfügige Verbesserung handeln muss, begründet. Dem Verfahren ist somit implizit, dass (auch statistisch signifikante) Effekte, die aber nur als geringfügig bewertet werden, zur Einstufung in die Kategorie „kein Zusatznutzen“ führen.

In der folgenden Tabelle NT5 sind die jeweiligen Schwellenwerte für alle Ausmaßkategorien und Zielgrößenkategorien abgebildet.

Tabelle NT5: Inferenzstatistische Schwellenwerte (Hypothesengrenzen) für relative Effektmaße

		Zielgrößenkategorie		
		Gesamt-mortalität	Schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen und gesundheitsbezogene Lebensqualität[a]	Nicht schwerwiegende (bzw. nicht schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen
Ausmaßkategorie	Erheblich	0,85	0,75 und Risiko ≥ 5%[b]	Nicht besetzt
	Beträchtlich	0,95	0,90	0,80
	Gering	1,00	1,00	0,90

a: Voraussetzung ist die Verwendung eines validierten bzw. etablierten Instruments sowie eines validierten bzw. etablierten Responsekriteriums. Werte gelten für Non-Response.

b: Risiko muss für mindestens eine der beiden zu vergleichenden Gruppen mindestens 5% betragen.

Detaillierte methodische Rationale für die Festlegung der Schwellenwerte

Den Ausgangspunkt bildete die (fiktive) Planung einer Studie zur Testung der üblichen Hypothesen

$H_0: RR \geq RR_0 \quad vs. \quad H_1: RR < RR_0$

anhand des relativen Risikos mit $RR_0 = 1$. Durch die Festlegung des Signifikanzniveaus, der Power, des Risikos in der Kontrollgruppe und des tatsächlichen Effekts (RR_1) ergibt sich die benötigte Fallzahl.

Eine solche Studie hätte für alle gegenüber 1 verschobenen Hypothesengrenzen ($RR_0 < 1$) eine geringere Power. Um für eine interessierende verschobene Hypothesengrenze (die oben genannten Schwellenwerte) dieselbe Power zu erhalten, die zur Testung der üblichen (nichtverschobenen) Hypothesen festgelegt wurde, muss die Fallzahl erhöht werden, und zwar entweder innerhalb der Studie oder durch Kombination mehrerer Studien. Vom Regelfall des Vorliegens von zwei (z. B. pivotalen) Studien ausgehend, wurde eine Verdoppelung der Fallzahl angenommen. Die Hypothesengrenze für die verschobenen Hypothesen wurde dann gerade so gewählt, dass die Power aus zwei Einzelstudien zu den üblichen Hypothesen der Power der gemeinsamen (gepoolten) Analyse zu den verschobenen Hypothesen entspricht. Diese Hypothesengrenze diente als Schwellenwert für die obere Grenze des zweiseitigen 95 %-Konfidenzintervalls für das relative Risiko. Bei Vorgabe eines Signifikanzniveaus von 5 % (zweiseitig) und einer Power von 90 % (sowohl für die übliche als auch für die verschobene Hypothesengrenze), einer Verdopplung der Fallzahl für die verschobene Hypothesengrenze ergab sich z. B. für den für die Zielgröße „Mortalität“ und die Ausmaßkategorie „erheblich“ postulierten tatsächlichen Effekt von 0,5 ein Schwellenwert von (gerundet) 0,85.

Die im Anhang A der Nutzenbewertung zu Ticagrelor [N7] aufgeführte Formel für den Zusammenhang des tatsächlichen Effekts und des Schwellenwerts ist unabhängig von den sonstigen Vorgaben und beruht auf dem Algorithmus, der in der Prozedur „Power“ der Software SAS verwendet wird. In der entsprechenden Dokumentation für diesen Algorithmus [N9] wird auf die Arbeit von Fleiss et al. [N4] verwiesen. Ein Austausch mit Herrn Röhmel (damals Sprecher der Arbeitsgruppe Pharmazeutische Forschung der Deutschen Region der Internationalen Biometrischen Gesellschaft) sowie direkt mit dem Technical Support von SAS ergab, dass die Gültigkeit dieses Algorithmus offensichtlich nicht publiziert ist. Es stellte sich die Frage, welche tatsächlichen Effekte bei genauerer Berechnung notwendig sind, um mit einer hohen Wahrscheinlichkeit die jeweilige Ausmaßkategorie zu erreichen.

