



Allgemeine Methoden

Version 6.0 vom 05.11.2020

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

ISBN 978-3-9815265-3-0

© 2020 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

www.iqwig.de

Anschrift des Herausgebers

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

Tel.: +49 221 35685-0

Fax: +49 221 35685-1

E-Mail: methoden@iqwig.de

Internet: www.iqwig.de

Präambel

Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) ist eine Einrichtung der Stiftung für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Das IQWiG ist ein fachlich unabhängiges wissenschaftliches Institut. Informationen zur Struktur und Organisation von Stiftung und Institut finden sich auf der Website www.iqwig.de.

Die „Allgemeinen Methoden“ erläutern die gesetzlichen und wissenschaftlichen Grundlagen des Instituts. Die Aufgaben des Instituts werden hier ebenso dargelegt wie die wissenschaftlichen Werkzeuge, die für die Bearbeitung der Institutsprodukte verwendet werden. Somit leistet das Methodenpapier des Instituts einen wichtigen Beitrag dazu, die Arbeitsweise des Instituts transparent zu machen.

Die „Allgemeinen Methoden“ richten sich vor allem an Fachwissenschaftlerinnen und Fachwissenschaftler. Um aber möglichst allen Interessierten einen Zugang zu Informationen über die Arbeitsweise des Instituts zu verschaffen, haben sich die Autorinnen und Autoren um Verständlichkeit bemüht. Wie jeder fachwissenschaftliche Text setzen allerdings auch die „Allgemeinen Methoden“ ein bestimmtes Maß an Vorwissen voraus.

Die „Allgemeinen Methoden“ sollen die Vorgehensweise des Instituts allgemein beschreiben. Welche konkreten einzelnen Schritte das Institut bei der Bewertung einer bestimmten medizinischen Maßnahme unternimmt, hängt unter anderem von der jeweiligen Fragestellung und von der vorhandenen wissenschaftlichen Evidenz ab. Insofern sind die „Allgemeinen Methoden“ als eine Art Rahmen zu verstehen. Wie der Bewertungsprozess im konkreten Einzelfall ausgestaltet ist, wird projektspezifisch detailliert dargelegt.

Die Methoden des Instituts werden in der Regel jährlich auf eine notwendige Überarbeitung hin überprüft, es sei denn, Fehler im Dokument oder wesentliche Entwicklungen legen eine vorzeitige Aktualisierung nahe. Die projektspezifische Methodik wird auf Basis der zum jeweiligen Zeitpunkt gültigen Methodenversion festgelegt. Ergeben sich im Projektverlauf Änderungen des allgemeinen methodischen Vorgehens, wird geprüft, ob das projektspezifische Vorgehen entsprechend anzupassen ist. Um seine Arbeitsweise kontinuierlich weiterzuentwickeln und zu verbessern, stellt das Institut seine „Allgemeinen Methoden“ öffentlich zur Diskussion. Für die jeweils gültige Fassung gilt das ebenso wie für Entwürfe der folgenden Versionen.

Was ist neu?

In diesem Entwurf für die Version 6.0 wurden im Vergleich zur Version 5.0 der „Allgemeinen Methoden“ des Instituts vom 10.07.2017 kleinere Fehler beseitigt und redaktionelle Änderungen sowie ein Literaturupdate durchgeführt. Es haben folgende inhaltliche und strukturelle Änderungen stattgefunden:

- Ergänzung des Abschnitts 1.2 zum Generalauftrag
- Änderungen in Abschnitt 2.1.1 bezüglich des Wegfalls der Anhörung zum Berichtsplan bei Aufträgen zur Bewertung von Untersuchungs- und Behandlungsmethoden nach § 139a Abs. 3 Nr. 1 SGB V sowie zur Anpassung von Veröffentlichungsfristen nach Fertigstellung von Institutsprodukten
- Ergänzungen in Abschnitt 3.1.3 zur Darstellung von Schadensaspekten in Nutzenbewertungen
- Ergänzungen in Abschnitt 3.2.1 zur Auswirkung nicht publizierter Studienergebnisse auf das Fazit
- Ergänzungen in Abschnitt 3.4 zu nicht medikamentösen therapeutischen Interventionen
- Ergänzungen in Abschnitt 3.5 zu diagnostischen Verfahren
- Ergänzungen in Abschnitt 3.8 zur Potenzialbewertung und §-137h-Bewertung
- Umbenennung von Kapitel 5 in „Bewertungen zur Versorgung“ und grundlegende Überarbeitung des Kapitels
- Änderung des Abschnitts 5.1.4 hinsichtlich der Bewertung der methodischen Leitlinienqualität
- Ergänzung des Abschnitts 5.1.5 zur strukturierten Aufbereitung von Empfehlungen
- Überarbeitung des Abschnitts 5.1.6 zur strukturierten Informationssynthese
- Ergänzung des Abschnitts 5.2 zu Mindestmengenregelungen
- Streichung des Abschnitts zur Versorgungsanalyse aus Kapitel 5
- Ergänzung des Abschnitts 5.3 zur Analyse von Versorgungsdaten
- Ergänzungen in Abschnitt 6.3.2 zur Vorabrecherche zu den für eine HTA-Berichterstellung geeigneten Themen
- Ergänzungen in Abschnitt 6.3.4 zur Diskussion der Themen im erweiterten Fachbeirat
- Ergänzung des Abschnitts 6.3.5 zum Umgang mit vom Auswahlbeirat nominierten, aber vom Institut nicht für die Erstellung eines HTA-Berichts ausgewählten Themen
- Ergänzung in Abschnitt 8.1 zu Spracheinschränkungen bei umfassender Informationsbeschaffung

- Aktualisierung zur Datenbank AMIS (nun AMIce) in Abschnitt 8.1.2
- Ergänzung der HTA Database in Abschnitt 8.2.1
- Löschung der Datenbank PharmNet.Bund in Abschnitt 8.2.5
- Änderungen in Abschnitt 9.3.3 zur Beurteilung der klinischen Relevanz
- Löschung des Abschnitts 9.3.4 zur Bewertung subjektiver Endpunkte bei offenem Studiendesign
- Ergänzungen in Abschnitt 9.3.7 zur Anwendung der Knapp-Hartung-Methode bei Metaanalysen mit zufälligen Effekten und zu Metaanalysen von Studien zur diagnostischen Güte
- Ergänzung des Abschnitts 9.3.11 zum Umgang mit unvollständigen Daten
- Ergänzung des Abschnitts 9.3.12 zum Umgang mit variablen Beobachtungszeiten

Inhaltsverzeichnis

Präambel	ii
Was ist neu?	iii
Tabellenverzeichnis	xii
Abbildungsverzeichnis	xiii
Abkürzungsverzeichnis.....	xiv
1 Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen	1
1.1 Gesetzliche Aufgaben.....	1
1.2 Generalauftrag des G-BA.....	4
1.3 Evidenzbasierte Medizin	5
1.3.1 Praktische evidenzbasierte Medizin	6
1.3.2 Bedeutung von EbM für das Institut	6
1.3.3 Die Strategien der EbM	7
1.3.4 Die Bedeutung der Ergebnissicherheit	8
1.3.5 Die Verbindung zwischen Ergebnissicherheit und Alltagsnähe	9
1.3.6 Nutzen im Einzelfall.....	11
1.4 Gesundheitsökonomie	11
1.4.1 Bedeutung von Gesundheitsökonomie für das Institut.....	12
1.4.2 Die internationalen Standards der Gesundheitsökonomie.....	13
1.4.3 Methodische Standards in der Gesundheitsökonomie.....	14
1.5 Evidenzbasierte Entscheidungsfindung im Gesundheitswesen	15
2 Produkte des Instituts	16
2.1 Produktspezifische Verfahrensabläufe	16
2.1.1 Bericht	18
2.1.2 Rapid Report.....	21
2.1.3 Dossierbewertung	23
2.1.4 Kosten-Nutzen-Bewertung nach § 35b SGB V	24
2.1.5 Potenzialbewertung	27
2.1.6 Bewertung gemäß § 137h SGB V	28
2.1.7 Addendum	28
2.1.8 Gesundheitsinformationen.....	29
2.1.9 Arbeitspapier	32
2.1.10 HTA-Bericht.....	33
2.2 Generelle Aspekte bei der Produkterstellung	35
2.2.1 Einbindung von Betroffenen in die Erstellung von IQWiG-Produkten	35

2.2.2	Auswahl externer Sachverständiger	37
2.2.3	Gewährleistung der fachlichen Unabhängigkeit	38
2.2.4	Review der Produkte des Instituts	39
2.2.5	Stellungnahmeverfahren.....	40
2.2.6	Veröffentlichung der Produkte des Instituts.....	41
2.2.7	Wissenschaftliche Beratung	41
3	Nutzenbewertung medizinischer Interventionen.....	42
3.1	Patientenrelevanter medizinischer Nutzen und Schaden.....	42
3.1.1	Definition des patientenrelevanten medizinischen Nutzens bzw. Schadens	42
3.1.2	Surrogate patientenrelevanter Endpunkte.....	44
3.1.3	Ermittlung des Schadens medizinischer Interventionen.....	46
3.1.4	Endpunktbezogene Bewertung	49
3.1.5	Zusammenfassende Bewertung.....	54
3.2	Spezielle Aspekte der Nutzenbewertung.....	55
3.2.1	Auswirkung nicht publizierter Studienergebnisse auf das Fazit	55
3.2.2	Dramatischer Effekt	57
3.2.3	Studiendauer	58
3.2.4	Patientenberichtete Endpunkte	59
3.2.5	Nutzen und Schaden in kleinen Populationen.....	59
3.3	Nutzenbewertung von Arzneimitteln	60
3.3.1	Stellenwert des Zulassungsstatus	61
3.3.2	Studien zur Nutzenbewertung von Arzneimitteln	62
3.3.3	Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V	63
3.4	Nicht medikamentöse therapeutische Interventionen	67
3.5	Diagnostische Verfahren	69
3.6	Früherkennung und Screening.....	74
3.7	Prävention.....	75
3.8	Potenzialbewertung und §-137h-Bewertung	76
3.8.1	Potenzialbewertung	76
3.8.2	§-137h-Bewertung.....	78
4	Kosten-Nutzen-Bewertung medizinischer Interventionen.....	80
4.1	Einführung.....	80
4.1.1	Rechtliche Grundlagen für die Kosten-Nutzen-Bewertung nach SGB V	80
4.1.2	Perspektive	80
4.1.3	Zeithorizont	81
4.1.4	Auswahl der Komparatoren.....	81
4.1.5	Versorgungspfad	81

4.1.6	Modell	82
4.1.7	Spezifische Datenquellen der Gesundheitsökonomie	82
4.1.8	Unsicherheit.....	83
4.1.9	Interpretation der Ergebnisse.....	83
4.2	Modellierung.....	85
4.2.1	Grundlagen	85
4.2.2	Grundsätzliche Aspekte der Modellentwicklung	86
4.2.3	Einflussdiagramm und Modellkonzept	87
4.2.4	Datengrundlage	87
4.2.5	Auswahl der Modellierungstechnik.....	87
4.2.6	Modelldokumentation und Modellvalidierung.....	88
4.3	Nutzen	90
4.3.1	Übertragung und Darstellung des Nutzens.....	90
4.3.2	Endpunkte.....	91
4.3.3	Maß des Gesamtnutzens	91
4.3.4	Datengrundlage	95
4.3.5	Unsicherheit und Verteilung von Nutzendaten	95
4.4	Kosten.....	95
4.4.1	Perspektive und zu berücksichtigende Kosten	95
4.4.2	Abgrenzung der Kosten.....	97
4.4.3	Schritte zur Kostenbestimmung	99
4.4.4	Datengrundlage	102
4.4.5	Unsicherheit und Verteilung von Kostendaten	102
4.4.6	Inflationsbereinigung und Diskontierung.....	103
4.5	Epidemiologische Daten	103
4.5.1	Daten	103
4.5.2	Datengrundlage	103
4.5.3	Unsicherheit und Verteilung von epidemiologischen Daten.....	104
4.6	Ergebnisdarstellung als Effizienzgrenze.....	104
4.6.1	Definition.....	104
4.6.2	Verfahrensverlauf	104
4.6.3	Konstruktion der Effizienzgrenze	107
4.6.4	Sonderkonstellationen	109
4.7	Unsicherheit (Sensitivitätsanalysen).....	110
4.7.1	Quantifizierung von Unsicherheit	110
4.7.2	Sensitivitätsanalysen	110
4.7.3	Darstellung von Unsicherheit mittels des Net Health Benefit.....	111

4.8 Ausgaben-Einfluss-Analyse (Budget-Impact-Analyse)	111
4.8.1 Perspektive in der Ausgaben-Einfluss-Analyse	111
4.8.2 Zeithorizont in der Ausgaben-Einfluss-Analyse	112
4.8.3 Szenarien in der Ausgaben-Einfluss-Analyse	112
4.8.4 Population in der Ausgaben-Einfluss-Analyse	112
4.8.5 Einzuschließende Kosten in der Ausgaben-Einfluss-Analyse	112
4.8.6 Ergebnisdarstellung in der Ausgaben-Einfluss-Analyse	113
4.9 Besonderheiten der Kosten-Nutzen-Bewertung nach § 35b SGB V	113
4.9.1 Gesetzliche Anforderungen und Verfahrensablauf	113
4.9.2 Der Net Health Benefit zur Berechnung zusatznutzenbasierter Erstattungspreise	114
4.9.3 Sensitivitätsanalysen zur Berechnung zusatznutzenbasierter Erstattungspreise	115
4.9.4 Interquartilsregion als Maß der Streuung für Preisverhandlungen	115
5 Bewertungen zur Versorgung	117
5.1 Leitliniensynopsen	117
5.1.1 Hintergrund	117
5.1.2 Evidenzbasierte Leitlinien	117
5.1.3 Übertragbarkeit auf das deutsche Gesundheitssystem	117
5.1.4 Bewertung der methodischen Leitlinienqualität	118
5.1.5 Strukturierte Aufbereitung von Empfehlungen: Evidenz- und Empfehlungsstufen	119
5.1.6 Strukturierte Informationssynthese: Extraktion und Analyse der Empfehlungen	120
5.2 Mindestmengenregelungen	121
5.2.1 Hintergrund	121
5.2.2 Informationsgrundlage und Bewertung	122
5.3 Analyse von Versorgungsdaten	122
5.3.1 Hintergrund	122
5.3.2 Ziele einer Analyse von Versorgungsdaten	123
5.3.3 Inhaltliche Aspekte einer Analyse von Versorgungsdaten	123
5.3.4 Datenquellen	124
5.3.5 Methodische Besonderheiten der Analyse von Versorgungsdaten	125
5.3.6 Ergebnisdarstellung	126
6 HTA-Berichte	127
6.1 Hintergrund und Ziel	127
6.2 Themensammlung	127
6.3 Auswahl der Themen für die HTA-Berichte	127
6.3.1 Auswahlkriterien	127

6.3.2	Prüfung der Fragestellung und Themenaufbereitung	128
6.3.3	1. Stufe des Auswahlverfahrens: Nominierung von Themen.....	128
6.3.4	2. Stufe des Auswahlverfahrens: Auswahl von Themen, zu denen HTA-Berichte erstellt werden	129
6.3.5	Umgang mit nominierten, aber nicht für eine HTA-Erstellung ausgewählten Themen	129
6.4	Sicherstellung der Qualität der HTA-Berichte	129
6.5	Bearbeitung der Themen (HTA-Berichte).....	129
6.5.1	Nutzenbewertung.....	130
6.5.2	Gesundheitsökonomie	130
6.5.3	Ethik	130
6.5.4	Soziale Aspekte	130
6.5.5	Rechtliche Aspekte	131
6.5.6	Organisatorische Aspekte	131
7	Evidenzbasierte Gesundheitsinformation für Bürgerinnen und Bürger	132
7.1	Ziele der Gesundheitsinformationen.....	132
7.2	Themenauswahl und Identifizierung der Informationsbedürfnisse	134
7.2.1	Themenkatalog gemäß dem gesetzlichen Auftrag	134
7.2.2	Identifizierung der Informationsbedürfnisse / Aufbereitung von Krankheitserfahrungen	134
7.2.3	Patientenwege.....	135
7.3	Informationsbeschaffung zur Erstellung von Gesundheitsinformationen	137
7.4	Auswahl der Evidenz	138
7.5	Auswahl der dargestellten Ergebnisse (Endpunkte)	139
7.6	Wahl und Darstellung von Vergleichen	139
7.7	Umgang mit Zahlen und Risikoangaben	139
7.8	Berücksichtigung von Alters- und Geschlechtsunterschieden.....	141
7.9	Anpassung an die Zielgruppe	142
7.9.1	Beteiligung von Betroffenen	142
7.9.2	Nicht öffentliches Stellungnahmeverfahren	143
7.9.3	Nutzertestung.....	143
7.9.4	Kommentare von Nutzerinnen und Nutzern	144
7.9.5	Informationszugang und Barrierefreiheit	144
7.10	Sachlich angemessene Darstellung	144
7.11	Ableitung von Bewertungen und Empfehlungen	145
7.12	Vorgehen bei der Erstellung von Entscheidungshilfen	145
7.13	Transparenz bezüglich der Verfasserin / des Verfassers und der Herausgeberin / des Herausgebers	145

7.14 Offenlegung von Beziehungen.....	145
7.15 Benennung von Wirkstoffen, medizinischen Methoden und Geräten	145
7.16 Beschreibung der typischen Formate und Inhalte.....	146
7.16.1 Ergänzende Formate	146
7.16.2 Erfahrungsberichte	147
7.16.3 Website.....	149
7.17 Aktualisierung der Inhalte	149
7.18 Aktualisierung der Methoden von gesundheitsinformation.de	150
8 Informationsbeschaffung	151
8.1 Umfassende Informationsbeschaffung.....	151
8.1.1 Recherche in bibliografischen Datenbanken	152
8.1.2 Suche in Studienregistern	155
8.1.3 Herstelleranfragen	156
8.1.4 Weitere Informationsquellen und Suchtechniken	157
8.2 Fokussierte Informationsbeschaffung.....	159
8.2.1 Recherche nach systematischen Übersichten	159
8.2.2 Recherche nach qualitativer Forschung.....	160
8.2.3 Recherche für gesundheitsökonomische Fragestellungen.....	160
8.2.4 Recherchen zu Addenda von §-137e- oder §-137h-Bewertungen	161
8.2.5 Überprüfung der Vollständigkeit eines Studienpools	161
8.3 Orientierende Recherchen	161
8.4 Recherche nach Leitlinien zur Erstellung von Leitliniensynopsen	162
8.5 Prüfung der Informationsbeschaffung.....	162
9 Informationsbewertung.....	164
9.1 Qualitätsbewertung von Einzelstudien	164
9.1.1 Kriterien für den Einschluss von Studien.....	164
9.1.2 Zusammenhang zwischen Studientyp / -art und Fragestellung.....	165
9.1.3 Rangordnung verschiedener Studienarten / Evidenzgrade.....	166
9.1.4 Aspekte der Bewertung des Verzerrungspotenzials	167
9.1.5 Interpretation von kombinierten Endpunkten.....	169
9.1.6 Bewertung der Konsistenz von Daten	170
9.2 Berücksichtigung systematischer Übersichten.....	171
9.2.1 Einordnung systematischer Übersichten	171
9.2.2 Nutzenbewertung auf Basis systematischer Übersichten	172
9.2.3 Berücksichtigung publizierter Metaanalysen	175
9.3 Spezielle biometrische Aspekte	175
9.3.1 Darstellung von Effekten und Risiken	175

9.3.2	Beurteilung statistischer Signifikanz	176
9.3.3	Beurteilung klinischer Relevanz.....	177
9.3.4	Nachweis der Verschiedenheit	180
9.3.5	Nachweis der Gleichheit	181
9.3.6	Adjustierung und multifaktorielle Verfahren	182
9.3.7	Metaanalysen.....	183
9.3.8	Indirekte Vergleiche	188
9.3.9	Subgruppenanalysen.....	190
9.3.10	Umgang mit nicht oder nicht vollständig publizierten Daten	193
9.3.11	Umgang mit unvollständigen Daten.....	193
9.3.12	Umgang mit variablen Beobachtungszeiten.....	195
9.3.13	Darstellung von Verzerrungsarten.....	196
9.3.14	Auswertung abhängiger Daten	199
9.4	Qualitative Methoden	199
9.4.1	Qualitative Forschung bei der Erstellung von Gesundheitsinformationen.....	200
9.4.2	Qualitative Studien bei der Erstellung anderer IQWiG-Produkte.....	201
Anhang A – Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens.....		202
Literaturverzeichnis.....		214

Tabellenverzeichnis

	Seite
Tabelle 1: Übersicht über die Produkte des Instituts.....	18
Tabelle 2: Institutsprodukte und Formen der möglichen Einbindung Betroffener	37
Tabelle 3: Regelmäßig abgeleitete Aussagesicherheiten für verschiedene Evidenzsituationen beim Vorliegen von Studien derselben qualitativen Ergebnissicherheit	53
Tabelle 4: Szenarien der Datenvollständigkeit und Konsequenzen für das Fazit einer Nutzenbewertung.....	56
Tabelle 5: Schwellenwerte zur Ausmaßbestimmung für das relative Risiko.....	66
Tabelle 6: Konzepte von Unsicherheit in der gesundheitsökonomischen Entscheidungsanalyse	83
Tabelle 7: Perspektive und relevante zu berücksichtigende Kosten	96
Tabelle 8: Mögliche Datenquellen für die Analyse von Versorgung.....	125
Tabelle 9: Verschiedene Dimensionen eines Patientenwegs	137
Tabelle 10: Feststellung des Ausmaßes des Zusatznutzens – Kriterien gemäß AM- NutzenV.....	203
Tabelle 11: Feststellung des Ausmaßes des Zusatznutzens – Kriterien gemäß AM- NutzenV mit Ergänzungen	205
Tabelle 12: Feststellung des Ausmaßes des Zusatznutzens – hierarchisierte Kriterien gemäß AM-NutzenV mit Ergänzungen	206
Tabelle 13: Inferenzstatistische Schwellenwerte (Hypothesengrenzen) für relative Effektmaße.....	210
Tabelle 14: Tatsächliche Effekte für das relative Risiko	213

Abbildungsverzeichnis

	Seite
Abbildung 1: Ablauf der Erstellung eines Berichts	20
Abbildung 2: Ablauf der Erstellung eines Rapid Reports.....	22
Abbildung 3: Ablauf der Erstellung einer Dossierbewertung.....	23
Abbildung 4: Ablauf der Kosten-Nutzen-Bewertung nach § 35b SGB V	26
Abbildung 5: Ablauf der Erstellung einer Potenzialbewertung	27
Abbildung 6: Ablauf der Erstellung eines Addendums	29
Abbildung 7: Ablauf der Neuerstellung einer Gesundheitsinformation	31
Abbildung 8: Ablauf der Erstellung eines Arbeitspapiers	33
Abbildung 9: Ablauf der Erstellung eines HTA-Berichts.....	34
Abbildung 10: Darstellung der entscheidungsrelevanten Bereiche	84
Abbildung 11: Interpretation der Steigung der theoretischen Effizienzgrenze.....	105
Abbildung 12: Absolute versus erweiterte Dominanz	106
Abbildung 13: Darstellung der Effizienzgrenze.....	109
Abbildung 14: Darstellung eines $NHB > 0$	115
Abbildung 15: Interquartilsregion möglicher zusatznutzenbasierter Erstattungspreise (basierend auf PSAs) als Maß der Streuung für Preisverhandlungen.....	116
Abbildung 16: Schematische Darstellung des mehrstufigen Auswahlverfahrens	127
Abbildung 17: Tatsächliche Effekte in Abhängigkeit des Basisrisikos	212

Abkürzungsverzeichnis

Abkürzung	Definition
AEA	Ausgaben-Einfluss-Analyse
AGREE	Appraisal of Guidelines Research and Evaluation
AHP	Analytic-Hierarchy-Process
AHRQ	Agency for Healthcare Research and Quality
AMG	Arzneimittelgesetz
AMIce	Arzneimittel-Informationssystem
AMNOG	Arzneimittelmarktneuordnungsgesetz
AM-NutzenV	Arzneimittel-Nutzenbewertungsverordnung
AMSTAR	A Measurement Tool to Assess Systematic Reviews
ANP	Analytic Network Process (analytischer Netzwerkprozess)
AOK	Allgemeine Ortskrankenkasse
AWMF	Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V.
BfArM	Bundesinstitut für Arzneimittel und Medizinprodukte
BITV	Barrierefreie-Informationstechnik-Verordnung
BMG	Bundesministerium für Gesundheit
CA	Conjoint-Analyse
CMA	Canadian Medical Association
CONSORT	Consolidated Standards of Reporting Trials
CTCAE	Common Terminology Criteria for Adverse Events
DALY	Disability-adjusted Life Year (um Behinderungen adjustiertes Lebensjahr)
DCE	Discrete-Choice-Experiment
DIPEx	Database of Individual Patient Experience
DMP	Disease-Management-Programm
DMP-A-RL	DMP-Anforderungen-Richtlinie
DRG	Diagnosis-related Group
EbM	evidenzbasierte Medizin
EBM	Einheitlicher Bewertungsmaßstab
EMA	European Medicines Agency (Europäische Arzneimittel-Agentur)
EU-CTR	EU Clinical Trials Register
FDA	Food and Drug Administration (behördliche Lebensmittelüberwachungs- und Arzneimittelzulassungsbehörde der USA)
G-BA	Gemeinsamer Bundesausschuss

Abkürzung	Definition
GCP	Good clinical Practice (gute klinische Praxis)
GEP	Gute Epidemiologische Praxis
GKV	gesetzliche Krankenversicherung
GKV-Spitzenverband	Spitzenverband Bund der Krankenkassen
GoR	Grade of Recommendation
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HCQI	Health Care Quality Indicator
HON	Health on the Net Foundation
HONcode	HON Code of Conduct
HTA	Health Technology Assessment
HVPI	harmonisierter Verbraucherpreisindex
HYE	Healthy Years Equivalent
ICD	International Statistical Classification of Diseases and Related Health Problems (Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme)
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
ICTRP	International Clinical Trials Registry Platform Search Portal
IFA	Informationsstelle für Arzneispezialitäten
IKNV	inkrementelles Kosten-Nutzen-Verhältnis
IPD	Individual Patient Data (individuelle Patientendaten)
IPDAS	International Patient Decision Aid Standards
IQR	Interquartilsregion
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ISOQOL	International Society of Quality of Life Research
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
KNB	Kosten-Nutzen-Bewertung
LoE	Level of Evidence
MAUI	Multi-Attribute-Utility-Instrument
MBVerfV	Methodenbewertungsverfahrensverordnung
MDR	Medical Device Regulation (EU-Medizinprodukte-Verordnung)
MedDRA	Medical Dictionary for Regulatory Activities (Medizinisches Wörterbuch für Aktivitäten im Rahmen der Arzneimittellzulassung)
MID	Minimal important Difference

Abkürzung	Definition
MORE	McMaster Online Rating of Evidence
MPG	Medizinproduktegesetz
MTC	Mixed Treatment Comparison
MTM	Multiple-Treatments Meta-Analysis
NHB	Net Health Benefit
NHC	National Health Committee
NICE	National Institute for Health and Care Excellence
NUB	neue Untersuchungs- und Behandlungsmethoden
NVL	Nationale VersorgungsLeitlinie
NVL-Programm	Programm für Nationale VersorgungsLeitlinien
OECD	Organisation for Economic Co-operation and Development (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung)
OR	Odds Ratio
PICOS	Population, Intervention, Comparison, Outcome, Study Design
POP	Planned and Ongoing Projects
PRESS	Peer Review of Electronic Search Strategies
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRO	Patient-reported Outcome (patientenberichteter Endpunkt)
PROSPERO	Prospective Register of Systematic Reviews
PSA	probabilistische Sensitivitätsanalyse
PT	Preferred Term (bevorzugter Begriff)
QALY	Quality-adjusted Life Year (qualitätsadjustiertes Lebensjahr)
RCT	Randomized controlled Trial (randomisierte kontrollierte Studie)
RD	Risikodifferenz
RKI	Robert Koch-Institut
ROBINS-I	Risk of Bias in non-randomized Studies – of Interventions
ROBIS	Risk of Bias in Systematic Reviews
ROC	Receiver Operating Characteristic
RR	relatives Risiko
SAP	statistischer Analyseplan
SAVE	Saved young Life Equivalent
SGB V	Sozialgesetzbuch Fünftes Buch – Gesetzliche Krankenversicherung

Abkürzung	Definition
SMD	Standardized Mean Difference (standardisierte Mittelwertdifferenz)
SMQ	Standardized MedDRA Query
SMDM	Society for Medical Decision Making
STARD	Standards for Reporting of Diagnostic Accuracy
STE	Surrogate-Threshold-Effekt
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
STROSA	Standardisierte Berichtsroutine für Sekundärdatenanalysen
TREND	Transparent Reporting of Evaluations with Non-randomized Designs
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
UK NSC	United Kingdom National Screening Committee
US PSTF	United States Preventive Services Task Force
WHO	World Health Organization (Weltgesundheitsorganisation)
WIdO	Wissenschaftliches Institut der AOK
ZE	Zusatzentgelt

Eine Hauptursache der Armut in den Wissenschaften ist meist eingebildeter Reichtum. Es ist nicht ihr Ziel, der unendlichen Weisheit eine Tür zu öffnen, sondern eine Grenze zu setzen dem unendlichen Irrtum.

Bertolt Brecht. Leben des Galilei. Frankfurt: Suhrkamp. Uraufführung, erste Version 1943, Schauspielhaus Zürich.

1 Das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

1.1 Gesetzliche Aufgaben

Das Institut wurde im Zuge der Gesundheitsreform 2004 [171] als Einrichtung der Stiftung für Qualität und Wirtschaftlichkeit im Gesundheitswesen gegründet. Die gesetzlichen Grundlagen und Aufgaben des Instituts sind im Sozialgesetzbuch Fünftes Buch – Gesetzliche Krankenversicherung (SGB V) [3] verankert und wurden im Zuge weiterer Gesundheitsreformen mehrfach angepasst und erweitert. Informationen zur Struktur und Organisation des Instituts sind auf der Website www.iqwig.de verfügbar.

Das Institut wird zu Fragen von grundsätzlicher Bedeutung für die Qualität und Wirtschaftlichkeit der im Rahmen der gesetzlichen Krankenversicherung erbrachten Leistungen tätig. Die spezifischen Aufgaben sind in § 139a SGB V näher benannt:

- Recherche, Darstellung und Bewertung des aktuellen medizinischen Wissensstandes zu diagnostischen und therapeutischen Verfahren bei ausgewählten Krankheiten,
- Erstellung von wissenschaftlichen Ausarbeitungen, Gutachten und Stellungnahmen zu Fragen der Qualität und Wirtschaftlichkeit der im Rahmen der gesetzlichen Krankenversicherung erbrachten Leistungen unter Berücksichtigung alters-, geschlechts- und lebenslagenspezifischer Besonderheiten,
- Bewertungen evidenzbasierter Leitlinien für die epidemiologisch wichtigsten Krankheiten,
- Abgabe von Empfehlungen zu Disease-Management-Programmen,
- Bewertung des Nutzens und der Kosten von Arzneimitteln,
- Bereitstellung von für alle Bürgerinnen und Bürger verständlichen allgemeinen Informationen zur Qualität und Effizienz in der Gesundheitsversorgung sowie zur Diagnostik und Therapie von Krankheiten mit erheblicher epidemiologischer Bedeutung,
- Beteiligung an internationalen Projekten zur Zusammenarbeit und Weiterentwicklung im Bereich der evidenzbasierten Medizin.

Die Beauftragung und Wahrnehmung der Aufgaben werden in § 139b SGB V näher geregelt. Danach können nur der Gemeinsame Bundesausschuss (G-BA) und das Bundesministerium für Gesundheit (BMG) das Institut beauftragen. Das Institut kann einen Antrag des BMG als

unbegründet ablehnen, es sei denn, das Ministerium übernimmt die Finanzierung der Bearbeitung.

Das Institut hat darauf zu achten, dass externe Sachverständige in die Auftragsbearbeitung einbezogen werden. Diese haben zur Sicherstellung der fachlichen Unabhängigkeit des Instituts alle Beziehungen zu Interessenverbänden, Auftragsinstituten, insbesondere der pharmazeutischen Industrie und der Medizinprodukteindustrie, einschließlich der Art und Höhe von Zuwendungen offenzulegen (siehe Abschnitt 2.2.3).

Die Arbeitsergebnisse zu Aufträgen des G-BA leitet das Institut dem G-BA als Empfehlungen zu. Der G-BA hat laut Gesetz die Empfehlungen im Rahmen seiner Entscheidungen zu berücksichtigen.

Der G-BA kann Aufträge an das Institut zurücknehmen, ruhend stellen oder ändern. Das Vorgehen ist im 1. Kapitel, § 16b Abs. 3 der Verfahrensordnung des G-BA beschrieben [276]. Demnach erfolgen diese Entscheidungen in Abstimmung mit dem Institut.

Das Institut wird zum weitaus überwiegenden Teil aus den Beiträgen der Mitglieder der gesetzlichen Krankenversicherungen finanziert. Dazu werden nach § 139c SGB V vom G-BA Zuschläge festgelegt. Diese Zuschläge werden von allen deutschen Arztpraxen und Krankenhäusern abgeführt, die gesetzlich versicherte Patientinnen und Patienten behandeln.

Im Rahmen des Arzneimittelmarktneuordnungsgesetzes (AMNOG) wurden Anfang 2011 die Aufgaben des Instituts um die Bewertung des Nutzens von Arzneimitteln mit neuen Wirkstoffen kurz nach der Markteinführung erweitert [172]. Dazu müssen Herstellerinnen und Hersteller Dossiers einreichen, die die Ergebnisse aus Studien zusammenfassen. Zuständig für die frühe Nutzenbewertung ist der G-BA; er hat jedoch die Möglichkeit, das Institut oder Dritte mit der Prüfung und Bewertung der Dossiers zu beauftragen.

Grundlage sind die neu gefassten Regelungen des § 35a SGB V. Sie werden ergänzt durch eine ebenfalls seit Anfang 2011 geltende Rechtsverordnung des BMG [106] und die Verfahrensordnung des G-BA [276].

Verbunden mit einer Nutzenbewertung kann der G-BA das Institut auch mit einer Kosten-Nutzen-Bewertung beauftragen. Die Rahmenbedingungen dieser Kosten-Nutzen-Bewertungen geben § 35b SGB V und § 139a SGB V vor. Dabei werden Kosten-Nutzen-Verhältnisse medizinischer Technologien gegenübergestellt mit dem Ziel, Informationen bereitzustellen, auf deren Grundlage die Angemessenheit und die Zumutbarkeit einer Kostenübernahme durch die Versichertengemeinschaft berücksichtigt werden können.

Die Kosten-Nutzen-Bewertung selbst basiert auf einem Vergleich mit anderen Arzneimitteln oder nicht medikamentösen Behandlungsformen. Als Kriterien für den Patientennutzen nennt das Gesetz insbesondere die Verbesserung des Gesundheitszustands, eine Verkürzung der Krankheitsdauer, eine Verlängerung der Lebensdauer, eine Verringerung der Nebenwirkungen

sowie eine Verbesserung der Lebensqualität. Aus diesen Vorgaben des Gesetzes leitet sich die für das Institut geltende Definition des patientenrelevanten Nutzens ab (siehe Abschnitt 3.1).

Im Rahmen des Versorgungsstrukturgesetzes erfolgten 2012 Änderungen des § 137c SGB V und die Hinzufügung des § 137e SGB V. Hiermit erhielt der G-BA die Möglichkeit, klinische Studien zu neuen Untersuchungs- oder Behandlungsmethoden zu initiieren (Erprobung), sofern der Nutzen einer Methode noch nicht hinreichend belegt ist, die Methode jedoch das Potenzial einer erforderlichen Behandlungsalternative erkennen lässt. Auch externe Antragsteller, beispielsweise Medizinproduktehersteller, können eine Erprobung beantragen, indem sie dem G-BA aussagekräftige Unterlagen zum Potenzial der Methode vorlegen. Die Feststellung des Potenzials einer Methode obliegt dem G-BA, der hierfür Kriterien festgelegt hat [276]. Der G-BA beauftragt in der Regel das Institut damit, Erprobungsanträge gemäß § 137e Abs. 7 SGB V dahin gehend zu prüfen, ob sich aus den Antragsunterlagen ein Potenzial ergibt.

Mit dem Versorgungsstärkungsgesetz wurde 2015 über den § 137h SGB V eine Bewertung neuer Untersuchungs- und Behandlungsmethoden (NUB) mit Medizinprodukten hoher Risikoklasse eingeführt. Dieser Bewertung unterfallen Methoden, die a) ein neues theoretisch-wissenschaftliches Konzept verfolgen [107,276], b) einen besonders invasiven Charakter aufweisen [107,276] und c) zu einer erstmaligen Anfrage nach § 6 Krankenhausentgeltgesetz (NUB-Anfrage) führen. Der G-BA erhält zu solchen Methoden Unterlagen von Krankenhäusern und auch Medizinprodukteherstellern. Der G-BA beauftragt in der Regel das Institut damit, Unterlagen gemäß § 137h SGB V dahin gehend zu prüfen, ob sich darin ein Nutzen oder aber Schädlichkeit oder Unwirksamkeit erkennen lässt.

Aufgrund des Versorgungsstärkungsgesetzes wurde 2015 der § 139b SGB V um den Abs. 5 erweitert. In § 139b Abs. 5 SGB V ist vorgesehen, dass Versicherte und sonstige interessierte Einzelpersonen Bewertungen zu medizinischen Untersuchungs- und Behandlungsmethoden bei ausgewählten Krankheiten sowie zu Fragen der Qualität und Wirtschaftlichkeit der im Rahmen der gesetzlichen Krankenversicherung erbrachten Leistungen vorschlagen können. Ausgenommen sind gemäß § 139b Abs. 5 SGB V Themenvorschläge, bei denen die eigenständige Bewertung eines Arzneimittels im Vordergrund steht. Aufgabe des IQWiG ist es, aus diesen Vorschlägen Themen auszuwählen, die für die Versorgung von Patientinnen und Patienten von besonderer Bedeutung sind und zu denen dann Health-Technology-Assessment(HTA)-Berichte im Auftrag des IQWiG erstellt werden (siehe Kapitel 6).

Das Institut ist gemäß § 139a Abs. 4 Satz 1 SGB V gesetzlich verpflichtet, die „Bewertung des medizinischen Nutzens nach den international anerkannten Standards der evidenzbasierten Medizin und die ökonomische Bewertung nach den hierfür maßgeblichen international anerkannten Standards, insbesondere der Gesundheitsökonomie“ zu gewährleisten. Das Institut bestimmt auftragsbezogen die Methoden und Kriterien für die Erarbeitung von Bewertungen auf der Grundlage der in den jeweiligen Fachkreisen anerkannten internationalen Standards der evidenzbasierten Medizin und der Gesundheitsökonomie. In Abschnitt 1.3 werden der Begriff der evidenzbasierten Medizin, seine Entwicklung und das dahinterliegende Konzept näher

erläutert. In Abschnitt 1.4 werden der Begriff der Gesundheitsökonomie und das dahinterliegende Konzept näher erläutert.

Das Institut gewährleistet während der Bearbeitung seiner Berichte eine hohe Verfahrenstransparenz und eine angemessene Beteiligung. Das Gesetz verpflichtet das Institut dazu, in allen wichtigen Abschnitten der Berichterstellung Fachexpertinnen und Fachexperten, Herstellerinnen und Herstellern sowie den „für die Wahrnehmung der Interessen der Patientinnen und Patienten und die Selbsthilfe chronisch kranker und behinderter Menschen maßgeblichen Organisationen sowie der oder dem Beauftragten der Bundesregierung für die Belange der Patientinnen und Patienten Gelegenheit zur Stellungnahme zu geben“. Das Institut geht über diese Verpflichtung hinaus, indem es allen interessierten Personen und Institutionen die Möglichkeit zur Stellungnahme zu seinen Berichten einräumt.

Diese Stellungnahmen bezieht das Institut in seine Bewertungen ein. Die Umsetzung dieser Regelungen ist in Abschnitt 2.2.5 beschrieben.

Außerdem dokumentiert das Institut seine Arbeitsergebnisse und ergänzende Informationen auf seiner frei zugänglichen Website. Interessierte können darüber hinaus den E-Mail-Dienst (Infodienst) des Instituts abonnieren. Er bietet Abonnentinnen und Abonnenten die Möglichkeit, selbst festzulegen, welche Informationen des Instituts sie erhalten möchten.

1.2 Generalauftrag des G-BA

Der Gemeinsame Bundesausschuss (G-BA) hat dem IQWiG im Dezember 2004 einen sogenannten Generalauftrag erteilt [277]. Mit dem Generalauftrag wurde das Institut beauftragt, „durch die Erfassung und Auswertung des relevanten Schrifttums eine kontinuierliche Beobachtung und Bewertung medizinischer Entwicklungen von grundlegender Bedeutung und ihrer Auswirkungen auf die Qualität und Wirtschaftlichkeit der medizinischen Versorgung in Deutschland vorzunehmen und den G-BA hierüber regelmäßig zu informieren. Der G-BA geht bei diesem Auftrag davon aus, dass das Institut auf den ihm gemäß § 139a Abs. 3 SGB V übertragenen Arbeitsfeldern nicht nur Einzelaufträge des G-BA bearbeitet, sondern aus der eigenverantwortlichen wissenschaftlichen Arbeit heraus dem G-BA für dessen gesetzliche Aufgaben notwendige Informationen über versorgungsrelevante Entwicklungen in der Medizin zur Verfügung stellt und konkrete Vorschläge für Einzelaufträge erarbeitet, die aus Sicht des Instituts vor dem Hintergrund dieser Informationen relevant sind.“

Der Generalauftrag wurde im Juli 2006 sowie März 2008 in Hinblick auf die Erstellung von Gesundheitsinformationen weiter konkretisiert bzw. angepasst [278]. Das IQWiG „erstellt und veröffentlicht die Informationsinhalte eigenverantwortlich im Rahmen seiner wissenschaftlichen Arbeit, ohne dass es jeweils einer Einzelbeauftragung bedarf. Die Inhalte der Informationen einschließlich sich daraus ergebender Konsequenzen sind allein vom Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen zu verantworten.“

1.3 Evidenzbasierte Medizin

Der Begriff evidenzbasierte Medizin (EbM) ist eine etwas unscharfe Eindeutschung des englischen Begriffs Evidence-based Medicine. Gemeint ist damit eine medizinische Betreuung von Patientinnen und Patienten, die sich nicht allein auf Meinungen und Übereinkünfte stützt, sondern Evidenz einbezieht – Belege, die mit möglichst objektiven wissenschaftlichen Methoden erhoben wurden. Evidenzbasierte Medizin umfasst Werkzeuge und Strategien, die vor Fehlentscheidungen und falschen Erwartungen schützen sollen. Eine Fehlentscheidung kann in diesem Zusammenhang bedeuten, dass nützliche Interventionen nicht oder erst verspätet in die Versorgung kommen oder dass nutzlose oder gar schädliche Interventionen weite Verbreitung finden [23,235,306,307].

Instrumente, die einer subjektiv geprägten und damit häufig verzerrten Bewertung vorbeugen sollen (siehe u. a. Abschnitt 7.13), wurden allerdings nicht erst mit der Einführung des Begriffs evidenzbasierte Medizin erfunden, sondern ihre Wurzeln reichen weit in die Vergangenheit zurück. In Deutschland war es bereits 1932 Paul Martini, der in seiner Monografie „Methodenlehre der therapeutischen Untersuchung“ die wesentlichen Elemente einer fairen Überprüfung der Wirksamkeit von Arzneimitteln beschrieb [503]. Die Methode der zufälligen Zuteilung (Randomisierung) von Probandinnen und Probanden auf Vergleichsgruppen wurde mit Beginn der 1960er-Jahre international akzeptierter Standard, um die Wirksamkeit und Sicherheit von medizinischen Interventionen zu prüfen [356]. Beginnend in den USA wurden zu dieser Zeit entsprechende Studien zur Voraussetzung für die behördliche und in Gesetzen und Verordnungen geregelte Zulassung von Arzneimitteln und (teilweise) Medizinprodukten gemacht [40]. Etwa 20 Jahre später bemühten sich klinische Epidemiologinnen und Epidemiologen, die Methodenlehre auch in der klinischen Praxis zu etablieren [241]. Begleitet von teilweise heftig geführten Kontroversen gelang dies tatsächlich erst in den 1990er-Jahren, zeitgleich mit der Benennung des Konzepts als evidenzbasierte Medizin. Klinische Studien und das systematische Auffinden und Bewerten von Studien bilden seit dieser Zeit auch den internationalen wissenschaftlichen Standard für die Technologiebewertung im Gesundheitswesen (Health Technology Assessment) [38].

EbM ist kein starres Konzept. Welches Standardinstrument wann eingesetzt werden sollte, hängt von der Frage ab, die es zu beantworten, und von der Entscheidung, die es zu treffen gilt. Trotz der Anwendung von Standards sind bei der Suche, Aufarbeitung und Bewertung von Studien immer wieder Entscheidungen zu treffen, für die es (noch) keine internationalen Festlegungen gibt. EbM beinhaltet auch die Freiheit, in solchen Situationen eigene Festlegungen vorzunehmen. Allerdings ist diese Freiheit mit der Verpflichtung verbunden, solche eigenen Festlegungen nach Möglichkeit vorab zu definieren und Bewertungen transparent zu erläutern, sodass die Begründung nachvollziehbar ist. Die folgenden Abschnitte erläutern, dass eine Institution wie das IQWiG bei der Anwendung der EbM und den Definitionen von Festlegungen in einer anderen Situation ist als Ärztinnen und Ärzte, die Unterstützung bei einer Therapieentscheidung suchen.

1.3.1 Praktische evidenzbasierte Medizin

EbM ist von der Idee her als Strategie für Ärztinnen und Ärzte gedacht, die für ihre Patientinnen und Patienten unter möglichen Interventionen die vielversprechendsten und deren Bedürfnissen am ehesten entsprechenden Alternativen herausfinden und die Erfolgsaussichten neutral darstellen wollen. Diese Anwendung der evidenzbasierten Medizin in der täglichen Praxis für individuelle Patientinnen und Patienten haben im Jahr 1996 David Sackett und Kollegen [615] folgendermaßen definiert: „EbM ist der gewissenhafte, ausdrückliche und vernünftige Gebrauch der gegenwärtig besten externen wissenschaftlichen Evidenz für Entscheidungen in der medizinischen Versorgung individueller Patienten. Die Praxis der EbM bedeutet die Integration individueller klinischer Expertise mit der bestverfügbaren externen Evidenz aus systematischer Forschung.“

Oft erweist sich die „bestverfügbare Evidenz“ aber als lückenhaft oder nicht zuverlässig. Die EbM hat Instrumente entwickelt, um die Unsicherheit einzuschätzen; oft werden sogenannte Evidenzlevel zur Illustration verwendet. Auf diese Weise hilft EbM Ärztinnen und Ärzten sowie Patientinnen und Patienten dabei, die Art und das Ausmaß der Unsicherheit zu erkennen. Ärztinnen und Ärzte sowie Patientinnen und Patienten können dann besprechen, wie man mit der Unsicherheit umgehen soll. Gerade in unsicheren Situationen kommt es auf die persönlichen Präferenzen an, die darüber entscheiden, welche Option Patientinnen und Patienten wählen. Im Idealfall basieren Entscheidungen außer auf der Evidenz auch auf dem klinischen Zustand und den Umständen, in denen sich eine Person befindet, und auf ihren Präferenzen und Handlungen [346]. Gleichzeitig schafft die Beschreibung der identifizierten Wissenslücken erst die Voraussetzung für zielgerichtete medizinische Forschung, die sich nach den Bedürfnissen von Patientinnen und Patienten richtet.

Hinter EbM steht eine hinterfragende Grundhaltung [442]. Wie notwendig Skepsis häufig ist, zeigt sich daran, dass in den letzten Jahrzehnten eine Reihe von unzureichend erprobten, aber bereits breit eingesetzten Therapien mit Methoden der EbM überprüft wurde und diese Überprüfung gezeigt hat, dass vorschneller Optimismus gefährliche Folgen für Patientinnen und Patienten haben kann [198,596]. Aufgabe des Instituts ist es, objektiv zu bewerten, mit welcher Sicherheit der Nutzen medizinischer Maßnahmen nachgewiesen ist, um nicht sachgerechten Urteilen entgegenzuwirken.

1.3.2 Bedeutung von EbM für das Institut

Die Hauptaufgabe des Instituts ist es, die vorgegebene Frage seiner Auftraggeber, ob ein Nutzen oder Schaden durch eine Maßnahme nachzuweisen ist, möglichst zuverlässig zu beantworten. Ziel ist es, ausreichend zuverlässige Belege dafür vorzulegen, dass Therapie A für Patientinnen und Patienten mit einer bestimmten Krankheit besser ist als Alternative B. Oder kürzer: Was ist der Nutzen von A im Vergleich zu B?

Der Auftrag des Instituts zielt also ganz bewusst nicht auf die Behandlung einzelner Personen mit ihren möglichen Besonderheiten, sondern darauf, für welche Gruppen von Patientinnen und

Patienten es Belege für einen Nutzen gibt. Der G-BA berücksichtigt dann bei seinen Entscheidungen Aspekte der Patientenversorgung, die außerhalb der Nutzenbewertung liegen [276].

1.3.3 Die Strategien der EbM

Ein charakteristisches Standardelement der evidenzbasierten Medizin ist die strukturierte und systematische Art, mit der Antworten auf eine medizinische Frage gesucht werden.

- 1) Die medizinische Frage muss präzise formuliert werden. In der Medizin geht es (fast) immer um die Entscheidung zwischen mindestens 2 Alternativen: Diese können Therapien, Diagnoseverfahren oder komplexe Lebensstiländerungen betreffen. Daraus leitet sich stets die Frage ab: Ist Option A besser als Option B? Dabei kann zum Beispiel auch der Verzicht auf eine Therapie eine Option sein, die ernsthaft überprüft werden muss. Zu betonen ist allerdings, dass der Verzicht auf eine Therapie nicht gleichbedeutend mit „nichts tun“ ist (zum Beispiel abwartendes Beobachten, Watchful Waiting [392]).
- 2) Es muss definiert sein, woran der Nutzen einer Therapie (Diagnose, Lebensstiländerung) gemessen werden soll. Standardelement der EbM ist die Frage nach relevanten Konsequenzen für Patientinnen und Patienten: Kann das Leben verlängert werden, bessern sich Beschwerden und die Lebensqualität?
- 3) EbM formuliert explizit, dass in der Medizin bezüglich des Nutzens von Therapie, Diagnose und Lebensstil in der Regel nur Wahrscheinlichkeitsaussagen oder Aussagen über Gruppen von Patientinnen und Patienten möglich sind. Nutzen wird dadurch nachgewiesen, dass sich die Wahrscheinlichkeit eines günstigen Ausgangs erhöht und / oder das Risiko eines ungünstigen Ausgangs verringert. Um den Nutznachweis zu führen, sind Studien an ausreichend großen Gruppen von geeigneten Personen nötig.
Zur Planung, Durchführung und Auswertung solcher Studien haben internationale Wissenschaftlerinnen und Wissenschaftler eine Vielzahl von Regeln und Instrumenten entwickelt. Wichtigstes Ziel ist es, Einflüsse, die die Ergebnisse eines Vergleichs verzerren können, zu minimieren oder – falls das unmöglich ist – zumindest zu erfassen. Die Effekte solcher verzerrenden Einflüsse werden auch im Deutschen häufig mit dem englischen Begriff „bias“ bezeichnet. Diese Regeln und Werkzeuge, die international als Standard akzeptiert sind und laufend weiterentwickelt werden, sind die methodische Basis der EbM und der Arbeit des Instituts.
- 4) Eine weitere zentrale Strategie der EbM besteht darin, alle bezüglich der Qualität ihres Designs und ihrer Durchführung angemessenen Studien zu einer Frage zu identifizieren und so den Stand des zuverlässigen Wissens zusammenzufassen. Zeigen sich dabei große Unterschiede zwischen den Ergebnissen einzelner Studien (eine sogenannte Heterogenität), muss versucht werden, diese Unterschiede zu erklären. Die Ergebnisse dieser Zusammenfassungen und Bewertungen werden als systematische Übersicht bezeichnet, die statistische Auswertung als Metaanalyse.

1.3.4 Die Bedeutung der Ergebnissicherheit

Eine Besonderheit der EbM ist, dass sie eine Einschätzung erlaubt, inwieweit das vorhandene Wissen zuverlässig ist. Entscheidungen des G-BA benötigen eine hohe wissenschaftliche Zuverlässigkeit, da sie für gesetzlich Krankenversicherte weitreichende Konsequenzen wie zum Beispiel den Ausschluss von Leistungen haben können.