Die tatsächlichen Effekte wurden daher per Monte-Carlo-Simulationen ermittelt. In Abbildung NA1 sind die resultierenden (genaueren) tatsächlichen Effekte in Abhängigkeit des Risikos in der Kontrollgruppe für alle oben festgelegten Schwellenwerte aufgetragen (Signifikanzniveau 5 %; Power 90 %, Kurven geglättet).

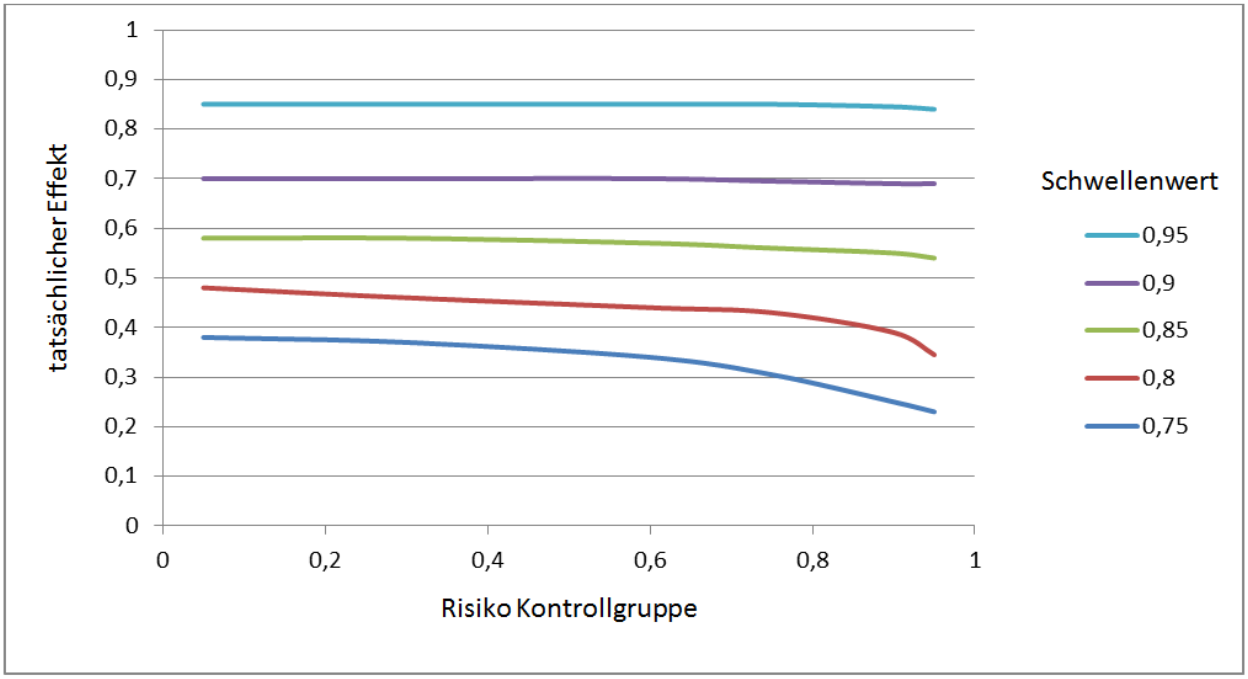


Abbildung NA1: Tatsächliche Effekte in Abhängigkeit des Basisrisikos

In Tabelle NT6 sind die Bereiche, in denen sich die tatsächlichen Effekte (in Abhängigkeit des Risikos der Kontrollgruppe) realisieren, pro Zielgrößen- und Ausmaßkategorie noch einmal eingetragen.

Tabelle NT6: Tatsächliche Effekte für das relative Risiko

		Zielgrößenkategorie		
		Gesamt-mortalität	Schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen und gesundheitsbezogene Lebensqualität	Nicht schwerwiegende (bzw. nicht schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen
Ausmaßkategorie	Erheblich	0,53 – 0,58	0,24 – 0,38	Entfällt
	Beträchtlich	0,84 – 0,85	0,69 – 0,71	0,34 – 0,48
	Gering	Entfällt	Entfällt	0,69 – 0,71

Bezogen auf die Gesamtmortalität sind für ein erhebliches Ausmaß tatsächliche relative Risiken im Bereich 0,55 – also weiterhin etwa einer Halbierung des Risikos entsprechend – zu veranschlagen. Für das Ausmaß „beträchtlich“ muss der tatsächliche Effekt bei etwa 0,85 liegen. Für schwerwiegende Symptome und gleichwertige Zielgrößen bedarf es für ein erhebliches Ausmaß tatsächlich einer Risikoreduktion auf etwa ein Viertel bis ein Drittel. Verglichen mit den ursprünglich veranschlagten tatsächlichen Effekten [N7] ergibt sich für die nahe an 1 liegenden Schwellenwerte eine gute Übereinstimmung. Bei den weiter von 1 entfernten Schwellenwerten zeigen die Simulationsergebnisse etwas moderatere Anforderungen an die Stärke der tatsächlichen Effekte. Die in Tabelle NT5 veranschlagte Rasterung der Schwellenwerte erscheint vernünftig und praktikabel.

Literatur

Literatur mit Nummerierung aus den Allgemeinen Methoden Version 4.0:

25. Atkins D, Best D, Briss PA, Eccles MP, Falck-Ytter Y, Flottorp S et al. Grading quality of evidence and strength of recommendations. BMJ 2004; 328(7454): 1490.

51. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol 2009; 62(8): 825-830, 830.e1-830.e10.

58. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007; 26(1): 53-77.

61. Bridges JF, Kinter ET, Kidane L, Heinzen RR, McCormick C. Things are looking up since we started listening to patients: trends in the application of conjoint analysis in health 1982-2007. Patient 2008; 1(4): 273-282.

64. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001; 20(6): 825-840.

65. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. J Clin Epidemiol 2008; 61(8): 763-769.

66. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive: trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2009; 38(1): 287-298.

72. 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.

110. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008. S. 243-296.

122. Deutscher Ethikrat (Ed). Nutzen und Kosten im Gesundheitswesen: zur normativen Funktion ihrer Bewertung; Stellungnahme. Berlin: Deutscher Ethikrat; 2011. URL: http://www.ethikrat.org/dateien/pdf/stellungnahme-nutzen-und-kosten-im-gesundheitswesen.pdf.

126. Devillé WL, Buntinx F, Bouter LM, Montori VM, De Vet HCW, Van der Windt DAWM et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002; 2: 9.

135. Dolan JG. Shared decision-making: transferring research into practice; the Analytic Hierarchy Process (AHP). Patient Educ Couns 2008; 73(3): 418-425.

136. Dolan JG, Isselhardt BJ Jr, Cappuccio JD. The Analytic Hierarchy Process in medical decision making: a tutorial. Med Decis Making 1989; 9(1): 40-50.

137. Dolan P, Shaw R, Tsuchiya A. The relative societal value of health gains to different beneficiares [online]. 01.2008 [Zugriff: 15.04.2013]. URL: http://www.hta.ac.uk/nihrmethodology/reports/1577.pdf.

138. Dolan P, Shaw R, Tsuchiya A, Williams A. QALY maximisation and people's preferences: a methodological review of the literature. Health Econ 2005; 14(2): 197-208.

150. Egger M, Davey Smith G, Altman DG. Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group; 2001.

154. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med 2000; 19(13-35): 1707-1728.

157. European Medicines Agency. 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.

188. Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol 2002; 31(1): 72-76.

193. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses [online]. 06.12.2012 [Zugriff: 15.04.2013]. URL: http://www.g-ba.de/downloads/62-492-667/VerfO_2012-12-06.pdf.

198. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. Stat Med 2002; 21(11): 1503-1511.

206. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989; 18(1): 269-274.

220. Hamza TH, Van Houwelingen HC, Heijenbrok-Kal MH, Stijnen T. Associating explanatory variables with summary receiver operating characteristic curves in diagnostic meta-analysis. J Clin Epidemiol 2009; 62(12): 1284-1291.

221. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 2008; 61(11): 1095-1103.

242. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21(11): 1539-1558.

243. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327(7414): 557-560.

253. Hummel JM, IJzerman MJ. The use oft the Analytic Hierarchy Process in health care decision making. Enschede: University of Twente; 2009.

273. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994; 120(8): 667-676.

274. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. Stat Med 2006; 25(15): 2688-2699.

300. Koch A, Ziegler S. Metaanalyse als Werkzeug zum Erkenntnisgewinn. Med Klin 2000; 95(2): 109-116.

326. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001; 322(7300): 1479-1480.

384. Nüesch E, Jüni P. Commentary: which meta-analyses are conclusive? Int J Epidemiol 2009; 38(1): 298-303.

409. Raum E, Perleth M. Methoden der Metaanalyse von diagnostischen Genauigkeitsstudien. Köln: Deutsches Institut für Medizinische Dokumentation und Information; 2003. (Health Technology Assessment; Band 2). URL: http://portal.dimdi.de/de/hta/hta_berichte/hta025_bericht_de.pdf.

410. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005; 58(10): 982-990.

425. Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. Stat Med 2009; 28(5): 721-738.

427. Saaty TL. A scaling method for priorities in hierarchical structures. J Math Psychol 1977; 15(3): 234-281.

428. Saaty TL. Decision making with the Analytic Hierarchy Process. International Journal of Services Sciences 2008; 1(1): 83-98.

429. Saaty TL, Vargas LG. The Analytic Hierarchy Process: wash criteria should not be ignored. International Journal of Management and Decision Making 2006; 7(2/3): 180-188.

457. Senn SJ. The many modes of meta. Drug Inf J 2000; 34(2): 535-549.

458. Senn SJ. Trying to be precise about vagueness. Stat Med 2007; 26(7): 1417-1430.

466. Simmonds MC, Higgins JPT. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. Stat Med 2007; 26(15): 2982-2999.

468. Skipka G, Bender R. Intervention effects in the case of heterogeneity between three subgroups: assessment within the framework of systematic reviews. Methods Inf Med 2010; 49(6): 613-617.

492. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21(11): 1559-1573.

493. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JPA, Thabane L et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009; 38(1): 276-286.

496. Tian L, Cai T, Pfeffer MA, Piankov N, Cremieux PY, Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 x 2 tables with all available data but without artificial continuity correction. Biostatistics 2009; 10(2): 275-281.

504. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 2000; 19(24): 3417-3432.

509. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002; 21(4): 589-624.

515. Villar J, Mackey ME, Carroli G, Donner A. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. Stat Med 2001; 20(23): 3635-3647.

520. Wailoo A, Tsuchiya A, McCabe C. Weighting must wait: incorporating equity concerns into cost-effectiveness analysis may take longer than expected. Pharmacoeconomics 2009; 27(12): 983-989.

528. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008; 61(1): 64-75.

Neue Literatur:

N1. Da Costa BR, Rutjes AWS, Johnston BC, Reichenbach S, Nüesch E, Tonia T et al. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. Int J Epidemiol 2012; 41(5): 1445-1459.

N2. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002; 21(11): 1575-1600.

N3. Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008; 168(6): 632-642.

N4. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 1980; 36(2): 343-346.

N5. Guddat C, Grouven U, Bender R, Skipka G. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Syst Rev 2012; 1: 34.

N6. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc 2009; 172(1): 137-159.

N7. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Ticagrelor: Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung; Auftrag A11-02 [online]. 29.09.2011 [Zugriff: 11.03.2013]. (IQWiG-Berichte; Band 96). URL: https://www.iqwig.de/download/A11-02_Ticagrelor_Nutzenbewertung_%C2%A735a_SGB_V_.pdf.

N8. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011; 342: d549.

N9. SAS Institute. SAS/STAT 9.2 user's guide: second edition [online]. 2009 [Zugriff: 11.04.2013]. URL: http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf.

N10. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses: sometimes informative, usually misleading. BMJ 1999; 318(7197): 1548-1551.

N11. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials 2007; 8: 16.