Die Bewertung der Ergebnissicherheit spielt daher in den Berichten des Instituts eine zentrale Rolle. Zahlreiche Details, wie Studien geplant, ausgeführt, ausgewertet und veröffentlicht wurden, haben einen Einfluss darauf, wie verlässlich die vorhandenen Ergebnisse sind. Es ist internationaler Standard der EbM, diese Aspekte kritisch zu prüfen und zu bewerten. Wie die zur Beantwortung einer Frage nötige Ergebnissicherheit erreicht werden kann, hängt aber auch von der Krankheit und von der Größe des Effekts einer Intervention ab: Wenn 2 Läufer in einem fairen Rennen mit großem Abstand über die Ziellinie gehen, braucht man keine Uhr, um den Sieger zu erkennen. So kann der Nutzen einer neuen Therapie, die zur Heilung einer bislang immer tödlichen Krankheit führt, schon durch eine relativ kleine Zahl von überlebenden Patientinnen und Patienten bewiesen sein. Auch hier beruht das Urteil letztlich auf einem Vergleich, doch bei solch dramatischen Effekten kann schon der Vergleich von früheren mit heutigen Patientinnen und Patienten genügend Sicherheit bieten. Allerdings sind Therapien, die einen so dramatischen Nutzen aufweisen, in der heutigen Medizin sehr selten [343].

Gerade bei chronischen Krankheiten sind Unterschiede zwischen 2 Therapiealternativen meist kleiner und leicht von variablen und schwankenden Krankheitsverläufen überlagert. Hier sind sehr präzise Methoden und geeignete Studiendesigns nötig, um unter den Schwankungen einen Effekt der Therapie erkennen zu können.

Es kann davon ausgegangen werden, dass das Institut gerade mit Vergleichen beauftragt wird, bei denen eben nicht auf den ersten Blick zu erkennen ist, welche Alternative Vorteile bietet. Je kleiner aber die zu erwartenden Unterschiede zwischen 2 Alternativen sind, umso verlässlichere Studien sind nötig, um hinreichend sicher sein zu können, dass ein beobachteter Unterschied nicht auf einem Zufall oder Messfehlern beruht. Der Weltrekord im 100-Meter-Lauf lässt sich heute nicht mehr mit einer Sanduhr messen. Bei kleinen Unterschieden muss zudem auch die klinische Relevanz des Unterschieds beurteilt werden.

Die Anforderungen an die Präzision und Verlässlichkeit bestimmen die Arbeitsweise des Instituts:

- 1) Es gehört zu den internationalen Standards der evidenzbasierten Medizin, für jede Frage den Typ von Studien als Messinstrument zu benennen, der das Risiko minimiert, dass eine der Alternativen unberechtigt benachteiligt wird.

- 2) Das Institut stützt sich bei der Bewertung des Nutzens bzw. Schadens in der Regel nur auf Studien mit ausreichender Ergebnissicherheit. Damit ist gewährleistet, dass die auf Empfehlungen des Instituts beruhenden Entscheidungen des G-BA auf einer wissenschaftlich gesicherten Basis stehen. Zudem wäre eine Untersuchung, die eine Recherche nach Studien unzureichender Ergebnissicherheit einschließt, zeit- und kostenintensiv.
- 3) Falls sich herausstellt, dass Studien mit der nötigen Qualität und Präzision generell fehlen, ist es die Kernaufgabe des Instituts, den Sachverhalt zu beschreiben mit der Schlussfolgerung, dass auf Basis der „gegenwärtig besten“ Evidenz keine sicheren Empfehlungen möglich sind.
- 4) Der G-BA hat die Aufgabe, diese Unsicherheit bei seiner Entscheidungsfindung zu beachten. Neben der wissenschaftlichen Sachlage bezieht er auch andere Aspekte in seine Entscheidungen ein, wozu neben der Wirtschaftlichkeit auch die Bedürfnisse und Werte der Menschen gehören [297]. In einer wissenschaftlich unsicheren Situation gewinnen diese Aspekte an Gewicht. Zudem hat der G-BA die Möglichkeit, selbst Studien zu fordern oder zu initiieren, um die identifizierten Evidenzlücken zu schließen.

1.3.5 Die Verbindung zwischen Ergebnissicherheit und Alltagsnähe

Großen Wert auf die Abschätzung der Ergebnissicherheit zu legen, wird oft kritisiert. Ein Argument ist, dass Studien mit hoher Ergebnissicherheit, insbesondere randomisierte kontrollierte Studien, zwar eine hohe interne Validität besäßen, jedoch oft die Versorgung von Patientinnen und Patienten im Alltag nicht abbildeten und deshalb nicht „übertragbar“ seien, also nur eine geringe externe Validität vorhanden sei. Dabei ist zu prüfen, wie gut die in Studien untersuchte Patientenpopulation, die eingesetzten Interventionen und die Zielkriterien dem Versorgungsalltag entsprechen. Diese Kritik ist dann oft mit der Forderung verbunden, andere Studientypen ohne Randomisierung hinzuzunehmen, um Alltagsbedingungen besser einfließen zu lassen.

Diese Kritik vermischt jedoch Argumentationsebenen, die klar voneinander zu trennen sind. Folgende Aspekte müssen Beachtung finden:

- 1) Grundlage einer Nutzenbewertung ist der Nachweis von Kausalität. Unverzichtbare Bedingung für den Nachweis von Kausalität ist ein vergleichendes Experiment, das so angelegt sein muss, dass ein Unterschied zwischen Interventionsgruppen – ein Effekt – nur auf eine einzige Einflussgröße – die geprüfte Intervention – zurückgeführt werden kann. Dieses Ziel macht für klinische Studien erhebliche Anstrengungen nötig, weil es zahlreiche unerwünschte Einflüsse gibt, die einen Effekt vortäuschen oder auch verbergen können (Verzerrung). Der stärkste dieser verzerrenden Einflüsse sind ungleiche Ausgangsbedingungen zwischen den Vergleichsgruppen. Die Randomisierung zusammen mit einer sorgfältigen Verdeckung der Zuteilung (Concealment) ist das beste gegenwärtig verfügbare Instrument, um diese Verzerrung zu minimieren. Die zufällige Zuteilung der Teilnehmer zu den Gruppen stellt sicher, dass es keine systematischen Unterschiede zwischen den Gruppen gibt, und zwar weder bezüglich bekannter Eigenschaften (z. B.

Alter, Geschlecht, Schwere der Erkrankung) noch bezüglich unbekannter Faktoren. Damit ist in randomisierten kontrollierten Studien (Randomized controlled Trials, RCTs) eine grundlegende Voraussetzung für einen Kausalitätsnachweis gegeben. Allerdings ist auch eine Randomisierung allein kein Garant für eine hohe Ergebnissicherheit. Dazu sind z. B. auch eine unverzerrte Erhebung, Zusammenfassung und Veröffentlichung der Ergebnisse nötig.

- 2) Andere Studientypen als RCTs sind in der Regel für einen Kausalitätsnachweis nicht geeignet. In nicht randomisierten vergleichenden Studien kann grundsätzlich nicht von einer Strukturgleichheit der Gruppen ausgegangen werden. Sie liefern daher immer ein potenziell verzerrtes Ergebnis und können zumeist die maßgebliche Frage nicht hinreichend sicher beantworten, ob ein Unterschied auf der geprüften Intervention beruht. Daher bedarf die Verwendung nicht randomisierter Studien als Nachweis der Kausalität einer Intervention einer besonderen Begründung bzw. besonderer Voraussetzungen und spezieller Qualitätsanforderungen.
- 3) Es trifft zu, dass viele randomisierte Studien Aspekte des Alltags der Patientenversorgung nicht abbilden, z. B. Patientinnen und Patienten mit Begleiterkrankungen ausschließen, die im Alltag häufig sind. Dies ist jedoch keine Konsequenz der Technik der Randomisierung, sondern die Folge anderer Faktoren wie z. B. der Festlegung enger Ein- und Ausschlusskriterien der Studie, der Auswahl der Interventionen oder der Zielkriterien. Zudem werden Patientinnen und Patienten in randomisierten Studien oft anders (intensiver und enger) betreut als im Alltag. Das alles sind aber bewusste Entscheidungen derjenigen, die in einer Studie eine bestimmte Frage beantworten wollen. Durch einen Verzicht auf die Randomisierung werden diese Entscheidungen nicht geändert. Auch in nicht randomisierten Studien findet eine Selektion der Teilnehmer durch Ein- und Ausschlusskriterien und ggf. andere Designcharakteristika statt, sodass externe Validität ebenso wenig per se gegeben ist wie bei RCTs.
- 4) Aber selbst wenn sich Patientengruppen einer RCT vom Versorgungsalltag unterscheiden, muss das die externe Validität des Ergebnisses nicht infrage stellen. Entscheidend ist vielmehr die Frage, ob zu erwarten ist, dass der in einer Population festgestellte Therapieeffekt in einer anderen Population anders ist.
- 5) Es hängt vom Einzelfall ab, wie sich die Intensität der Betreuung in einer Studie auswirken könnte. Zum Beispiel ist denkbar, dass ein Nutzen tatsächlich nur in den Händen besonders qualifizierter Ärztinnen und Ärzte vorhanden ist, weil unter Alltagsbedingungen ansonsten zu häufig Komplikationen auftreten. Es ist aber auch möglich, dass eine enge Betreuung Unterschiede zwischen Gruppen eher verringert. Zum Beispiel könnten Unterschiede in der Therapieadhärenz in Studien kleiner ausfallen, in denen Patientinnen und Patienten grundsätzlich sehr eng betreut werden.
- 6) Die Festlegung der Studienbedingungen liegt jedoch in der Hand der Initiatorinnen und Initiatoren einer klinischen Studie. Sie können die Fragestellung und Endpunkte definieren, die sie als so wichtig einschätzen, dass sie in der Studie beobachtet werden

sollen. Wenn eine Herstellerin oder ein Hersteller eines Arzneimittels z. B. Therapieadhärenz für einen wichtigen Aspekt des Nutzens seines Produkts hält, wäre die nahe liegende Konsequenz, dann Studien aufzulegen, die diesen Aspekt mit möglichst großer Ergebnissicherheit und Nähe zu Alltagsbedingungen messen können und gleichzeitig die Patientenrelevanz demonstrieren.

Diese Ausführungen zeigen, dass Ergebnissicherheit und Alltagsnähe (oder interne und externe Validität) nicht fest miteinander verknüpft sind: Hohe Ergebnissicherheit und Alltagsnähe schließen einander nicht aus, sondern erfordern nur eine geeignete Kombination aus Studientyp, Design und Durchführung.

Auch wenn die Kritik mangelnder Alltagsnähe tatsächlich für viele Studien zutreffen mag, wäre nichts dadurch gewonnen, wenn man eine hohe Ergebnissicherheit für eine größere Alltagsnähe aufgeben würde. Man würde einen Mangel durch Inkaufnahme eines zweiten, noch schwerwiegenderen Mangels zu kompensieren versuchen [344].

Es sind Studien wünschenswert und machbar, die Alltagsnähe und hohe Ergebnissicherheit vereinigen. RCTs, in denen weder über den Behandlungsalltag hinausgehende Anforderungen an Patientinnen und Patienten gestellt noch definierte Studienvisiten vorgegeben werden, sind sehr wohl möglich. Sie werden international als Real-World Trials, Practical Trials oder Pragmatic Trials diskutiert [261,263,288,499,720]. Allerdings können solche sogenannten pragmatischen Studien ihrerseits ebenfalls zu Interpretationsproblemen führen. Werden z. B. sehr breite Einschlusskriterien gewählt, stellt sich die Frage, ob die (Gesamt-)Studienergebnisse auf die gesamte Studienpopulation anwendbar sind [762], was letztlich über adäquate Subgruppenanalysen zumindest ansatzweise zu beantworten wäre.

1.3.6 Nutzen im Einzelfall

Ziel der Nutzenbewertung ist es, anhand der Ergebnisse von Studien, die geeignet sind, kausal begründete Effekte nachzuweisen, belastbare Voraussagen für zukünftige Patientinnen und Patienten zu treffen. Diese Aussagen betreffen immer Gruppen von Patientinnen und Patienten mit bestimmten Eigenschaften. Aussagen über den Nutzen im Sinne von Erfolgsaussagen für den Einzelfall sind grundsätzlich nicht möglich. Umgekehrt sind auf Einzelfällen basierende Erfahrungen – abgesehen von Ausnahmen (z. B. dramatischer Effekt) – für eine Nutzenbewertung ungeeignet, da es nicht möglich ist, Ergebnisse im Einzelfall (d. h. ohne Vergleich) einer Intervention zuzuschreiben.

Für bestimmte Fragestellungen (Therapieoptimierung bei einzelnen Personen) können sogenannte (randomisierte) N-of-1-Studien durchgeführt werden [308,315,424,635]. Diese sind jedoch für die Nutzenbewertung einer Behandlungsmethode für zukünftige Patientinnen und Patienten in aller Regel nicht geeignet.

1.4 Gesundheitsökonomie

Mit dem Begriff Gesundheitsökonomie können 2 Sachverhalte zum Ausdruck gebracht werden.

Im weiten Sinne geht es um „die Analyse der wirtschaftlichen Aspekte des Gesundheitswesens unter Verwendung von Konzepten der ökonomischen Theorie“ [638]. Dazu werden u. a. Konzepte aus den Bereichen der mikroökonomischen Verhaltenstheorie, der Wettbewerbstheorie, der ökonomischen Theorie der Politik und der Managementtheorie herangezogen [638]. Gegenstand einer solchen Untersuchung könnte z. B. sein, wie Akteurinnen und Akteure im Gesundheitswesen durch eine Anreizsetzung, z. B. Zuzahlungen, ihr Verhalten ändern oder ob die Verhandlungslösung nach dem Arzneimittelmarktneuordnungsgesetz (AMNOG) tatsächlich vor überhöhten Preisen neuer Arzneimittel schützt. Inwieweit solche Untersuchungen zur Steuerung des Gesundheitswesens genutzt werden können und sollen, kann sowohl methodisch als auch ethisch diskutiert werden, ist aber nicht weiter Gegenstand dieser kurzen Darstellung.

Unter Gesundheitsökonomie im engen Sinne werden hier gesundheitsökonomische Evaluationen verstanden. Nachfolgend werden die Kosten-Nutzen-Bewertung (KNB) als vergleichende Analyse und die Ausgaben-Einfluss-Analyse als nicht vergleichende Analyse betrachtet [201]. Diese Analysen dienen u. a. der Information der Entscheidungsträger zu Kosten-Nutzen-Verhältnissen von Interventionen und stellen damit neben der Nutzenbewertung einen Bereich in sogenannten Health Technology Assessments dar.

1.4.1 Bedeutung von Gesundheitsökonomie für das Institut

Mit Gründung des Instituts 2004 stand es dem G-BA und dem BMG offen, eine KNB zu beauftragen. Eine KNB von Arzneimitteln war bis zur Gesetzesänderung 2007 nicht vorgesehen. Mit dem GKV-Wettbewerbsstärkungsgesetz wurde in § 35b SGB V die Kosten-Nutzen-Bewertung von Arzneimitteln verankert, um Informationen für die Empfehlung eines sogenannten Höchstbetrags zu gewinnen. Bis zu diesem Höchstbetrag sollten neue Arzneimittel erstattet werden, da dieser Höchstbetrag die angemessenen Kosten für den Zusatznutzen eines neuen Arzneimittels im Vergleich zu anderen Arzneimitteln und Behandlungsformen in einer Indikation darstellen sollte. Bedingung für die Beauftragung einer KNB musste also ein Nachweis des Zusatznutzens für ein Arzneimittel sein, der in Form einer Nutzenbewertung des IQWiG vorliegen musste. Die Entwicklung der aus dieser gesundheitsökonomischen Fragestellung resultierenden Methoden ist umfassend dokumentiert [377,380-384,386,389].

Mit dem AMNOG, das zum 01.01.2011 in Kraft trat, verlagerte sich die Bedeutung der KNB im Verfahren der frühen Nutzenbewertung von Arzneimitteln; sie ist vornehmlich für den Fall des Scheiterns der Preisverhandlungen zwischen dem GKV-Spitzenverband und pharmazeutischen Unternehmen und der Nichteinigung im anschließenden Schiedsverfahren vorgesehen. Die Fragestellung ist allerdings geblieben: Nach § 35b Abs. 1 Satz 4 SGB V in Verbindung mit dem 5. Kapitel § 32 Abs. 3 der Verfahrensordnung des G-BA [276] sind die Angemessenheit und Zumutbarkeit einer Kostenübernahme durch die Versichertengemeinschaft zu berücksichtigen. Damit der G-BA diese angemessen berücksichtigen kann, müssen ihm entsprechende Informationen vorgelegt werden. Diese Informationen liefern die KNB (Angemessenheit) und die Ausgaben-Einfluss-Analyse (Zumutbarkeit). Die Bewertung der

Angemessenheit und der Zumutbarkeit einer Kostenübernahme erfolgt im Hinblick darauf, ob unter Beachtung des Grundsatzes der Verhältnismäßigkeit eine begründbare Relation zwischen den Kosten und dem Nutzen des Arzneimittels besteht. Dabei soll das IQWiG nach dem 5. Kapitel § 32 Abs. 2 und 3 der Verfahrensordnung des G-BA eine Empfehlung vorlegen, auf deren Basis der G-BA einen Beschluss fällen soll [276]. Die Darstellung einer begründbaren Relation zwischen den Kosten und dem Nutzen muss also aus der KNB hervorgehen.

1.4.2 Die internationalen Standards der Gesundheitsökonomie

Wie in jeder Wissenschaft gibt es auch in der Gesundheitsökonomie internationale Standards. Dazu gehören die Einteilung der KNB in die Studienformen Kosten-Effektivitäts- / Kosten-Wirksamkeits-Analyse (Cost-Effectiveness Analysis), Kosten-Nutzwert-Analyse (Cost-Utility Analysis) und Kosten-Nutzen-Analyse (im engeren Sinne: Cost-Benefit Analysis) [201].

Auch zum Vorgehen bei KNBs gibt es internationale Standards. Hinsichtlich der Nutzenbewertung folgt das Institut mit seiner Orientierung an der evidenzbasierten Medizin und dem daraus sich ergebenden methodischen Vorgehen den entsprechenden internationalen Standards. Bevor man in der Gesundheitsökonomie von internationalen Standards spricht, muss man unterscheiden zwischen klar methodischen Fragen und Fragen, die auf Wertentscheidungen, Einschätzungen oder Umfragen basieren. Dies kann am Beispiel der Diskontierungsrate verdeutlicht werden. Mit einer Diskontierungsrate werden Nutzen und Kosten, die in verschiedenen Perioden anfallen, auf eine Periode abgezinst, um damit für eine Entscheidung vergleichbar zu sein. Die reine Durchführung der Diskontierung ist mathematisch klar geregelt, somit eine methodische Frage. Die Wahl der Diskontierungsrate und insbesondere die Entscheidung, ob Kosten und Nutzen mit derselben Rate oder ggf. sogar einer nicht konstanten Rate abgezinst werden sollen, unterliegt unter anderem Fragen der Einschätzung der zukünftigen wirtschaftlichen Entwicklung und der Generationengerechtigkeit und ist damit eine Wertentscheidung [138,340,536,547,551,563,581].

Wie international anerkannte Instrumente der Bewertung von gesundheitsökonomischen Analysen zeigen [134,200,372,570], gibt es eine Reihe von Schritten und Aspekten, für die methodische Vorgaben bestehen und die transparent und nachvollziehbar abgearbeitet werden müssen. Dazu gehören u. a.:

- Definition der zu bewertenden Interventionen und ihrer Komparatoren. Eine Auswahl muss begründet werden, um Fehlentscheidungen auf Grundlage einer von Interessen gesteuerten Komparatorenauswahl vorzubeugen.
- Perspektive der KNB
- Zeithorizont der KNB
- Art der Studienform (siehe oben) mit Begründung
- Kosten mit Darstellung des Ressourcenverbrauchs und der Ressourcenbewertung
- Inflationsbereinigung und Währungsumrechnung (falls notwendig)

- Entwicklung und Erläuterung des Modells und möglichst auch Begründung der Modellwahl, z. B. Entscheidungsbaum, Markov-Modell
- Diskontierungsrate
- Darstellung der Ergebnisse, z. B. in aggregierter und disaggregierter Form
- Untersuchung der Unsicherheit der Ergebnisse mittels deterministischer und probabilistischer Sensitivitätsanalysen
- Darstellung der Unsicherheit, z. B. mit sogenannten Kosten-Effektivitäts-Akzeptanz-Kurven oder dem sogenannten Net Benefit

Hierzu gibt es in Lehrbüchern und auch z. B. in den Leitlinien der International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Vorgaben für eine gute methodische Praxis.

1.4.3 Methodische Standards in der Gesundheitsökonomie

Im Weiteren werden auch andere Aspekte einer KNB als internationaler Standard verstanden. So wird oft eine an einem Schwellenwert je gewonnenes qualitätsadjustiertes Lebensjahr (QALY) orientierte Entscheidung im Gesundheitssystem als internationaler Standard der Gesundheitsökonomie ausgegeben. Dies ist kritisch zu sehen. Zum einen hat die überwiegende Mehrheit der Länder, in denen KNBs zur Entscheidungsfindung beitragen, keinen (festen) Schwellenwert [260,644]. Zum anderen wäre dies eine Wertentscheidung und fiel somit nicht in die internationalen methodischen Standards, um die es sich nach § 139a Abs. 4 Satz 1 SGB V nur handeln kann, wenn es um Bewertungen des Instituts geht.

Darüber hinaus stellt sich die Frage nach dem Maß des Gesamtnutzens nicht nur als methodische Frage, sondern immer auch unter dem Aspekt einer Wertentscheidung. Dabei haben Fragestellung und Ziel einer gesundheitsökonomischen Analyse einen Einfluss darauf, welches Instrument man als Maß des Gesamtnutzens einsetzen soll. Das heißt, auch die Frage, ob das QALY herangezogen werden soll, muss unter ethischen, rechtlichen und kulturellen Aspekten beleuchtet werden. Wissenschaftlich bzw. methodisch wiederum kann diskutiert werden, welche Annahmen in das QALY-Konzept einfließen, z. B. die Annahme, dass die Bewertung eines Zustands unabhängig von seiner Dauer ist („constant proportional [time] trade-off“), und ob diese Annahmen tragfähig sind. Ebenfalls kann methodisch untersucht werden, ob die verschiedenen Erhebungsmethoden, z. B. indirekt vs. direkt, zu unterschiedlichen Ergebnissen führen und wie diese Differenz interpretiert werden kann. Eine Frage der Wertentscheidung auf Basis rechtlicher Vorgaben, z. B. des SGB V, ist wieder, bei wem die Nutzwerte zur Generierung von QALYs erhoben werden sollen: bei den tatsächlich von einer Krankheit Betroffenen oder bei der Allgemeinbevölkerung.

Letztlich ist die Frage, auf Grundlage welcher Rechte, Ansprüche oder Bedürfnisse Ressourcen im Gesundheitswesen mit welchem Ziel und mit welcher Wirkung auf die Allokation und Distribution von Gütern bzw. Leistungen verteilt werden sollen, nur als Wertentscheidung

aufzufassen, und sie bestimmt wiederum, welche wissenschaftlichen Standards und Methoden eingesetzt werden sollen.

1.5 Evidenzbasierte Entscheidungsfindung im Gesundheitswesen

Die Berichte des Instituts sollen dem G-BA als eine Grundlage für Entscheidungen dienen, die im Grundsatz für alle gesetzlich Krankenversicherten gelten. Weitere Institutsprodukte dienen dem G-BA zum Beispiel als Information. Die Art der Entscheidungen, die Institutionen wie der G-BA zu treffen haben, hat Auswirkungen auf den Einsatz der Methoden der EbM und der Gesundheitsökonomie.

2 Produkte des Instituts

Gemäß seinem gesetzlichen Auftrag erstellt das Institut verschiedene Produkte in Form von wissenschaftlichen Berichten und allgemein verständlichen Informationen für Bürgerinnen und Bürger sowie Patientinnen und Patienten. Dieses Kapitel beschreibt die Abläufe und allgemeinen Methoden der Erstellung von Institutsprodukten. Dabei werden zunächst die einzelnen Produkte benannt und produktspezifische Verfahrensabläufe dargestellt (Abschnitt 2.1). Im darauf folgenden Abschnitt werden weitere, produktunabhängige Aspekte beschrieben (Abschnitt 2.2).

2.1 Produktspezifische Verfahrensabläufe

Zu den Produkten des Instituts gehören:

- Bericht,
- Rapid Report (Schnellbericht),
- Dossierbewertung,
- Kosten-Nutzen-Bewertung nach § 35b SGB V,
- Potenzialbewertung,
- Bewertung gemäß § 137h SGB V,
- Addendum,
- Gesundheitsinformation,
- Arbeitspapier,
- HTA-Bericht.

Die Erstellung von Berichten und Rapid Reports erfolgt auf Basis eines Einzelauftrags durch den G-BA oder das BMG. Grundlage hierfür sind die in § 139a SGB V beschriebenen Aufgaben des Instituts (siehe Abschnitt 1.1). Der wesentliche Unterschied zwischen Berichten und Rapid Reports besteht darin, dass nur bei Berichten, nicht aber bei Rapid Reports ein öffentliches Stellungnahmeverfahren (Anhörung) durchgeführt wird. Entsprechend sind Rapid Reports insbesondere für zeitnahe Empfehlungen gedacht, zu denen aus Sicht des Auftraggebers keine Anhörungen durch das Institut erforderlich sind.

Dossierbewertungen werden vom G-BA beauftragt. Grundlage hierfür ist § 35a SGB V, in dem die Bewertung des Nutzens neuer Wirkstoffe auf Basis eines Dossiers des pharmazeutischen Unternehmers geregelt ist (siehe auch Abschnitt 3.3.3). Für Dossierbewertungen nach § 35a SGB V ist keine Anhörung durch das Institut vorgesehen. Ein Stellungnahmeverfahren erfolgt im weiteren Verfahren durch den G-BA.

Weiterhin kann das Institut vom G-BA nach § 35b SGB V mit Kosten-Nutzen-Bewertungen von Arzneimitteln beauftragt werden. Für Kosten-Nutzen-Bewertungen nach § 35b SGB V

sind Anhörungen durch das Institut vorgesehen. Beim G-BA wird ein weiteres Stellungnahmeverfahren durchgeführt.

Potenzialbewertungen erfolgen im Auftrag des G-BA und beziehen sich auf Erprobungsanträge gemäß § 137e SGB V. Eine Anhörung durch das Institut ist nicht vorgesehen. Sofern es zu einer Erprobung kommt, führt der G-BA ein Stellungnahmeverfahren zur Erprobungsrichtlinie durch.

Bewertungen gemäß § 137h SGB V erfolgen im Auftrag des G-BA und beziehen sich auf neue Untersuchungs- und Behandlungsmethoden mit Medizinprodukten hoher Risikoklasse. Es ist keine Anhörung durch das Institut vorgesehen. Sofern eine Richtlinie beschlossen wird, führt der G-BA hierzu ein Stellungnahmeverfahren durch.

Addenda können vom G-BA oder BMG beauftragt werden, wenn sich nach Erstellung eines Produkts im Zuge der Beratung ein zusätzlicher Bearbeitungsbedarf ergibt.

Gesundheitsinformationen können auf Basis eines Einzelauftrags erstellt werden, sie können aber auch Folge eines Auftrags zu anderen Aufgabenbereichen sein (allgemein verständliche Version anderer Institutsprodukte, z. B. eines Berichts) oder im Rahmen der allgemeinen gesetzlichen Aufgabe der Bereitstellung von Gesundheitsinformationen erarbeitet werden.

Arbeitspapiere entstehen in Eigenverantwortung des Instituts, ohne dass es hierzu eines Auftrags durch den G-BA oder das BMG bedarf. Dies geschieht entweder auf Grundlage des Generalauftrags (siehe Abschnitt 1.2) mit dem Ziel, zu versorgungsrelevanten Entwicklungen im Gesundheitswesen Auskunft zu erteilen, oder im Rahmen des gesetzlichen Auftrags zur Entwicklung der Institutsmethoden. Die „Allgemeinen Methoden“ des Instituts sind nicht als Arbeitspapier in diesem Sinne zu verstehen und unterliegen einem gesonderten Verfahren der Erstellung und Aktualisierung, das in der Präambel dieses Dokuments dargestellt ist.

HTA-Berichte werden zu von Versicherten und interessierten Einzelpersonen vorgeschlagenen Themen erstellt. Aus den Vorschlägen wählt das Institut Themen aus, die für die Patientenversorgung in Deutschland besonders bedeutsam sind. Dabei wird sowohl die Bürger- und Patientensicht als auch die wissenschaftliche Perspektive miteinbezogen. Zu den ausgewählten Themen werden HTA-Berichte erarbeitet. Dies geschieht auf Grundlage von § 139b Abs. 5 SGB V. Zu den HTA-Berichten ist eine Anhörung durch das Institut vorgesehen.

In der nachfolgenden Tabelle 1 sind die verschiedenen Produkte des Instituts in einer Übersicht dargestellt. Die produktspezifischen Verfahrensabläufe sind in den sich daran anschließenden Abschnitten 2.1.1 bis 2.1.10 beschrieben.

Tabelle 1: Übersicht über die Produkte des Instituts

Produkt	Zielsetzung	Ablauf	Beauftragung durch
Bericht	Empfehlungen zu den in § 139a SGB V beschriebenen Aufgaben, inklusive Anhörung	beschrieben in Abschnitt 2.1.1	G-BA, BMG
Rapid Report	Empfehlungen zu den in § 139a SGB V beschriebenen Aufgaben, sofern keine Anhörung erforderlich ist, insbesondere zeitnahe Informationen zu aktuellen Themen	beschrieben in Abschnitt 2.1.2	G-BA, BMG
Dossierbewertung	Bewertung des Nutzens von Arzneimitteln mit neuen Wirkstoffen gemäß § 35a SGB V	beschrieben in Abschnitt 2.1.3	G-BA
Kosten-Nutzen-Bewertung nach § 35b SGB V	Bewertung des Kosten-Nutzen-Verhältnisses von Arzneimitteln gemäß § 35b SGB V, inklusive Anhörung	beschrieben in Abschnitt 2.1.4	G-BA
Potenzialbewertung	Bewertung des Potenzials neuer Untersuchungs- und Behandlungsmethoden gemäß § 137e SGB V	beschrieben in Abschnitt 2.1.5	G-BA
Bewertung gemäß § 137h SGB V	Bewertung von Nutzen, Schädlichkeit oder Unwirksamkeit neuer Untersuchungs- und Behandlungsmethoden mit Medizinprodukten hoher Risikoklasse gemäß § 137h SGB V	beschrieben in Abschnitt 2.1.6	G-BA
Addendum	zeitnahe Stellungnahme zu Fragestellungen, die sich im Zuge der Beratung abgeschlossener Institutsprodukte ergeben haben	beschrieben in Abschnitt 2.1.7	G-BA, BMG
Gesundheitsinformation	allgemein verständliche Informationen für Bürgerinnen und Bürger sowie Patientinnen und Patienten, breit gefächertes Themenspektrum	beschrieben in Abschnitt 2.1.8	G-BA-Generalauftrag
Arbeitspapier	Information zu versorgungsrelevanten Entwicklungen in der Medizin oder zu methodischen Aspekten	beschrieben in Abschnitt 2.1.9	G-BA-Generalauftrag
HTA-Bericht	Bewertung medizinischer Untersuchungs- und Behandlungsmethoden gemäß § 139b Abs. 5 SGB V, inklusive Anhörung	beschrieben in Abschnitt 2.1.10	Initiierung durch das Institut auf Basis von Vorschlägen interessierter Einzelpersonen
BMG: Bundesministerium für Gesundheit; G-BA: Gemeinsamer Bundesausschuss; HTA: Health Technology Assessment; SGB V: Sozialgesetzbuch Fünftes Buch – Gesetzliche Krankenversicherung			

2.1.1 Bericht

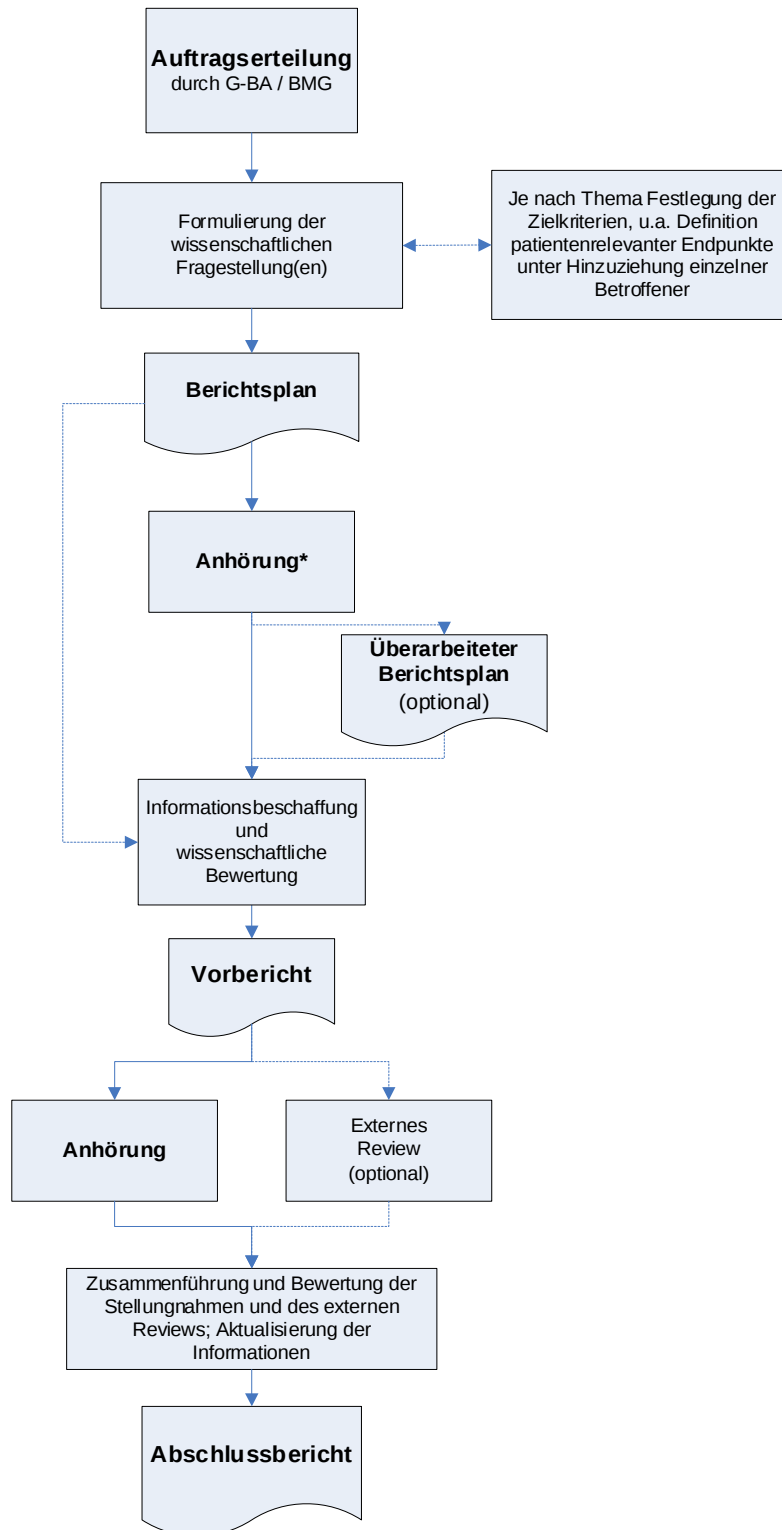
Der Ablauf der Berichterstellung ist Abbildung 1 in schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt. Dabei werden regelhaft externe Sachverständige beteiligt (siehe Abschnitt 2.2.2). Bei Bedarf wird auch der Wissenschaftliche

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, Betroffene beteiligt. Betroffene können z. B. sein: Patientinnen und Patienten (ggf. vertreten durch ihre Eltern oder sonstige Angehörige) sowie potenzielle Teilnehmerinnen und Teilnehmer an Präventionsmaßnahmen. Anschließend wird der Berichtsplan erstellt.

Der **Berichtsplan** bildet die Grundlage für die Erstellung des Vorberichts und 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. Der Berichtsplan 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. 5 Arbeitstage später.

Eine Anhörung des Berichtsplans erfolgt entsprechend den gesetzlichen Regelungen in § 139a Abs. 5 SGB V. Die Anhörung erfolgt mittels **schriftlicher Stellungnahmen**, die innerhalb einer Frist von mindestens 4 Wochen ab Bekanntmachung auf der Website des Instituts abgegeben werden können. Gegenstand der Anhörung ist insbesondere das projektspezifische methodische Vorgehen zur Beantwortung der Fragestellung. Die Fragestellung selbst ist i. d. R. durch den Auftrag vorgegeben und nicht Gegenstand der Anhörung. Die Stellungnahmen werden ausgewertet und zwecks Dokumentation der Anhörung veröffentlicht. Falls keine Änderung der Berichtsmethodik erforderlich ist, wird die Dokumentation der Anhörung zusammen mit dem Vorbericht veröffentlicht. Falls eine Änderung der Berichtsmethodik erforderlich ist, wird eine entsprechend angepasste neue Version des Berichtsplans erstellt, diese 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. 5 Arbeitstage später auf der Website des Instituts veröffentlicht. Die überarbeitete Version des Berichtsplans bildet die Grundlage für die Erstellung des Vorberichts.



* Ein Anhörungsverfahren zum Berichtsplan erfolgt entsprechend den gesetzlichen Regelungen im § 139a Abs. 5 SGB V.

Abbildung 1: Ablauf der Erstellung eines Berichts

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 Berichtsplan formulierten Kriterien. Das Ergebnis der Anhörung wird damit jedoch explizit nicht vorweggenommen, da diese Kriterien sich durch die Anhörung zum Berichtsplan 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. 5 Arbeitstage nach Versendung an den Auftraggeber.

Der Vorbericht wird zur Anhörung gestellt. Die Anhörung erfolgt grundsätzlich mittels **schriftlicher Stellungnahmen**, die innerhalb einer Frist von mindestens 4 Wochen ab Bekanntmachung auf der Website des Instituts abgegeben werden können. Optional ist eine mündliche wissenschaftliche Erörterung mit Stellungnehmenden vorgesehen. Diese Erörterung dient der ggf. notwendigen Klärung von Inhalten der schriftlichen Stellungnahmen. Gegenstand der Anhörung sind insbesondere die im Vorbericht dargestellten Ergebnisse der Informationsbeschaffung und -bewertung.

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 Berichterstellung dar. Der Abschlussbericht und die Dokumentation der Anhörung zum Vorbericht werden zunächst dem Auftraggeber, dem Vorstand der Stiftung, dem Stiftungsrat sowie dem Kuratorium der Stiftung zugeleitet. In der Regel 4 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.

Sofern eine Untersuchungs- oder Behandlungsmethode nach § 139a Abs. 3 Nr. 1 SGB V bewertet wird, ist der Abschlussbericht gemäß § 4 der Methodenbewertungsverfahrensverordnung (MBVerfV) innerhalb 1 Jahres nach Erteilung des Auftrags fertigzustellen [105].

2.1.2 Rapid Report

Der Ablauf der Erstellung eines **Rapid Reports** ist in Abbildung 2 schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt. Dabei können externe

Sachverständige beteiligt werden (siehe Abschnitt 2.2.2). Bei Bedarf wird auch der Wissenschaftliche Beirat des Instituts einbezogen. Der interne Qualitätssicherungsprozess ist in diesem Flussdiagramm nicht dargestellt.

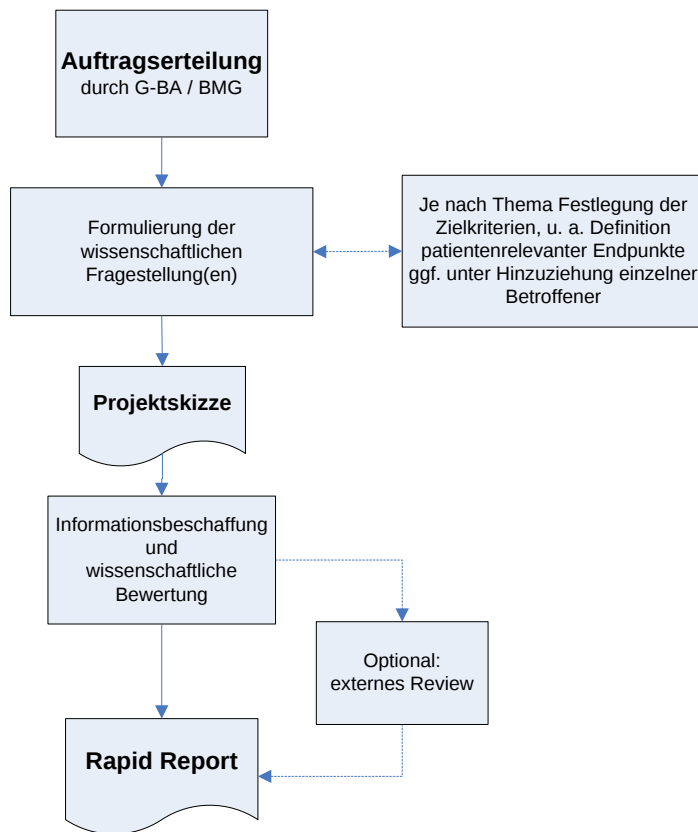


Abbildung 2: Ablauf der Erstellung eines Rapid Reports

Rapid Reports werden vorrangig mit dem Ziel einer zeitnahen Information zu relevanten Entwicklungen im Gesundheitswesen (z. B. neue Technologien, Publikation wichtiger Meilensteinstudien) erstellt. Dabei ist zumeist eine kürzere Erarbeitungszeit notwendig. Daher finden keine Anhörungen im Projektverlauf statt.

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. Dabei kann auch die Meinung einzelner Betroffener eingeholt werden, insbesondere zur Definition patientenrelevanter Endpunkte. Anschließend erfolgt die Erstellung der Projektskizze.

In der **Projektskizze** werden die wesentlichen Schritte der Informationsbeschaffung und der wissenschaftlichen Bewertung zusammengefasst. Sie bildet die Grundlage für die Erstellung des Rapid Reports. Die Projektskizze wird nicht veröffentlicht.

Im **Rapid Report** werden die Ergebnisse der Informationsbeschaffung und wissenschaftlichen Bewertung dargestellt. Vor der Fertigstellung kann optional ein Entwurf des Rapid Reports

einem oder mehreren externen Reviewern (siehe Abschnitt 2.2.4) mit ausgewiesener methodischer und / oder fachlicher Kompetenz als weiterer Schritt der Qualitätssicherung vorgelegt werden. Nach der Fertigstellung wird der Rapid Report an den Auftraggeber, den Vorstand der Stiftung sowie den Stiftungsrat und anschließend (i. d. R. 1 Woche später) an das Kuratorium versandt. Die Veröffentlichung auf der Website des Instituts erfolgt i. d. R. 4 Wochen nach Versand an den Auftraggeber und Vorstand. Sollten Stellungnahmen zu Rapid Reports eingehen, die substanzielle nicht berücksichtigte Evidenz enthalten, oder erlangt das Institut auf andere Weise Kenntnis von solcher Evidenz, wird dem Auftraggeber begründet mitgeteilt, ob eine Neubeauftragung zu dem Thema (ggf. eine Aktualisierung des Rapid Reports) aus Sicht des Instituts erforderlich erscheint. Der Auftraggeber entscheidet über die Beauftragung des Instituts. Eine solche Aktualisierung unterliegt den generellen methodischen und verfahrenstechnischen Anforderungen an Produkte des Instituts.

2.1.3 Dossierbewertung

Der Ablauf der Erstellung einer Dossierbewertung ist in Abbildung 3 schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt. Dabei werden regelhaft externe Sachverständige beteiligt (siehe Abschnitt 2.2.2). Bei Bedarf wird auch der Wissenschaftliche Beirat des Instituts einbezogen. Der interne Qualitätssicherungsprozess ist in diesem Flussdiagramm nicht dargestellt.

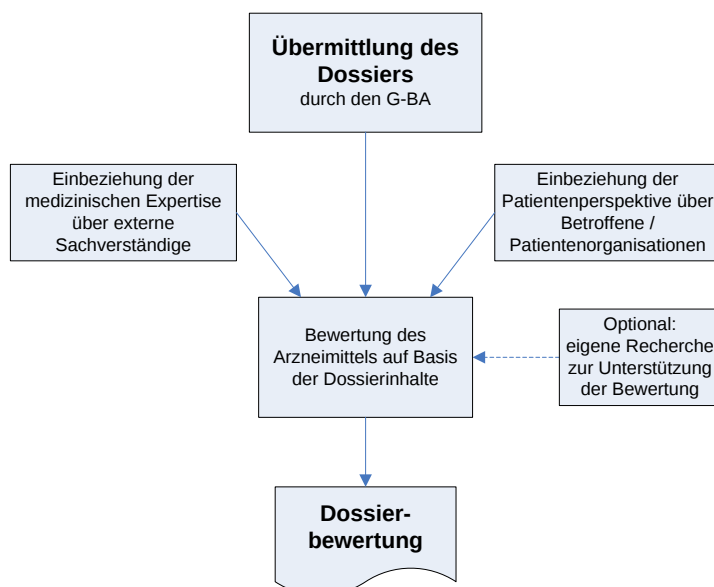


Abbildung 3: Ablauf der Erstellung einer Dossierbewertung

Nach **Übermittlung des Dossiers** durch den G-BA erfolgt die inhaltliche Bewertung der Dossierinhalte in Verantwortung des Instituts. Dabei werden regelhaft die medizinische Expertise über externe Sachverständige sowie die Patientenperspektive über Betroffene bzw. Patientenorganisationen einbezogen.

Die **Einbeziehung der medizinischen Expertise** erfolgt primär auf Basis eines Fragenkatalogs, der zu Beginn der Bewertung an die externen Sachverständigen verschickt wird. Das Institut berücksichtigt bei seiner Bewertung die Rückmeldungen der externen Sachverständigen. Darüber hinaus können die externen Sachverständigen im Bedarfsfall zur Klärung spezieller Fragen im Verlauf der Bewertung herangezogen werden. Externe Sachverständige werden über eine institutseigene Sachverständigendatenbank identifiziert (siehe Abschnitt 2.2.2).

Die **Einbeziehung der Patientenperspektive** erfolgt auf Basis eines Fragebogens, der zu Beginn der Bewertung an Betroffene bzw. Patientenorganisationen verschickt wird. Das Institut berücksichtigt bei seiner Bewertung die in diesem Fragebogen getroffenen Angaben z. B. zu relevanten Endpunkten und zu wichtigen Subgruppen. Betroffene bzw. Patientenorganisationen werden über die in § 140f SGB V benannten maßgeblichen Organisationen identifiziert.

Grundlage für die Bewertung ist das vom pharmazeutischen Unternehmer an den G-BA und dann weiter an das Institut übermittelte Dossier. Optional kann zur Unterstützung der Bewertung eine **eigene Recherche** des Instituts durchgeführt werden.

Den Abschluss des Prozesses bildet die Erstellung einer **Dossierbewertung**. Nach § 35a SGB V muss die Bewertung spätestens 3 Monate nach dem für die Einreichung des Dossiers maßgeblichen Zeitpunkt abgeschlossen werden. Die Dossierbewertung wird nach ihrer Fertigstellung dem G-BA zugeleitet. Zeitnah im Anschluss daran erfolgen die Information des Vorstands der Stiftung, des Stiftungsrats und des Kuratoriums der Stiftung über den Versand sowie die Publikation der Dossierbewertung auf der Website des Instituts.

2.1.4 Kosten-Nutzen-Bewertung nach § 35b SGB V

Der Ablauf der Kosten-Nutzen-Bewertung nach § 35b SGB V ist in Abbildung 4 schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt. Dabei werden regelhaft externe Sachverständige beteiligt. Bei Bedarf wird auch der Wissenschaftliche Beirat des Instituts einbezogen. Der interne Qualitätssicherungsprozess ist in diesem Flussdiagramm nicht dargestellt.

Vor der **Auftragserteilung** durch den G-BA bereitet der G-BA wesentliche Auftragsinhalte vor (im Zuge des sog. Scoping, siehe Abschnitt 4.9.1) und gibt den Stellungnahmeberechtigten die Gelegenheit zur Stellungnahme. Zeitgleich mit der Auftragserteilung teilt der G-BA in seinem Beschluss mit, ob Versorgungsstudien zu berücksichtigen sind, die der G-BA mit dem pharmazeutischen Unternehmer vereinbart hat.

Parallel dazu fordert der G-BA den pharmazeutischen Unternehmer zur **Einreichung des Dossiers** auf. Das Dossier des pharmazeutischen Unternehmers fließt in diese Bewertung ein.

Im **Vorbericht** werden die Ergebnisse der Informationsbeschaffung und der wissenschaftlichen Bewertung dargestellt. Bei der inhaltlichen Bewertung werden regelhaft die medizinische

Expertise über externe Sachverständige und die Patientenperspektive über Betroffene bzw. Patientenorganisationen einbezogen.

Die **medizinische Expertise** wird primär auf Basis eines Fragenkatalogs eingeholt, der zu Beginn der Bewertung an die externen Sachverständigen verschickt wird. Bei der Bewertung werden die Rückmeldungen der externen Sachverständigen berücksichtigt. Darüber hinaus können die externen Sachverständigen im Bedarfsfall zur Klärung spezieller Fragen im Verlauf der Bewertung herangezogen werden. Externe Sachverständige werden über eine instituts-eigene Sachverständigendatenbank identifiziert (siehe Abschnitt 2.2.2).

Die **Patientenperspektive** wird auf Basis eines Fragebogens erhoben, der zu Beginn der Bewertung an Betroffene bzw. Patientenorganisationen verschickt wird. Bei der Bewertung werden die in diesem Fragebogen getroffenen Angaben z. B. zu relevanten Endpunkten und zu wichtigen Subgruppen berücksichtigt. Betroffene bzw. Patientenorganisationen werden über die in § 140f SGB V benannten maßgeblichen Organisationen identifiziert.

Der **Vorbericht** enthält die vorläufige Empfehlung an den G-BA. Er wird nach Fertigstellung zunächst dem G-BA, dem Vorstand der Stiftung, dem Stiftungsrat und dem Kuratorium zugeleitet. Er wird auf der Website des Instituts zeitnah nach Versendung an den G-BA veröffentlicht.

Der Vorbericht wird zur öffentlichen Anhörung gestellt. Die Anhörung erfolgt grundsätzlich mittels schriftlicher Stellungnahmen, die innerhalb einer Frist von 3 Wochen ab Bekanntmachung auf der Website des Instituts abgegeben werden können. Optional kann eine mündliche wissenschaftliche Erörterung mit Stellungnehmenden durchgeführt werden. Diese Erörterung dient der ggf. notwendigen Klärung von Inhalten der schriftlichen Stellungnahmen. Gegenstand der Anhörung sind insbesondere die im Vorbericht dargestellten Ergebnisse der Informationsbeschaffung und -bewertung.

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 Berichterstellung dar und ist dem G-BA innerhalb von 3 Monaten nach Einleitung des Stellungnahmeverfahrens zum Vorbericht zu übermitteln (siehe Verfahrensordnung des G-BA, 5. Kapitel § 31 [276]). Der Abschlussbericht und die Dokumentation der Anhörung zum Vorbericht werden zunächst dem G-BA, dem Vorstand der Stiftung sowie dem Stiftungsrat und anschließend dem Kuratorium der Stiftung zugeleitet. Die Dokumente werden danach auf der Website des Instituts veröffentlicht. Sollten Stellungnahmen zu Abschlussberichten eingehen, die substanzielle nicht berücksichtigte Evidenz enthalten, oder erlangt das Institut auf andere Weise Kenntnis von solcher Evidenz, wird dem G-BA begründet mitgeteilt, ob eine Neubeauftragung zu dem Thema (ggf. Aktualisierung des Berichts) erforderlich erscheint. Der G-BA entscheidet über die Beauftragung des Instituts. Eine solche Aktualisierung unterliegt den allgemeinen methodischen und verfahrenstechnischen Anforderungen an Produkte des Instituts.

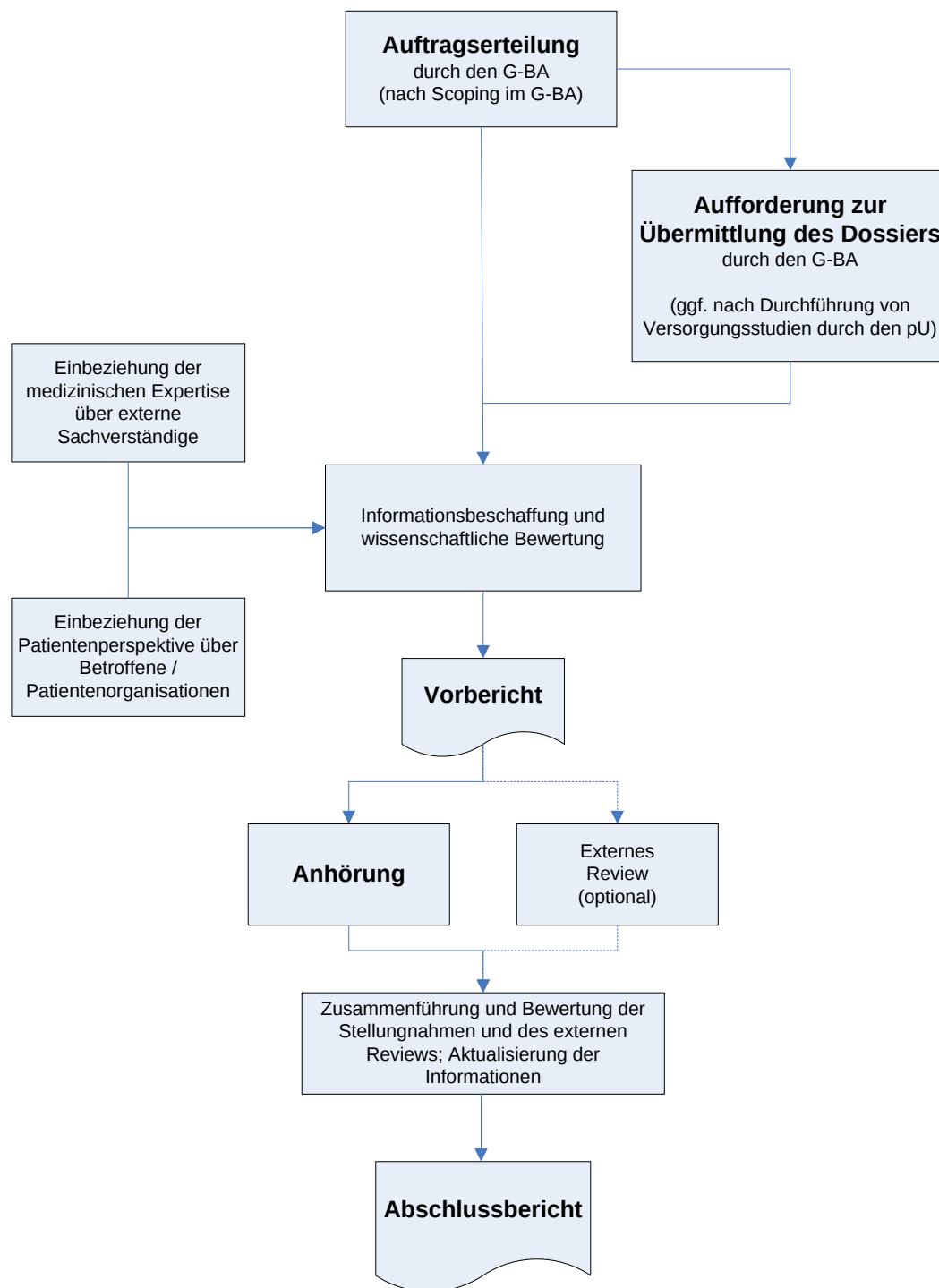


Abbildung 4: Ablauf der Kosten-Nutzen-Bewertung nach § 35b SGB V

2.1.5 Potenzialbewertung

Der Ablauf der Erstellung einer Potenzialbewertung ist in Abbildung 5 schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt. Dabei können externe Sachverständige beteiligt werden (siehe Abschnitt 2.2.2). Der interne Qualitätssicherungsprozess ist in diesem Flussdiagramm nicht dargestellt.

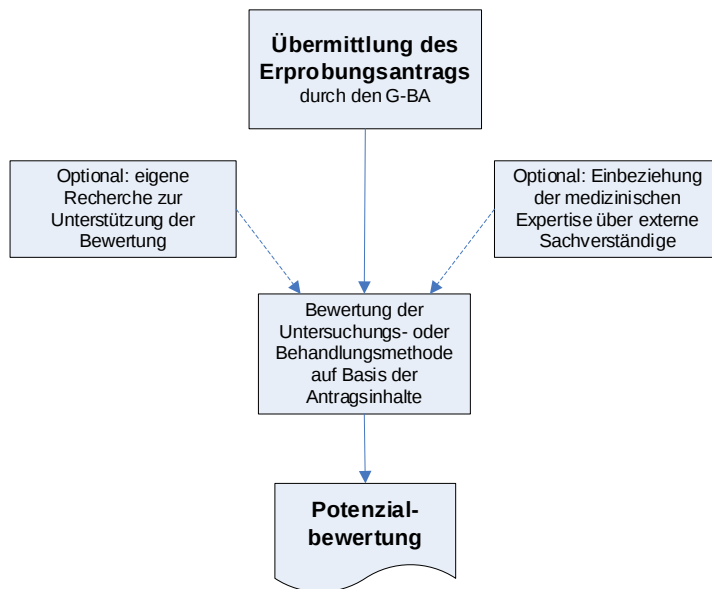


Abbildung 5: Ablauf der Erstellung einer Potenzialbewertung

Nach **Übermittlung des Erprobungsantrags** durch den G-BA erfolgt die inhaltliche Bewertung der Antragsinhalte in Verantwortung des Instituts. Hierbei kann externe medizinische Expertise miteinbezogen werden. Dies erfolgt in gleicher Weise wie bei Dossierbewertungen, allerdings unter Berücksichtigung der spezifischen Anforderungen an die Wahrung der hohen Vertraulichkeit im Rahmen von Potenzialbewertungen.

Grundlage für die Bewertung ist der vom Antragsteller an den G-BA und dann weiter an das Institut übermittelte Antrag. Optional kann zur Unterstützung der Bewertung eine **eigene Recherche** des Instituts durchgeführt werden. Da die Eckpunkte einer Erprobungsstudie ein optionaler Antragsinhalt sind, konkretisiert das Institut diese Eckpunkte, wenn der Antragsteller hierzu keine Angaben macht.

Den Abschluss des Prozesses bildet die Erstellung der **Potenzialbewertung**. Nach § 137e SGB V muss der G-BA innerhalb von 3 Monaten eine Entscheidung zum Potenzial der beantragten Untersuchungs- oder Behandlungsmethode treffen. Daher werden Potenzialbewertungen durch das Institut regelhafte binnen 6 Wochen fertiggestellt. Nach Fertigstellung wird die Potenzialbewertung dem G-BA zugeleitet. Eine Publikation der Potenzialbewertung erfolgt nicht, weil das Antragsverfahren gemäß § 137e SGB V der hohen Vertraulichkeit unterliegt. Erst wenn im weiteren Verlauf der G-BA eine Erprobungsrichtlinie erlässt, wird auch die Potenzialbewertung veröffentlicht.

2.1.6 Bewertung gemäß § 137h SGB V

Der Ablauf der Erstellung einer Bewertung gemäß § 137h SGB V entspricht im Wesentlichen dem der Potenzialbewertung (Abbildung 5): Alle Arbeitsschritte werden in Verantwortung des Instituts getätigt; hierbei können externe Sachverständige beteiligt werden (siehe Abschnitt 2.2.2). Anders als bei Potenzialbewertungen sind jedoch bei Bewertungen gemäß § 137h SGB V weder das Thema der Bewertung noch die zentralen Unterlagen vertraulich.

Der G-BA erhält von einem Krankenhaus Unterlagen zu einer neuen Untersuchungs- und Behandlungsmethode (NUB), die maßgeblich auf dem Einsatz eines Medizinprodukts hoher Risikoklasse beruht. Diese Unterlagen werden durch den G-BA öffentlich bekannt gemacht. Nachdem weitere Krankenhäuser und betroffene Medizinproduktehersteller die Unterlagen ergänzen konnten, übermittelt der G-BA dem Institut die Gesamtheit der der Bewertung zugrunde zu legenden Unterlagen.

Das Institut führt eine Bewertung hinsichtlich Nutzen, Schädlichkeit oder Unwirksamkeit durch. Grundlage für die Bewertung sind die von einem Krankenhaus bzw. Medizinproduktehersteller an den G-BA übermittelten Unterlagen. Optional kann zur Unterstützung der Bewertung eine eigene Recherche des Instituts durchgeführt werden. Zur Klärung spezieller Fragen kann externe medizinische Expertise miteinbezogen werden. Hierfür werden externe Sachverständige über eine institutseigene Sachverständigendatenbank identifiziert.

Das Institut prüft, ob sich anhand der Unterlagen entweder ein Nutzen oder aber eine Schädlichkeit oder Unwirksamkeit erkennen lässt. Da der G-BA gesetzlich verpflichtet ist, innerhalb von 3 Monaten über den Nutzen, die Schädlichkeit oder die Unwirksamkeit der Methode zu entscheiden, erstellt das Institut seine Bewertungen gemäß § 137h SGB V innerhalb von 6 Wochen. Der jeweilige Bericht wird dem Auftraggeber zugeleitet. Die Veröffentlichung der Bewertung erfolgt in der Regel 4 bis 6 Wochen nach Übermittlung an den G-BA.

2.1.7 Addendum

Der Ablauf der Erstellung eines Addendums ist in Abbildung 6 schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts, bei Bedarf unter Einbeziehung des Wissenschaftlichen Beirats des Instituts, getätigt. Der interne Qualitätssicherungsprozess ist in diesem Flussdiagramm nicht dargestellt.

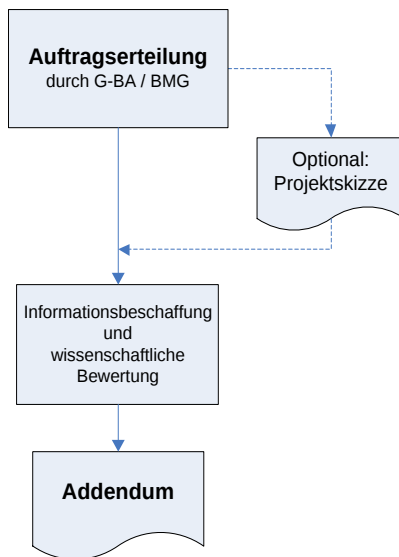


Abbildung 6: Ablauf der Erstellung eines Addendums

Ein Addendum kann in Auftrag gegeben werden, wenn sich im Zuge der Beratungen zu fertiggestellten Institutsprodukten zusätzlicher Bearbeitungsbedarf ergibt. Je nach Art und Umfang der Fragestellung kann es sinnvoll sein, eine **Projektskizze** zu erstellen, in der die wesentlichen Schritte der Informationsbeschaffung und der wissenschaftlichen Bewertung zusammengefasst werden. Die Projektskizze wird nicht veröffentlicht.

Bei dem **Addendum** kann es je nach Art und Umfang der Fragestellung sinnvoll sein, diejenigen externen Sachverständigen zu beteiligen, die an der Erstellung des zugrunde liegenden Institutsprodukts beteiligt waren. Das Verfahren zur Veröffentlichung eines Addendums orientiert sich an dem des ursprünglichen Institutsprodukts. So wird beispielsweise ein Addendum zu Berichten zunächst an den Auftraggeber sowie an den Stiftungsrat und Vorstand versandt. In der Regel 1 Woche später wird es dem Kuratorium der Stiftung übermittelt und weitere 3 Wochen später auf der Website des Instituts veröffentlicht.

2.1.8 Gesundheitsinformationen

Das Institut erstellt **Gesundheitsinformationen** für Bürgerinnen und Bürger in verschiedenen Formaten, die in Abschnitt 7.16 ausführlicher vorgestellt werden.

Grundlagen der Erstellung sind der gesetzliche Auftrag zur Bereitstellung von verständlichen Informationen für Bürgerinnen und Bürger nach § 139a SGB V (siehe Abschnitt 1.1) sowie der Generalauftrag des G-BA (siehe Abschnitt 1.2).

Diese Informationen werden der Öffentlichkeit primär über die Website www.gesundheitsinformation.de zur Verfügung gestellt (auf Englisch unter www.informedhealth.org). Den Kern der Website bilden gesundheits- oder krankheitsbezogene Themen. Je nach Breite und Tiefe können innerhalb eines Themas unterschiedliche Textformate kombiniert werden.

Der Ablauf der Neuerstellung von Gesundheitsinformationen ist in Abbildung 7 schematisch dargestellt.

Nach Festlegung der innerhalb eines Themas zu behandelnden Aspekte erfolgt die systematische **Informationsbeschaffung**, gefolgt von der **wissenschaftlichen Bewertung** der identifizierten Publikationen.

In der Regel werden Gesundheitsinformationen im Anschluss an die **Texterstellung, Redaktion** und ressortinterne Qualitätssicherung im Entwurf zum **externen Review** verschickt.

Details zum Prozess der Themenwahl, der Informationsbeschaffung für Gesundheitsinformationen, der wissenschaftlichen Bewertung und darüber hinaus der Einbeziehung von Betroffenen werden in Kapitel 7 beschrieben.

Der Entwurf einer neuen Gesundheitsinformation wird zur nicht öffentlichen **Stellungnahme** an den Auftraggeber, das Kuratorium, den Vorstand der Stiftung und den Stiftungsrat sowie den Wissenschaftlichen Beirat verschickt. Eine Gesundheitsinformation durchläuft außerdem vor der Veröffentlichung – in der Regel zeitgleich mit dem Stellungnahmeverfahren – eine standardisierte externe Nutzertestung. Die während der in der Regel 4-wöchigen Frist eingegangenen Stellungnahmen und das Ergebnis der Nutzertestung werden zusammengefasst und im Hinblick auf einen daraus resultierenden inhaltlichen oder redaktionellen Änderungsbedarf der Gesundheitsinformation geprüft.

Korrekturen, Verbesserungen und Aktualisierungen der veröffentlichten Gesundheitsinformationen werden in der Regel intern vorgenommen. Bei umfassenden oder grundlegenden inhaltlichen Änderungen erfolgt wiederum ein externes Review. In solchen Fällen können auch eine erneute nicht öffentliche Stellungnahme und externe Nutzertestung erfolgen.

Im Falle der Direktbeauftragung durch den G-BA oder das BMG wird die Gesundheitsinformation in Form eines Berichts, Rapid Reports oder Addendums erarbeitet. Die Erstellung und Publikation unterliegen dem im IQWiG üblichen Verfahren. Anschließend wird in der Regel die betreffende Gesundheitsinformation auch auf www.gesundheitsinformation.de publiziert. Dabei wird, soweit relevant, das Inkrafttreten der entsprechenden Richtlinien des G-BA abgewartet.

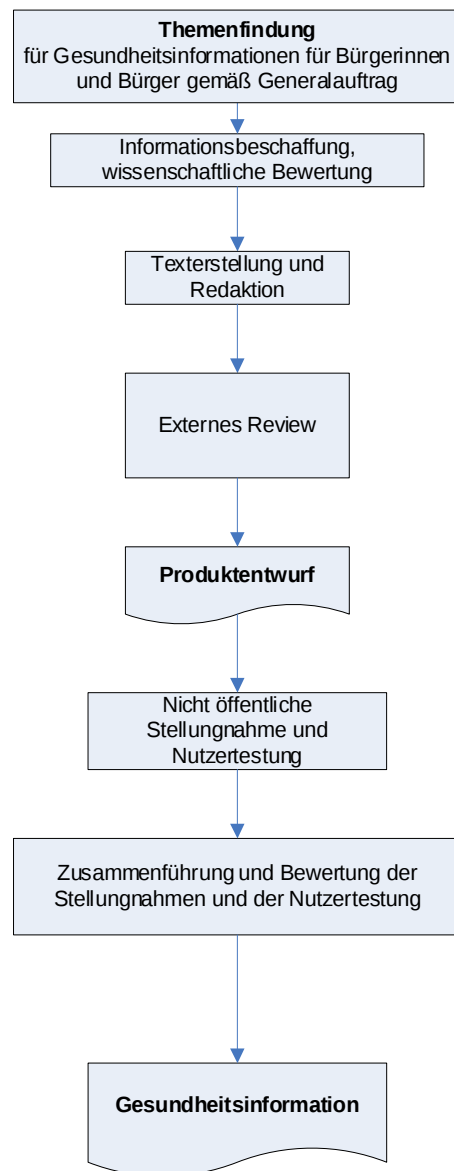


Abbildung 7: Ablauf der Neuerstellung einer Gesundheitsinformation

Eine Variante der Gesundheitsinformationen sind solche, die begleitend zu allen anderen Produkten des Instituts erstellt werden können, um deren Ergebnisse allgemein verständlich für Bürgerinnen und Bürger darzustellen. Für diese gelten folgende Anpassungen im Erstellungsprozess:

- Es findet kein externes Review durch Sachverständige statt, da entsprechende Expertise bereits bei Erstellung des IQWiG-Produkts eingeflossen ist. Es findet ein fachliches Review der Gesundheitsinformation im Institut durch das Ressort statt, das das Produkt erstellt hat.

- Bei Gesundheitsinformationen zu „frühen“ Nutzenbewertungen von Arzneimitteln nach § 35a SGB V (Dossierbewertungen und dazugehörige Addenda) findet aufgrund der engen Fristvorgaben wie in der Satzung des Instituts verankert kein Stimmabgabeverfahren statt.

2.1.9 Arbeitspapier

Der Ablauf der Erstellung eines **Arbeitspapiers** ist in Abbildung 8 schematisch dargestellt. Alle Arbeitsschritte werden in Verantwortung des Instituts, bei Bedarf unter Einbeziehung externer Expertise bzw. des Wissenschaftlichen Beirats des Instituts, getätigt. Der interne Qualitätssicherungsprozess ist in diesem Flussdiagramm nicht dargestellt.

Die Erstellung von Arbeitspapieren erfolgt u. a. im Rahmen des **Generalauftrags**, der am 21.12.2004 durch den G-BA erteilt und im Juli 2006 sowie im März 2008 in Hinblick auf die Erstellung von Gesundheitsinformationen weiter konkretisiert bzw. angepasst wurde. Mit dem Generalauftrag wurde das Institut beauftragt, „durch die Erfassung und Auswertung des relevanten Schrifttums eine kontinuierliche Beobachtung und Bewertung medizinischer Entwicklungen von grundlegender Bedeutung und ihrer Auswirkungen auf die Qualität und Wirtschaftlichkeit der medizinischen Versorgung in Deutschland vorzunehmen und den G-BA hierüber regelmäßig zu informieren. Der G-BA geht bei diesem Auftrag davon aus, dass das Institut auf den ihm gemäß § 139a Abs. 3 SGB V übertragenen Arbeitsfeldern nicht nur Einzelaufträge des G-BA bearbeitet, sondern aus der eigenverantwortlichen wissenschaftlichen Arbeit heraus dem G-BA für dessen gesetzliche Aufgaben notwendige Informationen über versorgungsrelevante Entwicklungen in der Medizin zur Verfügung stellt und konkrete Vorschläge für Einzelaufträge erarbeitet, die aus Sicht des Instituts vor dem Hintergrund dieser Informationen relevant sind“.

Aus dem gesetzlichen Auftrag des Instituts und dem Generalauftrag folgt das Erfordernis eigenständiger wissenschaftlicher Arbeit. Hierunter sind u. a. auch Projekte zur methodischen Weiterentwicklung zu verstehen, deren Ergebnisse ebenfalls als Arbeitspapiere veröffentlicht werden können.

Die **Themenauswahl** erfolgt innerhalb des Instituts, insbesondere auf Basis der im Generalauftrag formulierten Kriterien. An der Formulierung der wissenschaftlichen Fragestellung können auch Patientenorganisationen beteiligt sein oder die Meinung einzelner Betroffener kann eingeholt werden, insbesondere zur Definition patientenrelevanter Endpunkte. Anschließend wird die Projektskizze erstellt.

In der **Projektskizze** werden die wesentlichen Schritte der Informationsbeschaffung und der wissenschaftlichen Bewertung zusammengefasst. Sie bildet die Grundlage für die Erstellung des Arbeitspapiers. Die Projektskizze wird nicht veröffentlicht.

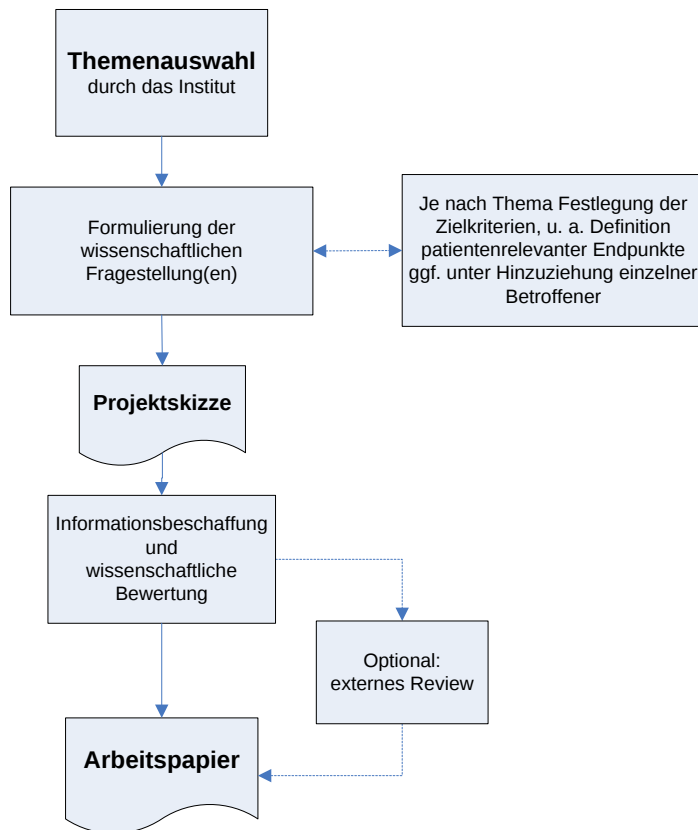


Abbildung 8: Ablauf der Erstellung eines Arbeitspapiers

Im Arbeitspapier werden die Ergebnisse der Informationsbeschaffung und wissenschaftlichen Bewertungen dargestellt. Die Qualitätssicherung kann optional ein externes Review beinhalten. Das Arbeitspapier wird nach seiner Fertigstellung zunächst an den G-BA, den Vorstand der Stiftung sowie den Stiftungsrat versandt, i. d. R. 1 Woche später dem Kuratorium der Stiftung übermittelt und weitere 3 Wochen später auf der Website des Instituts veröffentlicht. Sollten Stellungnahmen zu Arbeitspapieren eingehen, die substanzielle nicht berücksichtigte Evidenz enthalten, oder erlangt das Institut auf andere Weise Kenntnis von solcher Evidenz, wird geprüft, ob eine Aktualisierung des Arbeitspapiers aus Sicht des Instituts erforderlich erscheint. Eine solche Aktualisierung unterliegt den generellen methodischen und verfahrenstechnischen Anforderungen an Produkte des Instituts.

2.1.10 HTA-Bericht

Der Ablauf der Erstellung der HTA-Berichte gemäß § 139b Abs. 5 SGB V ist in Abbildung 9 schematisch dargestellt. Der HTA-Bericht wird von den externen Sachverständigen (siehe Abschnitt 2.2.2) unter Anwendung der Methodik des IQWiG erstellt und vom IQWiG durch einen Herausgeberkommentar ergänzt. Die internen Qualitätssicherungsprozesse sind in diesem Flussdiagramm nicht dargestellt. Nach dem Abschluss der Themenauswahl für die HTA-Berichte durch das Institut (siehe Abschnitt 6.3) erfolgt die Formulierung der wissenschaftlichen HTA-Fragestellung.

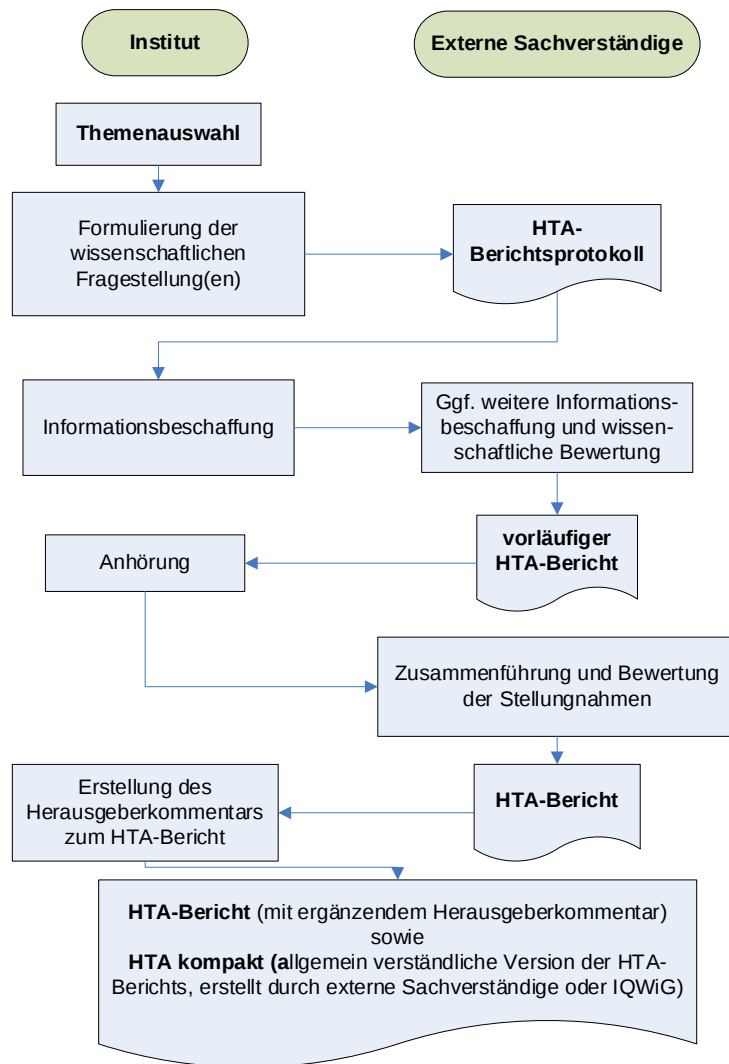


Abbildung 9: Ablauf der Erstellung eines HTA-Berichts

Das **HTA-Berichtsprotokoll** wird durch externe Sachverständige erstellt. Es 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 entsprechend der Methodik des Instituts. Es bildet die Grundlage für die Erstellung des vorläufigen HTA-Berichts. Das HTA-Berichtsprotokoll wird auf der Website des Instituts veröffentlicht und der Vorstand der Stiftung, der Stiftungsrat, das Kuratorium sowie der G-BA und das BMG werden hierüber informiert.

Im **vorläufigen HTA-Bericht** stellen die externen Sachverständigen die Ergebnisse der Informationsbeschaffung und der wissenschaftlichen Bewertung einschließlich eines eigenen Fazits dar. Er wird nach Fertigstellung ebenfalls auf der Website des Instituts veröffentlicht und der Vorstand der Stiftung, der Stiftungsrat, das Kuratorium sowie der G-BA und das BMG werden hierüber informiert.

Der vorläufige HTA-Bericht wird zur Anhörung gestellt. Die Anhörung erfolgt grundsätzlich mittels schriftlicher Stellungnahmen, die innerhalb einer Frist von mindestens 4 Wochen ab Bekanntmachung auf der Website des Instituts abgegeben werden können. Optional wird eine mündliche wissenschaftliche Erörterung mit Stellungnehmenden durchgeführt. Diese Erörterung dient der gegebenenfalls notwendigen Klärung von Inhalten der schriftlichen Stellungnahmen. Gegenstand der Anhörung sind neben dem projektspezifischen methodischen Vorgehen zur Beantwortung der Fragestellung insbesondere die im vorläufigen HTA-Bericht dargestellten Ergebnisse der Informationsbeschaffung und -bewertung. Die Anhörung wird vom IQWiG administriert und durchgeführt. Es erfolgt eine Würdigung der Stellungnahmen durch die externen Sachverständigen im HTA-Bericht.

Der **HTA-Bericht** enthält aufbauend auf dem vorläufigen HTA-Bericht die Bewertung der wissenschaftlichen Erkenntnisse unter Berücksichtigung der Ergebnisse der Anhörung zum vorläufigen HTA-Bericht und stellt das abschließende Produkt den externen Sachverständigen dar. Der HTA-Bericht wird durch einen **Herausgeberkommentar** eingeleitet, in dem eine Einordnung der Ergebnisse durch das Institut erfolgt. Zum HTA-Bericht gehört zudem **HTA kompakt**, eine allgemein verständliche Zusammenfassung des HTA-Berichts.

Der HTA-Bericht, HTA kompakt und die Dokumentation der Anhörung zum vorläufigen HTA-Bericht werden zunächst dem Vorstand der Stiftung, dem Stiftungsrat, dem Kuratorium der Stiftung sowie dem G-BA und dem BMG zugeleitet. In der Regel weitere 2 Wochen später erfolgt die Veröffentlichung dieser Dokumente auf der Website des Instituts.

2.2 Generelle Aspekte bei der Produkterstellung

In diesem Kapitel werden folgende für alle Produkte geltenden Abläufe und Aspekte dargestellt:

- die Einbindung von Betroffenen in die Erstellung von IQWiG-Produkten,
- die Auswahl externer Sachverständiger für die Mitarbeit an der Produkterstellung,
- die Gewährleistung der fachlichen Unabhängigkeit bei der Produkterstellung,
- das Review der Produkte,
- das Stellungnahmeverfahren,
- die Veröffentlichung der Produkte.

Darüber hinaus wird die Durchführung wissenschaftlicher Beratung (Scientific Advice) im Verhältnis zur Produkterstellung beschrieben.

2.2.1 Einbindung von Betroffenen in die Erstellung von IQWiG-Produkten

Die Einbindung von Betroffenen im Rahmen der Erstellung von systematischen Übersichten und HTA-Berichten ist mittlerweile etablierter internationaler Standard der Nutzenbewertung

[144,237,451]. Am IQWiG findet die Einbindung Betroffener primär zu Beginn der Projektbearbeitung im Rahmen der Festlegung patientenrelevanter Endpunkte und relevanter Subgruppen statt. Darüber hinaus haben Betroffene im Rahmen der Anhörungsverfahren die Möglichkeit, sich in die Berichterstellung einzubringen. Im Rahmen der Erstellung von Gesundheitsinformationen können Selbsthilfeorganisationen zum Informationsbedarf von Betroffenen und zu den Herausforderungen bei der Bewältigung der Erkrankung befragt werden. Die Entwürfe der Gesundheitsinformationen werden regelhaft einer externen Nutzer-
testung unterzogen. Darüber hinaus werden themenspezifisch Erfahrungsberichte von Betroffenen erhoben.

Betroffene können insbesondere sein: Patientinnen und Patienten (ggf. vertreten durch ihre Eltern oder sonstige Angehörige) sowie potenzielle Teilnehmerinnen und Teilnehmer an Präventionsmaßnahmen.

Bei der Auswahl der Teilnehmenden wird der Fokus auf tatsächlich Betroffene gerichtet, denn Patientenvertreterinnen und Patientenvertreter oder Repräsentantinnen und Repräsentanten von Selbsthilfegruppen sind selbst teilweise keine Patientinnen oder Patienten und können nicht in vergleichbarer Weise beurteilen, wie Betroffene Symptome empfinden, ihre Funktionen und Aktivitäten wahrnehmen können oder in ihrer Lebensqualität beeinträchtigt sind.

Zur Findung Betroffener wird regelhaft die Patientenvertretung im G-BA gebeten, selbst oder über ihre Mitgliedsorganisationen Betroffene zu einem Thema zu benennen. Ergänzend können vom Institut auch über nationale oder lokale Selbsthilfeorganisationen oder -gruppen, über Krankenhäuser oder Arztpraxen, externe Sachverständige oder über sonstige Wege Betroffene gesucht oder angefragt werden.

Zur Einbindung von Betroffenen existieren 2 verschiedene Wege: Zum Ersten kann ein persönliches Konsultationsgespräch geführt werden, bei dem Betroffene und Institutsmitarbeiterinnen sowie -mitarbeiter in kleinem Kreis über eine Erkrankung sprechen. Zum Zweiten können die Sichtweise und die Erfahrungen von Betroffenen schriftlich erfragt werden. In beiden Fällen werden potenzielle Interessenkonflikte der Betroffenen erfasst (analog zur Vorgehensweise wie in Abschnitt 2.2.2 beschrieben). Betroffene mit Interessenkonflikten werden jedoch nicht von der Konsultation ausgeschlossen. Die Namen der Betroffenen, die am Konsultationsgespräch teilgenommen oder den Fragebogen zur Konsultation ausgefüllt haben, werden grundsätzlich nicht veröffentlicht, es sei denn, sie willigen explizit in die Namensnennung ein.

In welcher Form Betroffene in die Arbeit des Instituts eingebunden werden können, hängt vor allem davon ab, wie viel Zeit für die Auftragsbearbeitung zur Verfügung steht und ob das Auftragsthema vertraulich ist. Ferner ist eine (erneute) Einbindung von Betroffenen verzichtbar, wenn das Auftragsthema bereits im Rahmen einer früheren Konsultation bearbeitet wurde. Tabelle 2 zeigt für alle Produkte des Instituts die Form möglicher Betroffenenbindung; zum jeweiligen produktspezifischen Ablauf siehe Abschnitt 2.1.

Tabelle 2: Institutsprodukte und Formen der möglichen Einbindung Betroffener

Institutsprodukt	Form der Einbindung Betroffener
Bericht Rapid Report	mündliche Konsultation
Dossierbewertung Kosten-Nutzen-Bewertung nach § 35b SGB V	schriftliche Konsultation
Potenzialbewertung Bewertung gemäß § 137h SGB V Addendum	keine Einbindung
Gesundheitsinformation	mündliche Konsultation, Nutzertestungen, Erfahrungsberichte
Arbeitspapier	mündliche oder schriftliche Konsultation nach Bedarf
HTA-Bericht	mündliche Konsultation
HTA: Health Technology Assessment; SGB: Sozialgesetzbuch Fünftes Buch – Gesetzliche Krankenversicherung	

2.2.2 Auswahl externer Sachverständiger

Gemäß seinem gesetzlichen Auftrag bezieht das Institut externe Sachverständige in seine Arbeit ein. Als externe Sachverständige gelten dabei Personen, an die wissenschaftliche Forschungsaufträge im Rahmen der Erstellung oder Begutachtung von Institutsprodukten vergeben werden oder die das Institut in medizinisch-fachlichen Fragestellungen beraten. Das Institut vergibt diese Aufträge nach den allgemeinen Grundsätzen des Beschaffungswesens in einem transparenten und nicht diskriminierenden Wettbewerb.

Auftragsbekanntmachungen für Forschungsaufträge nach § 139b Abs. 3 SGB V sowie § 139b Abs. 5 SGB V werden auf der Website des Instituts veröffentlicht. Ausnahmen sind insbesondere bei eilbedürftigen Aufträgen möglich. Aufträge mit einem Auftragsvolumen oberhalb des aktuell geltenden Schwellenwerts der Vergabebestimmungen der Europäischen Union (EU) werden EU-weit ausgeschrieben. Die konkreten Eignungsanforderungen an die Bewerber sind den jeweiligen Bekanntmachungen bzw. den Vergabeunterlagen zu entnehmen.

Die Beauftragung externer Sachverständiger bei Dossierbewertungen, Kosten-Nutzen-Bewertungen nach § 35b SGB V, Potenzialbewertungen, Bewertungen gemäß § 137h SGB V und bei der Erstellung von Gesundheitsinformationen erfolgt auf der Basis von Angaben Interessierter in einer Sachverständigendatenbank. Für die Aufnahme in die Sachverständigendatenbank ist auf der Website des Instituts ein Zugang eingerichtet, über den die Interessentinnen und Interessenten unter Angabe ihres Fachgebiets und ihrer fachlichen Expertise ihr Sachverständigenprofil eingeben können. Für die zu vergebenden Projekte wird anhand einer Kriterienliste die oder der jeweils am besten geeignete Bewerberin oder Bewerber des entsprechenden Fachgebiets aus der Sachverständigendatenbank ausgewählt und beauftragt. Weitere Angaben zum Auswahlverfahren sind auf der Website des Instituts veröffentlicht.

2.2.3 Gewährleistung der fachlichen Unabhängigkeit

Die wissenschaftliche und fachliche Unabhängigkeit des Instituts und der von ihm vertretenen und veröffentlichten Inhalte ist gesetzlich in § 139a SGB V sowie in der Satzung der Stiftung verankert.

A) Gewährleistung der internen fachlichen Unabhängigkeit

Den wissenschaftlichen Mitarbeiterinnen und Mitarbeitern des Instituts sind Nebentätigkeiten, die grundsätzlich geeignet sind, ihre fachliche Unabhängigkeit infrage zu stellen, untersagt. Einzelheiten sind in den Arbeitsverträgen und internen Regelungen festgelegt.

B) Gewährleistung der Unabhängigkeit externer Sachverständiger

Vor dem Abschluss eines Vertrages zwischen dem Institut und einem externen Sachverständigen oder einer externen Institution zur Erstellung eines Produkts müssen dem Institut gemäß § 139b SGB V „alle Beziehungen zu Interessenverbänden, Auftragsinstituten, insbesondere der pharmazeutischen Industrie und der Medizinprodukteindustrie, einschließlich Art und Höhe von Zuwendungen“ offengelegt werden.

- Aufbauend auf der in der Wissenschaft üblichen Offenlegung solcher Beziehungen [475,481] interpretiert das Institut diese Regelung als Aufgabe, im Rahmen der Auswahl von externen Sachverständigen die Offenlegungen im Hinblick auf die fachliche Unabhängigkeit und Unbefangenheit der Bewerberinnen und Bewerber zu beurteilen bzw. dahin gehend, ob aufgrund von Interessenkonflikten gravierende Bedenken hinsichtlich einer sachgerechten Mitarbeit bestehen. Sollte das der Fall sein, kommt eine Mitarbeit für dieses Auftrags-thema i. d. R. nicht oder nur unter besonderen Vereinbarungen infrage. Da diese Bewertung auftragsbezogen stattfindet, ist eine Mitarbeit bei anderen Auftrags-themen sehr wohl möglich. Der weitere Prozess der Auswahl von externen Sachverständigen ist in Abschnitt 2.2.2 beschrieben.
- Die Hauptgrundlage der Bewertung von Beziehungen sind Selbstauskünfte unter Verwendung des „Formblatts zur Offenlegung von Beziehungen“. Dieses Formblatt steht auf der Website des Instituts zur Verfügung. Die Selbstauskunft bezieht sich auf folgende Arten von Beziehungen:
 - Anstellungsverhältnisse / selbstständige Tätigkeiten / ehrenamtliche Tätigkeiten,
 - Beratungstätigkeiten,
 - Honorare zum Beispiel für Vorträge, Stellungnahmen, die Ausrichtung von und / oder Teilnahme an Kongressen und Seminaren,
 - finanzielle Unterstützung für Forschungsaktivitäten, andere wissenschaftliche Leistungen oder Patentanmeldungen,
 - sonstige finanzielle oder geldwerte Zuwendungen (z. B. Ausrüstung, Personal, Reisekostenunterstützung ohne wissenschaftliche Gegenleistung),

- Aktien, Optionsscheine oder sonstige Geschäftsanteile, Patente, Gebrauchsmuster,
- sonstige Umstände, die aus Sicht eines unvoreingenommenen Betrachters als Interessenkonflikt bewertet werden können (z. B. Aktivitäten in gesundheitsbezogenen Interessengruppierungen bzw. Selbsthilfegruppen).

Das Institut behält sich vor, zusätzliche Informationen heranzuziehen und Angaben auf ihre Vollständigkeit und Richtigkeit zu überprüfen.

Die Namen der externen Sachverständigen, die an der Erstellung der Produkte des Instituts beteiligt waren, werden in der Regel in den entsprechenden Produkten veröffentlicht. Diese Veröffentlichungen sind grundsätzlich über die Website frei zugänglich. Die Angaben zu Beziehungen werden nur zusammenfassend veröffentlicht. Dabei wird für die im Formblatt erfassten Arten von Beziehungen lediglich genannt, ob diese Art der Beziehung bestand. Konkrete Einzelheiten, wie zum Beispiel die Namen von Partnerinnen oder Partnern oder die Höhe von Zuwendungen, werden nicht veröffentlicht.

2.2.4 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 aller produktspezifischen Veröffentlichungen vor Auftragsabschluss unterliegen einem umfangreichen mehrstufigen internen Qualitätssicherungsverfahren. Darüber hinaus kann im Verlauf der Produkterstellung ein externes Reviewverfahren als optionaler weiterer Schritt der Qualitätssicherung durchgeführt werden. Die Auswahl der internen und externen Reviewerinnen und Reviewer erfolgt primär auf Basis ihrer methodischen und / oder fachlichen Expertise.

Die Identifikation externer Reviewerinnen und 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 Reviewerinnen und Reviewer ist die Darlegung potenzieller Interessenkonflikte erforderlich.

Die Auswahl der externen Reviewerinnen und Reviewer erfolgt durch das Institut. Eine Höchstgrenze für Reviewerinnen und Reviewer 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 Reviewerinnen und 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 Reviewerinnen und Reviewer ist durch die Veröffentlichung der Institutsprodukte und die damit verbundene Möglichkeit zur Stellungnahme ein offenes und unabhängiges Reviewverfahren gewährleistet.

2.2.5 Stellungnahmeverfahren

A) 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 durchgeführt werden und der Personenkreis der Stellungnahmeberechtigten dabei nicht begrenzt wird. Darüber hinaus werden alle Institutsprodukte nach § 139a SGB V vor der Publikation dem Kuratorium des Instituts zugeleitet. Im Kuratorium sind Patientenorganisationen, der oder die Beauftragte der Bundesregierung für die Belange der Patientinnen und Patienten, Organisationen der Leistungserbringer und der Sozialpartner und die Selbstverwaltungsorgane der Trägerorganisationen des G-BA vertreten.

B) Ablauf des öffentlichen Stellungnahmeverfahrens (Anhörungen)

Der Beginn und die Stellungnahmefrist werden auf der Website des Instituts bekannt gemacht. Interessierte können anschließend schriftliche Stellungnahmen beim Institut einreichen. Die Einreichung ist elektronisch (bevorzugt) oder auf dem Postweg möglich. Optional wird ergänzend eine wissenschaftliche Erörterung mit Stellungnehmenden durchgeführt mit dem Ziel, Inhalte schriftlicher Stellungnahmen zu klären. Auf Anfrage können Mitarbeiterinnen und Mitarbeiter der Geschäftsstelle des G-BA an der Erörterung teilnehmen.

Um die Arbeit des Instituts nicht unangemessen zu verzögern, müssen die Stellungnahmen bestimmten formalen Anforderungen genügen. Die Fristen sind in den jeweiligen Abschnitten zu produktspezifischen Verfahrensabläufen beschrieben (siehe Abschnitt 2.1). Weiterführende Angaben zum Stellungnahmeverfahren einschließlich der Bedingungen für die Teilnahme an einer wissenschaftlichen Erörterung finden sich auf der Website des Instituts.

Stellungnahmen, die den formalen Anforderungen genügen, werden auf der Website des Instituts in einem gesonderten Dokument (Dokumentation der Anhörung) veröffentlicht. Zur Gewährleistung der Transparenz werden mit den Stellungnahmen eingereichte Unterlagen, die nicht öffentlich zugänglich sind (z. B. Manuskripte), sowie ein Wortprotokoll der mündlichen wissenschaftlichen Erörterung – sofern sie stattgefunden hat – ebenfalls veröffentlicht.

Im Rahmen einer Anhörung besteht die Möglichkeit, qualitativ angemessene Unterlagen jeglicher Art, die aus Sicht der oder des jeweiligen Stellungnehmenden für die Beantwortung der Fragestellung geeignet sind, vorzulegen. Falls beispielsweise die in einem Berichtsplan definierte Suchstrategie 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.

2.2.6 Veröffentlichung der Produkte des Instituts

Wesentliche Aufgabe des Instituts ist es, die vorhandene Evidenz durch eine sorgfältige Prüfung von verfügbaren Informationen zu ergründen und über die Ergebnisse dieser Prüfung zu informieren. Es ist gesetzlich festgelegt, dass das Institut „in regelmäßigen Abständen über die Arbeitsprozesse und -ergebnisse einschließlich der Grundlagen für die Entscheidungsfindung öffentlich zu berichten“ hat (§ 139a Abs. 4 SGB V).

Zur Wahrung der Unabhängigkeit des Instituts muss ausgeschlossen werden, dass die Auftraggeber oder interessierte Dritte Einfluss auf die Inhalte der Berichte nehmen können. Dies könnte zu einer Vermengung der wissenschaftlichen Ergebnisse mit politischen oder wirtschaftlichen Aspekten oder Interessen führen. Gleichzeitig muss vermieden werden, dass das Institut seinerseits bestimmte Ergebnisse zurückhält. Deshalb werden alle Ergebnisse, die das Institut im Rahmen seiner gesetzlichen Aufgaben erarbeitet (mit Ausnahme von Potenzialbewertungen, siehe hierzu 2. Kapitel § 19 der Verfahrensordnung des G-BA [276]), möglichst zeitnah veröffentlicht. Dazu gehört im Falle von Berichten auch der Berichtsplan. Produktspezifische Besonderheiten sind in den jeweiligen Abschnitten, in denen die Verfahrensabläufe beschrieben werden, angegeben. Von den dargestellten Regelfällen (Latenzzeit zwischen Fertigstellung und Veröffentlichung) kann in begründeten Ausnahmefällen abgewichen werden.

Die Rechte an der Publikation liegen, falls nicht anders vereinbart, beim Institut.

2.2.7 Wissenschaftliche Beratung

Das Institut beteiligt sich in speziellen Fällen an der wissenschaftlichen Beratung (Scientific Advice) von Studiensponsoren, z. B. im Rahmen der Mitarbeit im europäischen Netzwerk EUnetHTA. Primäres Ziel ist dabei die Unterstützung des Designs von Studien, die aussagekräftige Daten für die Nutzenbewertung liefern. Um die Unabhängigkeit der Bewertung auch in den Fällen zu gewährleisten, in denen die Nutzenbewertung Studien enthält, zu denen das Institut eine Beratung durchgeführt hat, wird eine angemessene organisatorische Trennung von Beratung und Bewertung sichergestellt.

3 Nutzenbewertung medizinischer Interventionen

3.1 Patientenrelevanter medizinischer Nutzen und Schaden

3.1.1 Definition des patientenrelevanten medizinischen Nutzens bzw. Schadens

Mit dem Begriff **Nutzen** werden kausal begründete positive Effekte, mit dem Begriff **Schaden** kausal begründete negative Effekte einer medizinischen Intervention auf patientenrelevante Endpunkte (s. u.) bezeichnet. Kausal begründet meint in diesem Zusammenhang, dass ausreichend Sicherheit besteht, dass beobachtete Effekte allein auf die zu prüfende Intervention zurückgeführt werden können [761].

Sofern ein Vergleich nicht explizit genannt ist, beziehen sich die Begriffe Nutzen und Schaden auf einen Vergleich mit Placebo (oder einer andersartigen Scheinintervention) oder keiner Behandlung. Im Falle eines Vergleichs der zu evaluierenden medizinischen Intervention mit einer anderen eindeutig definierten medizinischen Intervention werden für die vergleichende Bewertung der Nutzen- bzw. Schadenaspekte folgende Begriffe verwendet; die Beschreibung erfolgt dabei immer ausgehend von der zu evaluierenden Intervention:

- Nutzenaspekte:
 - Im Falle eines Vorteils wird von einem höheren Nutzen im Vergleich zur anderen Intervention gesprochen. Eine Ausnahme bilden Dossierbewertungen, für die in diesem Fall der Begriff Zusatznutzen anstelle des Begriffs höherer Nutzen verwendet wird.
 - Im Falle eines Nachteils wird von einem geringeren Nutzen gesprochen.
 - Bei vergleichbaren Effekten wird von einem vergleichbaren Nutzen gesprochen.
- Schadenaspekte:
 - Es werden die Begriffe höherer Schaden, vergleichbarer Schaden und geringerer Schaden verwendet.

Die Bewertung der Evidenz soll nach Möglichkeit in die eindeutige Feststellung münden, dass entweder das Vorliegen eines (Zusatz-)Nutzens (bzw. Schadens) einer Maßnahme oder das Fehlen eines (Zusatz-)Nutzens (bzw. Schadens) belegt ist oder das Vorliegen oder Fehlen eines (Zusatz-)Nutzens (bzw. Schadens) nicht belegt und daher unklar ist, ob ein (Zusatz-)Nutzen (bzw. Schaden) durch die Maßnahme erzielt wird. Darüber hinaus kann es sinnvoll sein, für den Fall eines nicht eindeutig belegten (Zusatz-)Nutzens (bzw. Schadens) eine weitere Untergliederung vorzunehmen, ob nämlich zumindest Hinweise auf oder auch nur Anhaltspunkte für einen (Zusatz-)Nutzen (bzw. Schaden) bestehen (siehe Abschnitt 3.1.4).

Da sich der Nutzen einer Maßnahme auf Patientinnen und Patienten beziehen soll, beruht diese Bewertung auf Ergebnissen wissenschaftlicher Untersuchungen zur Beeinflussung patientenrelevanter Endpunkte. Als patientenrelevant soll in diesem Zusammenhang verstanden werden,

wie eine Patientin oder ein Patient fühlt, ihre oder seine Funktionen und Aktivitäten wahrnehmen kann oder ob sie oder er überlebt [59]. Dabei werden sowohl die beabsichtigten als auch die unbeabsichtigten Effekte der Interventionen berücksichtigt, die eine Bewertung der Beeinflussung insbesondere folgender patientenrelevanter Zielgrößen zur Feststellung krankheits- und behandlungsbedingter Veränderungen erlauben:

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

Diese Zielgrößen sind auch im SGB V als vorrangig zu berücksichtigende Zielgrößen genannt, z. B. in § 35 Abs. 1b SGB V. Ergänzend kann der interventions- und erkrankungsbezogene Aufwand der Behandlung berücksichtigt werden. Auch die Patientenzufriedenheit kann ergänzend einbezogen werden, sofern hierbei gesundheitsbezogene Aspekte abgebildet werden. Ein Nutzen oder Zusatznutzen kann sich allein auf Basis dieser beiden Zielgrößen jedoch nicht ergeben.

Für alle genannten Zielgrößen gilt, dass es erforderlich sein kann, diese jeweils nur im Zusammenhang mit Informationen zur vergleichenden Beeinflussung anderer Zielgrößen zu bewerten. Bei besonders schwerwiegenden oder gar lebensbedrohlichen Erkrankungen ist es beispielsweise in der Regel nicht allein ausreichend, eine Verbesserung der Lebensqualität durch die Anwendung einer zu bewertenden Intervention nachzuweisen, wenn nicht gleichzeitig mit ausreichender Sicherheit ausgeschlossen werden kann, dass die schwerwiegende Morbidität oder gar Mortalität in einem nicht mehr akzeptablen Ausmaß ungünstig beeinflusst wird. Dies entspricht im Grundsatz der höchstrichterlichen Rechtsprechung, dass bestimmte (Nutzen-)Aspekte erst dann notwendigerweise zu bewerten sind, wenn die therapeutische Wirksamkeit hinreichend belegt ist [108]. Darüber hinaus wird umgekehrt in vielen Bereichen, insbesondere in palliativen Therapiesituationen, eine Beeinflussung der Mortalität nicht ohne Kenntnis damit einhergehender (ggf. ungünstiger) Effekte auf die Lebensqualität adäquat zu bewerten sein.

Gemäß § 35b Abs. 1 Satz 4 SGB V sollen beim Patientennutzen insbesondere die Verbesserung des Gesundheitszustands, eine Verkürzung der Krankheitsdauer, eine Verlängerung der Lebensdauer, eine Verringerung der Nebenwirkungen sowie eine Verbesserung der Lebensqualität angemessen berücksichtigt werden. Diese Nutzendimensionen sind durch die oben aufgeführten Zielgrößen abgebildet. So sind die Verbesserung des Gesundheitszustands und die Verkürzung der Krankheitsdauer Aspekte der unmittelbaren krankheitsbedingten Morbidität, die Verringerung von Nebenwirkungen ist ein Aspekt der therapiebedingten Morbidität. Es werden in erster Linie Endpunkte berücksichtigt, die zuverlässig und direkt konkrete Änderungen des Gesundheitszustands abbilden. Dabei werden insbesondere einzelne Betroffene in die themenbezogene Definition patientenrelevanter Endpunkte einbezogen. Zur Erfassung der Lebensqualität sollen nur Instrumente eingesetzt werden, die für den Einsatz in

klinischen Studien geeignet und entsprechend evaluiert sind [227]. Darüber hinaus können für die Nutzenbewertung valide Surrogatendpunkte berücksichtigt werden (siehe Abschnitt 3.1.2).

Sowohl Nutzen- als auch Schadenaaspekte können eine unterschiedliche Wichtigkeit für die Betroffenen haben, die sich ggf. durch qualitative Erhebungen oder bereits bei der Beratung durch Betroffene im Zusammenhang mit der Definition patientenrelevanter Endpunkte abzeichnet (entsprechende Methoden werden am Ende des Abschnitts 3.1.4 exemplarisch aufgeführt). In einer solchen Situation kann es sinnvoll sein, eine Hierarchisierung von Endpunkten vorzunehmen. Allgemeine Nutzen- bzw. Schadenaussagen werden sich dann in erster Linie auf Belege bezüglich der höher gewichteten Zielgrößen stützen. Geplante Subgruppen- und Sensitivitätsanalysen werden dann im Wesentlichen für die höher gewichteten Zielgrößen durchgeführt, während solche Analysen für die verbleibenden Zielgrößen nicht routinemäßig durchlaufen werden.

Diagnostische Maßnahmen können indirekt nutzen, indem sie eine notwendige Voraussetzung für therapeutische Interventionen sind, durch die das Erreichen eines Effekts auf die oben genannten patientenrelevanten Zielgrößen möglich wird. Voraussetzung für den Nutzen diagnostischer Maßnahmen sind also die Existenz und der belegte Nutzen einer Behandlung bei Patientinnen und Patienten in Abhängigkeit vom Testergebnis.

Interventionen können auch Auswirkungen auf indirekt Betroffene wie zum Beispiel Familienangehörige und Pflegepersonen haben. Diese Auswirkungen können im Rahmen der Institutsberichte gegebenenfalls auch berücksichtigt werden.

Der Begriff **Nutzenbewertung** bezeichnet den gesamten Prozess der Evaluation medizinischer Interventionen hinsichtlich ihrer kausal begründeten positiven und negativen Effekte im Vergleich mit einer klar definierten anderen Therapie, einem Placebo (oder einer andersartigen Scheinbehandlung) oder keiner Behandlung. Dabei werden Nutzen- und Schadenaaspekte zunächst endpunktbezogen evaluiert und dargestellt. Darüber hinaus ist eine gemeinsame Würdigung der endpunktbezogenen Nutzen- und Schadenaaspekte möglich (siehe Abschnitt 3.1.4), sodass beispielsweise ein endpunktbezogener geringerer Schaden (im Sinne einer Verringerung von Nebenwirkungen) bei Betrachtung der Effekte auf alle anderen Endpunkte in die abwägende Feststellung eines Zusatznutzens münden kann.

3.1.2 Surrogate patientenrelevanter Endpunkte

Surrogatendpunkte werden in der medizinischen Forschung häufig als Ersatz für patientenrelevante Endpunkte verwendet, meist um Aussagen zum patientenrelevanten (Zusatz-)Nutzen früher und einfacher zu erhalten [20,254,577]. Die meisten Surrogatendpunkte sind jedoch in dieser Hinsicht nicht verlässlich und können bei der Nutzenbewertung irreführend sein [135,294,302]. Daher werden Surrogatendpunkte im Rahmen der Nutzenbewertung des Instituts in der Regel nur dann in Betracht gezogen, wenn sie zuvor anhand geeigneter statistischer Methoden innerhalb einer hinreichend eingegrenzten Patientenpopulation und

innerhalb von vergleichbaren Interventionen (z. B. Arzneimittel mit vergleichbarem Wirkmechanismus) validiert wurden. Ein Surrogatendpunkt kann dann als valide gelten, wenn der Effekt auf den zu ersetzenden patientenrelevanten Endpunkt durch den Effekt auf den Surrogatendpunkt in einem ausreichenden Ausmaß erklärt wird [36,749]. Die Notwendigkeit der Betrachtung von Surrogatendpunkten kann im Rahmen der frühen Nutzenbewertung von Arzneimitteln (siehe Abschnitt 3.3.3) eine besondere Bedeutung haben, da in den Zulassungsverfahren primär die Wirksamkeit, aber nicht immer der patientenrelevante Nutzen oder Zusatznutzen untersucht wird.

Für die Validierung eines Surrogatendpunkts gibt es weder ein Standardverfahren noch eine allgemein beste Schätzmethode noch ein allgemein akzeptiertes Kriterium, dessen Erfüllung den Nachweis der Validität bedeuten würde [498]. Allerdings widmet sich die methodische Literatur sehr häufig korrelationsbasierten Verfahren zur Surrogatvalidierung mit Schätzung von Korrelationsmaßen auf Studien- und individueller Ebene [378]. In der Nutzenbewertung des Instituts werden daher bevorzugt Validierungen auf Basis eines solchen Verfahrens berücksichtigt. Diese Verfahren bedürfen in der Regel einer Metaanalyse von mehreren randomisierten Studien, in denen sowohl Effekte auf den Surrogatendpunkt als auch Effekte auf den interessierenden patientenrelevanten Endpunkt untersucht wurden [112,521]. Alternative Methoden [749] werden nur in begründeten Ausnahmefällen in Betracht gezogen.

Bei korrelationsbasierten Verfahren wird für den Nachweis der Validität in der Regel einerseits auf individueller Ebene eine hohe Korrelation zwischen Surrogat und patientenrelevantem Endpunkt sowie andererseits auf Studienebene eine hohe Korrelation zwischen den Effekten auf das Surrogat und den patientenrelevanten Endpunkt verlangt [112,114]. Da in der Nutzenbewertung des Instituts Aussagen bezogen auf Gruppen von Patientinnen und Patienten gemacht werden, stützt sich die Bewertung der Validität eines Surrogatendpunkts hier primär auf den Grad der Korrelation auf der Ebene der Behandlungseffekte. Zusätzlich zur Höhe der Korrelation wird für die Bewertung der Validität eines Surrogatendpunkts die Aussage-sicherheit der Ergebnisse zur Validierung berücksichtigt. Dazu werden verschiedene Kriterien herangezogen [378]. Zum Beispiel sind beobachtete Zusammenhänge zwischen einem Surrogat- und dem entsprechenden patientenrelevanten Endpunkt für eine Intervention mit einem bestimmten Wirkmechanismus nicht zwangsläufig auf Interventionen zur Behandlung derselben Erkrankung, aber mit einem anderen Wirkmechanismus anwendbar [252,294,302,498]. Die Studien, auf denen die Validierung basiert, müssen daher mit Patientenkollektiven und Interventionen durchgeführt worden sein, die Aussagen über das der Nutzenbewertung zugrunde liegende Anwendungsgebiet und die zu bewertende Intervention sowie die Vergleichsintervention erlauben. Zur Beurteilung der Übertragbarkeit sollten bei Validierungsstudien, die verschiedene Krankheitsentitäten oder Interventionen einschließen, mindestens geeignete Untersuchungen zur Heterogenität vorliegen.

Im Falle, dass ein Surrogatendpunkt nicht abschließend validiert werden konnte (zum Beispiel durch eine nicht ausreichend hohe Korrelation), besteht darüber hinaus die Möglichkeit der Anwendung des Konzepts eines sogenannten Surrogate-Threshold-Effekts (STE) [111,378].

Hierzu wird der Effekt auf das Surrogat, der sich aus den Studien zur Nutzenbewertung ergibt, dem STE gegenübergestellt [114,521].

Die Schlussfolgerungen für eine Nutzenbewertung des Instituts bezüglich patientenrelevanter Endpunkte können aus den Effekten auf das Surrogat in Abhängigkeit vom Validitätsnachweis des Surrogats bzw. der Betrachtung eines STE gezogen werden. Entscheidend für Ersteres sind der Grad der Korrelation der Effekte auf das Surrogat und den patientenrelevanten Endpunkt und die Aussagesicherheit der Validierung in den Validierungsstudien. Bei Betrachtung eines STE ist die Größe des Effekts auf das Surrogat in den Studien zur Nutzenbewertung im Vergleich zum STE das entscheidende Kriterium. Je nach Konstellation sind bei einem statistisch signifikanten Effekt auf die Surrogatendpunkte alle Abstufungen der Aussagen zum (Zusatz-)Nutzen hinsichtlich des zugehörigen patientenrelevanten Endpunkts gemäß Abschnitt 3.1.4 möglich.

Surrogatendpunkte, die nicht valide sind oder für die kein adäquates Validierungsverfahren durchgeführt wurde, können dennoch in den Berichten des Instituts dargestellt werden. Derartige Endpunkte sind aber unabhängig von den beobachtbaren Effekten nicht als Belege für einen Nachweis des (Zusatz-)Nutzens einer Intervention geeignet.

Je nach Nähe zu einem entsprechenden patientenrelevanten Endpunkt werden in der Literatur verschiedene andere Bezeichnungen für Surrogatendpunkte verwendet (z. B. intermediärer Endpunkt). Auf eine diesbezügliche Differenzierung wird an dieser Stelle allerdings verzichtet, da die Frage nach der notwendigen Validität davon unberührt bleibt. Weiterhin ist zu beachten, dass ein Endpunkt ein patientenrelevanter Endpunkt sein kann und darüber hinaus als Surrogat (also als Ersatz) für einen anderen patientenrelevanten Endpunkt betrachtet werden kann.

3.1.3 Ermittlung des Schadens medizinischer Interventionen

Die Anwendung einer medizinischen Intervention gleich welcher Art (medikamentös, nicht medikamentös, chirurgisch, diagnostisch, präventiv etc.) birgt per se das Risiko unerwünschter Wirkungen in sich. Der Begriff unerwünschte Wirkungen bezeichnet dabei all diejenigen Wirkungen, die einen individuell wahrgenommenen oder objektiv feststellbaren physischen oder psychischen Schaden darstellen, der zu einer mehr oder weniger starken kurz- oder langfristigen Verkürzung der Lebenserwartung, Erhöhung der Morbidität oder Beeinträchtigung der Lebensqualität führt oder führen kann. Für eine Begriffsklärung ist anzumerken, dass bei Verwendung des Begriffs unerwünschte Wirkungen ein kausaler Zusammenhang mit der Intervention angenommen wird, während dies bei Verwendung des Begriffs unerwünschte Ereignisse noch offenbleibt [143].

Der Begriff Schaden beschreibt das Auftreten unerwünschter Wirkungen bei der Anwendung einer medizinischen Intervention. Die Darstellung des Schadens ist ein wesentlicher und gleichberechtigter Bestandteil der Nutzenbewertung einer Intervention. Sie gewährleistet eine informierte populationsbezogene, aber auch individuelle Nutzen-Schaden-Abwägung [774].

Voraussetzung dafür ist, dass anhand der vorliegenden Daten die Effektstärken einer medizinischen Intervention sowohl für die erwünschten als auch für die unerwünschten Wirkungen im Vergleich z. B. zu therapeutischen Alternativen beschrieben werden können.

Die Darstellung, Analyse und Bewertung des Schadens einer medizinischen Intervention im Rahmen einer systematischen Übersicht sind jedoch in vielen Fällen ungleich schwerer als die des (Zusatz-)Nutzens. Dies betrifft insbesondere unerwartet auftretende unerwünschte Wirkungen [143]. Typischerweise sind Studien so konzipiert, dass sie den Einfluss einer medizinischen Intervention auf wenige, vorab definierte Endpunkte messen sollen. Zumeist handelt es sich dabei um Endpunkte zur Erfassung der Wirksamkeit, während Nebenwirkungen als unerwünschte Ereignisse begleitend erhoben werden. Die Ergebnisse bezüglich unerwünschter Ereignisse hängen dabei stark von der zugrunde gelegten Methodik der Erfassung ab. So werden z. B. bei expliziten Abfragen definierter unerwünschter Ereignisse i. d. R. höhere Ereigniszahlen ermittelt als bei einer generellen Abfrage [55,401]. Gerade für die Erfassung unerwartet auftretender unerwünschter Ereignisse ist jedoch eine generelle Abfrage des Befindens der Patientinnen und Patienten erforderlich. Des Weiteren ist die zielgerichtete Forschung zur Erhebung insbesondere der seltenen schwerwiegenden unerwünschten Ereignisse einschließlich der Darstellung eines kausalen Zusammenhangs mit der medizinischen Intervention erheblich unterrepräsentiert [64,211,400]. Darüber hinaus ist die Qualität der Berichterstattung unerwünschter Ereignisse in Einzelstudien mangelhaft, was auch zu einer Erweiterung des CONSORT-Statements für randomisierte klinische Studien geführt hat [399]. Schließlich wird die systematische Bewertung unerwünschter Wirkungen einer Intervention auch dadurch erschwert, dass deren Codierung in Literaturdatenbanken unzureichend ist und die zielgerichtete Suche nach relevanter wissenschaftlicher Literatur deshalb häufig ein unvollständiges Bild liefert [164].

Die oben genannten Hindernisse erschweren häufig die Untersuchung des Schadens. In Fällen, in denen für die Bewertung vollständige Studienberichte zur Verfügung stehen, ist häufig zumindest eine ausreichende Datentransparenz auch für unerwünschte Ereignisse gegeben. Zudem erfolgt die Erfassung der Daten insbesondere bei Arzneimitteln durch die Verwendung des Medizinischen Wörterbuchs für Aktivitäten im Rahmen der Arzneimittelzulassung (MedDRA) nach einem standardisierten Codierungssystem. Es ist dennoch notwendig, eine sinnvolle Balance zwischen der Vollständigkeit der Aufarbeitung von Schadenaspekten und dem Ressourcenaufwand zu finden. Daraus folgend ist es notwendig, die Aufarbeitung und Darstellung auf relevante unerwünschte Wirkungen zu beschränken. Als relevant können dabei insbesondere diejenigen unerwünschten Wirkungen bezeichnet werden, die

- den Nutzen der Intervention teilweise oder ganz aufwiegen könnten,
- sich zwischen 2 oder mehr ansonsten gleichwertigen Therapieoptionen erheblich unterscheiden könnten,
- insbesondere bei denjenigen Therapieoptionen auftreten, die besonders effektiv sein könnten,

- eine Dosis-Wirkungs-Beziehung haben könnten,
- von Patientinnen und Patienten als besonders wichtig angesehen werden und
- von schwerwiegender Morbidität oder gar erhöhter Mortalität begleitet sind oder in der Folge mit einer erheblichen Beeinträchtigung der Lebensqualität einhergehen.

Aufbereitung und Darstellung von Schadensaspekten in Nutzenbewertungen

Bei der Aufarbeitung und Darstellung unerwünschter Wirkungen werden daher folgende Grundsätze beachtet: Für die Nutzenbewertung wird zunächst angestrebt, eine Auswahl potenziell relevanter unerwünschter Wirkungen, die bei der Entscheidungsfindung für oder gegen den Einsatz der zu evaluierenden Intervention einen wesentlichen Stellenwert haben, zusammenzustellen. Dabei erfolgt eine Auswahl unerwünschter Wirkungen nach den oben genannten Kriterien.

- Grundsätzlich werden die Gesamtraten der schwerwiegenden unerwünschten Ereignisse, ggf. der schweren unerwünschten Ereignisse (z. B. solche mit Common-Terminology-Criteria-for-Adverse-Events[CTCAE]-Grad ≥ 3) und der Abbrüche wegen unerwünschter Ereignisse herangezogen.
- Zudem werden für die Bewertung potenziell relevante spezifische unerwünschte Wirkungen ausgewählt. Dies erfolgt auf 2 verschiedenen Wegen:
 - Zum einen werden solche spezifischen unerwünschten Wirkungen ausgewählt, die für das Krankheitsbild oder die in der Studie bzw. den Studien eingesetzten Interventionen von besonderer Bedeutung sind. Für Nutzenbewertungen gemäß § 139a SGB V ist dies der primäre Weg der Identifizierung solcher unerwünschten Wirkungen. Die Zusammenstellung erfolgt in diesem Fall im Rahmen der Vorabrecherche zur jeweiligen Fragestellung und durch die Einbindung von Betroffenen.
 - Zum anderen werden für die Nutzenbewertung relevante spezifische unerwünschte Wirkungen basierend auf den in der relevanten Studie bzw. den relevanten Studien aufgetretenen unerwünschten Ereignissen identifiziert. Dieser Weg ist insbesondere für Nutzenbewertungen von Arzneimitteln gemäß § 35a SGB V (siehe Abschnitt 3.3.3) von Bedeutung. Erstens werden aufgrund des Verfahrens keine Vorabrecherchen durchgeführt und zweitens werden regelhaft neue Arzneimittel bewertet, sodass die Identifizierung unbekannter bzw. unerwarteter unerwünschter Wirkungen hier einen höheren Stellenwert hat. Zur Auswahl werden in diesem Fall die unerwünschten Ereignisse, schwerwiegenden unerwünschten Ereignisse, ggf. schweren unerwünschten Ereignisse, zugrunde gelegt, die vom pharmazeutischen Unternehmer gemäß den Anforderungen in der Verfahrensordnung des G-BA im Dossier übermittelt werden [279]. Die Kriterien für die Auswahl stellen die Patientenrelevanz der Ereignisse und die Unterschiede zwischen den Behandlungsarmen dar. Für nicht schwerwiegende bzw. nicht schwere unerwünschte Ereignisse werden zudem ggf.

zusätzliche Anforderungen an die Mindesthäufigkeit der Ereignisse gestellt, sofern dies die Gesamtabwägung zum Zusatznutzen nicht beeinflusst.

Für die ausgewählten spezifischen unerwünschten Ereignisse wird jeweils die am besten geeignete Operationalisierung in Bezug auf Mess- und Ergebnissicherheit herangezogen. Beispielsweise kann es der Fall sein, dass eine bestimmte Nebenwirkung durch eine Standardized MedDRA Query (SMQ) oder durch einen außerhalb von MedDRA operationalisierten Endpunkt besser erfasst wird als durch einen einzelnen bevorzugten Begriff (Preferred Term [PT]). Auch wird geprüft, ob die ausgewählten Ereignisse zum selben inhaltlichen Konstrukt konsistent sind. Als nicht konsistent können Ergebnisse beispielsweise angesehen werden, wenn sich in einem PT (Beispiel: Nasopharyngitis) ein Vorteil der Intervention zeigt, in einem ähnlichen PT (Beispiel: Rhinitis) jedoch ein Nachteil. Liegt keine geeignete Operationalisierung vor oder sind die Ergebnisse nicht konsistent, werden diese in der Regel nicht für Aussagen zum Schaden einer Intervention herangezogen.

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 (GRADE: Grading of Recommendations Assessment, Development and Evaluation) erarbeitet werden [27].

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 Beleglage des (Zusatz-)Nutzens und Schadens in 4 Abstufungen bezüglich der jeweiligen Aussagesicherheit getroffen: Es liegt entweder ein Beleg (höchste Aussagesicherheit), ein Hinweis (mittlere Aussagesicherheit), ein Anhaltspunkt (schwächste Aussagesicherheit) oder keine dieser 3 Situationen vor. Der letzte Fall tritt ein, wenn keine Daten vorliegen oder die vorliegenden Daten keine der 3 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 9.3.5).

A) Sicherheit von Studienergebnissen

Ein wichtiges Kriterium für die 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 beeinträchtigt durch systematische

Fehler (Verzerrung, siehe Abschnitt 9.3.13) wie z. B. Informationsfehler, Selektionsfehler und Confounding. Die quantitative Ergebnissicherheit wird beeinflusst durch zufällige Fehler verursacht durch die Stichprobenziehung (statistische Unsicherheit).

Die qualitative Ergebnissicherheit wird somit bestimmt durch das Studiendesign, aus dem sich Evidenzgrade ableiten lassen (siehe Abschnitt 9.1.3), 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 9.1.4).

Die quantitative Ergebnissicherheit steht 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. Falls die zugrunde liegenden Daten dies zulassen, lässt sich die statistische Unsicherheit als Standardfehler bzw. Konfidenzintervall von Parameterschätzungen quantifizieren und beurteilen (Präzision der Schätzung).

Das Institut verwendet die folgenden 3 Kategorien zur Graduierung des Ausmaßes der qualitativen Ergebnissicherheit auf Einzelstudien- und Endpunktebene:

- **hohe qualitative Ergebnissicherheit:** Ergebnis bezüglich eines Endpunkts einer randomisierten Studie mit niedrigem Verzerrungspotenzial
- **mäßige qualitative Ergebnissicherheit:** Ergebnis bezüglich eines Endpunkts einer randomisierten Studie mit hohem Verzerrungspotenzial
- **geringe qualitative Ergebnissicherheit:** Ergebnis bezüglich eines Endpunkts einer nicht randomisierten vergleichenden Studie

B) Ableitung der Beleglage und Aussagesicherheit

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 Metaanalyse (siehe Abschnitt 9.3.7) sinnvoll ein gemeinsamer Effektschätzer bilden lässt oder nicht. Im Fall homogener Ergebnisse, die sich sinnvoll poolen lassen, muss der gemeinsame Effektschätzer statistisch signifikant sein, um gemäß der vorhandenen Ergebnissicherheit einen Beleg, einen Hinweis oder einen Anhaltspunkt abzuleiten.

Vorgehen bei Heterogenität

Sind die geschätzten Effekte zu heterogen, um sinnvoll einen gepoolten gemeinsamen Effektschätzer zu bilden, wird unterschieden zwischen Situationen mit nicht gleichgerichteten, mäßig gleichgerichteten und deutlich gleichgerichteten Effekten, die wie folgt definiert sind:

Falls das Prädiktionsintervall zur Darstellung der Heterogenität in einer Metaanalyse mit zufälligen Effekten (siehe Abschnitt 9.3.7) dargestellt wird und den Nulleffekt nicht überdeckt, liegen gleichgerichtete Effekte vor.

Falls keine Darstellung des Prädiktionsintervalls erfolgt oder dieses den Nulleffekt überdeckt, liegen gleichgerichtete Effekte zudem in der folgenden Situation vor. Die Effektschätzer von 2 oder mehr Studien zeigen in eine Richtung und für diese Studien gelten alle folgenden Bedingungen:

- Das Gesamtgewicht dieser Studien ist 80 % oder größer.
- Mindestens 2 dieser Studien zeigen statistisch signifikante Ergebnisse.
- Mindestens 50 % des Gewichts dieser Studien basieren auf statistisch signifikanten Ergebnissen.

Die Gewichte der Studien kommen hierbei in der Regel aus einer Metaanalyse mit zufälligen Effekten (siehe Abschnitt 9.3.7). Trifft keine der beiden beschriebenen Situationen zu, liegt eine heterogene Situation mit nicht gleichgerichteten Effekten vor.

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

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

- Das Prädiktionsintervall überdeckt den Nulleffekt: Die gleichgerichteten Effekte sind mäßig gleichgerichtet.

Einfluss der Ergebnissicherheit und weiterer Faktoren

Für den Fall, dass die vorhandenen Studien dieselbe qualitative Ergebnissicherheit aufweisen oder nur 1 Studie vorliegt, lassen sich mit diesen Definitionen die regelhaften Anforderungen an die Beleglage zur Ableitung von Aussagen mit unterschiedlichen Aussagesicherheiten definieren. Das Institut unterscheidet – wie oben erläutert – die 3 verschiedenen Aussagesicherheiten Beleg, Hinweis und Anhaltspunkt.

In der Regel wird an die Aussage eines Belegs die Anforderung zu stellen sein, dass eine Metaanalyse von Studien mit hoher qualitativer Ergebnissicherheit einen entsprechenden statistisch signifikanten Effekt zeigt. Falls eine Metaanalyse nicht durchführbar ist, sollten mindestens 2 voneinander unabhängig durchgeführte Studien mit hoher qualitativer Ergebnissicherheit und einem statistisch signifikanten 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 Metaanalyse von Studien mit mäßiger qualitativer Ergebnissicherheit oder eine einzelne Studie mit hoher qualitativer Ergebnissicherheit kann trotz statistisch signifikanten Effekts demnach in der Regel nur einen Hinweis liefern.

Aus lediglich 1 Studie kann im Ausnahmefall für einen Endpunkt ein Beleg für eine spezifische (Teil-)Population abgeleitet werden, wenn ein Studienbericht gemäß den ICH-Richtlinien vorliegt, die auch sonst für einen Beleg geforderten Bedingungen erfüllt sind und darüber hinaus die Studie folgende besondere Anforderungen erfüllt:

- Die Studie ist multizentrisch, in jeden Studienarm wurden mindestens 1000 Patientinnen oder Patienten eingeschlossen und es gibt mindestens 10 Zentren. Die Anzahl von 1000 Patientinnen oder Patienten sowie die Anzahl von 10 Zentren dienen dem Institut als Orientierung und bedeuten keine starren Grenzen.
- Der beobachtete Effektschätzer hat einen sehr kleinen zugehörigen p-Wert ($p < 0,001$).
- Das Ergebnis ist innerhalb der Studie konsistent: Für die interessierende (Teil-)Population liegen Auswertungen verschiedener weiterer Teilpopulationen (insbesondere Teilmengen von Studienzentren) vor, die jeweils bewertbare und ausreichend homogene Effektschätzer ergeben. Diese Konsistenzbewertung ist bei binären Daten nur möglich, falls eine gewisse Mindestanzahl von Ereignissen vorliegt.
- Die im vorherigen Punkt adressierten Auswertungen für Teilpopulationen liegen für alle relevanten Endpunkte vor, d. h., diese Auswertungen sind nicht auf einzelne ausgewählte Endpunkte beschränkt.

Es ist möglich, dass sich die Beleglage beim Vorliegen von lediglich 1 Studie, die für sich genommen nur einen Hinweis oder Anhaltspunkt ergibt, durch zusätzliche indirekte Vergleiche ändert. Hierfür sind allerdings hohe methodische Anforderungen an die indirekten Vergleiche zu stellen (siehe Abschnitt 9.3.8). Außerdem ist es bei entsprechend homogener Datenlage möglich, dass sich durch Ergänzung indirekter Vergleiche die Präzision der Effektschätzung erhöht, was bei der Feststellung des Ausmaßes des Zusatznutzens eine wichtige Rolle spielt (siehe Abschnitt 3.3.3).

Eine Metaanalyse 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 3 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 3: Regelhaft abgeleitete Aussagesicherheiten für verschiedene Evidenzsituationen beim Vorliegen von Studien derselben qualitativen Ergebnissicherheit

		Anzahl Studien				
		1 (mit statistisch signifikantem Effekt)	≥ 2			
			homogen	heterogen		
			Metaanalyse statistisch signifikant	gleichgerichtete Effekte ^a		
				deutlich	mäßig	nein
Qualitative Ergebnis- sicherheit	hoch	Hinweis	Beleg	Beleg	Hinweis	–
	mäßig	Anhaltspunkt	Hinweis	Hinweis	Anhaltspunkt	–
	gering	–	Anhaltspunkt	Anhaltspunkt	–	–
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 3 abgeleitet. Bei der Ableitung von Aussagen zur Beleglage für den gesamten Studienpool gelten dann folgende Grundsätze:

- Die Aussagen zur Beleglage bei einer 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 ein oder mehrere Ergebnisse mäßiger (jedoch nicht geringer) qualitativer Ergebnissicherheit im Rahmen einer gemeinsamen Metaanalyse erbracht werden. Dabei sollte das Gewicht der Studie hoher qualitativer Ergebnissicherheit eine angemessene Größe haben (zwischen 25 und 75 %).
- Ist das metaanalytische 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 niedrigste qualitative Ergebnissicherheit aller einbezogenen Studien bestimmt wird.

Nach diesen Definitionen und Grundsätzen wird für jeden Endpunkt einzeln eine entsprechende Nutzensaussage abgeleitet. Überlegungen zur endpunktübergreifenden Bewertung finden sich in nachfolgendem 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 bezüglich 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 lebenslagenpezifische Besonderheiten berücksichtigt.

Eine Möglichkeit der gleichzeitigen Würdigung des Nutzens und Schadens ist die Gegenüberstellung der endpunktbezogenen Nutzen- und Schadenaspekte. Dabei werden die Effekte auf alle Endpunkte (qualitativ oder semiquantitativ wie in der frühen Nutzenbewertung nach § 35a SGB V) 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 oder über ihre Gewichtung zu einer Gesamtaussage zu kommen. Die gleichzeitige Würdigung des Nutzens und Schadens wird themenspezifisch konkretisiert (siehe auch Abschnitt 4.3.3).

3.2 Spezielle Aspekte der Nutzenbewertung

3.2.1 Auswirkung nicht publizierter Studienergebnisse auf das Fazit

Eine wesentliche Voraussetzung für die Validität einer Nutzenbewertung ist die vollständige Verfügbarkeit der Ergebnisse der durchgeführten Studien. Eine Bewertung, die auf unvollständigen, ggf. sogar selektiv zusammengestellten Daten beruht, kann ein verzerrtes Ergebnis liefern [236,389] (siehe auch Abschnitt 9.3.13). Die Verzerrung der publizierten Evidenz durch Publikationsbias und Outcome Reporting Bias wurde umfangreich beschrieben [205,512,674,726]. Um die Auswirkungen dieser Verzerrung zu minimieren, führt das Institut über die bibliografische Recherche hinaus zusätzliche Suchen in Studienregistern sowie Anfragen an Dritte zur Datenübermittlung, insbesondere Herstelleranfragen, durch (siehe Abschnitte 8.1.2 und 8.1.3).

Die Übermittlung ansonsten unpublizierter Informationen kann das Problem der Verzerrung durch unpublizierte Evidenz nur lösen, wenn sie selbst nicht selektiv, sondern vollständig ist. Eine unvollständige Datenübermittlung birgt das Risiko, dass das Ergebnis der Nutzenbewertung verzerrt ist. Dieses Risiko muss im Fazit einer Nutzenbewertung berücksichtigt werden.

Die folgende Tabelle 4 beschreibt, in welchen Konstellationen ein Risiko für eine Verzerrung von Bewertungsergebnissen besteht und welche Konsequenzen sich daraus für das Fazit der Nutzenbewertung ergeben. Die Szenarien 1 bis 4 beziehen sich auf Nutzenbewertungen mit Anfragen an Dritte zur Datenübermittlung (z. B. Herstelleranfragen), die Szenarien 5 bis 7 auf Nutzenbewertungen ohne Anfragen zur Datenübermittlung.

Tabelle 4: Szenarien der Datenvollständigkeit und Konsequenzen für das Fazit einer Nutzenbewertung

Szenario	Datenübermittlung durch Dritte (z. B. Daten von Herstellerinnen / Herstellern)	Nachweis, dass insgesamt Daten fehlen	Verzerrung durch Publikationsbias	Bewertung / Einfluss auf das Fazit
Nutzenbewertungen mit Anfragen an Dritte zur Datenübermittlung				
1	vollständig	nein oder in vernachlässigbarem Umfang	unwahrscheinlich	keine Einschränkung des Fazits
2	vollständig	ja, in mäßigem Umfang	möglich	Einschränkung der Aussage-sicherheit ^a
3	vollständig	ja, in erheblichem Umfang	wahrscheinlich	kein Beleg (und kein Hinweis oder Anhaltspunkt) für einen Nutzen bzw. Schaden
4	unvollständig	ja oder unklar ^b	wahrscheinlich oder möglich	kein Beleg (und kein Hinweis oder Anhaltspunkt) für einen Nutzen bzw. Schaden
Nutzenbewertungen ohne Anfragen an Dritte zur Datenübermittlung				
5	keine Datenübermittlung vorgesehen	nein oder in vernachlässigbarem Umfang	unwahrscheinlich	keine Einschränkung des Fazits
6	keine Datenübermittlung vorgesehen	ja, in mäßigem Umfang	möglich	Einschränkung der Aussage-sicherheit ^a
7	keine Datenübermittlung vorgesehen	ja, in erheblichem Umfang	wahrscheinlich	kein Beleg (und kein Hinweis oder Anhaltspunkt) für einen Nutzen bzw. Schaden
a: Bei Nutzenbewertungen mit eigenem Stellungsverfahren kann das Fazit der Bewertung auch unter Vorbehalt gestellt werden.				
b: Sofern im Ausnahmefall vom Datenübermittelnden nachgewiesen wird, dass die Menge der fehlenden Daten sicher irrelevant ist, wird das Fazit nicht eingeschränkt.				

Nutzenbewertungen mit Datenübermittlung durch Dritte

War in einer Nutzenbewertung mit Datenübermittlung durch Dritte diese Datenübermittlung vollständig und liegt kein Nachweis vor, dass weitere Daten in einem mehr als vernachlässigbaren Umfang fehlen (Szenario 1), wird eine Verzerrung als unwahrscheinlich eingestuft. Die Schlussfolgerungen aus der Bewertung der Daten werden ohne Einschränkung in das Fazit übernommen.

Wird bei einer vollständigen Datenübermittlung durch die angefragte Herstellerin bzw. den angefragten Hersteller durch andere Recherchen der Nachweis erbracht, dass Daten in einem mäßigen oder erheblichen Umfang (z. B. aus investigatorinitiierten Studien oder Studien anderer Herstellerinnen und Hersteller) fehlen, auf die die Herstellerin oder der Hersteller keinen Zugriff hat, liegt zwar keine selektive Datenübermittlung durch die Herstellerin oder den Hersteller vor. Dennoch ist eine Verzerrung durch die anderweitig fehlenden Daten abhängig vom Umfang der fehlenden Daten möglich (Szenario 2) oder sogar wahrscheinlich (Szenario 3). Eine Nutzaussage im Fazit (im Sinne von Beleg, Hinweis oder Anhaltspunkt für

Nutzen oder Schaden) wird deshalb in ihrer Aussagesicherheit entsprechend abgeschwächt oder gar nicht abgeleitet.

Bei einer unvollständigen Datenübermittlung (Szenario 4) wird von einer selektiven Datenübermittlung ausgegangen. Weitere relevante unpublizierte Daten, die durch die Rechenschritte nicht zugänglich sind, können existieren. Eine weitere Auswertung der verfügbaren limitierten Daten und eine daraus abgeleitete Aussage zum Nutzen oder Schaden sind in dieser Situation wahrscheinlich hochgradig verzerrt und stellen damit keine valide Bewertungsgrundlage dar. In dieser Situation wird deshalb kein Beleg (und kein Hinweis oder Anhaltspunkt) für einen Nutzen oder Schaden der zu bewertenden Intervention festgestellt, unabhängig von den beobachteten Ergebnissen in den (selektiv) vorgelegten Daten.

Nutzenbewertungen ohne Datenübermittlung durch Dritte

Lässt sich in einer Nutzenbewertung ohne vorgesehene Datenübermittlung durch Dritte ein Fehlen von Daten nicht erkennen (Szenario 5), ergibt sich keine Einschränkung für das Fazit.

Lassen sich jedoch fehlende Daten nachweisen (z. B. durch Abgleich zwischen registrierten und publizierten Studien), so ist eine Verzerrung durch Publikationsbias möglich (Szenario 6) oder sogar wahrscheinlich (Szenario 7). Die Abgrenzung zwischen beiden Szenarien erfolgt anhand der eingeschätzten Höhe des Anteils fehlender Daten. Bei Szenario 6 (Publikationsbias möglich) wird das Fazit in seiner Aussagesicherheit herabgestuft. Alternativ kann das Fazit unter Vorbehalt gestellt werden, wenn im Rahmen eines vom IQWiG durchgeführten Stellungnahmeverfahrens zu einer Nutzenbewertung die Möglichkeit besteht, dass fehlende Daten nachgeliefert werden. Bei Szenario 7 (Publikationsbias wahrscheinlich) wird kein Beleg (und kein Hinweis oder Anhaltspunkt) für einen Nutzen bzw. Schaden abgeleitet. In diesem Fall werden lediglich die vorhandenen und die fehlenden Informationen dargestellt.

In bestimmten Fällen (z. B. je nach Datenübermittlungsverfahren und Datenanteil, Aufteilung der Nutzenbewertung in Teilfragestellungen) kann es trotz Nachweises fehlender Daten möglich sein, das Ausmaß des Publikationsbias zu verringern, indem ein nachweislich unverzerrter Teil-Studienpool bewertet wird. Falls zu einer Fragestellung Studien unterschiedlicher Hersteller bzw. Herstellerinnen vorliegen, kann die Beschränkung auf die Daten eines einzelnen Herstellers bzw. einer einzelnen Herstellerin hierfür jedoch ungeeignet sein [347].

3.2.2 Dramatischer Effekt

Ist der Verlauf einer Erkrankung sicher oder nahezu sicher vorhersagbar und bestehen keine Behandlungsoptionen zur Beeinflussung dieses Verlaufs, so kann der Nutzen einer medizinischen Intervention auch durch die Beobachtung einer Umkehr des (quasi)deterministischen Verlaufs bei einer gut dokumentierten Serie von Patientinnen und Patienten abgeleitet werden. Wenn z. B. von einer Erkrankung bekannt ist, dass sie nach Diagnosestellung innerhalb kurzer Zeit mit hoher Wahrscheinlichkeit zum Tode führt, und in Fallserien beschrieben wird, dass

nach Anwendung einer bestimmten Maßnahme die meisten Betroffenen eine längere Zeitspanne überlebt haben, so wird ein solcher dramatischer Effekt ausreichend für die Ableitung eines Nutzens sein können. Beispielhaft sei hier die Substitution lebenswichtiger Hormone bei Erkrankungen mit Ausfall der Produktion dieser Hormone (z. B. Gabe von Insulin bei Patientinnen und Patienten mit Diabetes mellitus Typ 1) genannt. Wesentliche Voraussetzung für die Einordnung als dramatischer Effekt ist die ausreichend sichere Dokumentation des schicksalhaften Verlaufs der Erkrankung in der Literatur und der Diagnosestellung bei den Patientinnen und Patienten in der zu bewertenden Studie. Dabei sollten auch mögliche Schäden der Maßnahme berücksichtigt werden. Glasziou et al. [284] versuchen, die Einordnung als dramatischen Effekt zu operationalisieren. In einer ersten Näherung wird vorgeschlagen, einen beobachteten Effekt dann als nicht mehr allein durch die Einwirkung von Störgrößen erklärbar anzusehen, wenn er auf dem Niveau von 1 % signifikant ist und als relatives Risiko ausgedrückt den Wert 10 übersteigt [284]. Diese Größenordnung dient dem Institut als Orientierung und bedeutet keine starre Grenze. Glasziou et al. [284] haben ihrer Empfehlung Ergebnisse von Simulationsstudien zugrunde gelegt, wonach ein beobachtetes relatives Risiko von 5 bis 10 nicht mehr plausibel allein durch Störgrößeneinflüsse erklärt werden könne. Damit wird deutlich, dass eine entsprechende Grenze auch von den Begleitumständen (u. a. der Qualität der Studien, die für das Vorliegen eines dramatischen Effekts bemüht werden, oder auch konsistenten Ergebnissen zu einer Endpunktkategorie) abhängig ist. Diese Abhängigkeit schlägt sich auch in Empfehlungen anderer Arbeitsgruppen (z. B. der GRADE-Gruppe) nieder [311].

Falls im Vorfeld der Bearbeitung einer spezifischen Fragestellung (z. B. aufgrund einer Vorabrecherche) ausreichende Informationen dazu vorliegen, dass ein durch die zu bewertende Intervention bewirkter dramatischer Effekt erwartet werden kann, wird bei der Informationsbeschaffung auch nach solchen Studien gesucht, die aufgrund ihres Designs eine höhere Ergebnisunsicherheit aufweisen.

3.2.3 Studiendauer

Ein wesentliches Kriterium bei der Auswahl der für die Nutzenbewertung relevanten Studien ist deren Dauer. Für die Evaluation einer Intervention zur Behandlung einer akuten Erkrankung, deren primäres Ziel z. B. die Verkürzung der Krankheitsdauer und die Reduktion der Akutsymptomatik in den Vordergrund stellt, ist die Forderung nach Langzeitstudien in der Regel nicht sinnvoll, es sei denn, es sind Spätkomplikationen zu erwarten. Auf der anderen Seite sind für die Evaluation von Interventionen zur Behandlung chronischer Erkrankungen Kurzzeitstudien in der Regel nicht für die vollständige Nutzenbewertung geeignet. Dies trifft insbesondere dann zu, wenn die Behandlung über mehrere Jahre oder sogar lebenslang erforderlich ist. In solchen Fällen sind insbesondere Studien, die einen mehrjährigen Behandlungszeitraum abdecken, sinnvoll und wünschenswert. Da sowohl der Nutzen als auch der Schaden über die Zeit unterschiedlich verteilt sein kann, ist bei Langzeittherapien eine aussagekräftige Gegenüberstellung des Nutzens und Schadens nur bei Vorliegen von Studien mit einer entsprechend langen Beobachtungsdauer hinreichend sicher möglich. Einzelne Aspekte des Nutzens bzw. Schadens können aber durchaus in kürzeren Studien betrachtet werden.

Bezüglich des Auswahlkriteriums Mindeststudiendauer orientiert sich das Institut primär an Standards bezüglich des Nachweises der Wirksamkeit. Für die Arzneimittelbewertung wird insbesondere auf Angaben in den entsprechenden indikationsspezifischen Leitlinien der Zulassungsbehörden zurückgegriffen (z. B. [229]). Da die Nutzenbewertung auch die Aspekte des Schadens beinhaltet, sind bei der Festlegung der Mindeststudiendauer auch die diesbezüglich allgemein konsentierten Anforderungen relevant. Bei Langzeitinterventionen wie oben beschrieben wird darüber hinaus auch auf das Kriterium Langzeitbehandlung aus entsprechenden Guidelines zurückgegriffen [374]. Im Einzelfall kann von diesem Vorgehen begründet abgewichen werden, z. B. falls es unter inhaltlichen Gesichtspunkten erforderlich ist, eine Nachbeobachtung über einen längeren Zeitraum zu fordern, spezielle (Unter-)Fragestellungen sich auf einen kürzeren Zeitraum beziehen oder wenn bei der Evaluation neu verfügbarer / zugelassener Interventionen und / oder der Evaluation von Technologien ohne adäquate Behandlungsalternative auch kurzfristige Wirkungen Gegenstand der Evaluation sind.

3.2.4 Patientenberichtete Endpunkte

Für die in Abschnitt 3.1.1 beschriebenen patientenrelevanten Nutzendimensionen können auch patientenberichtete Endpunkte (Patient-reported Outcomes [PROs]) zum Einsatz kommen. Neben der gesundheitsbezogenen Lebensqualität können auch andere Nutzendimensionen mittels PROs erfasst werden, wie z. B. Symptome der Erkrankung. Wie auch für die Erfassung der Lebensqualität sind hierfür Instrumente zu fordern, die für den Einsatz in klinischen Studien geeignet sind [227,258]. Bei der Auswahl der zu berücksichtigenden Evidenz (insbesondere der Studientypen) für den Nachweis eines Effekts gelten i. d. R. die gleichen Prinzipien wie bei anderen Endpunkten [258]. D. h., dass auch im Falle von PROs einschließlich der gesundheitsbezogenen Lebensqualität, Symptome und Behandlungszufriedenheit randomisierte kontrollierte Studien am besten für den Nachweis eines Effekts geeignet sind.

Da Angaben zu PROs aufgrund ihrer Natur subjektiv sind, sind offene, d. h. nicht verblindete Studien in diesem Bereich nur von eingeschränkter Validität. Für die Frage, ob sich aus offenen Studien ein Hinweis auf einen Nutzen einer Intervention bezüglich PROs ableiten lässt, ist die Größe des beobachteten Effekts ein wichtiges Entscheidungskriterium. Es gibt empirische Evidenz für ein hohes Verzerrungspotenzial bezüglich subjektiver Endpunkte in offenen Studien [770]. Dies ist bei der Interpretation solcher Studien zu beachten (siehe Abschnitt 9.1.4). Allerdings sind Situationen denkbar, in denen eine Verblindung von Ärztinnen und Ärzten sowie Patientinnen und Patienten nicht möglich ist. In solchen Situationen sind – sofern möglich – andere Anstrengungen zur Verzerrungsminimierung bzw. -einschätzung (z. B. verblindete Endpunkterhebung und -bewertung) zu fordern. Weitere Aspekte der Qualitätsbewertung von Studien, in denen PROs erhoben wurden, finden sich in einer FDA-Guideline [258].

3.2.5 Nutzen und Schaden in kleinen Populationen

Es gibt kein überzeugendes Argument dafür, bei kleinen Populationen (z. B. Patientinnen und Patienten mit seltenen Erkrankungen oder Untergruppen häufiger Erkrankungen) grundsätzlich

von der Hierarchie der Evidenzgrade abzuweichen. In diesem Zusammenhang ist problematisch, dass keine international einheitliche Definition dessen existiert, was unter einer seltenen Erkrankung zu verstehen ist [764]. Davon unberührt haben auch Patientinnen und Patienten mit seltenen Erkrankungen einen Anspruch auf eine möglichst sichere Information über Behandlungsoptionen [220]. Nicht randomisierte Studien benötigen aufgrund der Notwendigkeit der Adjustierung nach Störgrößen größere Patientenzahlen als randomisierte Studien. Allerdings kann es aufgrund der Seltenheit einer Erkrankung manchmal unmöglich sein, so viele Patientinnen und Patienten in eine Studie einzuschließen, dass diese eine ausreichende statistische Macht besitzt. Eine metaanalytische Zusammenfassung kleinerer Studien kann in solchen Fällen besonders sinnvoll sein. Im Allgemeinen führen kleinere Stichproben zu einer geringeren Präzision einer Effektschätzung und damit einhergehend zu breiteren Konfidenzintervallen. Aufgrund der Bedeutung des vermuteten Effektes einer Intervention, seiner Größe, der Verfügbarkeit von Alternativen und der Häufigkeit und Schwere potenzieller therapiebedingter Schäden kann es durchaus sinnvoll sein, bei kleinen Stichproben – ähnlich wie für andere problematische Konstellationen empfohlen [226] – größere p-Werte als 5 % (z. B. 10 %) für den Nachweis statistischer Signifikanz zu akzeptieren und somit die quantitative Unsicherheit zu erhöhen. Dies muss aber ex ante geschehen und nachvollziehbar begründet werden. Ebenso kann bei kleineren Stichproben eher die Notwendigkeit bestehen, einen zu selten auftretenden patientenrelevanten Endpunkt durch Surrogate zu ersetzen. Solche Surrogatendpunkte müssen allerdings auch bei kleinen Stichproben valide sein [228].

Für den Fall extrem seltener Erkrankungen oder extrem spezifischer Krankheitskonstellationen kann die Forderung nach (parallel) vergleichenden Studien unangemessen sein [764]. Nichtsdestotrotz sind auch hier zumindest adäquate Dokumentationen des Verlaufs solcher Patientinnen und Patienten einschließlich des erwarteten Verlaufs ohne die ggf. zu prüfende Intervention (z. B. anhand historischer Verläufe von Patientinnen und Patienten) möglich und bewertbar [109]. Die Spezifizierung und explizite Kennzeichnung, dass es sich um eine solche Situation einer extrem seltenen Erkrankung oder extrem spezifischen Krankheitskonstellation handelt, erfolgen im Berichtsplan.

3.3 Nutzenbewertung von Arzneimitteln

Ein wesentliches Ziel der Berichte zur Nutzenbewertung von Arzneimitteln ist die Unterstützung von Richtlinienentscheidungen des G-BA zur Erstattungsfähigkeit von Arzneimitteln innerhalb der gesetzlichen Krankenversicherung. Dafür ist es notwendig darzustellen, ob der Nutzen eines Arzneimittels nachgewiesen ist bzw. ob für ein Arzneimittel ein höherer Nutzen (Zusatznutzen) nachgewiesen ist als für eine andere medikamentöse oder nicht medikamentöse Alternative.

Richtlinienentscheidungen des G-BA betrachten i. d. R. nicht den Einzelfall, sondern den Regelfall. Auch die Berichte des Instituts beziehen sich deshalb i. d. R. nicht auf die Einzelfallentscheidung.

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

3.3.1 Stellenwert des Zulassungsstatus

Die Beauftragung des Instituts zur Nutzenbewertung von Arzneimitteln durch den G-BA erfolgt i. d. R. im Rahmen der Zulassung (Indikation, Dosierung, Gegenanzeigen, Begleitbehandlungen etc.) der zu untersuchenden Arzneimittel. Aus diesem Grund bezieht sich auch die Empfehlung des Instituts an den G-BA, die im Fazit des Berichts zur Nutzenbewertung formuliert wird, i. d. R. auf den Gebrauch der bewerteten Arzneimittel innerhalb der bestehenden Zulassung.

Projektspezifisch ist zu klären, wie mit Studien (bzw. der daraus abzuleitenden Evidenz) umgegangen wird, die nicht innerhalb der durch die Zulassung beschriebenen Anwendung eines Arzneimittels durchgeführt werden. Grundsätzlich ist denkbar, dass Studien, in denen Arzneimittel außerhalb der in den Fachinformationen beschriebenen Zulassung eingesetzt werden, den Nutzen und / oder Schaden des Arzneimittels über- oder unterschätzen. Dies kann zu einer Fehleinschätzung des Nutzens und Schadens für Patientinnen und Patienten, die innerhalb der Zulassung behandelt werden, führen. Ist für solche Studien allerdings hinreichend sicher plausibel oder gar nachgewiesen, dass die in diesen Studien gewonnenen Ergebnisse auf Patientinnen und Patienten, die gemäß der Zulassung behandelt werden, anwendbar sind, können die Ergebnisse dieser Studien in die Nutzenbewertung einbezogen werden.

Daher wird für Studien, die allein deshalb ausgeschlossen werden, weil sie nicht den aktuellen Zulassungsstatus widerspiegeln bzw. weil die Erfüllung des Zulassungsstatus unklar ist, jeweils geprüft, inwieweit die Studienergebnisse auf Patientinnen und Patienten anwendbar sind, die innerhalb der Zulassungsbedingungen behandelt werden.

Als anwendbar sind die Ergebnisse von Studien außerhalb des Zulassungsstatus dann anzusehen, wenn hinreichend sicher plausibel oder nachgewiesen ist, dass die Effektschätzer patientenrelevanter Endpunkte nicht wesentlich durch das betreffende Merkmal der Zulassung (z. B. die geforderte Vorbehandlung) beeinflusst werden. Die Gleichwertigkeit der Effekte ist i. d. R. mit adäquaten wissenschaftlichen Untersuchungen zu belegen. Diese Untersuchungen sollen auf den Nachweis der Gleichwertigkeit des Effekts zwischen der Gruppe mit und derjenigen ohne das Merkmal ausgerichtet sein. Ergebnisse, die auf Patientinnen und Patienten, die gemäß der Zulassung behandelt werden, anwendbar sind, können in das Fazit der Bewertung einbezogen werden.

Als nicht anwendbar sind die Ergebnisse aus Studien anzusehen, wenn kein Nachweis der Anwendbarkeit vorliegt und wenn plausible Gründe gegen die Übertragbarkeit der Ergebnisse

sprechen. Als in der Regel nicht anwendbar werden Studienergebnisse z. B. dann angesehen, wenn die behandelte Altersgruppe außerhalb der Zulassung liegt, nicht zugelassene Kombinationen mit einem anderen Wirkstoff eingesetzt werden, Patientinnen und Patienten mit einem außerhalb der Zulassung liegenden Schweregrad der Erkrankung behandelt werden oder Studien bei Patientinnen und Patienten mit einer Kontraindikation gegen die untersuchte Intervention durchgeführt werden. Die Ergebnisse dieser Studien werden im Bericht nicht dargestellt, da sie zur Bewertung des Arzneimittels nicht herangezogen werden können.

Werden Ergebnisse aus Studien, in denen Patientinnen und Patienten außerhalb der Zulassung behandelt wurden, als anwendbar angesehen, so wird dies im Berichtsplan spezifiziert. Unabhängig von der Anwendbarkeit von Studienergebnissen auf den durch die Zulassung spezifizierten Gebrauch werden in der Regel die Ergebnisse von Studien, die die Behandlung von Patientinnen und Patienten mit dem zu bewertenden Arzneimittel bei der im Auftrag spezifizierten Erkrankung untersuchen und die z. B. aufgrund der Größe, der Dauer oder der untersuchten Endpunkte von besonderer Bedeutung sind, diskutiert.

3.3.2 Studien zur Nutzenbewertung von Arzneimitteln

Die Ergebnisse der Nutzenbewertung von Arzneimitteln durch das Institut können Einfluss auf die Versorgung von Patientinnen und Patienten in Deutschland haben. Aus diesem Grund müssen hohe Anforderungen an die Ergebnissicherheit der Studien, die in die Nutzenbewertung einfließen, gestellt werden (siehe Abschnitt 3.1.4).

Erheblichen Einfluss auf die Ergebnissicherheit hat das Studiendesign. Denn mit Beobachtungsstudien, prospektiv oder retrospektiv, kann ein kausaler Zusammenhang zwischen Intervention und Effekt in der Regel nicht dargestellt werden, während die kontrollierte Interventionsstudie grundsätzlich hierfür geeignet ist [301]. Dies gilt vor allem dann, wenn andere, das Ergebnis beeinflussende Faktoren ganz oder weitgehend ausgeschaltet werden. Aus diesem Grund stellt die randomisierte kontrollierte Studie den Goldstandard bei der Bewertung medikamentöser und nicht medikamentöser Interventionen dar [550].

In der Regel sind RCTs zur Evaluation von Arzneimitteln möglich und praktisch durchführbar. Das Institut lässt daher in der Regel RCTs in die Nutzenbewertung von Arzneimitteln einfließen. Nur in begründeten Ausnahmefällen wird auf nicht randomisierte Interventionsstudien oder Beobachtungsstudien zurückgegriffen. Die Begründung kann darin liegen, dass für die vorliegende Fragestellung auch andere Studienarten eine hinreichende Ergebnissicherheit liefern könnten. Bei Erkrankungen, die ohne Intervention nach kurzer Zeit sicher zum Tod führen, können beispielsweise mehrere konsistente Fallbeobachtungen darüber, dass eine bestimmte Intervention diesen zwangsläufigen Verlauf verhindert, hinreichende Ergebnissicherheit liefern [474] (dramatischer Effekt, siehe auch Abschnitt 3.2.2). Die besondere Begründungspflicht für ein nicht randomisiertes Design bei der Prüfung von Arzneimitteln findet sich auch im Rahmen des Arzneimittelzulassungsrechts in den Arzneimittelprüfrichtlinien (Richtlinie 2001/83/EG, Abschnitt 5.2.5 [447]).

Das Institut legt im Rahmen der Erstellung des Berichtsplans (siehe auch Abschnitt 2.1.1) vorab fest, welche Studienarten aufgrund der Fragestellung als durchführbar und hinreichend ergebnissicher (bei hoher interner Validität) angesehen werden. Studien, die nicht mindestens diesem Qualitätsanspruch entsprechen (siehe auch Abschnitt 9.1.4), werden in den Bewertungsprozess primär nicht einfließen.

Neben der Charakterisierung der Ergebnissicherheit der berücksichtigten Studien ist es notwendig zu beschreiben, ob und in welchem Umfang sich das Studienergebnis auf lokale Gegebenheiten (zum Beispiel Population, Versorgungsbereich) übertragen lässt beziehungsweise welche lokalen Besonderheiten bei den Studien eine Auswirkung auf die Ergebnisse selbst beziehungsweise deren Interpretation haben oder haben könnten. Unter diesem Gesichtspunkt sind insbesondere Studien relevant, in denen die tatsächliche Versorgungsrealität Deutschlands so weit wie möglich abgebildet wird, ohne dass jedoch die oben beschriebenen Kriterien der Ergebnissicherheit außer Acht gelassen werden dürfen. Die Übertragbarkeit des Studienergebnisses (Generalisierbarkeit bzw. sogenannte externe Validität) selbst muss schließlich in einem gesonderten Prozess überprüft werden, der vom Design und von der Qualität der Studie zunächst unabhängig ist.

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 Dossievorlagen beschrieben, die Bestandteil der Verfahrensordnung des G-BA sind [276]. 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) [106].

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 für die 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 1. 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 entsprechend den 6 vorgegebenen Kategorien festzustellen: erheblicher, beträchtlicher, geringer, nicht quantifizierbarer Zusatznutzen, kein Zusatznutzen belegt, Nutzen des zu bewertenden Arzneimittels geringer als Nutzen der zweckmäßigen Vergleichstherapie.

Für die 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“ sowie bei Skipka et al. [669]. 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 quantifizierbaren 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 (Überlebenszeitanalysen),
- stetig oder quasistetig 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 kleiner 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 2-seitige 95 %-Konfidenzintervall für das relative Risiko – ggf. selbst berechnet – herangezogen. Falls mehrere Studien quantitativ zusammengefasst wurden, findet das metaanalytische 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 3 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 (d. h. näher an 1), je schwerwiegender das Ereignis ist. Die Schwellenwerte sind umso kleiner (also weiter entfernt von 1), je höher das Ausmaß ist. Die folgende Tabelle 5 gibt die für die 3 Ausmaßkategorien (gering, beträchtlich, erheblich) zu unterschreitenden Schwellenwerte für jede der 3 Kategorien der Qualität der Zielgrößen wieder.

Tabelle 5: Schwellenwerte zur Ausmaßbestimmung für das relative Risiko

		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
Ausmaß-kategorie	erheblich	0,85	0,75 und Risiko $\geq 5\%$ ^b	nicht besetzt
	beträchtlich	0,95	0,90	0,80
	gering	1,00	1,00	0,90
a: Voraussetzung ist – wie für alle patientenberichteten Endpunkte – die Verwendung eines validierten bzw. etablierten Instruments sowie eines geeigneten Responsekriteriums.				
b: Das Risiko muss für mindestens 1 der beiden zu vergleichenden Gruppen mindestens 5 % betragen.				

Das relative Risiko kann generell auf 2 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 Metaanalyse eine untergeordnete Rolle spielt. Das gilt jedoch nicht für den Abstand der Konfidenzintervallgrenzen 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 der Zielgröße Zeit bis zu einem Ereignis wird das 2-seitige 95 %-Konfidenzintervall für das Hazard Ratio benötigt. Falls mehrere Studien quantitativ zusammengefasst wurden, wird das metaanalytische 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 [713]. Für die Ausmaßfeststellung werden dieselben Grenzen wie für das relative Risiko angelegt (siehe Tabelle 5).

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 quasistetige Zielgrößen mit jeweils vorliegenden Responderanalysen

Zur Feststellung des Ausmaßes des Zusatznutzens bei stetigen oder quasistetigen Zielgrößen werden Responderanalysen herangezogen. Dazu bedarf es eines geeigneten Responsekriteriums bzw. Cut-off-Werts (siehe Abschnitt 9.3.3). Anhand der Responderauswertungen (Vierfeldertafeln) werden die relativen Risiken daraus direkt berechnet. Mithilfe von Tabelle 5 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 [153], um die entsprechenden Schwellenwerte zur Ausmaßfeststellung anzulegen. Anderenfalls ist das Ausmaß als nicht quantifizierbar festzustellen.

Für den 3. Schritt der Operationalisierung der Gesamtaussage zum Ausmaß des Zusatznutzens bei gemeinsamer Betrachtung aller Endpunkte ist eine strenge Formalisierung nicht möglich, da für die hierauf zutreffenden 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.

3.4 Nicht medikamentöse therapeutische Interventionen

Nicht medikamentöse therapeutische Interventionen werden zum einen im Rahmen von Berichten (siehe Abschnitt 2.1.1) oder Rapid Reports (siehe Abschnitt 2.1.2) oder in Form von HTA-Berichten (siehe Abschnitt 2.1.10) detailliert bewertet. Ablauf und Methodik dieser Bewertungen erfolgen gemäß den allgemeinen Prinzipien des Methodenpapiers. Des Weiteren können neue Untersuchungs- und Behandlungsmethoden, die maßgeblich auf dem Einsatz eines Medizinprodukts hoher Risikoklasse beruhen, auch im Rahmen einer Bewertung gemäß § 137h SGB V geprüft werden (siehe Abschnitt 2.1.6). Weil §-137h-Bewertungen primär auf eingereichten Unterlagen (und nicht auf einer selbst durchgeführten Informationsbeschaffung) beruhen und in deutlich kürzerer Zeit erstellt werden müssen, sind die methodische Tiefe der Bearbeitung und die Sicherheit einer eventuellen Aussage zum Nutzen oder Schaden regelhaft geringer, als dies bei Berichten und Rapid Reports der Fall ist. Insbesondere wird in §-137h-Bewertungen nicht zwischen verschiedenen Aussagesicherheiten (Beleg, Hinweis, Anhaltspunkt) unterschieden. Die Anforderungen für einen Nutzen sind jedoch bei Berichten und Rapid Reports im Vergleich zu Bewertungen gemäß § 137h SGB V gleich und erfordern in der Regel das Vorhandensein von RCT-Ergebnissen zu patientenrelevanten Endpunkten (siehe Abschnitte 1.3.5, 3.1 und 3.2).

Wenngleich die regulatorischen Voraussetzungen für den Marktzugang von Arzneimitteln und nicht medikamentösen therapeutischen Interventionen unterschiedlich sind, gibt es dennoch keinen Grund, an die Bewertung des Nutzens und Schadens einen bezüglich der Ergebnissicherheit prinzipiell anderen Maßstab anzulegen. So sieht zum Beispiel die Verfahrensordnung des G-BA [276] die prioritäre Berücksichtigung von RCTs, soweit möglich, unabhängig vom Typ (medikamentös bzw. nicht medikamentös) der zu bewertenden medizinischen Maßnahme vor. Die Grundsätze von evidenzbasierter Medizin und Nutzenbewertung sind daher auch für spezielle Felder der Medizin gültig, so beispielsweise für die Chirurgie, die Zahnheilkunde, die Psychotherapie, die Alternativ- / Komplementärmedizin oder auch für digitale Gesundheitsanwendungen (E-Health). Für Medizinprodukte wird im Rahmen der Konformitätsbewertung gemäß der aktuellen EU-Medizinprodukte-Verordnung (MDR) [174] zwar nicht die randomisierte kontrollierte Studie als Design der Wahl dargestellt, allerdings sind klinische Prüfungen „nach einem angemessenen Prüfplan durchzuführen“ [221], sodass randomisierte kontrollierte Designs in diesem Bereich tendenziell immer häufiger eingesetzt werden [631]. Im Gegensatz zum europäischen Konformitätsbewertungsverfahren für neue Medizinprodukte beziehen sich die Bewertungen des Instituts stets auf Untersuchungs- und Behandlungsmethoden, nicht auf einzelne Medizinprodukte.

Studien im nicht medikamentösen Bereich sind im Vergleich zu Arzneimittelstudien häufig mit besonderen Herausforderungen und Schwierigkeiten verbunden [511], die aber meist zumindest teilweise überwindbar sind:

- Oft wird die Verblindung des die Intervention ausführenden Personals unmöglich und die der Patientinnen und Patienten schwierig oder ebenfalls unmöglich sein. Durch eine verblindete Endpunkterfassung oder durch die Verwendung objektiver Endpunkte lässt sich eine Verringerung der Aussagekraft einer Studie dennoch abmildern.
- Über welche Fähigkeiten und Fertigkeiten die Anwenderinnen und Anwender verfügen, ist für den in einer Studie erkennbaren Nutzen einer nicht medikamentösen Intervention meist wesentlich. Anwenderinnen und Anwender können sowohl Ärztinnen und Ärzte als auch Patientinnen und Patienten sein. Im Idealfall erlauben es eine adäquate Auswahl und entsprechende Schulung, den negativen Einfluss von Lernkurveneffekten auf das Studienergebnis zu minimieren.
- Viele nicht medikamentöse Interventionen verändern sich im Verlauf einer Studie. Zahlreiche kleine Veränderungen (oft als „Schrittinnovationen“ bezeichnet) sind gerade bei Medizinprodukten üblich. Veränderungen der Intervention im Verlauf einer Studie erfordern meist keine neue Studie, sondern lassen sich durch angemessene statistische Methoden in ihrem Einfluss auf das Studienergebnis hinreichend genau untersuchen.

Hinzu kommt, dass nicht medikamentöse Interventionen teilweise mehr oder weniger komplexe Interventionen sind [149]. Um den genannten Aspekten gerecht zu werden, bedarf es qualitativ besonders guter Studien, um eine ausreichende Ergebnissicherheit zu erzielen. Paradoxerweise war in der Vergangenheit eher das Gegenteil der Fall, d. h., gerade im nicht medikamentösen

Bereich fehlen oft aussagekräftige randomisierte Studien (z. B. in der Chirurgie [511]). Um überhaupt Aussagen zum Stellenwert einer bestimmten nicht medikamentösen therapeutischen Intervention treffen zu können, kann es deshalb erforderlich sein, auch nicht randomisierte Studien in die Bewertung einzubeziehen. Auch an diese sind allerdings Qualitätsanforderungen zu stellen, insbesondere bezüglich der Maßnahmen zur Sicherung der Strukturgleichheit. In aller Regel werden solche Studien jedoch aufgrund der ihnen immanenten geringeren Ergebnissicherheit allenfalls Anhaltspunkte für einen (Zusatz-)Nutzen oder Schaden liefern können. Die Einbeziehung von Studien eines niedrigeren Evidenzgrades steht im Einklang mit der entsprechenden Regelung im 2. Kapitel § 13 Abs. 2 der Verfahrensordnung des G-BA [276], wo aber dafür eine besondere Begründungspflicht betont wird. Dort heißt es: „Die Anerkennung des medizinischen Nutzens einer Methode auf Grundlage von Unterlagen einer niedrigeren Evidenzstufe bedarf jedoch – auch unter Berücksichtigung der jeweiligen medizinischen Notwendigkeit – zum Schutz der Patientinnen und Patienten umso mehr einer Begründung, je weiter von der Evidenzstufe 1 abgewichen wird. Dafür ist der potenzielle Nutzen einer Methode insbesondere gegen die Risiken der Anwendung beim Patienten abzuwägen, die mit einem Wirksamkeitsnachweis geringerer Aussagekraft einhergehen“ [276]. D. h., allein die Nichtverfügbarkeit von Studien höchster Evidenz kann im Allgemeinen nicht als ausreichende Begründung für eine Nutzenbewertung auf Basis von Studien niedrigerer Evidenz angesehen werden.

Auch bei der Bewertung von nicht medikamentösen therapeutischen Interventionen kann es notwendig sein, die Verkehrsfähigkeit bzw. CE-Kennzeichnung (gemäß MPG) und den Zulassungsstatus von Arzneimitteln (gemäß AMG) zu beachten, sofern die zu prüfenden Interventionen oder Vergleichsinterventionen den Einsatz von Medizinprodukten oder Arzneimitteln beinhalten (siehe Abschnitt 3.3.1). Entsprechende Konsequenzen sind dann im Berichtsplan (siehe Abschnitt 2.1.1) zu spezifizieren.

3.5 Diagnostische Verfahren

Diagnostische Verfahren zeichnen sich dadurch aus, dass ihr gesundheitsbezogener Nutzen (oder Schaden) im Wesentlichen erst dadurch zustande kommt, dass sich an ihnen therapeutische oder präventive Verfahren anschließen. Die alleinige Gewinnung diagnostischer Informationen (ohne medizinische Konsequenzen) hat regelhaft keinen sozialrechtlich relevanten Nutzen.

Dies gilt in gleicher Weise sowohl für diagnostische Informationen, die sich auf den augenblicklichen Gesundheitszustand beziehen, als auch für prognostische Informationen (oder Marker), die sich auf einen zukünftigen Gesundheitszustand beziehen. Im Folgenden werden daher Verfahren zur Bestimmung diagnostischer oder prognostischer Informationen gemeinsam als diagnostische Verfahren betrachtet.

Allgemein kann der Prozess zur Evaluierung diagnostischer Verfahren in Analogie zur Arzneimittelprüfung in verschiedene, hierarchisch angeordnete Phasen oder Stufen eingeteilt werden [1,267]. Prospektiv geplante kontrollierte Diagnosestudien der Phase 4 nach Köbberling et al.

[1] bzw. Studien der Stufe 5 nach Fryback und Thornbury [267] mit einer (idealerweise zufälligen) Zuteilung von Patientinnen und Patienten zu einer Strategie mit bzw. ohne Anwendung der zu prüfenden diagnostischen Maßnahme oder zu Gruppen mit bzw. ohne Offenlegung der (diagnostischen) Testergebnisse können dabei in Entsprechung zu (Arzneimittel-)Zulassungsstudien der Phase 3 („Wirksamkeitsstudien“) gesehen werden. Dementsprechend wird ihnen auch der höchste Evidenzgrad zugesprochen (siehe z. B. Verfahrensordnung des G-BA [276]). Auch für die Zulassung für bestimmte Indikationen von Arzneimitteln und biologischen Produkten, die im Zusammenhang mit bildgebenden diagnostischen Methoden entwickelt werden, werden von der FDA derartige Studien empfohlen [257], und Beispiele zeigen, dass sie auch mit vergleichsweise moderatem Aufwand durchführbar sind [21,729].

Das Institut folgt dieser Logik und führt Nutzenbewertungen zu diagnostischen Verfahren in erster Linie auf der Grundlage solcher Studien mit patientenrelevanten Endpunkten durch. Die Grundzüge der Bewertung entsprechen dabei den in den Abschnitten 3.1 bis 3.4 gemachten Ausführungen. Unter patientenrelevanten Endpunkten sind in diesem Zusammenhang dieselben Nutzenkategorien zu verstehen wie bei der Bewertung therapeutischer Interventionen, nämlich Mortalität, Morbidität und gesundheitsbezogene Lebensqualität. Deren Beeinflussung durch diagnostische Maßnahmen lässt sich dabei durch die Vermeidung risikobehafteter(er) bzw. komplikationsträchtiger(er) Interventionen oder durch den gezielt(er)en Einsatz von Interventionen erzielen. Sofern die Erhebung diagnostischer oder prognostischer Informationen selbst risikobehaftet oder komplikationsträchtig ist, kann ein weniger belastendes diagnostisches Verfahren patientenrelevante Vorteile haben, nämlich wenn bei vergleichbarer Testgüte die Testdurchführung selbst weniger Mortalität, Morbidität oder Einschränkungen der gesundheitsbezogenen Lebensqualität verursacht.

Aussagen zum Nutzen diagnostischer Verfahren stützen sich im Idealfall auf randomisierte Studien, die in verschiedenster Form durchgeführt werden können [67,68,247,476,496,626]. In einer Studie mit Strategiedesign kommen bei 2 (oder mehr) Patientengruppen jeweils unterschiedliche Strategien zur Anwendung, die sich jeweils aus einer diagnostischen Maßnahme und der therapeutischen Konsequenz zusammensetzen. Ebenfalls von hoher Aussagekraft sind randomisierte Studien, in denen bei allen Patientinnen und Patienten zunächst die konventionelle und die zu prüfende diagnostische Maßnahme durchgeführt werden und anschließend lediglich diejenigen Patientinnen und Patienten randomisiert werden, bei denen der zu prüfende Test ein anderes Ergebnis und damit eine andere therapeutische Konsequenz als der konventionelle Test ergeben hat (Diskordanzdesign). Ebenfalls von hoher Evidenz und prioritär für die Nutzenbewertung diagnostischer Verfahren heranzuziehen sind solche Studien, bei denen die Wechselwirkung zwischen diagnostischer oder prognostischer Information und therapeutischem Nutzen geprüft wird (Interaktionsdesign [626,701]). Viele diagnostische oder prognostische Merkmale – insbesondere genetische Marker – können in prospektiv vergleichenden Studien auch retrospektiv bestimmt und auf eine mögliche Wechselwirkung hin geprüft werden (sogenanntes prospektiv-retrospektives Design [667]). Die Validität solcher

prospektiv-retrospektiven Designs hängt insbesondere davon ab, dass eine prospektive Planung der Analysen (insbesondere auch der Grenzwertfestlegung) vorhanden ist. Bei allen Studien im Interaktionsdesign ist es darüber hinaus wichtig, dass die eingesetzten Therapien dem aktuellen Standard entsprechen, dass die Informationen (z. B. Gewebeproben) zum interessierenden Merkmal für alle Studienteilnehmerinnen und -teilnehmer oder zumindest für eine Stichprobe, die klar charakterisiert ist und für die die Strukturgleichheit zwischen den Gruppen weiterhin gegeben ist, vollständig vorliegen und dass bei einer Analyse mehrerer Merkmale das Problem einer multiplen Signifikanztestung adäquat berücksichtigt wird (siehe hierzu auch Abschnitt 9.3.2) [627].

Insgesamt ist weniger entscheidend, inwieweit eine diagnostische oder prognostische Information einen aktuellen oder zukünftigen Gesundheitszustand feststellen kann, sondern vielmehr, dass diese Information auch prädiktive Bedeutung hat, also den höheren (oder geringeren) Nutzen einer Folgebehandlung vorhersagen kann [247,668]. Wichtig bei dieser notwendigerweise verknüpften Betrachtungsweise von diagnostischer und therapeutischer Intervention ist, dass ein Nutzen insgesamt sich in der Regel nur dann ergeben kann, wenn beide Interventionen ihr Ziel erfüllen: Sowohl bei unzureichender prädiktiver Diskriminationsfähigkeit der diagnostischen Intervention als auch bei fehlender Effektivität der therapeutischen Intervention wird eine Studie keinen Nutzen der diagnostischen Intervention zeigen können.

Neben dem Strategie- und Interaktionsdesign existiert mit dem Anreicherungsdesign eine 3. wesentliche Form von RCTs zu diagnostischen Fragestellungen [497,701]. Bei diesem Design erfolgt allein auf der Basis des zu prüfenden diagnostischen Tests eine Randomisierung (und damit ein Einschluss) lediglich eines Teils der Patientinnen und Patienten, beispielsweise der Testpositiven, welche dann 1 von 2 Therapieformen erhalten. Einem solchen Design fehlt im Vergleich zum Interaktionsdesign die Prüfung eines möglichen Therapieeffektes beim Rest der Patientinnen und Patienten, also beispielsweise den Testnegativen. Daher lassen sich aus solchen Designs nur dann tragfähige Schlüsse ziehen, wenn aufgrund anderer Informationen ausgeschlossen werden kann, dass ein in der randomisierten Patientengruppe beobachteter Effekt nicht auch in der nicht randomisierten Gruppe vorhanden gewesen wäre.

In speziellen Fällen lässt sich eine Interaktion zwischen diagnostischem oder prognostischem Marker und Therapieeffekt mit hinreichender Sicherheit ableiten, auch wenn der Therapieeffekt nur für die Gesamtgruppe (also Testpositive und -negative gemeinsam) bekannt ist. Im (theoretischen) Extremfall erlaubt es ein Testergebnis, die Krankheit sicher auszuschließen, sodass die Behandlung der Krankheit unnütz ist und allenfalls Nebenwirkungen mit sich bringt. Statistisch kann jedoch nicht nachgewiesen werden, dass ein bestimmtes Testergebnis mit absoluter Sicherheit einen bestimmten Gesundheitszustand anzeigt oder ausschließt. Kann in dieser Situation aber für einen Test gezeigt werden, dass Testnegative ein hinreichend niedriges (oder Testpositive ein hinreichend hohes) Risiko aufweisen, zentrale Endpunkte zu erreichen, so kann der Test in der Abwägung des Nutzens und Schadens einer Therapie eine hinreichend sichere Entscheidung gegen (oder für) eine Therapie ermöglichen [562]. So kann beispielsweise eine Therapie, die in der Gesamtgruppe aller Patientinnen und Patienten ein positives Nutzen-

Schaden-Verhältnis aufweist, in einer Subgruppe von Testnegativen nicht sinnvoll sein, weil der (absolute) Therapieeffekt in dieser Niedrigrisikogruppe höchstens vernachlässigbar gering sein kann. Damit eine solchermaßen verknüpfte Betrachtung des Therapieeffekts in der Gesamtgruppe und des Endpunktrisikos in einer Subgruppe tragfähig ist, muss ausreichend sicher auszuschließen sein, dass der (relative) Therapieeffekt sich in der Subgruppe deutlich von demjenigen der Gesamtgruppe unterscheidet. Ferner können Daten zu Patientenpräferenzen einbezogen werden, um geeignete Grenzwerte zur Bewertung des Nutzen-Schaden-Verhältnisses festlegen zu können. Zudem kann es sinnvoll sein, themenspezifisch eine prozentuale Mindestgröße der Subgruppe der Testnegativen bzw. Testpositiven festzulegen.

Die bisherigen Ausführungen beziehen sich primär auf diagnostische Verfahren, die durch eine Erhöhung der Testgüte (also Sensitivität, Spezifität oder beides) mehr Patientinnen und Patienten einer bestimmten therapeutischen Konsequenz zuführen. In diesen Fällen ist es regelhaft notwendig, durch eine Erfassung der gesamten diagnostisch-therapeutischen Behandlungskette den Einfluss des diagnostischen Verfahrens auf patientenrelevante Endpunkte zu untersuchen. Es kann jedoch vorkommen, dass das zu prüfende diagnostische Verfahren ein anderes, bereits etabliertes diagnostisches Verfahren lediglich ersetzen soll, ohne dass der neue Test zusätzliche Patientinnen und Patienten identifiziert oder ausschließt. Wenn der neue Test direkte patientenrelevante Vorteile aufweist, wie zum Beispiel geringere Invasivität oder Strahlungsfreiheit, wird es nicht immer notwendig sein, erneut die gesamte diagnostisch-therapeutische Kette zu untersuchen, weil sich die therapeutischen Konsequenzen, die sich aus dem neuen Test ergeben, nicht von denen des bisherigen Tests unterscheiden [57,68,516]. Ausreichend für einen Nutznachweis können in diesen Fällen Testgütestudien sein, in denen gezeigt wird, dass das Testergebnis des zu prüfenden Tests (= Indextest) mit dem des bisherigen Tests (= Referenztest) bei einem hinreichend hohen Anteil der Patientinnen und Patienten übereinstimmt (1-seitige Äquivalenzfragestellung). Diese Fragestellung zur (Nicht-)Übereinstimmung von Testergebnissen bei individuellen Patientinnen und Patienten wird auch als Konkordanzfragestellung bezeichnet.

Für einen Vergleich von 2 oder mehr diagnostischen Tests hinsichtlich bestimmter Testgüteeigenschaften ergibt sich die höchste Ergebnissicherheit aus Kohorten- und Querschnittstudien, in denen die diagnostischen Tests voneinander unabhängig jeweils bei denselben Patientinnen und Patienten durchgeführt und unter wechselseitiger Verblindung befundet werden [477,755]. Bei rasch progredienten Krankheitszuständen kann zusätzlich eine zufällige Reihenfolge der Testdurchführung wichtig sein. Neben solchen Studien, die einen intra-individuellen Vergleich der Testergebnisse erlauben, sind auch RCTs denkbar, in denen jeweils ein Teil der Patientinnen und Patienten nur mit dem einen oder dem anderen Indextest untersucht wird, bevor mittels eines einheitlichen Referenztests möglichst alle Ergebnisse verifiziert werden. Auch dieses Studiendesign erlaubt eine Bestimmung von Testgüteeigenschaften mit der höchsten Ergebnissicherheit.

Wenn eine Studie aussagekräftige Daten zum Nutzen, zur diagnostischen Güte oder zur prognostischen Wertigkeit liefern soll, ist es wesentlich, dass hierbei ein Vergleich zur

bisherigen diagnostischen Vorgehensweise erfolgt [703]. Nur auf diese Weise lässt sich der Mehrwert der diagnostischen oder prognostischen Information sicher bestimmen. Bei Studien zur Testgüte bedeutet dies, dass neben der Sensitivität und Spezifität der neuen und der bisherigen Methode von besonderem Interesse ist, inwieweit die diagnostischen Maßnahmen pro Patientin oder pro Patient unterschiedliche Befunde ergeben (siehe oben Konkordanzfragestellung).

Bei Studien zu prognostischen Markern kommt dagegen häufig multifaktoriellen Regressionsmodellen eine zentrale Rolle zu, sodass Abschnitt 9.3.6 zu beachten ist. Bei der Auswahl nicht randomisierter Studiendesigns zu diagnostischen Methoden findet die in Abschnitt 9.1.3 dargestellte Rangordnung verschiedener Studiendesigns regelhaft Anwendung.

Bei der Bewertung der Ergebnissicherheit von Studien zur diagnostischen Güte orientiert sich das Institut in erster Linie an den QUADAS-2-Kriterien [755,756], die allerdings ggf. projektspezifisch angepasst werden. Die Kriterien des STARD-Statements [70,71] werden herangezogen, um im Einzelfall über den Ein- oder Ausschluss nicht im Volltext publizierter Studien zu entscheiden. Für die methodische Bewertung von Prognosestudien wird primär das PROBAST-Instrument verwendet [766]. Anhand der Kriterien des TRIPOD-Statements [141] wird über den Ein- oder Ausschluss nicht im Volltext publizierter Studien entschieden [15,19,341,342,515,666,742].

Studien der Stufe 3 und 4 nach Fryback und Thornbury [267] sollen den Einfluss der zu prüfenden (diagnostischen) Verfahren auf (differenzial-)diagnostische Überlegungen und / oder auf nachfolgende therapeutische (oder andere Management-)Entscheidungen untersuchen, d. h., es wird untersucht, ob das Ergebnis des diagnostischen Tests überhaupt zu einer Veränderung von Entscheidungen führt. Solche Studien bzw. Studienkonzepte haben aber den wesentlichen Nachteil, dass sie nicht scharf definiert sind, und sind daher eher theoretischer Natur. Wesentliches (Qualitäts-)Kennzeichen ist, dass erkennbar geplant die beteiligten Ärztinnen und Ärzte vor der Durchführung oder dem Bekanntwerden der Ergebnisse des zu prüfenden diagnostischen Verfahrens im Hinblick auf die Wahrscheinlichkeit des Vorliegens der Erkrankung und auf ihr weiteres diagnostisches und / oder therapeutisches Vorgehen befragt werden, um die durch das Ergebnis des zu prüfenden diagnostischen Verfahrens veränderte Einstellung ermitteln zu können. Retrospektive Ein- und theoretische Abschätzungen sind dagegen verzerrungsanfällig [267,316]. Der Stellenwert derartiger, letztlich unkontrollierter Studien im Rahmen von Nutzenbewertungen diagnostischer (oder auch prognostischer) Verfahren muss als weitgehend unklar angesehen werden. Informationen zu Managementänderungen allein können deshalb nicht für einen Nutznachweis herangezogen werden, solange keine Informationen über die patientenrelevanten Auswirkungen solcher Änderungen vorliegen.

Es ist weiterhin vorstellbar, dass ein neues diagnostisches Verfahren in eine bereits existierende diagnostische Strategie eingebettet werden soll, beispielsweise dergestalt, dass ein neuer Test einem etablierten Test vor- (Triage-Test) oder nachgeschaltet (Add-on-Test) werden soll, um

den Einsatz des jeweils anderen zu reduzieren [67]. Dabei muss allerdings vor dem Hintergrund der nachfolgenden therapeutischen (oder andersartigen) Konsequenzen bedacht werden, dass sich durch eine solche Kombination von Tests die Patientenpopulationen, die aus den jeweilig kombinierten Testergebnissen resultieren, von denjenigen unterscheiden, die aus den Einzeltestergebnissen resultieren. Dieser Unterschied könnte wiederum Einfluss auf nachfolgende therapeutische (oder andersartige) Konsequenzen und deren Effektivität haben. Ist ein solcher Einfluss nicht hinreichend sicher auszuschließen, sind – wie oben bereits ausgeführt – vergleichende Studien zu den diagnostischen Strategien mit und ohne neues Verfahren erforderlich [257,482].

Mehrere diagnostische Einzeltests / -informationen werden zum Teil über Algorithmen, Scores oder Ähnliches zu einem Gesamttest zusammengefasst. Bei der Bewertung solcher kombinierten Tests sind die gleichen Prinzipien wie bei Einzeltests anzuwenden. Insbesondere müssen die Validierung und die klinische Evaluation eines jeden neuen Tests unabhängig erfolgen von der Testentwicklung (beispielsweise Grenzwertfestlegung, Scoregewichtung oder Auswertungsalgorithmus) [687].

Auch Biomarker, die im Rahmen einer personalisierten oder besser stratifizierten Medizin eingesetzt werden, sind mit den hier charakterisierten Methoden zu evaluieren [357,701]. Dies gilt sowohl für Biomarker, die vor der Entscheidung über den Beginn einer (oder einer anderen) Therapie bestimmt werden, als auch für solche, die während einer Therapie bestimmt werden, um über Fortführung, Abbruch, Wechsel oder Anpassung der Therapie zu entscheiden [672,728]. Auch hier ist die Unterscheidung zwischen der prognostischen und der prädiktiven Wertigkeit eines Merkmals wesentlich. Prognostische Marker geben Informationen über einen zukünftigen Gesundheitszustand und beziehen sich hierbei in aller Regel auf den Verlauf unter Behandlung und nicht auf den natürlichen Verlauf der Erkrankung ohne Behandlung. Dass ein Biomarker prognostische Bedeutung hat, bedeutet nicht, dass er auch prädiktive Bedeutung hat; Gleiches gilt auch umgekehrt nicht.

Schließlich kann es auch bei der Bewertung von Diagnostika notwendig sein, das Ergebnis des Konformitätsbewertungsverfahrens zur CE-Kennzeichnung und den Zulassungsstatus zur Diagnostik eingesetzter Arzneimittel zu beachten (siehe Abschnitt 3.3.1). Entsprechende Konsequenzen sind dann im Berichtsplan (siehe Abschnitt 2.1.1) zu spezifizieren.

3.6 Früherkennung und Screening

Screeningprogramme setzen sich aus unterschiedlichen Bausteinen zusammen, die entweder als Ganzes oder in Teilen untersucht werden (können) [158,663]. Bei der Bewertung erfolgt eine Orientierung an international akzeptierten Standards bzw. Kriterien, wie zum Beispiel denjenigen des UK National Screening Committee (UK NSC [724]), der US Preventive Services Task Force (US PSTF [328,568,632]) oder des National Health Committee (NHC) aus Neuseeland [533].

Gemäß den oben genannten Kriterien bewertet das Institut den Nutzen von Screeningmaßnahmen in erster Linie anhand von prospektiv geplanten vergleichenden Interventionsstudien der gesamten Screeningkette mit einer (idealerweise zufälligen) Zuteilung von Personen zu einer Strategie mit bzw. ohne Anwendung der Screeningmaßnahme (ggf. auch zu unterschiedlichen Screeningstrategien) und der Betrachtung von patientenrelevanten Endpunkten. Die Grundzüge der Bewertung entsprechen dabei den in den Abschnitten 3.1 bis 3.4 gemachten Ausführungen.

Liegen solche Studien nicht oder in nicht ausreichender Quantität und Qualität vor, kann eine Bewertung der einzelnen Bausteine der Screeningkette erfolgen. Dabei wird die Güte des Screeningtests anhand allgemein verwendeter Testgütekriterien, die in ausreichend ergebnssicheren Studien ermittelt wurden (i. d. R. Phase 3 nach Köbberling et al. [1]), bewertet (siehe Abschnitt 3.5) und geprüft, inwieweit belegt ist, dass die aus den Testergebnissen resultierenden Konsequenzen mit einem Nutzen verbunden sind. Für den (zumeist anzunehmenden) Fall therapeutischer Konsequenzen lassen sich solche Belege aus randomisierten Interventionsstudien ableiten, in denen eine frühzeitige(re) mit einer späte(re)n Intervention verglichen wurde. Ggf. kann der Nutzen einer frühzeitige(re)n gegenüber einer späte(re)n Intervention auch anhand von Interventionsstudien geprüft werden, in denen sich eine Wechselwirkung zwischen der Frühzeitigkeit des Interventionsbeginns und dem Interventionseffekt untersuchen lässt. Dies kann entweder direkt innerhalb einer Studie oder indirekt durch den Vergleich von Studien mit unterschiedlichem Interventionsbeginn, aber ansonsten vergleichbarem Design erfolgen. Auch hier entsprechen die Grundzüge der Bewertung den in den Abschnitten 3.1 bis 3.4 gemachten Ausführungen.

Ein besonderer Schadenaspekt von Screening besteht darin, Überdiagnosen hervorzubringen. Eine Überdiagnose ist definiert als eine eigentlich richtig-positive Diagnose, die aber – ohne Screening – zu Lebzeiten keine Symptome verursacht hätte [125]. Überdiagnosen treten beispielsweise bei einem Screening auf langsam fortschreitende Erkrankungen auf, weil dort die Wahrscheinlichkeit hoch ist, dass eine Person vor einer Symptomentwicklung an einer anderen Ursache verstirbt. Da Überdiagnosen als schädliche Wirkung jedes Screenings unvermeidlich sind, jedoch nur indirekt erfasst werden können, sind zur Erfassung dieses Endpunkts spezielle Methoden notwendig [124,219].

3.7 Prävention

Prävention zielt darauf ab, eine gesundheitliche Schädigung zu verhindern, weniger wahrscheinlich zu machen oder zu verzögern [746]. Während die Primärprävention alle Maßnahmen vor Eintritt einer fassbaren biologischen Schädigung zur Vermeidung auslösender oder vorhandener Teilursachen umfasst, beinhaltet die sekundäre Prävention Aktivitäten zur Entdeckung klinisch symptomloser Krankheitsfrühstadien und ihre erfolgreiche Frühtherapie (siehe auch Abschnitt 3.6). Maßnahmen der Primär- und Sekundärprävention zeichnen sich dadurch aus, dass im Unterschied zu kurativen Maßnahmen häufig ganze Bevölkerungsgruppen im Fokus der Intervention stehen. Tertiärprävention im engeren Sinne bezeichnet spezielle

Interventionen zur Verhinderung bleibender, insbesondere sozialer Funktionseinbußen nach dem Eintreten einer Krankheit [345]. Sie ist weniger Gegenstand dieses Abschnitts, sondern findet ihre Entsprechung in den Abschnitten zur Nutzenbewertung von Arzneimitteln und nicht medikamentösen Verfahren (siehe Abschnitte 3.3 und 3.4).

Nutzenbewertungen zu (anderen als Screening-)Präventionsprogrammen führt das Institut in erster Linie ebenfalls anhand prospektiv geplanter vergleichender Interventionsstudien mit einer (idealerweise zufälligen) Zuteilung von Personen zu einer Strategie mit bzw. ohne Anwendung der Präventionsmaßnahme und der Betrachtung von patientenrelevanten Endpunkten durch. Unter anderem wegen einer möglichen Kontamination zwischen der Interventions- und Kontrollgruppe kommen alternativ auch Studien, bei denen Cluster den Interventionsarmen zugeteilt werden, infrage [714].

In Einzelfällen ist zu prüfen, inwieweit die Berücksichtigung weiterer Studiendesigns sinnvoll sein kann [407]. So werden beispielsweise massenmediale Kampagnen häufig im Rahmen eines sogenannten Interrupted Time Series Design evaluiert (z. B. Vidanapathirana et al. [736]) und die Anwendung dieses Studiendesigns wird auch für gemeindebezogene Präventionsprogramme empfohlen [58]. Das Institut orientiert sich bei der Qualitätsbewertung solcher Studien an den von der Cochrane Effective Practice and Organisation of Care Review Group entwickelten Kriterien [139].

Für den Nutzen auf Populationsebene ist neben der Effektivität eines solchen Programms auch die Höhe der Teilnahmerate entscheidend. Außerdem ist die Frage relevant, welche Personen erreicht werden; so gibt es Hinweise darauf, dass Präventionsprogramme gerade von Bevölkerungsgruppen mit erhöhtem Krankheitsrisiko weniger in Anspruch genommen werden [458]. Beide Aspekte stehen deshalb im besonderen Fokus von Bewertungen des Instituts.

3.8 Potenzialbewertung und §-137h-Bewertung

Im Folgenden wird zuerst auf Potenzialbewertungen gemäß § 137e SGB V eingegangen, bevor Bewertungen gemäß § 137h SGB V erläutert werden. Die Gemeinsamkeit beider Bewertungsverfahren besteht darin, dass jeweils bestimmte nicht medikamentöse Untersuchungs- oder Behandlungsmethoden anhand von Unterlagen bewertet werden, die dem Institut von externen Dritten vorgelegt werden. Als weitere Gemeinsamkeit können beide Bewertungsverfahren dazu führen, dass eine Erprobungsstudie durchgeführt wird, sodass die Eckpunkte der Studie geprüft oder vom Institut selbst entwickelt werden müssen.

3.8.1 Potenzialbewertung

Potenzialbewertungen gemäß § 137e Abs. 7 SGB V zielen im Gegensatz zu Nutzenbewertungen darauf ab zu prüfen, ob neue Untersuchungs- oder Behandlungsmethoden möglicherweise einen Nutzen aufweisen. Potenzial bedeutet hierbei, dass erstens die bisher vorliegenden Erkenntnisse einen möglichen Nutzen erkennen lassen und zweitens auf der Basis dieser Erkenntnisse eine Studie geplant werden kann, die eine Bewertung des Nutzens der

Methode auf einem ausreichend sicheren Erkenntnisniveau erlaubt (siehe Kapitel 2 § 14 Abs. 3 und 4 der Verfahrensordnung des G-BA [276]).

Eine Potenzialbewertung gemäß § 137e Abs. 7 SGB V beruht auf einem Antrag, für den der G-BA die Form und die notwendigen Inhalte definiert hat. Antragsberechtigt sind die Herstellerinnen und Hersteller eines Medizinprodukts, auf dessen Einsatz die technische Anwendung einer neuen Untersuchungs- oder Behandlungsmethode maßgeblich beruht, und Unternehmen, die in sonstiger Weise als Anbieterin oder Anbieter einer neuen Methode ein wirtschaftliches Interesse an einer Erbringung zulasten der Krankenkassen haben. Im Antrag sind aussagekräftige Unterlagen vorzulegen, die sich insbesondere auf den aktuellen Erkenntnisstand und den zu erwartenden Nutzen der neuen Untersuchungs- oder Behandlungsmethode beziehen (siehe § 20 Abs. 2 Nr. 5 der Verfahrensordnung des G-BA [276]). Optional kann auch ein Vorschlag zu den Eckpunkten einer Erprobungsstudie eingereicht werden. Ein Antrag zu einer Methode kann sich auf eine oder mehrere Indikationen beziehen.

Im Rahmen der Potenzialbewertung wird vom Institut die Plausibilität der Angaben des Antragstellers geprüft. Diese Prüfung bezieht sich insbesondere auf die Sinnhaftigkeit der im Antrag dargestellten medizinischen Fragestellung(en), die Güte der vom Antragsteller durchgeführten Informationsbeschaffung (siehe Abschnitt 8.5), die Einschätzung der Ergebnissicherheit der relevanten Studien und die Korrektheit der im Antrag dargestellten Ergebnisse. Eigene Recherchen können zur Unterstützung der Bewertung durchgeführt werden; es ist jedoch nicht Aufgabe oder Ziel des Instituts, die vorgelegten Unterlagen zu vervollständigen. Die Bewertung mündet in einer Aussage zum Potenzial der beantragten Untersuchungs- oder Behandlungsmethode. Wenn aus Sicht des Instituts ein Potenzial feststellbar ist, wird die vom Antragsteller vorgeschlagene Erprobungsstudie geprüft; sofern der Antrag keinen solchen oder einen ungeeigneten Vorschlag enthält, konkretisiert das Institut die Eckpunkte einer möglichen Erprobungsstudie.

Aufgrund der besonderen Zielsetzung sind bei der Potenzialbewertung im Vergleich zur Nutzenbewertung deutlich niedrigere Anforderungen an die Evidenz zu stellen. Letztendlich besteht erst das Ziel der Erprobung darin, eine adäquate Datengrundlage für eine zukünftige Nutzenbewertung herzustellen. Dementsprechend lässt sich ein Potenzial insbesondere auch auf der Basis nicht randomisierter Studien begründen. Darüber hinaus finden auch weitere methodische Grundsätze der Nutzenbewertung bei der Potenzialbewertung keine oder eine nur eingeschränkte Anwendung, wie im Folgenden dargestellt wird.

Im Gegensatz zur Nutzenbewertung wird im Rahmen von Potenzialbewertungen aufgrund der niedrigeren Anforderungen an die Evidenz eine erweiterte Bewertung der qualitativen Ergebnissicherheit nicht randomisierter Studien vorgenommen. Hierbei werden neben den in Abschnitt 3.1.4 für randomisierte Studien genannten Stufen (hohe oder mäßige Ergebnissicherheit) folgende Grade verwendet:

- **geringe qualitative Ergebnissicherheit:** Ergebnis einer höherwertigen nicht randomisiert vergleichenden Studie (beispielsweise quasirandomisierte kontrollierte Studien, nicht randomisierte kontrollierte Studien mit aktiver Zuteilung der Intervention nach vorab geplanter Regel, prospektive vergleichende Kohortenstudien mit passiver Zuteilung der Intervention) mit adäquater Kontrolle für Confounder,
- **sehr geringe qualitative Ergebnissicherheit:** Ergebnis einer höherwertigen nicht randomisiert vergleichenden Studie (siehe vorheriger Punkt), jedoch ohne adäquate Kontrolle für Confounder, oder Ergebnis einer sonstigen nicht randomisiert vergleichenden Studie (beispielsweise retrospektive vergleichende Kohortenstudien, historisch kontrollierte Studien, Fall-Kontroll-Studien),
- **minimale qualitative Ergebnissicherheit:** Ergebnis einer nicht vergleichenden Studie (beispielsweise 1-armige Kohortenstudien, Verlaufsbeobachtungen oder Fallserien, Querschnittsstudien oder sonstige nicht vergleichende Studien), das einen indirekten Vergleich mit den Ergebnissen anderer Studien (Literaturkontrollen) erlaubt.

Ein wesentlicher Aspekt der Ergebnissicherheit ist die Kontrolle für Confounder, die insbesondere durch multifaktorielle statistische Methoden – wie in Abschnitt 9.3.6 beschrieben – angestrebt werden kann. Auch weitere Faktoren werden bei der Einschätzung der Ergebnissicherheit berücksichtigt (siehe Abschnitt 9.1.4).

Abweichend vom Vorgehen, das auf Ableitung von Nutzenaussagen im Sinne von Belegen, Hinweisen oder Anhaltspunkten zielt (siehe Abschnitt 3.1.2), werden bei der Potenzialbewertung auch Surrogatendpunkte berücksichtigt, für die noch keine hinreichende Validität gezeigt werden konnte. Surrogatendpunkte sollen jedoch etabliert und plausibel sein, um hierüber ein Potenzial begründen zu können.

Sofern das Potenzial diagnostischer Methoden zu bewerten ist, finden auch Daten zur Testgüte Berücksichtigung. Hierbei ist die Ergebnissicherheit der zugrunde liegenden Studien zu prüfen (siehe Abschnitte 3.5 und 9.3.13). Im 2. Schritt erfolgt eine Plausibilitätsprüfung der diagnostischen Methode hinsichtlich der vom Antragsteller postulierten Effekte bezüglich patientenrelevanter Endpunkte, d. h. möglicher direkter Effekte der Methode sowie therapeutischer Konsequenzen, über die die diagnostische Methode patientenrelevante Zielgrößen beeinflussen könnte.

3.8.2 §-137h-Bewertung

Bei §-137h-Bewertungen ist die Grundlage der Bewertung nicht wie bei Verfahren nach § 137e SGB V ein Antrag eines Medizinprodukteherstellers oder eines sonstigen Unternehmens; Bewertungen gemäß § 137h SGB V fußen auf den von einem Krankenhaus zusammengestellten Unterlagen. Die Bewertung bezieht sich auf eine neue Untersuchungs- oder Behandlungsmethode, deren technische Anwendung maßgeblich auf dem Einsatz eines Medizinprodukts hoher Risikoklasse beruht. Mit dem Hersteller des Medizinprodukts hat das Krankenhaus Einvernehmen herzustellen, bevor es die Informationen zum Nutzen der Methode an den G-BA

übermittelt. Diese Informationen können dann beim G-BA durch weitere Krankenhäuser und Medizinproduktehersteller ergänzt werden, bevor sie dem IQWiG zur Bewertung übermittelt werden.

Eine §-137h-Bewertung kann gemäß § 137h Abs. 1 Satz 4 SGB V zu 3 möglichen Ergebnissen führen:

- Die zu prüfende Methode lässt einen Nutzen erkennen im Sinne von § 137h Abs. 1 Satz 4 SGB V. In diesem Fall erübrigt sich die Befassung mit einer möglichen Erprobungsstudie.
- Schädlichkeit oder Unwirksamkeit der zu prüfenden Methode sind im Sinne von § 137h Abs. 1 Satz 4 SGB V erkennbar. Auch in diesem Fall erübrigt sich die Befassung mit einer möglichen Erprobungsstudie.
- Es lassen sich weder der Nutzen noch die Schädlichkeit oder die Unwirksamkeit der Methode erkennen. In diesem Fall sind die Eckpunkte einer möglichen Erprobungsstudie zu prüfen oder neu auszuarbeiten.

Die Prüfung der ersten beiden Punkte (erkennbarer Nutzen, Schädlichkeit) erfolgt gemäß den in Abschnitt 3.1 beschriebenen Prinzipien. Schädlichkeit entspricht einem höheren Schaden der neuen Methode im Vergleich zur Nichtbehandlung. Eine Unwirksamkeit besteht darin, dass die neue Methode erkennbar nur Effekte zeigt, die denen einer Nichtbehandlung vergleichbar sind. Hierfür können neben Ergebnissen zu patientenrelevanten Endpunkten auch Ergebnisse zu Surrogatendpunkten ausreichend sein.

4 Kosten-Nutzen-Bewertung medizinischer Interventionen

4.1 Einführung

Gemäß SGB V bestimmt das Institut auftragsbezogen über die Methoden und Kriterien für die Erarbeitung von Kosten-Nutzen-Bewertungen (KNBs) auf der Grundlage der in den jeweiligen Fachkreisen anerkannten internationalen Standards der evidenzbasierten Medizin und der Gesundheitsökonomie. Für jede KNB müssen u. a. auch Entscheidungen zur Perspektive, zum Zeithorizont, zur Auswahl der Komparatoren, zum zugrunde liegenden Versorgungspfad, zum Modell, zur Datengrundlage und zur Darstellung der Unsicherheit gefällt werden. Vor dem Hintergrund der Beauftragung durch den G-BA werden diese grundlegenden Kriterien einer KNB kurz erläutert. Alle Abweichungen von den hier vorgelegten Methoden sind im Einzelfall zu begründen.

4.1.1 Rechtliche Grundlagen für die Kosten-Nutzen-Bewertung nach SGB V

In den folgenden Fällen kann das Institut mit gesundheitsökonomischen Fragestellungen beauftragt werden:

- Nach § 139a Abs. 3 Nr. 2 SGB V kann das Institut zu Fragen der Qualität und Wirtschaftlichkeit der im Rahmen der gesetzlichen Krankenversicherungen erbrachten Leistungen beauftragt werden.
- KNBs von Arzneimitteln können vom G-BA auch nach § 35b SGB V beauftragt werden.
- Weiterhin kann eine KNB nach § 139b Abs. 2 SGB V vom BMG beauftragt werden.

Im Folgenden werden zunächst methodische Aspekte berührt, die für KNBs im Allgemeinen gelten. In Abschnitt 4.9 werden Abweichungen erläutert, die sich bei KNBs von Arzneimitteln nach § 35b SGB V ergeben.

4.1.2 Perspektive

Auftragsabhängig können die (reine) Perspektive der gesetzlichen Krankenversicherung (GKV), die Perspektive der GKV-Versichertengemeinschaft, die Sozialversicherungsperspektive bzw. die Perspektive einzelner Sozialversicherungsträger sowie die gesellschaftliche Perspektive berücksichtigt werden. Im Unterschied zur reinen GKV-Perspektive werden bei der GKV-Versichertenperspektive auch Kosten berücksichtigt, die den Versicherten entstehen, z. B. aus Zuzahlungen (siehe Abschnitt 4.4.1). Je nach Auftrag für eine KNB kann es erforderlich sein, die Perspektive einzelner Sozialversicherungsträger zusätzlich zur Perspektive der GKV-Versichertengemeinschaft einzunehmen. Die Entscheidung, ob weitere Perspektiven in eine KNB aufgenommen werden sollen, hängt allein von der Frage ab, ob dies für die Entscheidungsträger relevant ist. Die Ergebnisse der Bewertung aus einer erweiterten Perspektive werden dem Entscheidungsträger gesondert ausgewiesen zur Verfügung gestellt.

4.1.3 Zeithorizont

Der Zeithorizont muss mindestens die durchschnittliche Studiendauer abbilden und somit für die Entscheidung relevante Kosten- und Nutzenunterschiede zwischen den Interventionen einer KNB einbeziehen. Insbesondere bei chronischen Erkrankungen sollte möglichst ein längerer Zeithorizont gewählt werden [88,201,495,715]. Für gesundheitsökonomische Evaluationen über längere Zeithorizonte werden häufig entscheidungsanalytische Modelle eingesetzt [660]. Die Kosten und der Nutzen sollen immer über denselben Zeithorizont modelliert werden.

Der angemessene Zeithorizont ist häufig länger als der Zeitraum, den die verfügbaren Primärdaten aus prospektiven Studien abdecken. Dann sollte unter Abwägung der Dauer der Studien ein der Krankheit angemessener Zeithorizont gewählt werden [349,715].

4.1.4 Auswahl der Komparatoren

Für die Herleitung einer Effizienzgrenze, der vom Institut gewählten Darstellungsform der Ergebnisse einer KNB (siehe Abschnitt 4.6), sind alle in einem Indikationsgebiet versorgungsrelevanten Interventionen in eine KNB einzubeziehen. Wirkstoffe z. B. können auch zu Wirkstoffklassen zusammengefasst werden, wenn dies medizinisch sinnvoll erscheint und ausreichende Homogenität vorliegt (siehe Abschnitt 9.3.7).

4.1.5 Versorgungspfad

Für jede KNB sollte/n zunächst ein/mehrere Versorgungspfad/e für das Indikationsgebiet entwickelt werden. Ein Versorgungspfad beschreibt Behandlungsabläufe für Patientinnen und Patienten mit einer / mehreren spezifischen Indikation/-en in einer zeitlichen Abfolge und strukturiert sie nach Sektoren, beteiligten Berufsgruppen, Stadien und ggf. weiteren Aspekten. Dieser Versorgungspfad dient als Basis, um das entscheidungsanalytische Modell (siehe Abschnitte 4.1.6 und 4.2) zu erstellen. Weiterhin basieren auf dem Versorgungspfad auch die Recherchen nach Daten zu Kosten und weiteren notwendigen Daten für das Modell.

Zunächst sind auftragsbezogen der Krankheitsverlauf und die Versorgung in Deutschland kurz für die relevante Indikation mit Angabe der Quellen zu beschreiben. Die relevanten Interventionen und Behandlungsschritte in verschiedenen Leistungsbereichen inklusive der Leistungserbringer müssen innerhalb der Grenzen der Zulassung und des Wirtschaftlichkeitsgebots erbracht werden. Darüber hinaus muss die Anwendung innerhalb der Vorgaben der Richtlinien und Therapiehinweise bewertet werden, die im System der gesetzlichen Krankenversicherung gelten. Weiterhin sind aktuelle Therapieempfehlungen für Deutschland darzustellen, wobei gültige Leitlinien herangezogen werden sollten. Aus dem beschriebenen Versorgungskontext sollen die für die KNB relevanten Komponenten abgegrenzt werden, sodass ein modellrelevanter Versorgungspfad beschrieben werden kann. Wenn einzelne Komponenten dezidiert nicht in den Versorgungspfad aufgenommen werden, sollte diese Entscheidung begründet werden.

Selbst wenn eine Piggy-back-Studie vorliegt – eine klinische Studie, in der neben dem Nutzen und Schaden einer Technologie zugleich auch die Kosten erhoben werden –, sollte begleitend ebenfalls ein Versorgungspfad abgebildet werden, damit die Kosten und weitere Daten, die in der Piggy-back-Studie erhoben wurden, anhand eines beigefügten Versorgungspfads nachvollzogen werden können.

4.1.6 Modell

Piggy-back-Studien liegen sehr selten vor. Darüber hinaus werden ökonomische Daten in klinischen Studien meist nicht erhoben. Häufig sind die Daten nicht ausreichend für eine umfassende Aufbereitung der Kosten einer Intervention. Denn zum einen liefern klinische Studien selten Informationen zu den langfristigen ökonomischen Konsequenzen, die mit der Einführung einer neuen Intervention einhergehen. Zum anderen greifen sie nicht immer adäquat und vollständig die für die Kostenseite in Deutschland relevanten Versorgungsaspekte auf. Darüber hinaus kann auch ein protokollinduzierter Ressourcenverbrauch im Rahmen klinischer Studien Fehleinschätzungen auf der Kostenseite induzieren. Aus diesen Gründen ist die Modellierung der Kosten einer Intervention eine wesentliche Komponente der KNB (siehe Abschnitt 4.4). Ebenso kann in einer KNB der Nutzen modelliert werden, wenn einer KNB ein längerer als der studienbelegte Zeithorizont zugrunde gelegt werden soll (siehe Abschnitt 4.3).

4.1.7 Spezifische Datenquellen der Gesundheitsökonomie

Daten, die für die Abbildung der Versorgung, der Epidemiologie und der Kosten in die KNB eingehen, können auf verschiedenen Wegen erhoben werden und aus unterschiedlichen Quellen stammen (siehe auch Abschnitte 4.4.4 und 4.5.2).

Auswertungen von Sekundärdaten sollten sich an den Leitlinien und Empfehlungen zur Guten Praxis Sekundärdatenanalyse ausrichten [24]. Insbesondere sollen die Auswahl der Datenbasis, die Größe und relevante Merkmale der Stichprobe und der Studienpopulation (inkl. Ein- und Ausschlusskriterien), die statistischen Methoden und die Kontrolle von Störgrößen transparent beschrieben und begründet werden. Die Generalisierbarkeit und die Repräsentativität der Ergebnisse sollten erläutert werden. Die einzelnen Auswertungsschritte müssen nachvollziehbar sein; Plausibilitätskontrollen sind sicherzustellen.

Werden Leitlinien verwendet, sollten diese evidenzbasiert sein und aus dem deutschen Gesundheitssystem oder aus einem OECD-Land mit in Deutschland vergleichbaren Versorgungsstrukturen stammen (siehe Abschnitte 5.1.2 und 5.1.3). Sind keine evidenzbasierten Leitlinien in dem zu untersuchenden Indikationsgebiet vorhanden, ist abzuwägen und transparent darzulegen, ob andere deutsche Leitlinien verwendet werden können oder ob auf Expertenbefragungen zurückgegriffen werden soll.

Befragungen von Expertinnen und Experten folgen den allgemein anerkannten Methoden und Vorgehensweisen der quantitativen Sozialforschung. Das heißt, dass bei der Befragung von Expertinnen und Experten explizit Angaben zur Rekrutierung, Anzahl und Expertise der

Expertinnen und Experten, Fragestellung, zu individuellen Antworten (nicht nur Mittelwerten), zur Art der Konsensfindung sowie zur Darstellung und zum Umgang mit den Ergebnissen zu machen sind. Preiskataloge bzw. -verzeichnisse müssen aktuell sein und die für die GKV relevanten Preise abbilden.

4.1.8 Unsicherheit

Nach international üblichem Vorgehen werden folgende Arten der Unsicherheit in der gesundheitsökonomischen Entscheidungsanalyse unterschieden [81]:

Tabelle 6: Konzepte von Unsicherheit in der gesundheitsökonomischen Entscheidungsanalyse

Bezeichnung	Konzept	Andere Bezeichnungen in der Literatur	Vergleichbares Konzept bei Regressionsmodellen
stochastische Unsicherheit	zufällige Variabilität im Ergebnis bei gleichen Probandinnen und Probanden	Variabilität, First-Order Uncertainty	Fehlerterm
Parameterunsicherheit	Unsicherheit in der Schätzung des interessierenden Parameters	Second-Order Uncertainty	Standardfehler des Schätzers
strukturelle Unsicherheit	zugrunde liegende Annahmen im entscheidungsanalytischen Modell	Modellunsicherheit	Art des Regressionsmodells, z. B. linear, loglinear
(adaptiert und übersetzt aus Briggs et al. [81])			

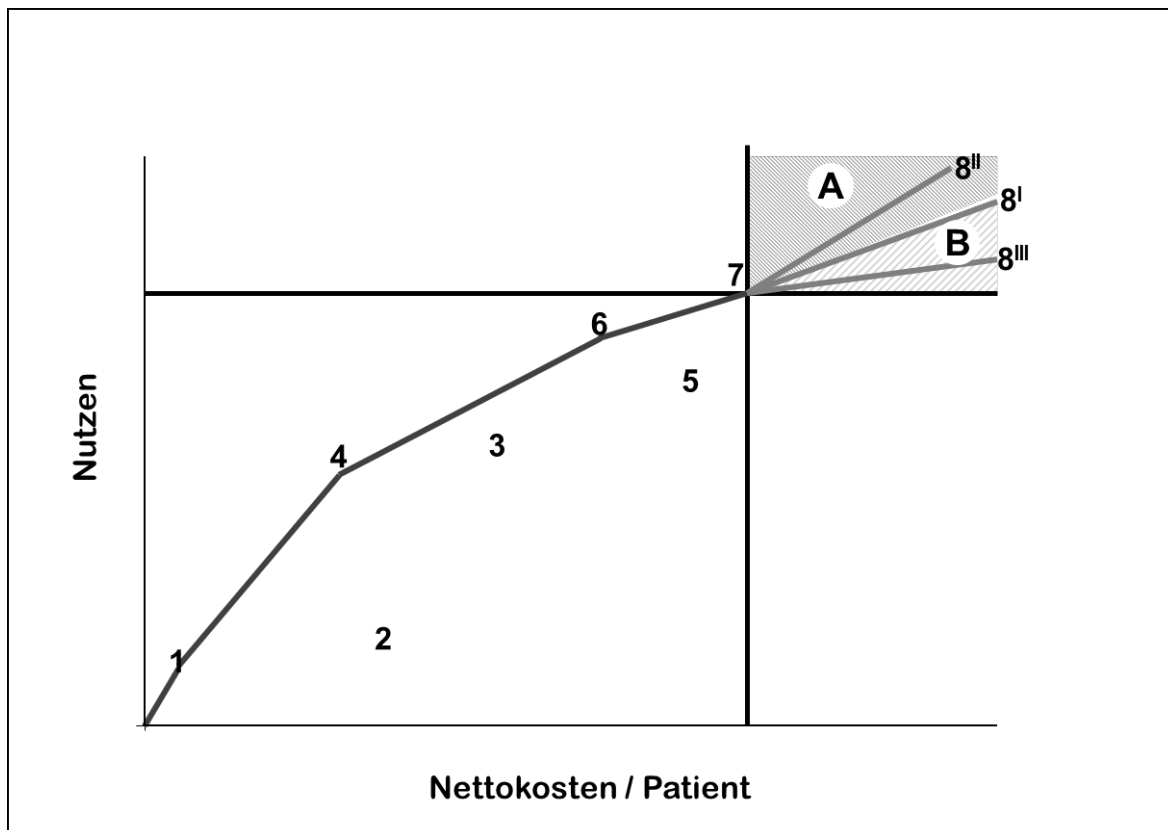
Aufgrund der Vielschichtigkeit einer KNB muss die Untersuchung der Unsicherheit in allen Bereichen bedacht werden. Dazu orientiert sich das Institut an der Einteilung der Unsicherheit (siehe Tabelle 6).

Dazu werden in den Abschnitten 4.2, 4.3, 4.4 und 4.5 zunächst grundsätzliche Bemerkungen zur Unsicherheit und zu den Verteilungsannahmen gemacht. In Abschnitt 4.7 werden dann die Durchführung und die Darstellung der Untersuchung von Unsicherheit dargestellt.

4.1.9 Interpretation der Ergebnisse

Die Ergebnisse werden tabellarisch und grafisch in Form einer Effizienzgrenze dargestellt (siehe Abbildung 10). Die Interventionen 1 bis 7 werden als Komparatoren mit ihren Kosten-Nutzen-Verhältnissen abgetragen. Die Interventionen 1, 4, 6 und 7 bilden eine Effizienzgrenze. Das letzte Segment der Effizienzgrenze kann unter der Annahme, dass es den Kehrwert der gegenwärtigen Zahlungsbereitschaft darstellt, linear extrapoliert werden. Dann gilt: Interventionen, die sich endpunktbezogen auf der Fläche A befinden (vgl. Intervention 8^{II}), haben nach ihrem Kosten-Nutzen-Verhältnis gegenüber dem extrapolierten letzten Segment der Effizienzgrenze ein besseres Kosten-Nutzen-Verhältnis und könnten somit zum angegebenen Preis erstattet werden. Interventionen auf der Fläche B (vgl. Intervention 8^{III}) haben nach ihrem Kosten-Nutzen-Verhältnis gegenüber dem extrapolierten letzten Segment der Effizienzgrenze

ein ungünstigeres Kosten-Nutzen-Verhältnis, sodass ihr Preis bezogen auf die Effizienzgrenze als nicht angemessen angesehen werden kann, da sich die vorhandene Effizienz verschlechtern würde. Der Entscheidungsträger kann unter Berücksichtigung des Kriteriums der Angemessenheit der Kosten von Interventionen auf der Fläche B einen Erstattungsbetrag verhandeln. Interventionen mit gleichbleibendem Kosten-Nutzen-Verhältnis (vgl. Intervention 8^I) erfüllen ebenfalls das Kriterium, dass ihr Preis im Vergleich zum extrapolierten letzten Segment der Effizienzgrenze angemessen wäre.



Nettokosten: Nettokosten sind die Kosten, die sich aus der Versorgung ergeben, bereinigt um Kosteneinsparungen, sogenannte Cost-Offsets.

Abbildung 10: Darstellung der entscheidungsrelevanten Bereiche

Die Effizienzgrenze zeigt auf, welche Interventionen den verhältnismäßig größten Nutzen im Verhältnis zu den entstehenden Kosten aufweisen. Ineffiziente Interventionen sind z. B. sowohl teurer als auch in Bezug auf einen Endpunkt von geringerem Nutzen als andere Interventionen. Wenn sowohl die Kosten als auch der durch die neue Intervention generierte Nutzen höher sind als diejenigen, die bereits in der Effizienzgrenze abgebildet sind, ist aus der Effizienzgrenze selbst nicht direkt ableitbar, welche Kosten dieser Intervention angemessen sind. Somit müssen weitere Kriterien herangezogen werden, um zu beurteilen, ob die Anwendung einer Zusatznutzen stiftenden, aber kostenintensiveren neuen Therapie angemessen ist. Das Institut geht davon aus, dass eine Verschlechterung der Effizienz in einer Indikation durch die Aufnahme neuer Interventionen unangemessen ist. Diese Effizienz wird durch die lineare Extrapolation

der Steigung des letzten Segments der Effizienzgrenze umgesetzt. Daher werden bei einem gegebenen Nutzen einer zu beurteilenden Intervention solche Kosten-Nutzen-Verhältnisse als angemessen angesehen, die – gemessen an der Effizienzgrenze – nicht zu einer Verschlechterung der Effizienz in einem gegebenen Indikationsgebiet führen (siehe Abbildung 10).

Wenn ein Maß des Gesamtnutzens festgelegt wird (siehe Abschnitt 4.3.3), ist dieses als primäres Ergebnis anzusehen. Ist die Bestimmung mehrerer Effizienzgrenzen für die Bewertung einer Intervention erforderlich, so bleibt es dem Entscheidungsträger unter Beachtung der Relevanz der patientenrelevanten Endpunkte vorbehalten, eine Gewichtung vorzunehmen. Eine ähnliche Vorgehensweise kann auch bei Zulassung einer Intervention in mehreren Indikationsgebieten erwogen werden.

Die Zumutbarkeit einer Kostenübernahme durch die GKV-Versichertengemeinschaft hängt einerseits von der Angemessenheit des Preises einer Intervention ab, andererseits aber auch von den damit verbundenen künftigen Gesamtausgaben in Abhängigkeit der finanziellen Leistungsfähigkeit und Zahlungsbereitschaft der GKV-Versichertengemeinschaft. Da weder die finanzielle Leistungsfähigkeit noch die Zahlungsbereitschaft der GKV-Versichertengemeinschaft bewertet wird, wird zur Zumutbarkeit der Kostenübernahme keine konkrete Empfehlung abgegeben. Um zukünftige finanzielle Auswirkungen einer Kostenübernahme darzustellen, soll eine Ausgaben-Einfluss-Analyse (Budget-Impact-Analyse) durchgeführt werden, die dem Entscheidungsträger als Informationsgrundlage für die Entscheidung über die Zumutbarkeit dienen kann.

4.2 Modellierung

4.2.1 Grundlagen

In einem gesundheitsökonomischen Entscheidungsmodell als Kernstück einer KNB werden Daten zum Nutzen und zu den Kosten aus verschiedenen Quellen zusammengeführt, um Kosten-Nutzen-Verhältnisse von Interventionen im Krankheitsverlauf zu berechnen. Eine Zusammenführung von Daten aus verschiedenen Quellen mithilfe eines Modells ist oft aus einer Vielzahl von Gründen notwendig. Zumeist werden in einer Studie nicht alle für die Entscheidung relevanten Variablen erhoben. Auch wird ein gesundheitsökonomisches Modell explizit dazu verwendet, Nutzen und Kosten über den studienbelegten Zeitraum hinaus zu extrapolieren.

Gesundheitsökonomische Modelle sind somit, wie mathematisch-formalisierte Modelle, auch eine vereinfachte Abbildung der Realität. Durch eine bewusste Reduktion der Komplexität der für das Entscheidungsproblem relevanten Entscheidungsfaktoren und -variablen wird zudem analytische Klarheit geschaffen.

Für gesundheitsökonomische Modelle ist eine gründliche Dokumentation von zentraler Bedeutung. Diese Dokumentation sollte grundsätzlich zweigeteilt sein: zum einen eine allgemeine anschauliche Dokumentation des Vorgehens mit Darstellung und Begründung der

getroffenen Entscheidungen und der Auswahl der Daten(-quellen); zum anderen eine technische Dokumentation, welche die funktionalen / mathematischen Beziehungen der Modellkomponenten darstellt, sodass eine sachverständige Dritte oder ein sachverständiger Dritter die Ergebnisse des Modells unabhängig von einer bestimmten Software replizieren kann.

Der Grad der Komplexität bzw. das Maß der Reduktion eines Modells hängt immer von der Fragestellung ab und lässt sich a priori nicht festlegen. Daher ist neben der internen Validität eines Modells die Anwendbarkeit zu beschreiben und zu belegen. Die Modellstruktur (z. B. Gesundheitszustände), die sich innerhalb der Zulassung und im Rahmen der Leistungserbringung nach SGB V bewegen muss, wird zur Sicherung der externen Validität regelhaft mit externen klinischen Sachverständigen abgestimmt.

4.2.2 Grundsätzliche Aspekte der Modellentwicklung

Die Modelle müssen im Ergebnis detailliert den Nutzen und die Kosten abbilden, die sich in Deutschland für die zu betrachtende Intervention ergeben. Dazu müssen folgende Informationen in das Modell eingehen:

- Ergebnisse zu den Effekten (Nutzen und Schaden) der Interventionen,
- vollständige Erfassung der Krankheitskosten und
- alle Krankheits- und Behandlungsaspekte, die relevanten Einfluss auf die Nutzen- oder Kostenkomponenten des Modells haben können, z. B. in den Bereichen Demografie, Epidemiologie, Versorgungspfad(e).

Da oft Daten zu einzelnen Aspekten fehlen, ist es besonders wichtig, den Einfluss dann getroffener Annahmen und Modellinputs auf die Ergebnisse durch Sensitivitätsanalysen zu explorieren.

Um die Validität und formale / inhaltliche Nachvollziehbarkeit von Modellierungen sicherzustellen, müssen folgende Bedingungen erfüllt sein:

- vollständige Transparenz mit klar beschriebenen und begründeten Modellinputs und Annahmen,
- hinreichende Tiefe zur ausreichenden Abbildung der modellierten Erkrankung, der mit ihr assoziierten Kosten und der betreffenden Versorgungspfade,
- ausreichende Flexibilität zur Berechnung multipler Szenarien bei variierenden Annahmen und Settings,
- Möglichkeit der Bestimmung von Unsicherheit in den vorhergesagten Kosten und Nutzenkomponenten.

4.2.3 Einflussdiagramm und Modellkonzept

Auf Basis der Überlegungen und Informationen, die zur Erstellung des Versorgungspfades geführt haben, werden die Grundlagen für das Modell in einem Einflussdiagramm und einem Modellkonzept dargelegt.

Ein Einflussdiagramm bildet grafisch die wesentlichen Zusammenhänge zwischen dem Krankheitsverlauf, den Patientencharakteristika, der Pathophysiologie und der Behandlung des Modells ab. Es zeigt die Faktoren auf, die Einfluss auf eine zu modellierende Fragestellung bzw. zu modellierende Fragestellungen haben oder haben können. Trotz seines Namens ist das Einflussdiagramm nicht per se eine kausale Darstellung.

Das Modellkonzept ergibt sich aus dem Einflussdiagramm und stellt das beabsichtigte Design in deutlich größerer Tiefe dar. Da auch die differenziertesten Modelle Vereinfachungen der Realität mit erforderlichen Annahmen und Einschränkungen in Bezug auf eingeschlossene Inhalte darstellen [6,201], kann das Modell nur dann richtig verstanden werden, wenn das Modellkonzept präzisiert und nachvollziehbar dokumentiert wird.

4.2.4 Datengrundlage

Für ein Modellkonzept können publizierte Modelle als Grundlage verwendet werden. Diese werden im Rahmen einer fokussierten Informationsbeschaffung zu gesundheitsökonomischen Evaluationen identifiziert. Zur Erstellung des Einflussdiagramms erfolgt eine orientierende Suche unter anderem nach deutschen Leitlinien im untersuchten Indikationsgebiet (siehe Abschnitt 4.1.7). Ergänzend können Ergebnisse von Expertenbefragungen oder eine Analyse von GKV-Routinedaten einfließen.

4.2.5 Auswahl der Modellierungstechnik

Die Wahl der geeigneten Modellierungstechnik hängt von der Fragestellung, den Charakteristika der zu bewertenden Intervention, der entsprechenden Erkrankung und den Rahmenbedingungen ab. Bei der Auswahl der Modellierungstechnik ist für das Institut der Grundgedanke leitend, dass das ökonomische Modell so differenziert und komplex sein sollte, wie es für die adäquate Beantwortung der gestellten Fragestellung(en) erforderlich ist. Die Datenlage an sich sollte nicht die Wahl der Modellierungstechnik bestimmen. Sollte durch die Wahl der Modellierungstechnik eine Modifikation des Modellkonzepts notwendig sein, sollte die Wahl der Modellierungstechnik nochmals überprüft werden [113].

Die gewählte Modellierungstechnik muss auch mit bereits durchgeführten / veröffentlichten Modellierungen zu gleichen bzw. eng verwandten Entscheidungsproblemen verglichen werden. Wenn von diesen schon vorliegenden Modellen abgewichen wird, sollte dies diskutiert und begründet werden. Da die angemessene Modellierungstechnik jedoch immer von der zugrunde liegenden Fragestellung abhängt, sind feststehende A-priori-Vorgaben nicht sinnvoll, zumal sich der internationale Standard der Gesundheitsökonomie kontinuierlich weiterentwickelt [397]. Die folgenden zentralen Problemkreise sind grundsätzlich zu berücksichtigen:

- Zeitliche Dimension: Für welchen Zeithorizont werden Aussagen getroffen, wird extrapoliert und wie wird Zeit innerhalb des Modells strukturiert (z. B. kontinuierlich / diskret, Zykluslänge)?
- Analyseeinheit: Welche Analyse- bzw. Beobachtungseinheit wird abgebildet (z. B. Individuen, Kohorten) und welche Merkmale werden berücksichtigt (Alter, Geschlecht etc.)?
- Interaktion: Welche Interaktion zwischen den Analyseeinheiten selbst (d. h. den Patientinnen und Patienten) oder anderen Elementen des Modells wird abgebildet?

Da für eine Modellierung oft Daten aus unterschiedlichen Quellen zusammengeführt werden, kann es notwendig sein, diese Daten in das gleiche Format zu transformieren, z. B. in denselben Zeitbezug.

4.2.6 Modelldokumentation und Modellvalidierung

A) Modellvalidierung und strukturelle Unsicherheit

Ein Simulationsmodell, das für eine Fragestellung valide ist, kann für eine andere nicht valide sein [463]. Der externe Validierungsprozess muss daher jede intendierte Verwendung des Modells abdecken und bei einem Einsatz für andere Fragestellungen muss die Validierung erneut erfolgen. Über das Vorgehen zur Validierung eines Modells herrscht Uneinigkeit, aber es gibt einige grundlegende Schritte, die eingehalten werden müssen [628].

Ein Schlüsselement der Validierung befasst sich mit der Frage, ob das Modell inhaltlich die Realität von Krankheitsverlauf und Behandlung hinreichend abbildet. Die Plausibilitätsprüfung (Augenscheinvalidität) bezieht sich auf das Einflussdiagramm, das Modellkonzept, die Datenbeschaffung, die Erarbeitung der funktionalen Beziehungen und die Auswahl der Modellierungstechnik.

Ein weiteres Schlüsselement der Validierung ist die korrekte technische Umsetzung des Modells (interne bzw. technische Validierung). Dieser Aspekt bezieht sich auf die Frage, ob die technische Implementation tatsächlich das Modellkonzept korrekt umsetzt, z. B. ob die Ergebnisse numerisch korrekt und stabil sind.

Ein drittes Element der Validität ist die Vorhersagevalidität: Inwieweit sagt das Modell die Zukunft voraus, d. h., finden die vorhergesagten Ergebnisse eine Entsprechung in der Realität? Diese ist sicherlich die wünschenswerteste Form der Validität, jedoch die am schwierigsten belegbare, wenn dies überhaupt möglich ist [732]. Jedoch ist ein Abgleich von Modellergebnissen mit vergangenen, vergleichbaren Beobachtungen sinnvoll und Unterschiede sollten erklärbar sein. Dies gilt auch im Vergleich zu anderen gesundheitsökonomischen Modellen (Cross Validity).

Eine spezielle Form der Unsicherheit bei der Modellerstellung bzw. -validierung ist die sogenannte strukturelle Unsicherheit. Bei der strukturellen Unsicherheit wird hinterfragt,

inwieweit die funktionalen Zusammenhänge, die dem Modell zugrunde liegen, tatsächlich valide sind und ob andere funktionale Formen nicht zutreffender wären. Falls bei der Planung und Erstellung eines Modells augenscheinlich wird, dass die strukturelle Unsicherheit bei der zugrunde liegenden Fragestellung von Relevanz ist, kann es notwendig sein, mehrere (alternative) Modelle zu erstellen, um die Folgen dieser Form der Unsicherheit für das Ergebnis zu quantifizieren [700].

B) Allgemeine Dokumentation

Es ist ein detaillierter technischer Bericht mit der Beschreibung aller Modellierungsschritte von der Entwicklung des Einflussdiagramms bis zur finalen Validierung zu erstellen. Zusätzlich ist eine voll ausführbare Version des Modells einschließlich eines Anwendermanuals zur Verfügung zu stellen. In Anlehnung an die Empfehlungen aus Leitlinien [140,536,747] sollte die Dokumentation des Modells Folgendes beinhalten:

- das zur Anleitung der Modellentwicklung verwendete Einflussdiagramm
- Details zum Modellkonzept
 - Beschreibung der Zielpopulation(en) im Rahmen der Evaluation einschließlich Subgruppen
 - Beschreibung der evaluierten Interventionen
 - Auswahl und Begründung der Modellsettings (Simulationsgröße, Zeithorizont, Diskontierungsraten etc.)
 - Überblick zu aktuellen gesundheitsökonomischen Evaluationen im untersuchten Indikationsgebiet
- Beschreibung aller Datenquellen. Eine Begründung der Auswahl von Datenquellen muss gegeben werden.
- Details zu allen im Modell eingesetzten funktionalen Beziehungen. Wenn diese speziell für das Modell entwickelt wurden, müssen detaillierte Angaben zu den angewendeten Methoden gegeben werden.
- Auflistung aller Annahmen hinsichtlich der Datenquellen und der Modellstruktur. Besonders wichtig ist eine detaillierte Darstellung jeglicher Annahme und verwendeten Technik zur Projektion über den Zeitraum hinaus, für welchen die Daten gelten.
- Begründung der verwendeten Modellierungstechnik
 - Beschreibung, in welcher Hinsicht die Technik den erforderlichen Eigenschaften entspricht
- Überblick über die eingesetzten Validierungstechniken und ihre Ergebnisse
- Detaillierte Ergebnisdarstellung einschließlich einer Bewertung des Einflusses einer:
 - Anwendung der Intervention in relevanten Subgruppen

- Unsicherheit der Inputdaten (siehe Abschnitt 4.7 zu Sensitivitätsanalysen)
- Interpretation der Ergebnisse einschließlich einer Beschreibung der Limitationen des angewendeten Ansatzes

C) Technische Dokumentation und elektronische Version des Modells

Die technische Dokumentation ist von zentraler Bedeutung für das Verständnis und die Bewertung des zugrunde liegenden gesundheitsökonomischen Modells. Alle verwendeten Variablen sind zu benennen und zu definieren. Die funktionalen / mathematischen Beziehungen der Modellkomponenten sind darzustellen und gegebenenfalls zu begründen. Die formal-mathematischen Beziehungen sollen alle in das Modell einfließenden Inputvariablen (z. B. Gesundheitszustände) mit den entsprechenden Operatoren (z. B. altersspezifischen Übergangswahrscheinlichkeiten) verbinden. Auch muss weiterhin die Herleitung von Zwischen- bzw. Endwerten dargestellt werden.

Alle kalkulatorischen Schritte innerhalb der Software sind nachvollziehbar zu dokumentieren. Dies kann in der Regel durch die Dokumentation des Programmcodes erfolgen, mit dem die elektronische Version des Modells implementiert ist. Bei Tabellenkalkulationsprogrammen (z. B. Excel) ist die Reihenfolge der Kalkulationsschritte nicht direkt der elektronischen Version zu entnehmen. Diese müssen dann ggf. schriftlich so dokumentiert werden, dass die Reihenfolge der Berechnungsschritte ersichtlich ist.

Eine elektronische Version des Modells muss zur Verfügung gestellt werden unter der Voraussetzung, dass das Modell öffentlich zugänglich gemacht wird und ggf. für zukünftige Evaluationen adaptiert werden kann. Die elektronische Modellversion muss vollständig zugänglich sein und den Reviewerinnen und Reviewern sowie der Öffentlichkeit ermöglichen, alle in der Analyse eingesetzten Formeln und Beziehungen einzusehen und das Modell mit unterschiedlichen Inputdaten laufen zu lassen. Um das Review des Modells zu erleichtern, sollte der elektronischen Version eine Anwenderanleitung beigelegt werden, die beschreibt, welche Software und Hardware benötigt werden, wie Modellinputs verändert werden können, wo diese Inputs im Modell gefunden werden können, wie das Modell ausgeführt werden kann und wie Ergebnisse extrahiert werden können.

4.3 Nutzen

Die Methoden, mit denen der Nutzen im Rahmen der Nutzenbewertung bestimmt wird, werden in Kapitel 3 beschrieben. In einer KNB kann der studienbelegte Nutzen vom modellierten Nutzen unterschieden werden, wenn einer KNB ein längerer als der studienbelegte Zeithorizont zugrunde gelegt wird.

4.3.1 Übertragung und Darstellung des Nutzens

Damit der Nutzen in die KNB mithilfe der Effizienzgrenze einfließen kann, muss er approximativ kardinalskaliert sein. In der KNB kann der approximativ kardinalskalierte Nutzen, der ggf. direkt aus den Studienergebnissen ableitbar ist, oder ein transformierter approximativ

kardinalskalierten Nutzen auf der vertikalen Achse abgetragen werden. Die Einschränkung der Bedingung, dass ein Nutzen „nur“ approximativ kardinalskaliert sein muss, ergibt sich aus folgender Überlegung: Eine in der Nutzenerhebung verwendete Skala muss nicht über ihre gesamte Spannbreite kardinalskaliert sein. Es genügt, wenn diese über den Bereich, der für die Definition des patientenrelevanten Zusatznutzens von Bedeutung ist, das Kriterium erfüllt, kardinalskaliert zu sein. Oft zeigen beispielsweise unterschiedliche Erhebungsinstrumente sogenannte Boden-(Floor-) oder Decken-(Ceiling-)Effekte an den Rändern ihrer Wertebereiche, sind jedoch im restlichen Wertebereich kardinalskaliert [76,240,590].

Hier wird keine spezielle Vorgehensweise zur Wertbestimmung des Nutzens auf einer Kardinalskala empfohlen, da jedes Indikationsgebiet unterschiedliche Möglichkeiten zur Bewertung des Nutzens bieten kann, die der Anforderung der Kardinalskalierung genügen.

4.3.2 Endpunkte

Der Nutzen kann auf der vertikalen Achse der Effizienzgrenze durch einzelne oder aggregierte patientenrelevante Endpunkte dargestellt werden (zur Definition des patientenrelevanten medizinischen Nutzens bzw. Schadens siehe Abschnitt 3.1.1). Wenn mehrere patientenrelevante Endpunkte nebeneinander dargestellt werden, wird für jeden patientenrelevanten Endpunkt eine eigene Effizienzgrenze erstellt. Andernfalls wird der Nutzen aggregiert zu einem einzigen Maß des Gesamtnutzens, das anschließend in einer Effizienzgrenze abgetragen wird. Ein Maß des Gesamtnutzens ist in einer ganz allgemeinen Definition eine Aggregation der Bewertung des Nutzens und Schadens in einer Größe, wobei unterschiedliche patientenrelevante Endpunkte zu einem einzigen Maß zusammengefasst werden. Es kann sowohl in der Nutzenbewertung als auch in der Kosten-Nutzen-Bewertung berücksichtigt werden. Die in diesem Kapitel dargestellten Anforderungen an die Erhebung eines Maßes des Gesamtnutzens gelten auch, wenn es im Rahmen der Nutzenbewertung eingesetzt wird.

4.3.3 Maß des Gesamtnutzens

International gibt es unterschiedliche Maße, den Gesamtnutzen auszudrücken bzw. zu erfassen. Dazu gehören das qualitätsadjustierte Lebensjahr (QALY = Quality-adjusted Life Year) und das behinderungsbereinigte Lebensjahr (DALY = Disability-adjusted Life Year). Maße wie das Saved Young Life Equivalent (SAVE) [542] oder das Healthy Years Equivalent (HYE) [269] sind mit dem Anspruch angetreten, Schwächen des QALY zu korrigieren, das am weitesten verbreitet ist.

In diesem Kontext werden je nach methodischem Ansatz oder ökonomischer Theorie in der wissenschaftlichen Literatur die Begriffe Präferenzen (Preferences), Nutzwerte bzw. Teilnutzenwerte (Utilities) und Werte (Values) verwendet [201]. Auf die weitere Debatte der Begriffe und des Stellenwerts der Erhebungsmethoden im Zusammenhang mit der Frage eines sogenannten Welfarist- versus Extra-Welfarist-Rahmens sei hier nur hingewiesen [89]. In Anlehnung an das SGB V wird im Weiteren von Gewichten gesprochen, mit deren Hilfe einzelne patientenrelevante Endpunkte in ein Maß des Gesamtnutzens überführt werden können.

Wenn der G-BA für eine KNB nach § 35b Abs. 1 Satz 2 SGB V das Maß des Gesamtnutzens festlegt (siehe Abschnitt 4.9), werden ein entsprechendes Instrument und ggf. die dafür festgelegten Erhebungsmethoden oder eine schon festgelegte Gewichtung von Endpunkten nach den Vorgaben im Auftrag eingesetzt. Die Ergebnisse sollten dem Entscheidungsträger zusammen mit der Endpunktgewichtung zur Verfügung gestellt werden. Hierdurch ergibt sich für den Entscheidungsträger die Option, einen aus mehreren zusatznutzenbasierten Erstattungspreisen gewichteten Erstattungsbetrag zu verhandeln.

A) Das QALY als Maß des Gesamtnutzens

Zur Berechnung des QALY werden Gewichte für Gesundheitszustände erhoben. Befragte wägen dabei ab, wie sie diese Gesundheitszustände empfinden bzw. einschätzen. Das Ergebnis ist dann eine Indexzahl für jeden Gesundheitszustand. Unter einer Integration der Dauer der entsprechenden Gesundheitszustände können diese zumeist als Nutzwerte (Utilities) bezeichneten Gewichte in QALYs überführt werden. Die Erhebung und Berechnung von Nutzwerten wird z. B. dargestellt in Puhan et al. [579], Lipscomb et al. [478] und Tierney et al. [713].

Das Institut schließt nicht aus, in Kosten-Nutzen-Bewertungen auf QALYs als Maß für den Gesamtnutzen zurückzugreifen. QALYs sollten nur verwendet werden, wenn die einfließenden Nutzwerte zu den Gesundheitszuständen bei den Betroffenen erhoben worden sind, die aktuell oder in der Vergangenheit diese Gesundheitszustände erlebt haben. Die Daten zu den vorliegenden Gesundheitszuständen, für die der Nutzwert bestimmt wird, sollten bei den an klinischen Studien Teilnehmenden erhoben worden sein. Wenn generische Indexinstrumente eingesetzt wurden, muss ein in Deutschland validierter Tarif bei der Ermittlung des Nutzwertes angewendet werden. Die Verwendung von QALYs sowie ihre Erhebung und Umrechnung in einen deutschen Tarif müssen in jedem Fall nachvollziehbar dargestellt und begründet werden. Im Übrigen gelten alle üblichen Standards, die an entsprechende Verfahren und Instrumente angelegt werden: D. h., Nachweise der Objektivität, Reliabilität, Validität und Reagibilität / Änderungssensitivität müssen vorliegen. Parallel zum Einsatz eines generischen Instruments sollten krankheitsspezifische Instrumente zur Erhebung der Lebensqualität in den klinischen Studien eingesetzt werden. Daher wird auch vom Mapping krankheitsspezifischer auf generische Instrumente abgeraten.

Angesichts der anhaltenden Diskussion um die Vor- und Nachteile verschiedener Instrumente, insbesondere der Multi-Attribute-Utility-Instrumente (MAUI), mit denen die Lebensqualität, das subjektive Wohlbefinden oder Nutzwerte erhoben bzw. abgebildet werden sollen / können, muss man sagen, dass keine allgemeine Empfehlung abgegeben werden kann. Die Wahl eines Instruments hängt davon ab, welches dieser 3 Konzepte im Vordergrund stehen soll und welche Dimensionen von Lebensqualität vorzugsweise erhoben werden sollen [588].

Die wissenschaftliche Debatte um die ethischen und methodischen Probleme des QALY-Konzepts an sich und deren Lösung oder einer daran geknüpften Zahlungsbereitschaftsschwelle in einer KNB sowie des Einsatzes des QALY zur reinen Abwägung des Nutzens und Schadens

soll hier nicht erneut aufgerollt werden. In diesem Zusammenhang sei auf eine Reihe von Publikationen verwiesen [173,193,194,327,478,493,513,543,737].

B) Erhebung von Präferenzen zur Erstellung eines Maßes des Gesamtnutzens

Neben den oben genannten krankheitsübergreifenden Maßen können auch Verfahren der multikriteriellen Entscheidungsfindung oder der Präferenzenerhebung eingesetzt werden, wenn ein Maß des Gesamtnutzens zum Vergleich von Interventionen erhoben werden soll. Für Endpunkte, die mittels dieser Verfahren gewichtet werden, gelten alle Vorgaben nach dem SGB V und der AM-NutzenV. Surrogate können nur bei nachgewiesener Validität herangezogen werden. Im Bereich der Gesundheitsversorgung setzen sich als Methoden der multikriteriellen Entscheidungsfindung oder der Präferenzenerhebung weitgehend das Analytic-Hierarchy-Process(AHP)-Verfahren und die Conjoint-Analyse (CA) durch [80,137,369,500,607]. Das Institut kann daher indikationsbezogen auf diese Verfahren zur Generierung eines Maßes des Gesamtnutzens zurückgreifen. Allerdings gibt es noch ungelöste methodische Probleme beim Einsatz dieser Verfahren, sodass gegenwärtig eine routinemäßige Anwendung dieser Methoden nicht vorgesehen ist.

Für das AHP-Verfahren [191,192] wird ein Entscheidungsproblem in sogenannte Kriterien zerlegt. Diese werden dann in eine Hierarchie gebracht. So kann z. B. ein Arzneimittel nach den Kriterien Mortalität, Morbidität und Lebensqualität beurteilt werden. Die Kriterien können in weitere Subkriterien zerlegt werden, die Endpunkten entsprechen können [369]. Die am AHP-Verfahren 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 [609,610,612] können über den sogenannten rechten Eigenvektor die Gewichte für die Kriterien bzw. Subkriterien ermittelt werden, die sich zu 1 aufsummieren müssen. Eine methodische Weiterentwicklung, die auch zulässt, dass Kriterien voneinander abhängig sind, ist das Analytic-Network-Process(ANP)-Verfahren [608,611].

Die CA gehört zur Gruppe der Stated-Preference-Techniken [80]. Eine Entscheidung wird in sogenannte Attribute zerlegt, die Endpunkten entsprechen können. Für jedes Attribut werden Ausprägungen (Level) angegeben. Die Wahlalternativen (Stimuli) werden für ein Discrete-Choice-Experiment (DCE = wahlbasierte CA) aus den Attributen mit unterschiedlichen Ausprägungen zusammengesetzt. Die Befragten werden dann mit einem Set von (theoretischen) Szenarien konfrontiert (Wahlszenario = Choice Set), die aus mindestens 2 Wahlalternativen (Stimuli) bestehen. Auf Basis der Wahl der Szenarien werden dann in einem Regressionsmodell Koeffizienten für die Ausprägungen der Attribute ermittelt. Der Einfluss der Attribute auf die Entscheidung kann dargestellt werden, indem im Anschluss daran Gewichte für die Attribute gebildet werden. Diese Gewichte können wiederum auf 1 normiert werden.

Ausgehend von ihrer Entwicklung wurde das AHP-Verfahren ausgerichtet auf die Entscheidungsfindung bei widerstrebenden Zielen in Gremien, z. B. dem Management eines Unternehmens, und die CA auf die Erhebung von Präferenzen, um Kaufentscheidungen

voraussehen und Produkte anpassen zu können. Mittlerweile spielen beide Verfahren eine Rolle bei der Identifikation und Priorisierung patientenrelevanter Endpunkte, z. B. vor der Planung einer Studie, und der Bestimmung des Nettonutzens (Maß des Gesamtnutzens) von Interventionen [156,526].

Eine klare Zuordnung, wann welches Verfahren bevorzugt werden soll, lässt sich somit kaum ableiten. Ein AHP-Verfahren wäre wohl eher angebracht, wenn in einer geschlossenen Gruppe eine Entscheidung gefällt werden soll [370], wohingegen man eine CA durchführen würde, wenn man auch die Kompensationsleistung für entgangenen Nutzen berücksichtigen wollte, wenn eine Intervention nicht erstattet wird. Nebenbei bemerkt gibt es auch die Möglichkeit, mittels einer CA QALYs zu berechnen [256,303]. Für die Auswahl eines der beiden Verfahren sind aber folgende Kriterien heranzuziehen: Für die CA gilt, dass maximal 6 bis 7 Attribute einbezogen werden können. Beim AHP-Verfahren gibt es diese Grenze nicht. Weiterhin scheint das AHP-Verfahren die Befragten vor geringere kognitive Anforderungen zu stellen, was man je nach Indikation berücksichtigen könnte. Diese Einschätzungen können sich derzeit nur teilweise auf empirische Daten stützen, sodass eine evidenzgesteuerte Auswahl eines der beiden Verfahren derzeit nicht möglich ist. Außerdem besteht zu einigen Fragen wie der Reliabilität beider Verfahren Forschungsbedarf.

Die Stärken und Schwächen beider Methoden können hier nicht im Detail beschrieben werden [538]. Nachvollziehbarkeit in Bezug auf die Planung, Durchführung, Auswertung und Bewertung einer jeden Umsetzung ist daher zentral. Für die CA gibt es eine grundlegende Liste von Kriterien, um die hohe Qualität, Transparenz und Verlässlichkeit des Ergebnisses einer CA zu gewährleisten [79]; viele der Anforderungen gelten auch für die Durchführung eines AHP-Verfahrens.

Im Detail sollten bei der Planung, Durchführung, Auswertung und Bewertung der Ergebnisse von Erhebungen mithilfe eines der beiden Verfahren folgende Anforderungen erfüllt werden:

- Vollständigkeit der Kriterien oder Attribute,
- umfassende Dokumentation des Vorgehens bei der Auswahl der zu Befragenden und Darstellung, inwieweit sie aufgrund soziodemografischer und krankheitsspezifischer Faktoren repräsentativ für das Kollektiv von Betroffenen sind.

Nicht nur, wer befragt wird, sondern auch die Rekrutierungswege müssen berichtet werden. Weiterhin gilt es, eine Fallzahl zu planen. Für die CA gibt es Faustformeln für eine Fallzahlschätzung [418]. Für das AHP-Verfahren gibt es derzeit keine Methode, um eine Fallzahl zu schätzen. Für ein AHP-Verfahren können aber zumindest Kriterien der Repräsentativität herangezogen werden, die auch anderen Befragungen zugrunde gelegt werden (Größe der Stichprobe, Art der Ziehung etc.):

- Untersuchung der befragten Population auf Homogenität

- umfassende Dokumentation der Auswertung samt Übergabe der Rohdaten inklusive der Fragen im Originalwortlaut
- Sprache, Wahl und Steuerung der Umsetzung inklusive einer Einschätzung der Verzerrung durch die Art des Designs
Eine den Befragten angemessene Sprache sollte gewählt werden.
- Untersuchung der Konsistenz und der Unsicherheit der Ergebnisse und Durchführung von entsprechenden Analysen (z. B. Sensitivitätsanalysen)

4.3.4 Datengrundlage

Das Vorgehen bei der Informationsbeschaffung für Daten, die auf der Nutzenseite in eine KNB einfließen, wird in den Abschnitten 8.1 und 8.2 erläutert. Publikationen, aus denen sich Aussagen zum Maß des Gesamtnutzens ergeben, werden über eine fokussierte Informationsbeschaffung (siehe Abschnitt 8.2.3) identifiziert. Ergänzend können Ergebnisse aus Befragungen zur Herleitung von Gewichten und Nutzwerten berücksichtigt werden (siehe Abschnitt 4.1.7).

4.3.5 Unsicherheit und Verteilung von Nutzendaten

Für geschätzte Effekte im Rahmen einer Nutzenbewertung können im Allgemeinen Konfidenzintervalle bzw. Credible Intervals (falls bayessche Methoden gewählt werden, siehe Abschnitte 9.3.2 und 9.3.8) berechnet werden, die die Präzision bzw. Unsicherheit der Punktschätzer anzeigen. Für die weitere Untersuchung der Unsicherheit sind geeignete Annahmen zu treffen, da viele Effekte nicht normalverteilt sind.

Schätzer aus indirekten Vergleichen (siehe Abschnitt 9.3.8) sind mit höherer Unsicherheit behaftet als Schätzer aus direkten Vergleichen; darauf wird bei der Bewertung der Unsicherheit hingewiesen. Für Schätzer aus indirekten Vergleichen, die zum Beispiel aufgrund von unterschiedlichen Annahmen zu A-priori-Verteilungen voneinander abweichen, sind ggf. Szenarioanalysen durchzuführen.

Insbesondere auch beim Gesamtmaß des Nutzens gilt es, die in Abschnitt 4.7 geforderten Untersuchungen der Unsicherheit (Sensitivitätsanalysen) durchzuführen.

4.4 Kosten

4.4.1 Perspektive und zu berücksichtigende Kosten

Auftragsabhängig können die (reine) GKV-Perspektive, die Perspektive der GKV-Versicherungsgemeinschaft, die Sozialversicherungsperspektive bzw. die Perspektive einzelner Sozialversicherungsträger sowie die gesellschaftliche Perspektive berücksichtigt werden. Im Folgenden werden die relevanten zu berücksichtigenden Kosten nach Perspektiven abgegrenzt.

In der (reinen) GKV-Perspektive werden alle direkten erstattungsfähigen Kosten und Transferleistungen (z. B. Krankengeld) berücksichtigt. Weiterhin können, soweit für die KNB

relevant, die Anteile der Beitragszahlungen an die Rentenversicherung, Pflegeversicherung und Arbeitslosenversicherung, die die GKV bei einer Erkrankung nach 6 Wochen Arbeitsunfähigkeit übernehmen muss, und Beitragsausfälle (während der Krankengeldzahlung) berücksichtigt werden.

In der Perspektive der GKV-Versichertengemeinschaft sind zusätzlich zu den direkten erstattungsfähigen Kosten eigene Aufwendungen der Versicherten zu berücksichtigen, die nicht erstattungsfähig sind (siehe Abschnitt 4.4.2). Krankengeld wird hingegen nicht berechnet, da das Geld lediglich von der GKV zu den Versicherten umverteilt wird, sodass keine Mehrkosten für die Versichertengemeinschaft entstehen [606]. Beitragsausfälle zur gesetzlichen Krankenversicherung durch Krankheit werden damit ebenso nicht berücksichtigt.

Tabelle 7: Perspektive und relevante zu berücksichtigende Kosten ¹

Kostenkategorie Perspektive	Direkte medizinische Kosten		Direkte nicht medizinische Kosten		Indirekte Kosten	Transferleistungen
	Erstattungs-fähig	Nicht erstattungs-fähig	Erstattungs-fähig	Nicht erstattungs-fähig	-	-
Gesellschaft	ja	ja	ja	ja	ja	nein
Sozialversicherung	ja	nein	ja	nein	nein	ja
GKV-Versicherten-gemeinschaft	ja	ja	ja	ja	nein	nein
GKV	ja	nein	ja	nein	nein	ja
GKV: gesetzliche Krankenversicherung						

Anders als bei der Perspektive der GKV-Versichertengemeinschaft werden bei der Sozialversicherungsperspektive bzw. der Perspektive einzelner Sozialversicherungsträger keine Zuzahlungen der Versicherten berechnet. Berücksichtigt werden erkrankungsbedingte erstattungsfähige Ausgaben inklusive Transferleistungen.

In der gesellschaftlichen Perspektive werden Kostenkomponenten unabhängig davon berücksichtigt, wer sie trägt und wer von den Effekten einer Intervention betroffen ist. Generell sind die Kosten zu berücksichtigen, die bei allen Sozialversicherungsträgern und sonstigen Betroffenen anfallen (siehe Tabelle 7). Ein Zeitaufwand bei Patientinnen und Patienten und / oder ggf. bei Angehörigen, der einen Arbeitsausfall darstellt, wird nicht nochmals als Zeitaufwand berücksichtigt. Dies würde mit der Berücksichtigung von Produktivitätsverlusten zu Doppelzählungen führen. Ebenso werden Transferzahlungen und durch die GKV finanzierte Beitragszahlungen an die Sozialversicherungen nicht berücksichtigt, weil sie nur umverteilt werden und aus volkswirtschaftlicher Sicht keine Mehrkosten entstehen [606].

¹ Die Inhalte der jeweiligen Kostenkategorie können je nach eingenommener Perspektive unterschiedlich sein. In einer engeren Auslegung der GKV-Versichertengemeinschaft z. B. werden ggf. Zuzahlungen berücksichtigt, aber keine weiteren Ausgaben der Versicherten. Dies wird in den Aufträgen des G-BA konkretisiert.

Generell ist bei der Bestimmung der Kosten in jeder Perspektive zu prüfen, ob diese Kosten und ggf. Einsparungen für die betrachteten Interventionen bzw. Indikationsgebiete und Patientengruppen relevant sind.

4.4.2 Abgrenzung der Kosten

A) Direkte Kosten

Direkte Kosten beziehen sich auf den Ressourcenkonsum bei der aktuellen und zukünftigen Erstellung von Gesundheitsleistungen. Sie werden weiter differenziert in direkte medizinische und direkte nicht medizinische Kosten. Unter direkten medizinischen Kosten ist der Ressourcenverbrauch zu verstehen, der im Gesundheitssektor bei der Gesundheitsversorgung entsteht. Sie schließen Kosten z. B. für Krankenhausaufenthalte, ambulante Arztkontakte, Arzneimittel und Heil- und Hilfsmittel ein. Direkte nicht medizinische Kosten umfassen die Ressourcen, die die Erstellung medizinischer Leistungen im Gesundheitssektor unterstützen, z. B. Fahrtkosten bei medizinischen Interventionen oder den bewerteten Zeitaufwand von Betroffenen und ihren pflegenden Angehörigen, der durch die Erkrankungen entsteht.

Erstattungsfähige Kosten umfassen Ausgaben für Gesundheitsleistungen, die von der GKV bzw. von anderen Sozialversicherungsträgern finanziert werden. Nicht erstattungsfähige medizinische Kosten sind Leistungen, die von den Versicherten direkt getragen werden, wie Zuzahlungen für Arzneimittel, Heil- und Hilfsmittel und ambulante Arztkontakte. Nicht erstattungsfähige nicht medizinische Kosten sind z. B. krankheitsbedingte Nettoeinkommensverluste² (z. B. finanzielle Einbußen durch den Empfang von Krankengeld, das unterhalb des Nettoeinkommens liegt) oder der Zeitaufwand von Betroffenen und Angehörigen.

Die meisten empirischen Studien berücksichtigen die Auswirkungen auf die Freizeit von Betroffenen und Angehörigen nicht. Insofern bezieht das Institut nicht regelhaft den Zeitaufwand für Betroffene und ihre Angehörigen in die gesellschaftliche Perspektive ein. Für den Fall, dass dennoch repräsentative und valide Informationsquellen zum Zeitaufwand vorliegen, kann dieser Zeitaufwand in Sensitivitätsanalysen zur Perspektive der Gesellschaft berücksichtigt werden. Die Lebensqualität von Angehörigen wird im Allgemeinen auf der Nutzenseite nicht berücksichtigt. Sollten deren Freizeitverluste betrachtet werden, sollten sie auch auf der Kostenseite bewertet werden [88,417,552,748].

B) Indirekte Kosten

Indirekte Kosten bezeichnen den Produktivitätsausfall bei Arbeitsunfähigkeit, Erwerbsunfähigkeit (bei langfristiger Erkrankung oder Behinderung) und vorzeitigem Tod.

² Streng genommen bezeichnen krankheitsbedingte Nettoeinkommensverluste die Differenz zwischen dem Nettoeinkommen von Gesunden und dem Nettoeinkommen von Erkrankten unter Berücksichtigung von Zuzahlungen für Gesundheitsleistungen zur Behandlung der Erkrankung. Im Rahmen der Perspektive der GKV-Versichertengemeinschaft werden Zuzahlungen jedoch als nicht erstattungsfähige Kosten berücksichtigt, sodass die Nettoeinkommensverluste aus der Differenz zwischen dem gezahlten Krankengeld und dem Nettoeinkommen einer oder eines Gesunden ermittelt werden können.

Das Institut berücksichtigt Produktivitätsausfälle primär auf der Kostenseite. So empfiehlt es auch überwiegend die Literatur [90,91,119,201,417,645,646]. Produktivitätsverluste wegen vorzeitigen Todes (Mortalitätskosten) sind nicht auf der Kostenseite auszuweisen, wenn die Mortalität bereits auf der Nutzenseite berücksichtigt wird, um Doppelzählungen zu vermeiden. Nur in den Fällen, in denen der betrachtete Endpunkt nicht die Mortalität oder Lebenszeit betrifft, werden die Mortalitätskosten auf der Kostenseite abgebildet. Kosten für die Gesellschaft (Ausfall von Steuern und Sozialversicherungsbeiträgen) werden immer auf der Kostenseite abgebildet [417,645,646].

International wird diskutiert, auch unbezahlte Arbeit (z. B. Hausarbeit) in einer KNB zu berücksichtigen. Das Institut sieht dies im Regelfall vorerst nicht vor.

C) Transferleistungen

Transferleistungen können, soweit für die KNB relevant, berücksichtigt werden. Transferleistungen sind generell nicht zu berücksichtigen, wenn Zahlungen lediglich umverteilt werden und damit für die gewählte Perspektive keine Mehrkosten entstehen.

D) Intangible Kosten

Als intangible Kosten werden nicht direkt als Ressourcenverbrauch berechenbare bzw. in Geldeinheiten bewertbare Erfahrungen wie Schmerz oder Angst aufseiten der Behandelten bezeichnet. Nach internationalen Standards der Gesundheitsökonomie sollten sie auf der Nutzenseite berichtet werden, soweit Daten für diese Angaben vorhanden sind.

E) Zukünftige Kosten

Weiterhin wird in der gesundheitsökonomischen Literatur häufig eine Unterscheidung vorgeschlagen zwischen interventionsassoziierten und nicht interventionsassoziierten (zukünftigen) Kosten. Interventionsassoziierte Kosten sind z. B. die nach einem Herzinfarkt anfallenden Arzneimittel und Nachkontrollen, wohingegen nicht interventionsassoziierte Kosten z. B. die Behandlungskosten eines später auftretenden Karzinoms wären, dessen Behandlung mit der des Herzinfarkts nichts zu tun hat.

Die Berücksichtigung von nicht interventionsassoziierten Kosten wird kontrovers diskutiert [88,201,270,486]. Interventions- und nicht interventionsassoziierte Kosten werden auftragspezifisch gegeneinander abgegrenzt. Ist die Verlängerung des Lebens für die KNB relevant, werden im Basisfall die interventionsassoziierten zukünftigen Kosten berücksichtigt (sowohl bei gegebener Lebenserwartung als auch bei gewonnenen Lebensjahren). Nicht interventionsassoziierte zukünftige Kosten können in separaten Sensitivitätsanalysen erfasst werden (nicht bei gegebener Lebenserwartung, da identisch für alle Strategien, wohl aber bei gewonnenen Lebensjahren).

F) Investitions- und Implementierungskosten

Wenn explizit für die GKV oder die GKV-Versichertengemeinschaft einmalige Kosten zur Finanzierung der Bereitstellung oder Umsetzung entstehen, sind die Investitions- und

Implementierungskosten angemessen zu berücksichtigen. Dies sollte über Sensitivitätsanalysen untersucht werden.

4.4.3 Schritte zur Kostenbestimmung

Grundsätzlich sind die Kosten so präzise wie möglich zu bestimmen. Methoden, verwendete Quellen und Ergebnisse sind für die einzelnen Schritte der Kostenbestimmung zu beschreiben. Die Bestimmung der in das Modell eingehenden Kosten folgt üblicherweise einem 4-stufigen Prozess:

- Identifikation der Ressourcen,
- Mengenerfassung der Ressourcen,
- Bewertung der Ressourcen und
- Berechnung der in das Modell einfließenden Kosten nach Gesundheitszuständen und ggf. Zyklen.

A) Identifikation der Ressourcen

Im Rahmen der Identifizierung der Ressourcen müssen die Gesundheitsleistungen bestimmt werden, die zur Behandlung der Erkrankung angewendet werden (siehe Abschnitt 4.1.5). Die Angaben sollten möglichst aktuell sein und können aus den in Abschnitt 4.4.4 beschriebenen Quellen gezogen werden.

B) Mengenerfassung der Ressourcen

Die Anwendungshäufigkeit, der Anteil der relevanten Patientenpopulation, die die jeweilige Leistung in Anspruch genommen hat, und die Dauer der Inanspruchnahme müssen bestimmt werden. Kosten für Leistungen, die sehr selten erbracht werden und / oder nur einen geringen Einfluss auf die Ergebnisse ausüben, sollten zwar beschrieben werden, werden aber nicht zwangsläufig in die Berechnung einbezogen [201].

Um den Ressourcenkonsum zu quantifizieren, können sowohl der Mikro- als auch der Makro(Gross)-Costing-Ansatz [704,705] verwendet und kombiniert werden. Der Präzisionsgrad der Mengenerfassung wird also unter anderem durch das Vergütungssystem und den entsprechenden Aggregationsgrad der Leistungen bestimmt.

Beide Ansätze können als Bottom-up-Ansatz oder als Top-down-Ansatz [637,704,705] angewendet werden, wenn entweder ausgehend von den einzelnen Patientinnen und Patienten die verbrauchten Ressourcen gemessen werden oder ausgehend von hochaggregierten Daten (Ausgaben für eine Krankheit) eine (durchschnittliche) Aufteilung auf Patientinnen und Patienten vorgenommen wird.

C) Bewertung der Ressourcen

GKV-Versichertenperspektive

Regulierte und verhandelte Preise (also Preise, die sich nicht ausschließlich über Marktmechanismen entwickelt haben) bestimmen im Allgemeinen die Ausgaben und repräsentieren die Opportunitätskosten der GKV-Versichertengemeinschaft. Wie zuvor beschrieben, bestimmt das Vergütungssystem den maximalen Präzisionsgrad bei der Ausgabenbestimmung der erstattungsfähigen Kosten. So repräsentieren aus der Perspektive der GKV-Versichertengemeinschaft Diagnosis-related Groups (DRGs) die bestmögliche Bewertung für den stationären Sektor und der Einheitliche Bewertungsmaßstab (EBM) die bestmögliche Bewertung für den ambulanten Sektor.

Bei der Kostenbestimmung für Arzneimittel ist zwischen dem stationären und ambulanten Bereich zu unterscheiden. Im Bereich der stationären Versorgung sind die Arzneimittel in der Regel Teil der entsprechenden pauschalen Vergütung. Wenn für relevante Arzneimittel Zusatzentgelte (ZEs) vereinbart wurden oder diese über neue Untersuchungs- und Behandlungsmethoden (NUBs) abgerechnet werden können, sind diese Kosten entsprechend zu ermitteln und in der KNB anzusetzen. Im ambulanten Bereich werden zunächst die Apothekenabgabepreise zur Grundlage genommen. Sofern Festbeträge vorhanden sind, müssen diese angegeben werden. Diese werden um den Apotheken- und Herstellerrabatt reduziert. Kassenspezifische Rabatte werden nicht abgebildet. Grundsätzlich werden, dem Prinzip der Effizienz folgend, die günstigsten Vertreterinnen und Vertreter eines Wirkstoffs bzw. einer Wirkstoffklasse ausgewählt. Relevante Preisänderungen im Zeitverlauf müssen berücksichtigt werden.

Nicht erstattungsfähige Kosten sind teilweise reguliert, sodass hier auf eine entsprechende Standardisierung bei der Ressourcenbewertung zurückgegriffen werden kann (z. B. Zuzahlungsregelungen im stationären Bereich und für Arzneimittel). Diese Kosten werden in der GKV-Versichertenperspektive gesondert ausgewiesen.

Besonderheiten in weiteren Perspektiven

In der Sozialversicherungsperspektive sind je nach Versicherungszweig unter Umständen nur aggregierte Daten verfügbar. In diesem Fall sollten die Ressourcen mittels Top-down-Ansatz aus den jeweiligen Statistiken bewertet werden.

Bei der Berechnung der Kosten aus der gesellschaftlichen Perspektive wäre theoretisch zu beachten, dass die gesellschaftlichen Opportunitätskosten sich in der Regel von den administrativen Preisen unterscheiden, weil die administrativen Preise nur die Perspektive des Bezahlenden darstellen. So sind z. B. in den Fallpauschalen nicht die Kosten für den Bau von Krankenhäusern enthalten, die man jeder Fallpauschale aus gesellschaftlicher Perspektive zuschlagen müsste. Das Institut ist sich dieser theoretischen Diskussion bewusst, richtet sich aber nach den internationalen Standards anderer Health-Technology-Assessment(HTA)-Organisationen, die auch in der gesellschaftlichen Perspektive administrative Preise nutzen, da ein anderes Vorgehen – aufgrund fehlender Daten, z. B. zu den tatsächlichen Kosten, die für

den Bau von Krankenhäusern den Fallpauschalen zugeschlagen werden müssten – mit hoher Unsicherheit behaftet wäre. International üblich in gesundheitsökonomischen Evaluationen ist zumeist nur die zusätzliche Betrachtung der indirekten Kosten. Wird der Zeitaufwand von Betroffenen oder Angehörigen in der Kostenbestimmung berücksichtigt, wird dieser mit dem Nettolohn bewertet.

Bewertung von indirekten Kosten

Für Produktivitätsverluste berücksichtigt das Institut im Basisfall den Friktionskostenansatz [300,448], da der Humankapitalansatz auf einigen unrealistischen Annahmen (insbesondere Vollbeschäftigung am Arbeitsmarkt) beruht. In Sensitivitätsanalysen kann diese Schätzung dem Humankapitalansatz gegenübergestellt werden.

In der KNB orientiert sich die Bewertung von indirekten Kosten an den individuellen Arbeitskosten (d. h. am Bruttolohnsatz und an den Lohnnebenkosten – in Deutschland Arbeitgeberbeiträge zur Sozialversicherung) oder den durchschnittlichen Arbeitskosten. Die Berechnung der durchschnittlichen Arbeitskosten pro Arbeitstag basiert auf den gewichteten durchschnittlichen Arbeitskosten von in Vollzeit und Teilzeit erwerbstätigen Personen in Deutschland. Näherungsweise kann hierzu das Arbeitnehmerentgelt in Deutschland pro Jahr dividiert durch die Anzahl der Arbeitnehmer mal 365 herangezogen werden (wobei dann Sonn- und Feiertage bei den Arbeitsunfähigkeitstagen berücksichtigt werden müssen). Die Anwendung auf Selbstständige ist zu diskutieren [295]. Die Friktionskosten werden mit 80 % der Lohnkosten angenommen (analog zu den Niederlanden [448]). Die Friktionsperiode wird, sofern keine aktuellen Daten verfügbar sind, in Anlehnung an die durchschnittliche tatsächliche Besetzungsdauer in Deutschland für das Jahr 2016 mit 82 Tagen angesetzt [376]. Sollte der Humankapitalansatz in einer Sensitivitätsanalyse untersucht werden, werden die zukünftigen Produktivitätsverluste auf Basis des durchschnittlichen Alters der Patientinnen und Patienten bis zum Erreichen der Regelaltersgrenze berechnet.

D) Darstellung der in das Modell einfließenden Kosten nach Zuständen oder Zyklen

Bevor die Kosten in das Modell eingespeist werden können, müssen sie als durchschnittliche Kosten je Patientin oder Patient nach Gesundheitszuständen und je nach Modell auch nach Zyklen vorliegen.

Je nach Indikation, Intervention, Endpunkten und Modell liegt unter Umständen keine direkte Information zu den Kosten der jeweiligen Gesundheitszustände im Modell vor. Dann können die durchschnittlichen Kosten einer Intervention je Patientin oder Patient und Kostenkategorie (Leistungsbereiche und indirekte Kosten) für den Betrachtungszeitraum durch Annahmen aus weiteren Quellen (siehe Abschnitt 4.4.4) auf die verschiedenen Gesundheitszustände und Zyklen des Modells verteilt werden.

Für absorbierende Zustände in einem Markov-Modell kann es notwendig sein, Übergangskosten zu berechnen, die nur einmalig beim Übergang in diesen Gesundheitszustand anfallen.

Dies ist dann zu empfehlen, wenn anzunehmen ist, dass die Kosten in diesem Zustand im ersten Zyklus wesentlich höher ausfallen als in den Folgezyklen.

4.4.4 Datengrundlage

Kosten, die in das Modell einzuspeisen sind, müssen, wie oben beschrieben, für die verschiedenen Gesundheitszustände und ggf. Zyklen eines Modells berechnet werden. Das Vorgehen bei der Datenerhebung und -auswertung, alle Berechnungen und die Ergebnisse sind transparent darzustellen.

Über eine fokussierte Informationsbeschaffung (siehe Abschnitt 8.2.3) werden Publikationen bzw. Analysen zur Identifikation und zur Mengenerfassung der Ressourcen ermittelt.

Wenn aktuelle Analysen nicht der Literatur entnommen werden können, sind bevorzugt eigene Auswertungen durchzuführen. Sekundärdaten in Form von ausgewerteten GKV-Routinedaten auf Basis einer repräsentativen Stichprobe sind hierbei die Datenquelle erster Wahl (siehe Abschnitt 4.1.7).

Ergänzend können Leitlinien oder Ergebnisse aus Expertenbefragungen hinzugezogen werden, wenn Routinedaten nicht hinreichend die Versorgung in allen Zuständen des Modells darstellen (siehe Abschnitt 4.1.7). Expertenbefragungen sind nur dann eine Option, wenn die Daten nicht repräsentativeren Quellen entnommen werden können bzw. wenn Letztere den in den Gesundheitszuständen erforderlichen Detailgrad nicht vollständig abdecken (siehe auch Abschnitt 4.1.7).

Zur Preisbestimmung stützt sich das Institut mithilfe orientierender Recherchen auf die jeweils relevanten regulierten oder verhandelten Preise, z. B. aus der Datenbank der Informationsstelle für Arzneispezialitäten (IFA), dem EBM, dem DRG-Katalog oder aus Statistiken der Rentenversicherung oder des Statistischen Bundesamts.

Eine Übertragbarkeit von Versorgungspfaden und Kostendaten aus anderen Gesundheitssystemen ist aufgrund von Systemunterschieden selten gegeben und nur unter sehr strengen Voraussetzungen möglich [468,659]. Die Übertragbarkeit von Kostendaten aus folgenden Ländern wird nicht grundsätzlich ausgeschlossen, da sie ein dem deutschen ähnliches System von ambulanter und stationärer Versorgung haben: Österreich, Schweiz, Niederlande, Belgien, Frankreich. Eine Verwendung muss aber jeweils begründet und diskutiert werden. Kostendaten aus weiteren Ländern sollten in einer KNB nicht verwendet werden.

4.4.5 Unsicherheit und Verteilung von Kostendaten

Die Unsicherheit in Kostendaten sollte angemessen adressiert werden. Kostendaten sind inhärent stetig, positiv, ohne Obergrenze und im Allgemeinen nicht normalverteilt, sondern für gewöhnlich rechtsschief verteilt [35].

4.4.6 Inflationsbereinigung und Diskontierung

A) Inflationsbereinigung

Wenn Kostendaten aus unterschiedlichen Zeitperioden stammen, muss inflationsbereinigt werden. Als Quelle für die jährliche Inflation soll der Harmonisierte Verbraucherpreisindex (HVPI) des Statistischen Bundesamts verwendet werden [681]. Im Rahmen einer Sensitivitätsanalyse können weitere Preissteigerungsraten für einzelne Bereiche der Versorgung, z. B. Arzneimittel, aus anderen Quellen einbezogen werden.

B) Diskontierung

Wenn Kosten und Nutzen in Perioden anfallen, die länger als 1 Jahr dauern, werden sie nach dem 1. Jahr im Basisfall mit einer identischen konstanten Rate von 3 % auf die aktuelle Periode diskontiert [39,115,167,201,479]. In Sensitivitätsanalysen sollten ebenfalls identische konstante Raten von 0 und 5 % verwendet werden. Abweichungen hiervon müssen begründet werden.

4.5 Epidemiologische Daten

4.5.1 Daten

Für eine KNB sind aktuelle epidemiologische Daten unerlässlich. Daten zur Prävalenz und Inzidenz in Deutschland dienen neben der Abschätzung der Krankheitslast auch dazu, im Rahmen der Ausgaben-Einfluss-Analyse die Veränderungen im Budget der GKV zu quantifizieren. Daher sind Aussagen nötig, ob in den nächsten 5 Jahren Veränderungen in der Inzidenz, Prävalenz oder Mortalität zu erwarten sind. Weiterhin sind Daten zur Mortalität von Bedeutung, um krankheitsbedingtes Versterben und auch die sogenannte Hintergrundmortalität abzubilden.

Eine besondere Rolle für die Modellierung spielen die Basiswahrscheinlichkeiten für Ereignisse. In einem Modell sind für jeden Endpunkt Angaben zu den endpunktbezogenen Ereignishäufigkeiten bzw. -wahrscheinlichkeiten nötig, die als Ausgangswerte in das entscheidungsanalytische Modell eingehen.

4.5.2 Datengrundlage

Öffentliche Datensammlungen von epidemiologischen Daten, beispielsweise vom Robert Koch-Institut, sind aufgrund der hohen methodischen Konsistenz primär heranzuziehen, wenn sie vorhanden und in geeigneter Form (beispielsweise passende Altersklassen) erhältlich sind. Darüber hinaus können epidemiologische Kennzahlen aus Sekundärdaten wie GKV-Routinedaten sowie Registerdaten ermittelt werden (siehe Abschnitt 4.1.7). Die Registerdaten nehmen hierbei eine Sonderstellung ein. Unabhängig von der Einschätzung der Qualität eines Registers haben diese Daten oft nur einen regionalen Bezug. Daher muss die Übertragbarkeit geprüft werden. Ergänzend können Ergebnisse von Expertenbefragungen einfließen.

Sind epidemiologische Daten aus Deutschland nicht vorhanden, findet eine fokussierte Informationsbeschaffung (siehe Abschnitt 8.2.3) statt. Lassen sich darüber wissenschaftliche Publikationen identifizieren, in denen epidemiologische Kennziffern bestimmt wurden, können diese unter Umständen direkt verwendet werden. Die Verwendbarkeit muss im Einzelfall geklärt werden, da die Studien oft methodisch verschiedene Ansätze verwenden. Kohortenstudien bzw. hinreichend große und repräsentative Stichproben sind zu bevorzugen. Die methodische Qualität der zugrunde liegenden Studie kann u. a. anhand der Vorgaben der Guten Epidemiologischen Praxis bewertet werden.

4.5.3 Unsicherheit und Verteilung von epidemiologischen Daten

Die Unsicherheit in epidemiologischen Daten sollte angemessen adressiert werden. Insbesondere die Unsicherheit von Daten zum Basisrisiko und zur Mortalität muss in den Sensitivitätsanalysen ausreichend und den Verteilungen angemessen berücksichtigt werden.

4.6 Ergebnisdarstellung als Effizienzgrenze

Als Basis der ökonomischen Bewertung von Interventionen innerhalb eines Indikationsgebiets wird eine Effizienzgrenze gezeichnet. Sie wird aus den effizientesten Interventionen der verfügbaren Komparatoren erzeugt und kann der Ableitung von Handlungsempfehlungen für die zu prüfende(n) Intervention(en) dienen. Sie kann Informationen zur Verhandlung von Erstattungsbeträgen liefern, ohne auf einen in Deutschland zurzeit nicht konsentierten Schwellenwert für die Zahlungsbereitschaft zu rekurren.

4.6.1 Definition

Die Effizienzgrenze stellt den endpunktbezogenen Nutzen der verfügbaren Interventionen innerhalb eines Indikationsgebiets den Nettokosten dieser Interventionen grafisch gegenüber. Dabei wird, falls erforderlich, der Nutzen in ein approximativ kardinalskaliertes Maß überführt³. Diejenigen Interventionen, die nach dem Nutzen und den Kosten die effizientesten sind, bilden die Effizienzgrenze.

4.6.2 Verfahrensverlauf

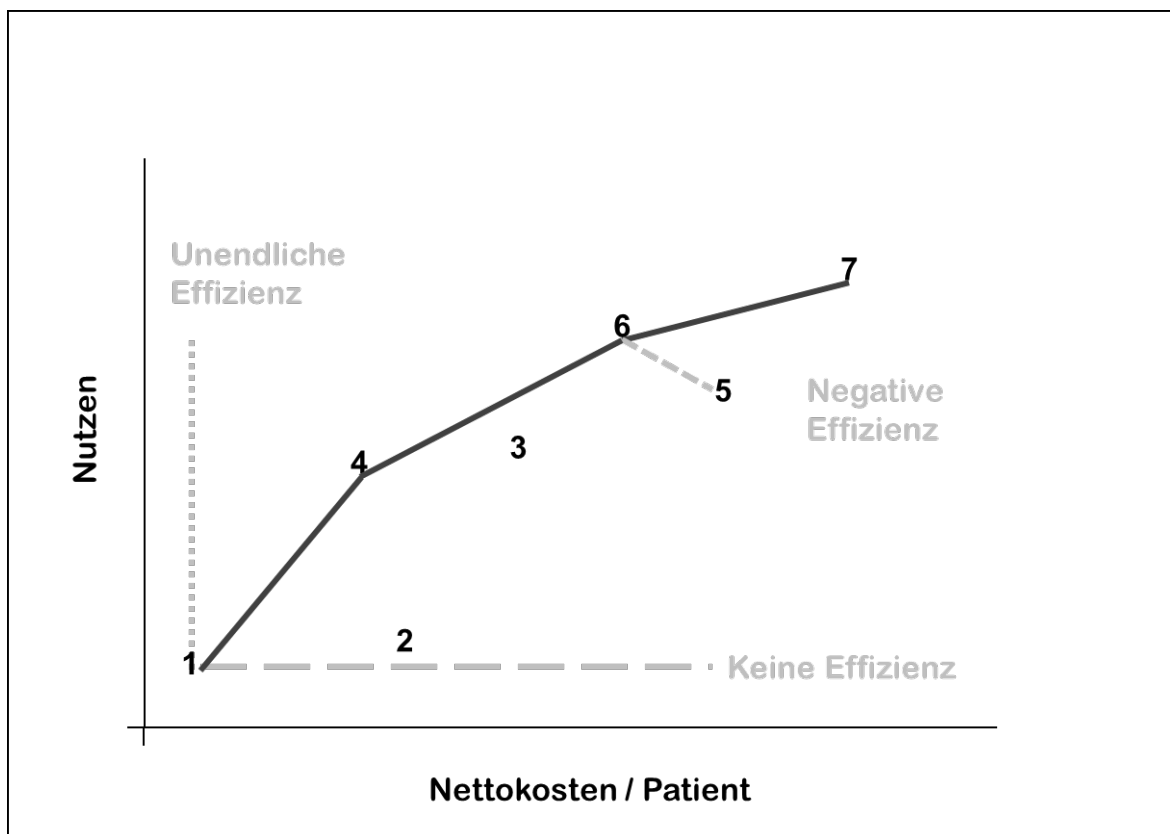
Im Verfahren muss unterschieden werden zwischen der (den) zu bewertenden Intervention(en) und den Interventionen, die die Effizienzgrenze bilden (Komparatoren). Letztere sind Interventionen, die derzeit in Deutschland in dem zur Bewertung anstehenden Indikationsgebiet angewendet und erstattet werden. Deren Kosten und Nutzen werden ermittelt und grafisch dargestellt. Erstere sind die zu prüfenden Interventionen.

In der Darstellung der Effizienzgrenze werden von links nach rechts die Interventionen mit jeweils höherer Effizienz aufgetragen. Die Steigung der theoretischen Verbindungslinie

³ Hat der aus der Nutzenbewertung ermittelte patientenrelevante Zusatznutzen bereits approximativ kardinalskalierte Eigenschaften, kann er direkt in die KNB überführt werden.

zwischen 2 Interventionen (das Liniensegment) gibt den zusätzlichen Nutzen pro zusätzliche Kosten an (siehe Abbildung 11).

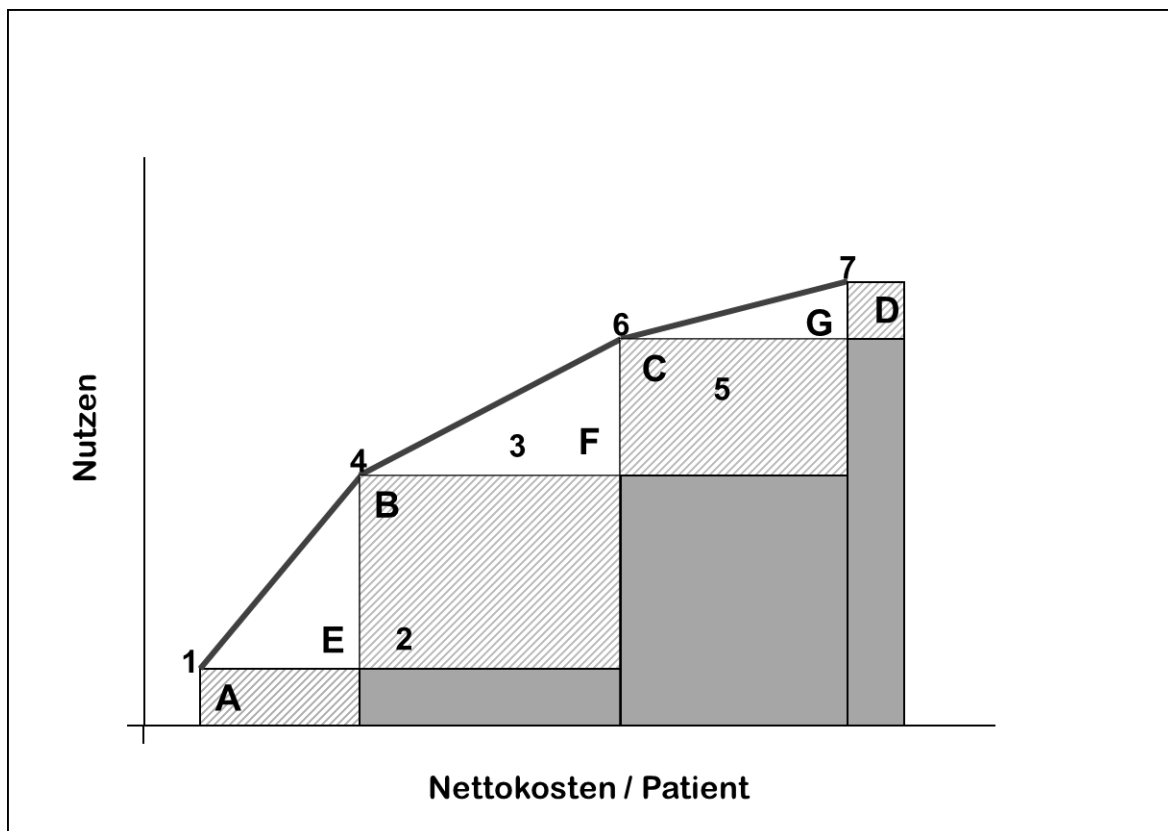
Die Positionen von Interventionen wie beispielsweise Intervention 3 in Abbildung 11 erfordern eine weitergehende Interpretation, weil sie keine negative Effizienz im Vergleich zu bereits eingeführten Interventionen (z. B. Intervention 4) haben. In Abbildung 12 ist die Fläche unter der theoretischen Effizienzgrenze durch eine Reihe von Rechtecken (A bis D) aufgeteilt. Jedes dieser Rechtecke enthält alle Interventionen, die eine negative Effizienz (höhere Kosten bei geringerem Nutzen) gegenüber mindestens einer bereits im Markt befindlichen Intervention auf der theoretischen Effizienzgrenze haben. Interventionen in diesen Teilbereichen (z. B. Intervention 2 bzw. Intervention 5 in Abbildung 12) sind eindeutig ineffizient. Es verbleiben die Dreiecke E, F und G, in welchen die Interventionen nicht eindeutig ineffizient sind. Üblicherweise sind Interventionen, die in diesen Dreiecken eingetragen werden, kein Bestandteil der Effizienzgrenze, weil die theoretische Kombination der beiden Interventionen, die die Hypotenuse des Dreiecks bilden, einen höheren Nutzen bei geringeren Kosten bietet (sogenannte erweiterte Dominanz).



Eine Horizontale (Steigungswinkel = 0°) bedeutet keine Effizienz, während eine Vertikale (Steigungswinkel = 90°) für unendliche Effizienz steht. Eine in aufsteigender Reihenfolge positive Steigung (z. B. zwischen Intervention 6 und 7) stellt einen zusätzlichen Nutzen bei erhöhten Kosten dar, während eine negative Steigung (z. B. zwischen Intervention 6 und 5) einen geringeren Nutzen bei höheren Kosten bedeutet.

Abbildung 11: Interpretation der Steigung der theoretischen Effizienzgrenze

Eine solche Kombination ist in der Praxis nicht immer möglich. Dies würde nämlich voraussetzen, dass bei einem fixen Preis von Intervention 3 die Leistungsempfänger auf Intervention 4 und 6 umverteilt werden müssten, um eine höhere Effizienz zu erreichen. Dies kann klinisch unerwünscht sein und schwierig zu rechtfertigen, weil es zu einer Schlechterstellung derjenigen führen würde, die Intervention 4 erhalten. Die Alternative, einen Wechsel der Leistungsempfänger zwischen beiden Therapien über die Zeit hinweg zu ermöglichen, ist eindeutig bei den meisten chirurgischen und vermutlich auch bei vielen medikamentösen Interventionen nicht möglich. So kann es viele Situationen geben, in welchen Interventionen, die sich in den dreieckigen Flächen befinden, einen Teil der praktischen Effizienzgrenze ausmachen. Wird das Kriterium der erweiterten Dominanz nicht angewendet, resultiert eine stufenförmige absolute Effizienzgrenze, welche sich aus der Verbindung der oberen Segmente der schraffierten Rechtecke gegenüber den dreieckigen Flächen ergibt. Hierbei ist jedoch zu beachten, dass die absolute Effizienzgrenze keine Steigung mehr im Sinne eines Kehrwerts der Zahlungsbereitschaft wiedergibt und somit keine Schwellenwerte zu erheben wären.



Die theoretische Effizienzgrenze (durchgezogene Linie) verbindet diejenigen Interventionen, die in Bezug auf jede andere Intervention bzw. deren Kombinationen effizient sind. Interventionen in den Rechtecken A bis D (z. B. Intervention 2 bzw. Intervention 5) sind eindeutig ineffizient. Intervention 3 befindet sich in einer der verbleibenden dreieckigen Flächen (E bis G) und ist nicht eindeutig ineffizient. Theoretisch würde sich durch die Kombination von Intervention 4 und Intervention 6 eine erweiterte Dominanz ergeben, jedoch kann sich diese in der Praxis als nicht durchführbar erweisen.

Abbildung 12: Absolute versus erweiterte Dominanz

4.6.3 Konstruktion der Effizienzgrenze

Die Effizienzgrenze wird so konstruiert, dass sie die relevanten Interventionen in einem vorgegebenen Indikationsgebiet darstellt. Dazu gehört:

- Vollständige und detaillierte Festlegung des Indikationsgebiets, das von Interesse ist. Dies kann die genaue Erkrankung, die Behandlungsgegebenheiten (z. B. stationäre Versorgung), die Zielpopulation, die Therapiesequenz (1., 2. Therapiewahl etc.) sowie die Angabe über eine Mono- oder Kombinationstherapie beinhalten.
- Positionierung der vorhandenen Therapien anhand ihres Nutzens und ihrer Kosten
- Eintragen der Interventionen in ein Koordinatensystem mit dem Nutzen auf der vertikalen (y-)Achse und den Kosten auf der horizontalen (x-)Achse.⁴ Hierbei ist gemäß guter wissenschaftlicher Praxis auf eine (zumindest je Endpunkt) gleichbleibende Skalierung der Achsen zu achten.
- Auftragen der Effizienzgrenze

Bei der Evaluation neuer Interventionen werden anschließend zusätzlich deren gesundheitliche Effekte und Kosten in dem betreffenden Indikationsgebiet dargestellt.

A) Vertikale Achse

- Auf der vertikalen Achse werden der Nutzen und der Schaden abgetragen. Hierbei ist auf einen positiven Wertebereich zu achten, sodass die Effizienzgrenze den vergrößerten Nutzen bzw. verringerten Schaden abbildet (ggf. ist z. B. eine Multiplikation mit -1 erforderlich oder die Umrechnung auf das Komplementärereignis $1 - \text{Schaden}$).
- Der Nutzen bzw. Schaden wird anhand von patientenrelevanten Endpunkten dargestellt, die geeignet operationalisiert werden müssen (z. B. Lebensqualitätsscores).
- Der Nutzen bzw. Schaden wird auf die vertikale Achse übertragen. Diese Übertragung kann unter Einbeziehung von Modellierungen erfolgen.

B) Horizontale Achse

- Auf der horizontalen Achse werden die gesamten Nettokosten pro Patientin oder Patient eingetragen.
- Die Kosten werden regelhaft aus der Perspektive der GKV-Versichertengemeinschaft berechnet und können auftragsabhängig zusätzliche Kosten aus ggf. erweiterten Perspektiven (z. B. Sozialversicherungsperspektive, gesellschaftliche Perspektive) beinhalten.
- Als Kosten werden die gegenwärtig zu erwartenden Kosten verwendet.

⁴ Eine tabellarische Darstellung ist auch möglich. Die Zusammenhänge werden aber dort nicht so anschaulich.

Um die Kosten für jede Intervention abzuschätzen und im Koordinatensystem zur Effizienzgrenze einzutragen, müssen mehrere Voraussetzungen eingehalten werden. Die Kosten sollten demjenigen Betrag entsprechen, der in der Praxis anfallen würde. Auf der Effizienzgrenze müssen die Gesamtnettokosten pro Patientin oder Patient abgetragen werden.

Um das Kosten-Nutzen-Verhältnis von (neuen) Interventionen mit mehr Nutzen und mehr Kosten als die Komparatoren zu bestimmen, wird das letzte Segment der Effizienzgrenze verlängert (siehe Abschnitt 4.1.9 sowie Abbildung 10 und Abbildung 13).

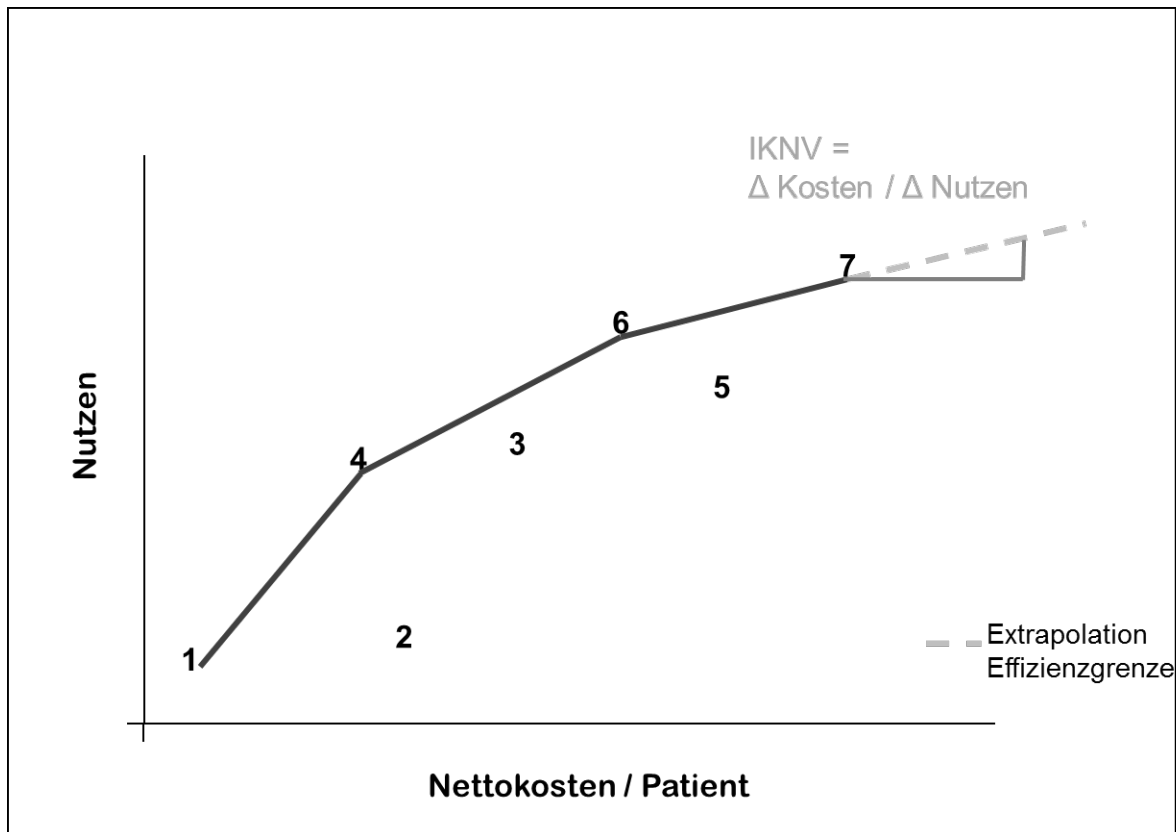
In Abhängigkeit von der Anzahl der Endpunkte, die aus der vorab durchgeführten Nutzenbewertung übernommen werden, können mehrere Effizienzgrenzen hergeleitet und präsentiert werden.⁵ Sofern eine Endpunktgewichtung vorgenommen wurde, wird diese ebenfalls vorgelegt. Wenn ein Maß des Gesamtnutzens festgelegt wurde, ist dieses als primäres Ergebnis anzusehen.

C) Definition des Ursprungs des Koordinatensystems

Der Punkt keine Intervention (d. h. der natürliche Verlauf) erfordert ebenfalls eine Bewertung. Obwohl er möglicherweise als Koordinatenursprung angesehen werden könnte (null Nutzen, null Kosten), ist dies selten angemessen, da auch eine nicht durchgeführte Intervention noch Kosten und Gesundheitseffekte verursachen kann, beispielsweise infolge der nicht behandelten Erkrankung, der Überwachung usw. Daher sollten auch Daten für den natürlichen Verlauf erhoben werden. Eine gängige Annahme ist hierbei, dass die Intervention mit Placebo am ehesten dem natürlichen Verlauf entspricht. Dies ist auftragsbezogen zu prüfen.

Wenn der Ursprung der Effizienzgrenze nicht dem Nullpunkt entspricht, müssen die Effizienzgrenzen (zumindest je Endpunkt) in gleichskalierte Koordinatensysteme eingetragen werden. Die am weitesten unten und links liegende Intervention wird im Allgemeinen zum Ursprung der Effizienzgrenze (siehe Abbildung 13). Eine Verschiebung des Nullpunkts (des Koordinatensystems) ist aus Gründen der Vergleichbarkeit der Darstellung verschiedener Effizienzgrenzen abzulehnen.

⁵ Dies bezieht sich auch auf die gesonderte Darstellung von divergierenden Schadenauspekten in Abgrenzung zum patientenrelevanten Zusatznutzen.



Die Effizienzgrenze beginnt in einem vom Nullpunkt des Koordinatensystems verschiedenen Ursprung. Die Verlängerung zeigt das inkrementelle Kosten-Nutzen-Verhältnis (IKNV), an dem eine (neue) Intervention mit mehr Nutzen und mehr Kosten als die Komparatoren gemessen wird.

Abbildung 13: Darstellung der Effizienzgrenze

4.6.4 Sonderkonstellationen

Es gibt 2 Sonderkonstellationen, in denen sich trotz vollständiger Information nicht unmittelbar eine Empfehlung anhand der Effizienzgrenze für eine neue Intervention ableiten lässt:

- 1) Die letzte Intervention auf der Effizienzgrenze dominiert alle anderen Interventionen und verursacht dieselben Kosten wie das Referenzszenario. Die Steigung wäre damit unendlich (siehe Darstellung in Abbildung 11).
- 2) Die letzte Intervention auf der Effizienzgrenze vor der Einführung der Innovation ist kostengünstiger und hat mehr Nutzen als alle Komparatoren einschließlich des Ursprungs.

In beiden Fällen würde sich ein neuer Ursprung ergeben, auf dem die jeweils letzte Intervention vor Einführung der innovativen Intervention liegen würde.

Hier kann die Ausgaben-Einfluss-Analyse weitere Daten liefern, indem die Auswirkungen auf das Budget dargestellt werden (siehe Abschnitt 4.1.9 und Abschnitt 4.8).

4.7 Unsicherheit (Sensitivitätsanalysen)

Die Arten der Unsicherheit wurden oben (siehe Abschnitt 4.1.8) dargestellt. Die Unsicherheit vieler Modellparameter ergibt sich daraus, dass ihr Wert aus Stichproben geschätzt wird. Diese Art von Unsicherheit wird oft durch Konfidenzintervalle oder andere statistische Ansätze zur Beschreibung von Variabilität erfasst.

4.7.1 Quantifizierung von Unsicherheit

Bei den Kosten kann Unsicherheit bezüglich der Annahmen zum Ressourcenverbrauch bestehen, z. B. zur Dosierung eines Arzneimittels über die Zeit. Auch kann das Modell stochastisch ausgelegt sein (d. h., es verwendet Zufallszahlen bei den Monte-Carlo-Ziehungen). Um diese Art von Unsicherheit einzugrenzen, können verschiedene Techniken angewendet werden [464,595,656].

Auch aus der in Abschnitt 4.2 beschriebenen Form von möglicher Variabilität in der Modellstruktur resultiert Unsicherheit, die bei der Untersuchung berücksichtigt werden muss. Zuletzt können selbst zuvor festgesetzte Inputparameter wie die Diskontierungsrate variiert werden, um die Unsicherheit abzubilden, die sich aus unterschiedlichen Diskontierungsraten ergibt (siehe Abschnitt 4.4.6).

4.7.2 Sensitivitätsanalysen

Parameterunsicherheit sowie andere Arten der Unsicherheit, die nicht reduzierbar sind, werden quantifiziert. Das Institut berücksichtigt sowohl univariate und multivariate deterministische als auch probabilistische Sensitivitätsanalysen und legt der Bearbeitung die Empfehlungen der gemeinsamen Modeling Good Research Practices Task Force Working Group der ISPOR und der SMDM zugrunde [81].

Alle dazu unternommenen Analysen sollten mit den minimalen und maximalen Werten für die eingesetzten Parameterwerte und den zugrunde liegenden Annahmen vollständig dokumentiert werden. Für probabilistische Sensitivitätsanalysen (PSAs) müssen folgende Aspekte präzisiert werden: eingesetzte Wahrscheinlichkeitsverteilungen und ihre Quellen, Korrelationen zwischen Inputparametern und jeglichen strukturellen Varianten.

Strukturelle Sensitivitätsanalysen werden durchgeführt, um den Einfluss einer Variierung der Annahmen in der Modellstruktur, z. B. zur Anzahl oder Art der Modellzustände, zu eruieren.

Darstellung der Ergebnisse der Sensitivitätsanalysen

Für die deterministische Sensitivitätsanalyse sind extreme Ausprägungen der Inputparameter anzugeben, für die die neue Intervention möglicherweise kostensparend ist bzw. oberhalb oder unterhalb der Effizienzgrenze liegt. Für univariate und multivariate Analysen müssen die Ergebnisse tabellarisch und in einem Tornadodiagramm dargestellt werden, in dem für die entsprechenden Intervalle der Inputparameter die Ausprägungen der Ergebnisse als Intervall dargestellt werden.

Für die PSAs wird der Anteil der Simulationen in Prozent angegeben, für die sich eine Kostenersparnis bzw. eine Lage oberhalb oder unterhalb der Effizienzgrenze ergibt. Im Falle von PSAs werden die Ergebnisse als kumulative Kostenverteilungen präsentiert.

4.7.3 Darstellung von Unsicherheit mittels des Net Health Benefit

Bei der Darstellung der Ergebnisse der Sensitivitätsanalysen ist zu beachten, dass die Berücksichtigung der Parameterunsicherheit zum einen die Lage mehrerer oder aller Interventionen verändern kann, die die Effizienzgrenze bilden. Zum anderen kann sich auch die Lage der zu prüfenden Intervention verändern, die dieser Effizienzgrenze gegenübergestellt wird.

Der Net Health Benefit (NHB), ein etabliertes Verfahren zur Darstellung von Ergebnissen aus PSAs [688], kann diesem Problem Rechnung tragen, da der NHB eine Funktion sowohl des Zusatznutzens und der Zusatzkosten als auch der Effizienzgrenze ist und die Position der zu bewertenden Intervention als Abstand zu der sich verlagernden Effizienzgrenze bzw. zu dem sich verlagernden letzten Segment der Effizienzgrenze abbildet. Daher sollten sowohl die Basisfallanalysen als auch die deterministischen und die probabilistischen Sensitivitätsanalysen basierend auf dem Konzept der NHB-Berechnung durchgeführt werden.

4.8 Ausgaben-Einfluss-Analyse (Budget-Impact-Analyse)

Eine Ausgaben-Einfluss-Analyse (AEA) ist eine Bewertung der direkten finanziellen Konsequenzen, die mit der Erstattung einer Intervention in einem Gesundheitssystem in Zusammenhang stehen [718]. In einem Berechnungsmodell für eine AEA werden der Anteil der Patientinnen und Patienten, die die neue Intervention möglicherweise erhalten werden, sowie die Verbreitung der Intervention im Gesundheitssystem einschließlich ihrer Anwendung bei zuvor unbehandelten Patientinnen und Patienten berücksichtigt. Eine AEA prognostiziert insbesondere, wie eine Veränderung im Mix der Interventionen für eine bestimmte Krankheit die Ausgaben für ein Indikationsgebiet zukünftig beeinflussen kann [508].

Zweck einer AEA ist es weniger, die finanziellen Konsequenzen einer Intervention exakt abzuschätzen. Vielmehr soll ein zuverlässiger Berechnungsrahmen zur Verfügung gestellt werden, der dem Entscheidungsträger ermöglicht, die möglichen Ausgabeneffekte einer neuen Intervention (oder der geänderten Nutzung von bereits vorhandenen Interventionen) zu verstehen [508]. Ein derartiges Modell ist erforderlich, weil viele der Ausgangsgrößen je nach Konstellation variieren können und darüber hinaus mit Unsicherheit behaftet sind. Daher ist das Ergebnis der AEA nicht ein einzelner Wert für die Ausgabenabschätzung, sondern vielmehr ein Wertebereich, der aus dem Modell resultiert.

4.8.1 Perspektive in der Ausgaben-Einfluss-Analyse

Die AEA sollte aus der Perspektive der GKV oder eines anderen relevanten Ausgabenträgers durchgeführt werden (siehe auch Abschnitt 4.4.1). Jegliche außerhalb dieser Perspektive anfallenden Ausgaben oder erzielten Einsparungen werden nicht einbezogen.

4.8.2 Zeithorizont in der Ausgaben-Einfluss-Analyse

Die AEA sollte den Zeithorizont umfassen, der für Kostenträger die höchste Relevanz für ihre Ausgaben hat [508]. Da sich der Einfluss auf das Ausgabengeschehen nach der Einführung der neuen Intervention im Laufe der Zeit wahrscheinlich verändert – sowohl aufgrund der Marktanpassung als auch aufgrund längerfristiger Auswirkungen auf die betreffende Krankheit –, sollte dieser für eine Periode von 1 und 3 Jahren abgeschätzt und dargeboten werden [507]. Das Ergebnis muss als Ausgaben und Einsparungen pro Jahr dargestellt werden anstatt in Form eines einzigen gegenwärtigen Nettowerts [508]. Daher darf in diesem Fall keine Diskontierung der Finanzflüsse vorgenommen werden. Sollte das Ergebnis als Gesamtkostenbetrag für 3 Jahre dargestellt werden, können die Kosten entsprechend diskontiert werden (siehe Abschnitt 4.4.6).

4.8.3 Szenarien in der Ausgaben-Einfluss-Analyse

Eine AEA vergleicht Versorgungsszenarien – jedes definiert durch eine Zusammenstellung von Interventionen – anstatt spezifische einzelne Interventionen [508]. Es müssen mindestens 2 Szenarien berücksichtigt werden: zum einen das Referenzszenario, definiert durch die aktuelle Kombination aus Interventionen, zum anderen die prognostizierte neue Kombination aus Interventionen.

4.8.4 Population in der Ausgaben-Einfluss-Analyse

Die Anzahl der Versicherten, die für die neue Intervention infrage kommen, ist einer der Schlüsselfaktoren zur Bestimmung der zu erwartenden Ausgaben für die neue Intervention. Die voraussichtliche Anzahl der Empfänger ergibt sich aus der prognostizierten Inanspruchnahme der Intervention innerhalb der Zielpopulation. Jede erwartete Off-Label-Anwendung der neuen Intervention sollte nicht in der primären AEA, sondern kann im Rahmen von Sensitivitätsanalysen berücksichtigt werden [554]. Bei der Vorhersage der Zahl der Anwender muss sowohl die Substitution vorhandener Interventionen als auch der induzierte Bedarf berücksichtigt werden.

4.8.5 Einzuschließende Kosten in der Ausgaben-Einfluss-Analyse

Die Kosten (Nettokosten, d. h. um Kosteneinsparungen, sogenannte Cost-Offsets, bereinigte Kosten) sollten entsprechend den in Abschnitt 4.4 beschriebenen Methoden abgeschätzt werden.

Für die AEA werden Investitions- und Implementierungskosten – soweit möglich und von der GKV getragen – identifiziert und quantifiziert. Sie sollten separat und nach Kostenkategorien geordnet dargestellt werden, wobei eine vollständige Erläuterung der Methode und der zur Kosteneinschätzung verwendeten Quellen enthalten sein muss.

4.8.6 Ergebnisdarstellung in der Ausgaben-Einfluss-Analyse

Die Ergebnisse (in €) sollten als Wertebereich präsentiert werden und nicht als einzelner Punktschätzer. Weiterhin sollte sowohl ein Gesamtbetrag als auch der Anteil an den Jahresausgaben ausgewiesen werden.

4.9 Besonderheiten der Kosten-Nutzen-Bewertung nach § 35b SGB V

4.9.1 Gesetzliche Anforderungen und Verfahrensablauf

Für die KNB nach § 35b SGB V gelten einige besondere Anforderungen. Standardmäßig gibt es im Rahmen der Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V 2 Konstellationen, die zu einer KNB führen können:

- 1) Ist ein pharmazeutischer Unternehmer mit dem Beschluss des G-BA nicht einverstanden, dass das zu bewertende Arzneimittel keinen Zusatznutzen hat oder keine therapeutische Verbesserung darstellt, kann der pharmazeutische Unternehmer nach § 35a Abs. 5a SGB V verlangen, dass der G-BA eine KNB nach § 35b SGB V oder nach § 139a Abs. 3 Nummer 5 SGB V in Auftrag gibt.
- 2) Nach einem Schiedsspruch der Schiedsstelle kann gemäß § 130b Abs. 8 SGB V sowohl der pharmazeutische Unternehmer als auch der GKV-Spitzenverband eine KNB nach § 35b SGB V beantragen.

Sollten ein pharmazeutischer Unternehmer und / oder der GKV-Spitzenverband beim G-BA eine KNB gemäß § 35b SGB V beantragen, ergeben sich weitere Besonderheiten im Ablauf des Verfahrens, die in Abschnitt 2.1.4 beschrieben werden.

Nach § 130b Abs. 8 Satz 3 SGB V dient eine KNB von Arzneimitteln nach § 35b SGB V dem Zweck der Verhandlung eines Erstattungsbetrags, der im Vergleich zu (einer) zweckmäßigen Vergleichstherapie(n) verhandelt werden soll. Nach § 35b SGB V legt der G-BA in seinem Auftrag auf eine KNB fest:

- zweckmäßige Vergleichstherapie und andere Arzneimittel und Behandlungsformen, mit denen das zu bewertende Arzneimittel verglichen werden soll,
- Patientengruppen,
- Zeitraum,
- Art des Nutzens und der Kosten und
- Maß des Gesamtnutzens.

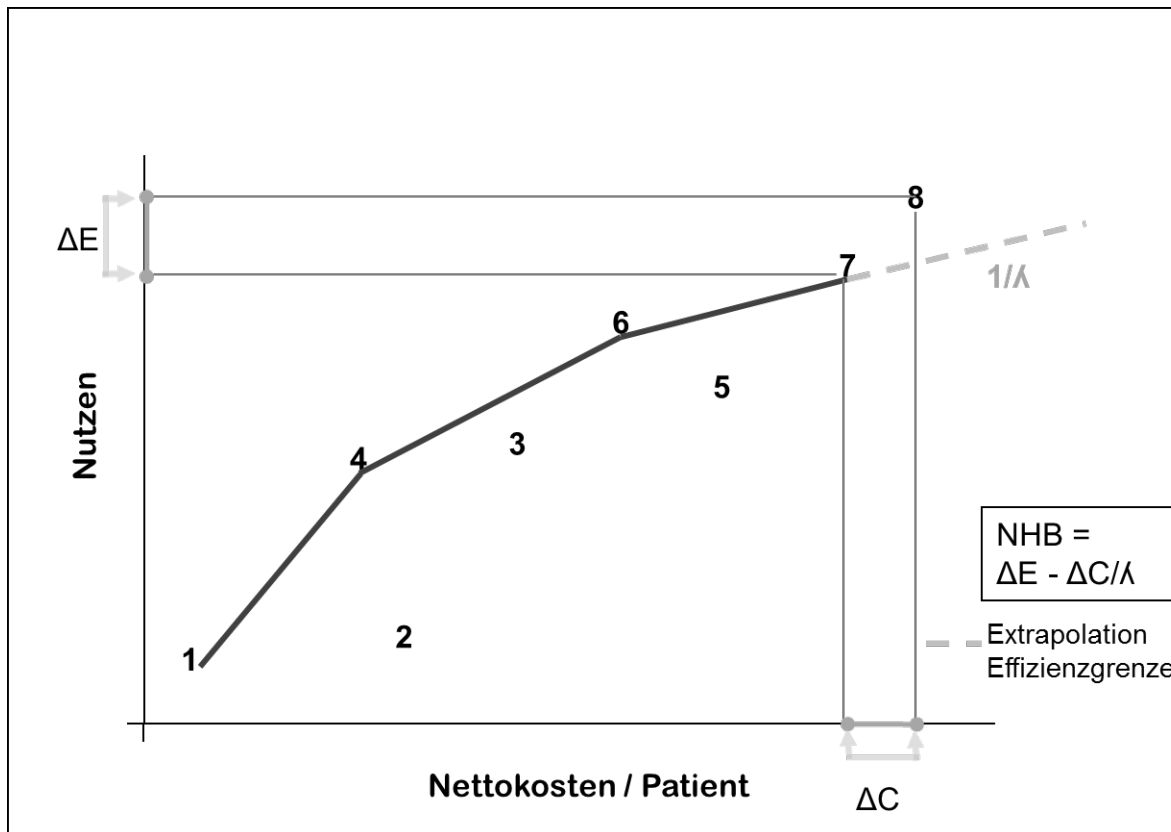
Basis für die KNB sind 1) die Ergebnisse klinischer Studien, 2) die Ergebnisse der Versorgungsstudien, die mit dem G-BA vereinbart wurden oder die der G-BA auf Antrag des pharmazeutischen Unternehmers anerkennt, und 3) die Nachweise des pharmazeutischen Unternehmers (vgl. § 35b Abs. 1 Satz 3 SGB V). Aufgrund der Gesetzeslage in Deutschland

(§ 35b Abs. 1 SGB V) wird zudem regelhaft die Perspektive der GKV-Versichertengemeinschaft eingenommen. Näheres beschreibt die Verfahrensordnung des G-BA ([276], siehe 5. Kapitel, 2. Abschnitt).

4.9.2 Der Net Health Benefit zur Berechnung zusatznutzenbasierter Erstattungspreise

Wie in Abschnitt 4.7.3 erläutert, kann der NHB zur Darstellung der Unsicherheit herangezogen werden. Auf Basis des Erwartungswerts des NHB der zu bewertenden Intervention kann über die weitere Berechnung des kostenbereinigten (Zusatz-)Nutzens der zu bewertenden Intervention auch ein zusatznutzenbasierter Erstattungspreis abgeleitet werden [689].

Der inkrementelle NHB wird mithilfe der Effektschätzer für den Nutzen und die Kosten der jeweiligen Interventionen sowie eines Schwellenwerts berechnet. In dieser Anwendung entspricht der Schwellenwert dem Kehrwert der Steigung des letzten (und ggf. extrapolierten) Segments der Effizienzgrenze für kosteneffektive Interventionen (siehe Abbildung 14). Wäre der NHB in etwa gleich 0, so läge Intervention 8 auf der Effizienzgrenze, die durch die Steigung ($1/\lambda$) des letzten Segments der Effizienzgrenze bestimmt wird, und kann im Vergleich zu den die Effizienzgrenze bildenden (per Definition kosteneffektiven) Interventionen ebenfalls als kosteneffektiv bewertet werden. Entsprechend wird ein zusatznutzenbasierter Erstattungspreis mittels des NHB durch Umformung und Berechnung der maximalen Interventionskosten ermittelt, die notwendig sind, um zu gewährleisten, dass der NHB mindestens gleich 0 ist. Praktisch kann der NHB mithilfe des Modells durch iterative Berechnungen geschätzt werden.

Abbildung 14: Darstellung eines $NHB > 0$

4.9.3 Sensitivitätsanalysen zur Berechnung zusatznutzenbasierter Erstattungspreise

Für den zusatznutzenbasierten Erstattungspreis können je Effizienzgrenze Preis-Akzeptanzkurven [245] und / oder NHB-Werte dargestellt werden (siehe Abschnitt 4.9.2).

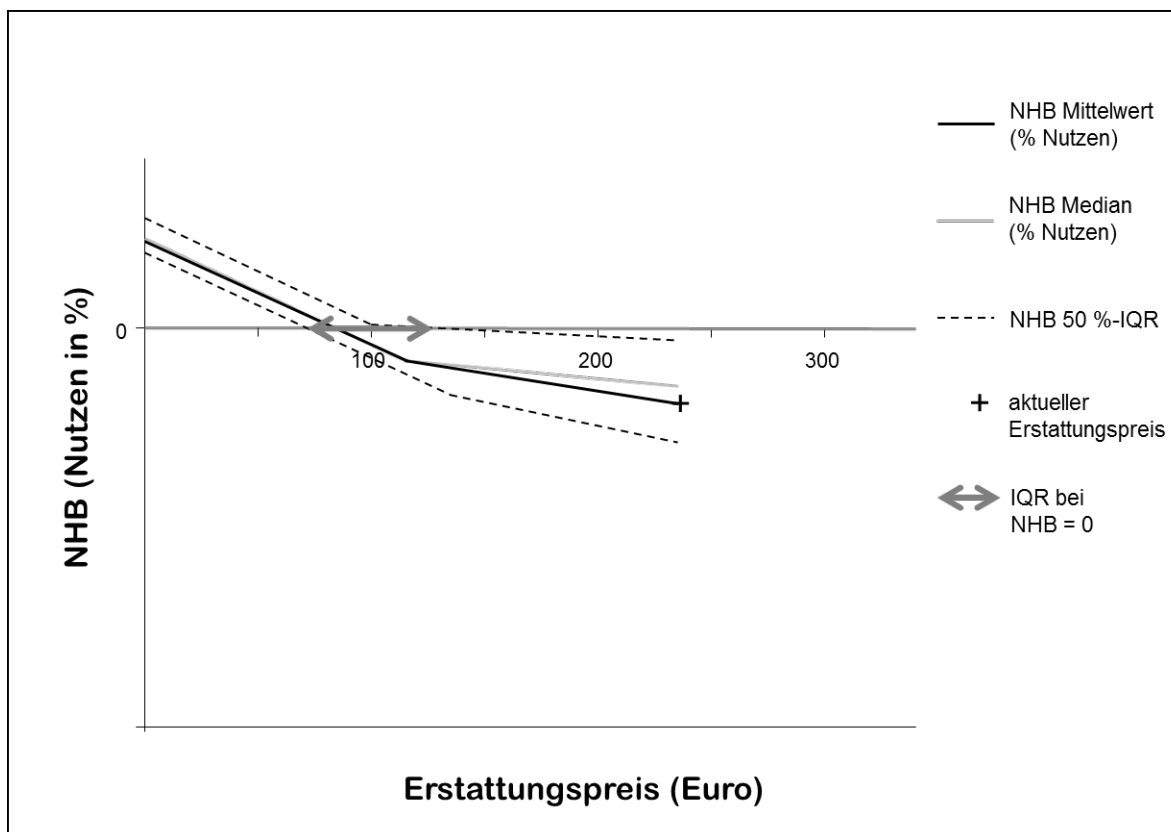
Bei Verwendung des NHB sollten die Ergebnisse der PSAs über die Berechnung und Mittelung der jeweiligen NHB-Erwartungswerte für die zu prüfende Intervention für eine hinreichend große Zahl von Durchläufen dargestellt werden. Bei jedem Durchlauf kann sich sowohl die Effizienzgrenze als auch die Lage der zu prüfenden Intervention relativ zu der Effizienzgrenze und somit der jeweilige NHB-Wert verändern. Hieraus lassen sich der gemittelte NHB-Wert der zu prüfenden Intervention sowie eine Interquartilsregion berechnen (siehe Abschnitt 4.9.4). Der NHB-Erwartungswert in Kombination mit der Interquartilsregion gibt an, wie groß erwartungsgemäß der kostenbereinigte (Zusatz-)Nutzen unter Berücksichtigung der Modellunsicherheit beim aktuellen zusatznutzenbasierten Erstattungspreis ist.

4.9.4 Interquartilsregion als Maß der Streuung für Preisverhandlungen

Um dem Spitzenverband Bund der Krankenkassen (kurz: GKV-Spitzenverband) und dem pharmazeutischen Unternehmer auf Basis der Ergebnisse der Sensitivitätsanalysen (siehe Abschnitt 4.7) ein Maß der Streuung für die Verhandlungen zu geben, wird eine Interquartilsregion (IQR) angegeben. Die IQR umschließt alle Werte des NHB aus den Simulationen,

die vom unteren und vom oberen Quartil begrenzt werden (siehe Abschnitt 4.9.3). Das heißt, dass die Interquartilsregion die 50 % der Simulationen in den PSAs abdeckt, die oberhalb der 25 % niedrigsten Ergebnisse und unterhalb der 25 % höchsten Ergebnisse liegen (siehe Abbildung 15). Grundsätzlich kann auch die Angabe von anderen Regionen mit anderen Maßen sinnvoll sein.

Die IQR erlaubt es, unter Berücksichtigung der gesamten Unsicherheit (umgesetzt durch PSAs) einen Spielraum für mögliche Verhandlungen um Erstattungsbeträge zu eröffnen, in dessen Grenzen auch die Unsicherheit der Effektschätzer und der Kosten berücksichtigt wird.



Die durchgezogene Linie gibt für jeden möglichen Erstattungspreis (x-Achse) den im Mittel zu erwartenden NHB an. An der Stelle, an der die durchgezogene Linie die x-Achse kreuzt, kann ein zusatznutzenbasierter Erstattungspreis abgelesen werden, bei dem der im Mittel zu erwartende NHB 0 ist, d. h. weder positiv noch negativ.

Abbildung 15: Interquartilsregion möglicher zusatznutzenbasierter Erstattungspreise (basierend auf PSAs) als Maß der Streuung für Preisverhandlungen

5 Bewertungen zur Versorgung

5.1 Leitliniensynopsen

5.1.1 Hintergrund

Leitlinien sind systematisch entwickelte Entscheidungshilfen für Leistungserbringer sowie Patientinnen und Patienten zur angemessenen Vorgehensweise bei speziellen Gesundheitsproblemen. Sie haben zum Ziel, die Patientenversorgung zu verbessern. Idealerweise liegen ihren Empfehlungen eine systematische Überprüfung der Evidenz und eine Bewertung des Nutzens und Schadens der alternativen Behandlungsoptionen zugrunde [249,296].

Leitlinien können, wenn sie in einem systematischen Verfahren erstellt werden und der Erstellungsprozess transparent dokumentiert ist, die normative Erwartungsbildung für alle Bereiche der Versorgungskette (Prävention, Diagnostik, Therapie, Rehabilitation und Nachsorge) fördern [330]. Diese normativ wirkenden Empfehlungen beinhalten wesentliche Informationen über die in einem Gesundheitssystem angestrebte Versorgungsqualität.

Die Identifikation und Beschreibung von Versorgungsstandards erfolgt in Leitliniensynopsen, in denen die Leitlinienempfehlungen aus systematisch recherchierten hochwertigen Leitlinien zusammengefasst werden [469,532,575] (siehe Abschnitte 5.1.5 und 5.1.6). Leitliniensynopsen eignen sich insbesondere für die überblicksartige Darstellung der normativen Anforderungen an die Versorgung für ausgewählte Krankheitsbilder. Sie dienen u. a. dem G-BA als Grundlage für die Entwicklung und Aktualisierung der Anforderungen an strukturierte Behandlungsprogramme für chronisch kranke Menschen (Disease-Management-Programme [DMPs]). Diese werden in der DMP-Anforderungen-Richtlinie (DMP-A-RL) unter verschiedenen Gliederungspunkten beschrieben [275,276].

5.1.2 Evidenzbasierte Leitlinien

Zur Beantwortung von Fragestellungen zum Versorgungsstandard im Rahmen der Leitliniensynopsen werden in der Regel evidenzbasierte Leitlinien herangezogen. Leitlinien werden als evidenzbasiert bezeichnet, wenn ihre Empfehlungen auf einer systematischen Literaturrecherche und -auswahl beruhen, wenn ihre Empfehlungen grundsätzlich mit einer Empfehlungs- und / oder Evidenzeinstufung (Grade of Recommendation [GoR] und / oder Level of Evidence [LoE]) versehen sind und mit den Referenzen der ihnen zugrunde liegenden Primär- und / oder Sekundärliteratur verknüpft sind (modifiziert nach Appraisal of Guidelines for Research and Evaluation [AGREE] [7,8]). Für den Einschluss von Leitlinien in die Leitliniensynopsen erfolgt eine Prüfung auf diese formalen Kriterien; eine inhaltliche Prüfung erfolgt nicht.

5.1.3 Übertragbarkeit auf das deutsche Gesundheitssystem

Werden für Leitliniensynopsen neben nationalen auch internationale Leitlinien recherchiert, wird grundsätzlich davon ausgegangen, dass Leitlinien aus Mitgliedsstaaten der Organisation

for Economic Co-operation and Development (OECD) von besonderer Relevanz sind. Ursprünglich ist die OECD 1961 als Nachfolgeeinrichtung der Organisation für europäische wirtschaftliche Zusammenarbeit gegründet worden [553]. Neben einem im globalen Vergleich hohen Pro-Kopf-Einkommen verfügen diese Mitgliedsländer jeweils über ein leistungsfähiges Gesundheitssystem. Seit 2003 wird eine gemeinsame Berichterstattung über ausgewählte Qualitätsindikatoren gesundheitlicher Versorgung innerhalb der OECD (Health Care Quality Indicator [HCQI] Project) angestrebt [506].

Im Grundsatz können alle Leitlinien aus OECD-Mitgliedsländern in Betracht gezogen werden. Allerdings kann die Übertragbarkeit von Empfehlungen aus internationalen Leitlinien problematisch sein, wenn es sich um spezifische Empfehlungen handelt, die beispielsweise mit den Finanzierungsprinzipien des deutschen Gesundheitswesens oder mit den berufs- und leistungsrechtlichen Besonderheiten des Systems der gesetzlichen Krankenversicherung in Deutschland nicht vereinbar sind. Dies wird in den Leitliniensynopsen bei Bedarf aufgegriffen.

5.1.4 Bewertung der methodischen Leitlinienqualität

Ein wichtiger Aspekt bei der Interpretation und Einschätzung von Leitlinienempfehlungen im Rahmen der Leitliniensynopse ist die Einschätzung, wie die methodische Qualität der zugrunde liegenden Leitlinien ist. Auf die Prüfung der inhaltlichen Qualität wird verzichtet.

International werden verschiedene Instrumente zur Bewertung der methodischen Qualität von Leitlinien eingesetzt [661,741]. Besonderes Augenmerk erhalten das AGREE-Instrument [7,492] bzw. seine Weiterentwicklung (AGREE-II-Instrument) [8,92-94]. Es wurde von einer international zusammengesetzten Gruppe von Wissenschaftlerinnen und Wissenschaftlern entwickelt und findet international am stärksten Verbreitung. Daher wird es auch in Leitliniensynopsen zur Bewertung der methodischen Qualität der eingeschlossenen Leitlinien herangezogen.

Das AGREE-II-Instrument [8] enthält 23 Beurteilungskriterien, die anhand einer mehrstufigen Skala bewertet werden. Diese Kriterien sind 6 voneinander unabhängigen Domänen zugeordnet, die jeweils eine separate Dimension der methodischen Leitlinienqualität beschreiben:

- Domäne 1: Geltungsbereich und Zweck
- Domäne 2: Beteiligung von Interessengruppen
- Domäne 3: Genauigkeit der Leitlinienentwicklung
- Domäne 4: Klarheit und Gestaltung
- Domäne 5: Anwendbarkeit
- Domäne 6: Redaktionelle Unabhängigkeit

Während das AGREE-Instrument die Berechnung standardisierter Domänenwerte für jede der Domänen vorsieht [8], wird die Anwendung im Rahmen der Leitliniensynopsen auf die Domänen 2, 3 und 6 begrenzt. Dies erfolgt mit Blick auf die Zielsetzung der Leitliniensynopsen

gemäß SGB V und Verfahrensordnung des G-BA [276], eine evidenzbasierte Grundlage für die Aktualisierung bestehender bzw. Entwicklung neuer DMPs zur Verfügung zu stellen. Die Begrenzung auf die Domänen 2,3 und 6 wird auch von anderen Autorentams vorgenommen [54].

Jede Leitlinienbewertung wird durch 2 Reviewerinnen beziehungsweise Reviewer unabhängig voneinander durchgeführt. Die jeweils berechneten standardisierten Domänenwerte werden in der Leitliniensynopse dargestellt.

AGREE gibt keine Schwellenwerte zur Abgrenzung zwischen methodisch guten und methodisch schwachen Leitlinien vor [8]. Jedoch geben einige Anwenderinnen und Anwender des AGREE-II-Instruments auf Basis der standardisierten Domänenwerte Anwendungsempfehlungen für Leitlinien, wobei 2- und 3-stufige Systeme zum Einsatz kommen. Im 3-stufigen System werden Leitlinien mit Domänenwerten unter einem spezifischen – aber je nach Anwendergruppe variierenden – Wert als schwach oder nicht empfehlenswert betrachtet [359]. In Anlehnung an dieses Verfahren werden in den Leitliniensynopsen Leitlinien, die in 1 oder mehreren der 3 betrachteten Domänen standardisierte Domänenwerte < 30 % erreichen, markiert. Dies gilt ebenso für Ergebnisse, die ausschließlich auf derartigen Leitlinien basieren.

Die Ergebnisse der AGREE-II-Bewertung sind somit zwar kein Kriterium für den Einschluss von Leitlinien in eine Leitliniensynopse, aber mithilfe der Markierung wird transparent dargestellt, ob die in eine Leitliniensynopse eingeschlossenen evidenzbasierten Leitlinien besondere methodische Stärken oder Schwächen aufweisen.

5.1.5 Strukturierte Aufbereitung von Empfehlungen: Evidenz- und Empfehlungsstufen

Eine Leitlinienempfehlung wird als ein Handlungsvorschlag für die klinische Entscheidung in einer spezifischen Situation bzw. für Systementscheidungen definiert. Adressatin bzw. Adressat ist in der Regel die oder der professionell Handelnde. Als Empfehlungen werden in den Leitlinien grundsätzlich jene Aussagen identifiziert, die von den Autorinnen und Autoren einer Leitlinie formal eindeutig als Empfehlung gekennzeichnet sind.

Die Autorinnen und Autoren von evidenzbasierten Leitlinien verwenden unterschiedliche Systeme zur Evidenz- und Empfehlungseinstufung ihrer Empfehlungen [28,208,313,456,643]. Empfehlungseinstufungen (GoRs) verleihen der Stärke einer Empfehlung Ausdruck. Sie beruhen in der Regel auf einer Abwägung des Nutzens und Schadens einer (medizinischen) Intervention und dem jeweils spezifischen Versorgungskontext auf der Grundlage einer Einschätzung der jeweiligen Evidenz. Evidenzeinstufungen (LoEs) fokussieren auf der internen Validität der zugrunde gelegten Studien, wobei systematische Übersichten zu randomisierten klinischen Studien (RCTs) üblicherweise den höchsten LoE erhalten. Die Systeme zur Evidenzeinstufung räumen klinischen und epidemiologischen Studien, den Merkmalen der Studiendurchführung und dem jeweiligen Verzerrungspotenzial gegebenenfalls einen unterschiedlichen Stellenwert innerhalb der LoE-Klassifikation ein [28,96-98,313,643].

International gibt es bisher keinen Konsens über eine Vereinheitlichung von Evidenz- und Empfehlungsgraduierungssystemen. Daher werden grundsätzlich die von den einzelnen Leitlinienerstellern verwendeten Evidenz- bzw. Empfehlungsstufen in der Leitliniensynopse angegeben und die entsprechenden Graduierungssysteme dokumentiert. Um die unterschiedlichen Systeme der verschiedenen Leitlinienautorengruppen besser miteinander vergleichen zu können, wird die Einstufung der GoRs und LoEs für die Leitliniensynopsen vereinfacht, indem diese in ein einheitliches Kategoriensystem überführt werden. Dabei werden jeweils die Kategorien hoch, nicht hoch und unklar unterschieden.

Für die Beurteilung der GoRs ist das Graduierungssystem aus dem Verfahren der Nationalen VersorgungsLeitlinie (NVL) maßgeblich [102], für die Einstufung der LoEs wird auf die vom G-BA verwendete Evidenzklassifizierung [276] zurückgegriffen. Ein hoher GoR wird dann vergeben, wenn die empfohlene Maßnahme der Empfehlungsstärke A (starke Empfehlung) zugeordnet werden kann. Alle anderen Empfehlungen werden der Kategorie nicht hoher bzw. unklarer GoR zugeordnet. Ein von den Autorinnen und Autoren von Leitlinien zugewiesener LoE wird dann als hoch eingestuft, wenn der LoE mindestens auf einer randomisierten kontrollierten Studie (RCT) beruht. Dies entspricht den Evidenzstufen Ia und Ib der vom G-BA verwendeten Evidenzklassifizierung.

Verwenden die Leitlinienautorinnen und -autoren ein Klassifizierungssystem entsprechend dem der GRADE Working Group [309,310,313], wird die höchste Evidenzstufe nach GRADE grundsätzlich der Kategorie hoher LoE zugeordnet. Alle weiteren von den Leitlinienautorinnen und -autoren angegebenen LoEs, die nicht der Kategorie hoch zugeordnet werden können, werden der Kategorie nicht hoher bzw. unklarer LoE zugeordnet.

Die Kategorie unklar wird vergeben, wenn ein von der Leitliniengruppe angegebener GoR / LoE nicht entsprechend dem Empfehlungsgraduierungssystem des Programms für Nationale VersorgungsLeitlinien (NVL-Programms) beziehungsweise der Evidenzklassifizierung des G-BA oder GRADE kategorisierbar ist, der angegebene GoR / LoE nicht eindeutig einer Empfehlung zugeordnet werden kann oder kein GoR / LoE angegeben ist.

5.1.6 Strukturierte Informationssynthese: Extraktion und Analyse der Empfehlungen

Basis für die Synthese von Empfehlungen im Rahmen der Leitliniensynopse sind aktuelle evidenzbasierte Leitlinien. Im 1. Schritt werden die Empfehlungen der Leitlinien mit den der aktuell gültigen DMP-A-RL inhaltlich abgeglichen. Dabei werden die Empfehlungen den jeweiligen Aussagen der DMP-A-RL (sogenannte Versorgungsaspekte) zugeordnet.

Im Falle bestehender DMPs wird bei der Zuordnung der Empfehlungen geprüft, ob diese inhaltlich diskrepant zur DMP-A-RL sind. Für die weitere Analyse werden nur die Versorgungsaspekte berücksichtigt, die zur DMP-A-RL inhaltlich diskrepante Empfehlungen enthalten.

Den Empfehlungen dieser Versorgungsaspekte werden die Evidenz- bzw. Empfehlungsstufen gemäß Abschnitt 5.1.5 zugeordnet. Darauf aufbauend erfolgt die methodische Beurteilung: Nur Versorgungsaspekte, die Empfehlungen mit mindestens einem hohen GoR enthalten oder bei denen alternativ ausschließlich Empfehlungen mit unklaren GoR vorliegen, wobei aber mindestens eine dieser Empfehlungen mit einem hohen LoE versehen sein muss, werden für die Entwicklung von Kernaussagen berücksichtigt und in der Leitliniensynopse dargestellt. Hierzu werden die zentralen Inhalte der Empfehlungen pro Versorgungsaspekte zusammengefasst (Kerninhalte).

In den Ergebnistabellen wird die Diskrepanz zwischen Versorgungsaspekt und DMP-A-RL kurz und prägnant erläutert. Ggf. werden diese Erläuterungen um methodische Hinweise und wissenschaftliche Aussagen, beispielsweise für die Versorgung in Deutschland abweichende, relevante Informationen oder wissenschaftliche Diskurse, ergänzt.

Im Fall von Leitliniensynopsen zu neuen DMPs erfolgt das Verfahren analog, mit Ausnahme des Abgleichs zur DMP-A-RL.

5.2 Mindestmengenregelungen

5.2.1 Hintergrund

In der Fachwelt wird seit mehr als 40 Jahren der potenzielle Zusammenhang zwischen definierten Leistungsmengen in (vorwiegend stationären) Einrichtungen und der Qualität untersucht [505]. Seit 2003 setzte der G-BA gemäß § 137 Abs. 3 Satz 1 Nr. 2 SGB V für bestimmte planbare stationäre Leistungen sogenannte verbindliche Mindestmengen für Kliniken fest: Kliniken dürfen die entsprechenden Leistungen nur dann erbringen und abrechnen, wenn die zuständigen Krankenhausträger gegenüber den Landesverbänden der Krankenkassen überzeugend darlegen, dass sie aufgrund der im Vorjahr erreichten Leistungsmenge die Mindestmenge im nächsten Jahr mindestens erfüllen werden.

Diese Mindestmengenregelungen sind für die nach § 108 SGB V zugelassenen Krankenhäuser verbindlich und legen fest, in welchem Fall ein Krankenhausstandort die Leistungen erbringen darf, zu denen Mindestmengen festgelegt [274] sind. Es gelten allerdings einige Ausnahmeregelungen. So bleiben zum Beispiel Notfälle grundsätzlich von der Mindestmengenregelung unberührt. Oder die für die Krankenhausplanung zuständigen Landesbehörden können Ausnahmeregelungen für solche Leistungen bestimmen, bei denen die Anwendung der Mindestmengenregelung die Sicherstellung einer flächendeckenden Versorgung der Bevölkerung gefährden könnte.

Für das Festlegen einer neuen bzw. das Überprüfen einer bestehenden Mindestmenge kann der G-BA das IQWiG mit der wissenschaftlichen Bewertung des Zusammenhangs zwischen Leistungsmenge und Qualität zu einem spezifischen medizinischen Verfahren beauftragen (§ 136 Abs. 1 Satz 2 SGB V und Verfahrensordnung des G-BA [276]).

5.2.2 Informationsgrundlage und Bewertung

Die Untersuchung des Zusammenhangs zwischen Leistungsmenge und Qualität des Behandlungsergebnisses erfolgt auf Grundlage geeigneter Beobachtungsstudien oder kontrollierter Interventionsstudien. Es werden Ergebnisse zu unterschiedlichen Zielgrößen betrachtet. Diese können sich auf die Mortalität, Morbidität, gesundheitsbezogene Lebensqualität einschließlich der Aktivitäten des täglichen Lebens und der Abhängigkeit von der Hilfe anderer Personen beziehen.

Da die Qualität des Behandlungsergebnisses einer medizinischen Intervention von zahlreichen individuellen Risikofaktoren der Patientinnen und Patienten sowie von der Qualität der Maßnahme selbst beeinflusst werden kann, werden nur Beobachtungsstudien eingeschlossen, in denen eine Kontrolle von relevanten Störgrößen (Risikoadjustierung) erfolgte.

Die Aussagekraft der Ergebnisse der Beobachtungsstudien wird auf Basis von Qualitätskriterien, die speziell für Studien zur Bewertung von Mengen-Ergebnis-Beziehungen entwickelt wurden, beurteilt [51,387,388,752]. Hierbei werden verschiedene Sachverhalte geprüft (verwendete Datenquellen, Angemessenheit der statistischen Modelle, geeignete Risikoadjustierung sowie Vollständigkeit der Berichterstattung). Basierend auf dieser qualitativen Beurteilung wird eine Qualitätseinstufung der Beobachtungsstudien in Studien mit hoher und niedriger Aussagekraft vorgenommen.

Zudem wird dargelegt, wie sich eine für die Versorgung verbindlich vorgegebene Leistungsmenge pro Klinik, pro Ärztin oder Arzt oder pro Kombination Klinik und Ärztin oder Arzt auf die Qualität des Behandlungsergebnisses ausgewirkt hat.

Sofern kontrollierte Interventionsstudien eingeschlossen werden, wird als prüfende Intervention die Vorgabe einer Mindestfallzahl betrachtet. Das Verzerrungspotenzial der eingeschlossenen kontrollierten Interventionsstudien wird entsprechend den Ausführungen in Kapitel 9 bewertet (siehe Abschnitt 9.1.4).

Die Ergebnisse zu den in den Studien berichteten Zielgrößen werden im Bericht vergleichend beschrieben. Wenn möglich werden über die Gegenüberstellung der Ergebnisse der Einzelstudien hinaus geeignete metaanalytische Verfahren eingesetzt (siehe Abschnitt 9.3.7).

5.3 Analyse von Versorgungsdaten

5.3.1 Hintergrund

Unter Versorgungsdaten werden Sekundärdaten verstanden, die die Gesundheitsversorgung beschreiben. Sekundärdaten charakterisiert, dass die Daten über ihren originären, vorrangigen Verwendungszweck hinaus ausgewertet werden. Die nachfolgende Nutzung über den primären Erhebungsanlass hinaus ist hier maßgeblich [696]. Hierunter fallen beispielsweise Routinedaten gesetzlicher Krankenkassen, Daten aus Registern zu epidemiologischen, klinischen oder

interventionsbezogenen Fragestellungen oder auch Daten aus bevölkerungsbezogenen Gesundheitsbefragungen, die z. B. vom Robert Koch-Institut durchgeführt werden. Die Analyse von Versorgungsdaten umfasst die – je nach Auftrag unterschiedlich detaillierte – aktuelle und systematische Beschreibung und Analyse von Versorgungsaspekten, oftmals unter den Gesichtspunkten der Bedarfsgerechtigkeit [613], Qualität und Wirtschaftlichkeit von im Rahmen der GKV erbrachten Leistungen einer definierten Bevölkerungsgruppe zu einer konkreten medizinischen oder systembezogenen Fragestellung (vgl. § 139a SGB V). Neben der Beschreibung einzelner Versorgungsaspekte des Versorgungsprozesses können Einzelinterventionen oder auch komplexe Interventionen untersucht werden. Für eine geplante Analyse von Versorgungsdaten können in einem modularen System verschiedene individualmedizinische, populations- und gesundheitssystembezogene Daten ggf. aus unterschiedlichen Quellen herangezogen werden.

5.3.2 Ziele einer Analyse von Versorgungsdaten

Das übergeordnete Ziel einer Analyse von Versorgungsdaten ist die Beschreibung der Gesundheitsversorgung.

Folgende Aspekte können Teilziele einer Analyse von Versorgungsdaten sein:

- Bestimmung epidemiologischer Kennzahlen,
- Ermittlung von Krankheitskosten,
- Bestimmung von gesundheitsökonomischen Kennzahlen (z. B. Ausgaben-Einfluss-Analyse, Kosten-Nutzen-Verhältnisse),
- Prüfung der Bedarfserfüllung und der Bedarfsgerechtigkeit, Hinweise auf eine mögliche Über-, Unter- oder Fehlversorgung [613],
- Identifizierung eines potenziellen Forschungsbedarfs (z. B. klinische Forschung, HTA, Gesundheitssystemforschung).

Im Rahmen eines Projekts wird aus Machbarkeitsgründen in der Regel auf einzelne oder einige wenige der oben genannten Ziele mit Bezug zu einer bestimmten Erkrankung oder einem bestimmten Versorgungsaspekt fokussiert.

5.3.3 Inhaltliche Aspekte einer Analyse von Versorgungsdaten

Zwingende Voraussetzung für eine Analyse von Versorgungsdaten ist, dass die zu untersuchende Fragestellung mit den jeweils vorliegenden Daten zu beantworten ist. Die Formulierung einer konkreten Fragestellung ist daher für die systematische Beschreibung und Untersuchung von Versorgungsbereichen notwendig. Zur Definition der Fragestellungen gehört die Festlegung folgender Punkte:

- Population (Alter, Geschlecht, Erkrankung, falls relevant Subgruppe bzw. Schweregrad der Erkrankung),

- ggf. die zu untersuchenden Interventionen (z. B. haus- oder fachärztliche Betreuung von Patientinnen und Patienten mit Diabetes mellitus),
- Setting, in dem die Versorgung stattfindet (z. B. ambulante Versorgung, akutstationäre Versorgung, Pflegeeinrichtungen oder sektorenübergreifende Versorgung),
- Zielgrößen (z. B. Anzahl der Patientinnen und Patienten, Kosten der GKV-Leistungen).

Die Analyse von Versorgungsdaten kann verschiedene Ebenen und / oder mehrere Versorgungsaspekte einbeziehen. Grundsätzlich werden dabei 3 Hauptbereiche unterschieden: ein epidemiologischer, ein gesundheitsökonomischer und ein Bereich der sozialen Organisation medizinischer Versorgung. Der 1. Bereich umfasst die Verteilung und Häufigkeit von Krankheiten in der Bevölkerung, auf deren Basis der Bedarf an medizinischen Leistungen abgeleitet werden kann. Besonderes Augenmerk kann dabei auf bestimmte Subgruppen der Bevölkerung gelegt werden (z. B. wie sieht die Altersstruktur einer Krankheitspopulation aus, die eine bestimmte Komorbidität aufweist). Der 2. Bereich beinhaltet die Ermittlung von Kosten bzw. dem Verhältnis von Kosten und Nutzen von Gesundheitsinterventionen. Schließlich werden auf einer 3. Ebene z. B. Fragen der Ausgestaltung der versorgungsbezogenen Strukturen und Prozesse der Leistungserbringung thematisiert.

Zusätzlich kann die Berücksichtigung von soziokulturellen und ethischen Aspekten bei der Beschreibung der Versorgung bestimmter Patientengruppen notwendig sein, z. B. die Berücksichtigung eines erschwerten Zugangs zur medizinischen Versorgung.

5.3.4 Datenquellen

In Abhängigkeit von der Fragestellung kommen unterschiedliche Datenquellen zur Beantwortung infrage.

Die für die jeweilige Fragestellung potenziell relevanten Datenquellen sollen vor Durchführung der Analyse definiert werden. Im Folgenden werden mögliche Datenquellen benannt (siehe Tabelle 8).

Tabelle 8: Mögliche Datenquellen für die Analyse von Versorgung

Beispiele für Daten	Öffentlich verfügbar	Vereinbarung mit dem Dateneigner erforderlich
Gesundheitsbefragungen, z. B.: ▪ Gesundheitsberichterstattung des Bundes und der Länder (z. B. Kinder- und Jugendgesundheits-survey des RKI) ▪ Berichte des Statistischen Bundesamts (z. B. Krankenhausentlassungsdiagnosen, Todesursachenstatistik)	x	
Registerdaten, z. B.: ▪ Daten des RKI ▪ epidemiologische und klinische Krebsregister	(x)	
Routinedaten, z. B.: ▪ der gesetzlichen Krankenkassen ▪ der Kassenärztlichen Vereinigungen		x
RKI: Robert Koch-Institut		

Wenn potenziell geeignete Daten für eine Fragestellung identifiziert wurden, erfolgen ggf. entsprechende vertragliche Vereinbarungen zwischen dem Dateneigner und dem IQWiG, damit die Daten für eine Analyse zur Verfügung gestellt werden können.

Grundsätzlich sollte eine Bewertung der Datenstruktur vorangehen, beispielsweise hinsichtlich der Datenqualität, d. h. Vollständigkeit und Plausibilität der Datensätze, Erhebungsmethodik, Vollzähligkeit, Aktualität.

Mit der Verwendung von Daten aus unterschiedlichen Quellen außerhalb klinischer Studien zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V hat sich das IQWiG im Rapid Report A19-43 [394] auseinandergesetzt.

5.3.5 Methodische Besonderheiten der Analyse von Versorgungsdaten

Die Komplexität der Analysen von Versorgungsdaten erfordert grundsätzlich eine interdisziplinäre Bearbeitung. Neben klinischer Expertise ist ausreichende Kenntnis in der Handhabung und Interpretation entsprechender Daten von Bedeutung.

Eine einheitliche Vorgehensweise lässt sich bei der Analyse von Versorgungsdaten nicht festlegen. Diese ist stark abhängig von den gegebenen Rahmenbedingungen, wie z. B. den zentralen Fragestellungen oder den verwendeten Datenquellen.

Die Bearbeitung der Fragestellungen erfolgt mittels quantitativer Methoden. Bei der Durchführung und Planung einer Analyse von Versorgungsdaten sind die Empfehlungen der Guten Praxis für Sekundärdatenanalyse zu beachten [24].

5.3.6 Ergebnisdarstellung

Eine transparente Darstellung des methodischen Vorgehens und der Ergebnisse ist daher zentral. Die angewendete Methode muss nachvollziehbar beschrieben sein, damit eine Replizierbarkeit gewährleistet ist. Die Berichterstattung soll der Standardisierten Berichtsroutine für Sekundärdatenanalysen entsprechen (STROSA) [695].

6 HTA-Berichte

6.1 Hintergrund und Ziel

Nach § 139b Abs. 5 SGB V können Versicherte und interessierte Einzelpersonen Bewertungen von medizinischen Untersuchungs- und Behandlungsmethoden bei ausgewählten Krankheiten sowie von Fragen der Qualität und Wirtschaftlichkeit vorschlagen. Ausgenommen sind gemäß § 139b Abs. 5 SGB V Themenvorschläge, bei denen die eigenständige Bewertung eines Arzneimittels im Vordergrund steht.

Aufgabe des Instituts ist es, aus diesen Vorschlägen Themen auszuwählen, die für die Versorgung von Patientinnen und Patienten von besonderer Bedeutung sind, und diese Themen in Form von Health-Technology-Assessment(HTA)-Berichten zu bearbeiten.

Im Folgenden wird der Prozess von der Themeneingabe bis hin zur Berichterstellung beschrieben.

6.2 Themensammlung

Themenvorschläge für HTA-Berichte können von Versicherten und interessierten Einzelpersonen über die Website www.themencheck-medizin.iqwig.de eingereicht werden.

6.3 Auswahl der Themen für die HTA-Berichte

Die Themen, die bis zu einem jährlichen Stichtag vorgeschlagen werden, durchlaufen ein mehrstufiges Auswahlverfahren (Abbildung 16). Dabei wird sowohl die Bürger- und Patientensicht als auch die wissenschaftliche Perspektive miteinbezogen.

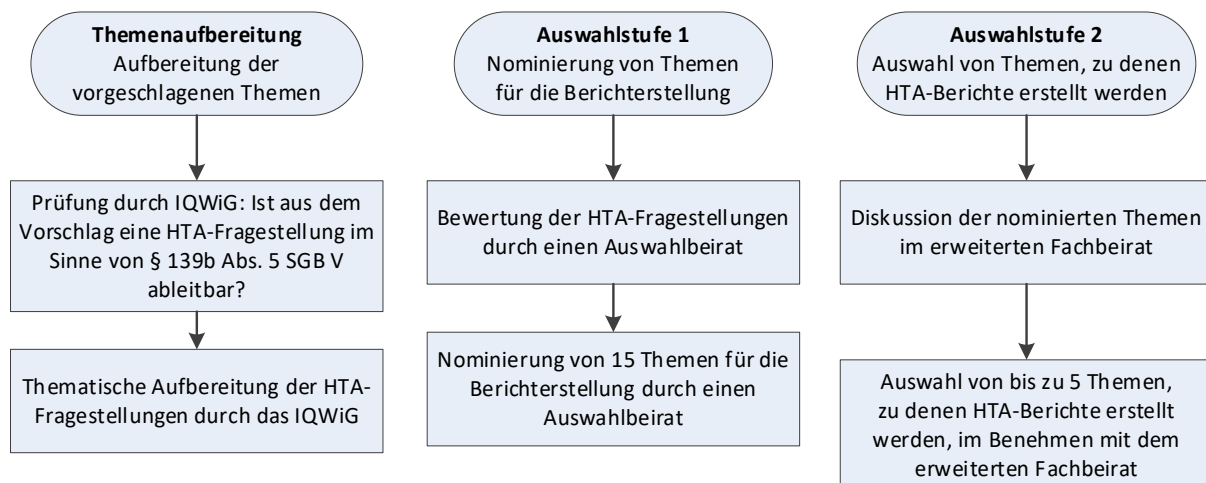


Abbildung 16: Schematische Darstellung des mehrstufigen Auswahlverfahrens

6.3.1 Auswahlkriterien

Für die Prüfung und Gewichtung der Themenvorschläge sowie auf jeder der 2 Auswahlstufen erfolgt eine Beurteilung der vorgeschlagenen Themen anhand vorab definierter Kriterien.

Ziel ist es, Themen auszuwählen, die für die Versorgung von Patientinnen und Patienten besonders bedeutsam sind. Ebenso ist grundsätzlich festgelegt, dass medizinische Untersuchungs- und Behandlungsmethoden bewertet werden. Weitere Kriterien und Fragen, die bei der Auswahl berücksichtigt werden, sind unter anderem:

- Wie groß ist die Anzahl der Betroffenen?
- Wie hoch ist die Krankheitslast oder der Schweregrad der Erkrankung?
- Wie umfangreich ist die zur Fragestellung vorhandene Evidenz?
- Wurde die Fragestellung bereits in aktuellen HTA-Berichten untersucht?
- Welche Kosten sind mit einer Intervention verbunden?

6.3.2 Prüfung der Fragestellung und Themenaufbereitung

Die eingereichten Vorschläge werden zeitnah durch das Institut geprüft und gegebenenfalls redaktionell und qualitativ bearbeitet.

Gemäß § 139b Abs. 5 SGB V sind Themenvorschläge, bei denen die eigenständige Bewertung eines Arzneimittels im Vordergrund steht, von der Bearbeitung ausgenommen. Daher werden nur die Themenvorschläge, aus denen sich im Sinne von § 139b Abs. 5 SGB V eine HTA-Fragestellung ableiten lässt, weiter bearbeitet.

Zu den oben genannten Auswahlkriterien (siehe Abschnitt 6.3.1) werden Informationen zu allen Themen mit einer HTA-Fragestellung im Sinne von § 139b Abs. 5 SGB V gesammelt und eine Vorabrecherche (siehe Abschnitt 8.1.1) wird durchgeführt. Die Ergebnisse werden in einer Themenaufbereitung zusammengefasst.

6.3.3 1. Stufe des Auswahlverfahrens: Nominierung von Themen

In der 1. Stufe des Auswahlverfahrens nominiert ein Auswahlbeirat Themen für die HTA-Berichterstellung. Dazu werden dem Auswahlbeirat die aufbereiteten Themenvorschläge zur Verfügung gestellt. Der Auswahlbeirat wählt auf dieser Basis 15 Themen für die HTA-Berichterstellung aus. Dabei wird sowohl die Bürger- und Patientensicht als auch die wissenschaftliche Perspektive miteinbezogen.

Der Auswahlbeirat wird erstens mit Vertreterinnen und Vertretern der auf Bundesebene als maßgeblich anerkannten Organisationen für die Wahrnehmung der Interessen der Patientinnen und Patienten und der Selbsthilfe chronisch kranker und behinderter Menschen, zweitens mit einer Vertreterin oder einem Vertreter der oder des Beauftragten der Bundesregierung für die Belange der Patientinnen und Patienten sowie drittens mit Bürgerinnen und Bürgern besetzt.

6.3.4 2. Stufe des Auswahlverfahrens: Auswahl von Themen, zu denen HTA-Berichte erstellt werden

Die vom Auswahlbeirat in der 1. Stufe des Auswahlverfahrens nominierten Themen werden in der 2. Stufe mit dem erweiterten Fachbeirat diskutiert. Dieser ist mit Vertreterinnen und Vertretern der den Stiftungsrat der Stiftung für Qualität und Wirtschaftlichkeit bildenden Organisationen sowie einer Vertreterin beziehungsweise einem Vertreter des BMG besetzt. Anschließend wählt das Institut aus den vom Auswahlbeirat nominierten Themen im Benehmen mit dem erweiterten Fachbeirat bis zu 5 aus, für die HTA-Berichte erstellt werden.

6.3.5 Umgang mit nominierten, aber nicht für eine HTA-Erstellung ausgewählten Themen

Die vom Auswahlbeirat nominierten, aber vom Institut nicht für eine HTA-Erstellung ausgewählten Themen gehen einmalig erneut in das nächste Auswahlverfahren im Folgejahr ein. Sie durchlaufen dort erneut die in den Abschnitten 6.3.3 und 6.3.4 beschriebenen Auswahlstufen.

6.4 Sicherstellung der Qualität der HTA-Berichte

Unter anderem die folgenden Maßnahmen stellen eine hohe Qualität der HTA-Berichte sicher:

- inhaltliche Strukturierung der HTA-Berichte durch das Institut durch die Bereitstellung von Templates für das HTA-Berichtsprotokoll sowie den vorläufigen HTA-Bericht,
- Konformitätsprüfung des HTA-Berichtsprotokolls, des vorläufigen HTA-Berichts und des HTA-Berichts durch das Institut,
- Erstellung des HTA-Berichts entsprechend der Methodik des Instituts,
- Durchführung der systematischen Recherchen für die HTA-Berichte im Institut,
- Durchführung eines Stellungnahmeverfahrens,
- Einordnung der Ergebnisse des HTA-Berichts in einem vom Institut erstellten Herausgeberkommentar.

6.5 Bearbeitung der Themen (HTA-Berichte)

Die HTA-Berichte enthalten regelhaft Inhalte zu allen HTA-relevanten Aspekten. Neben der obligatorischen Bewertung des Nutzens und Schadens von Interventionen werden in den HTA-Berichten in Anlehnung an internationale HTA-Definitionen auch ökonomische, ethische, soziale, rechtliche und organisatorische Aspekte der Intervention dargestellt [233,452,489,565].

Die Ausführlichkeit der Auseinandersetzung mit dem gesundheitsökonomischen Stellenwert der jeweiligen Untersuchungs- und Behandlungsmethode sowie deren ethischen, sozialen, rechtlichen, organisatorischen und sozialen Aspekten ist abhängig von der Fragestellung und wird im HTA-Berichtsprotokoll festgelegt. Nutzen und Schaden von Untersuchungs- und

Behandlungsmethoden sowie deren ökonomische, ethische, soziale, rechtliche und organisatorische Aspekte sind miteinander verknüpft. In der Bearbeitung und der zusammenfassenden Ergebnisdarstellung können daher die einzelnen Aspekte nicht losgelöst voneinander betrachtet werden.

6.5.1 Nutzenbewertung

Die Erstellung der Abschnitte der HTA-Berichte zur Bewertung des Nutzens und Schadens einer Untersuchungs- und Behandlungsmethode erfolgt unter Anwendung der in Kapitel 3, Abschnitt 7.13 und Kapitel 9 beschriebenen Vorgehensweise.

6.5.2 Gesundheitsökonomie

Sofern in den HTA-Berichten eine eigene gesundheitsökonomische Analyse erarbeitet werden soll, sind die Vorgaben in den Abschnitten 4.2 bis 4.5 und 4.7 sowie 4.8 zu beachten. Die Ergebnisse sind grafisch in einem Koordinatensystem darzustellen.

Eine weitere Möglichkeit ist die Erstellung einer systematischen Übersicht der verfügbaren Evidenz zu gesundheitsökonomischen Analysen. Die Analysen werden dabei insbesondere hinsichtlich der Übertragbarkeit der Ergebnisse auf Deutschland, der Einordnung der Ergebnisse unter Beachtung der methodischen Anforderungen, die in den Abschnitten 4.2 bis 4.5 beschrieben sind, sowie der Vollständigkeit der Datengrundlage beurteilt.

6.5.3 Ethik

In der Medizin, aber auch in der Public-Health- und Versorgungsforschung sind für eine ethische Bewertung die 4 Prinzipien des Wohltuns und Nutzens, des Nichtschadens, des Respekts vor der Autonomie der Patientin oder des Patienten und der Gerechtigkeit nach Beauchamp und Childress [42] weitverbreitet [207,317]. Für die Analyse ethischer Aspekte von medizinischen Interventionen nach den Methoden des HTA wurde dieser Ansatz ebenfalls häufig verwendet [41,502]. In den vergangenen Jahren wurden auch weitere methodische Ansätze eingesetzt. Dazu zählen z. B. der sokratische Ansatz, das Social Shaping of Technology, die Virtue Ethics oder auch der trianguläre Ansatz [25,363]. Je nach den zu bearbeitenden Fragestellungen eignen sich die verfügbaren methodischen Konzepte in unterschiedlicher Weise.

Der sokratische Ansatz liefert detaillierte Ergebnisse und wurde bei vielen unterschiedlichen medizinischen Interventionen eingesetzt [199,362,490]. Der Fragenkatalog von Hofmann 2005 [360,361] basiert auf dem sokratischen Ansatz und soll regelhaft für die HTA-Berichte eingesetzt werden. Bei besserer Eignung anderer methodischer Ansätze können bei entsprechender Begründung im HTA-Berichtsprotokoll auch diese eingesetzt werden.

6.5.4 Soziale Aspekte

Soziale und soziokulturelle Aspekte im HTA greifen die wechselseitigen Interaktionen zwischen der Untersuchungs- bzw. Behandlungsmethode und der sozialen Umwelt (z. B.

Verteilung von Ressourcen in einer Gesellschaft, Zugang zu Technologien, Patientenpräferenzen, gesellschaftliche Normen und Wertvorstellungen) auf. Es werden 4 Ansätze zur Identifikation von sozialen Aspekten in HTAs diskutiert: Checklisten, Literaturreviews, partizipatorische Ansätze und empirische Forschung [281,525]. Für die Aufbereitung bzw. orientierende Bewertung soziokultureller Aspekte in den HTA-Berichten wird das Framework von Mozygemba et al. [525] empfohlen. Gegebenenfalls kann auch ein generischer Fragebogen (z. B. Gerhardus und Stich [281]) oder die Checkliste aus dem HTA-Core-Modell von EUnetHTA [233] hilfreich sein.

6.5.5 Rechtliche Aspekte

Rechtliche Aspekte im HTA beziehen sich einerseits auf den rechtlichen Rahmen, in den die Untersuchungs- bzw. Behandlungsmethode und deren Bewertung eingebettet sind, und zum anderen auf die mit der Implementation und Nutzung der Gesundheitstechnologie verbundenen rechtlichen Aspekte. Unterschieden wird zwischen den technologie- und patientenbezogenen rechtlichen Aspekten [216,329,758].

6.5.6 Organisatorische Aspekte

Ein HTA-Bericht kann auch untersuchen, welche Wechselwirkungen durch eine Untersuchungs- bzw. Behandlungsmethode für die Organisation der Versorgung entstehen und welche Bedingungsfaktoren die Implementierung einer Untersuchungs- bzw. Behandlungsmethode beeinflussen können. Grundsätzlich kann hier zwischen der Wechselwirkung von organisatorischen Rahmenbedingungen, Abläufen und Prozessen und den Anforderungen an die Professionellen im Gesundheitswesen sowie deren Ent- bzw. Mehrbelastungen unterschieden werden [566]. Bisher existiert jedoch kein methodischer Standard, mit dem die organisatorischen Wechselwirkungen von Untersuchungs- bzw. Behandlungsmethoden im Gesundheitswesen untersucht werden können [566].

Eine Hilfestellung bei der Bearbeitung organisatorischer Fragestellungen kann das von Perleth et al. [566] vorgeschlagene Raster zur Einschätzung der organisatorischen Folgen von Untersuchungs- bzw. Behandlungsmethoden bieten.

7 Evidenzbasierte Gesundheitsinformation für Bürgerinnen und Bürger

Das Institut hat den gesetzlichen Auftrag, allgemein verständliche Gesundheitsinformationen für Bürgerinnen und Bürger zu erstellen. Es zählt nicht zu seinen Aufgaben, Einzelpersonen direkt zu beraten.

Für die Erstellung seiner Gesundheitsinformationen folgt das Institut den in Abschnitt 1.3 dargelegten Grundsätzen der evidenzbasierten Medizin. Dazu gehören die Anwendung systematischer Methoden zur wissenschaftlichen Bewertung von medizinischen Interventionen sowie die Einbindung der Betroffenenperspektive.

Methodische Grundlage der Erstellung von evidenzbasierten Gesundheitsinformationen ist ein systematisches Vorgehen mit dem Ziel, den aktuellen Stand des Wissens verständlich darzustellen, systematische Fehler (Bias) zu minimieren und Neutralität zu wahren.

Um das zu gewährleisten, beruhen evidenzbasierte Informationen auf folgenden Grundsätzen:

- systematische Recherche in Form einer fokussierten Informationsbeschaffung entsprechend den für die Zielgruppe relevanten Fragestellungen,
- begründete Auswahl der für die Fragestellung geeigneten Evidenz,
- nachvollziehbare und möglichst objektive Darstellung der für Betroffene relevanten Ergebnisse, wie zum Beispiel zur Sterblichkeit (Mortalität), zu den Beschwerden und Komplikationen (Morbidität) und zur gesundheitsbezogenen Lebensqualität,
- angemessene inhaltliche und sprachliche Darstellung von Unsicherheiten und ungeklärten Sachverhalten,
- Verzicht auf direktive Empfehlungen,
- Berücksichtigung der aktuellen Evidenz zur Risikokommunikation.

Diese Grundsätze gelten unabhängig vom Typ der Beauftragung einer Gesundheitsinformation. Es gibt 2 Möglichkeiten der Beauftragung:

- Direktaufträge des G-BA oder des BMG zur Erstellung von Patienteninformationen; diese Aufträge werden als Bericht, Rapid Report oder Addendum bearbeitet (siehe Abschnitte 2.1.1, 2.1.2 und 2.1.7),
- die aus der gesetzlichen Aufgabe des Instituts zur Bereitstellung von Gesundheitsinformationen für Bürgerinnen und Bürgern abgeleitete Beauftragung, konkretisiert durch einen Generalauftrag des G-BA (siehe Abschnitte 1.2 und 7.2).

7.1 Ziele der Gesundheitsinformationen

Primäres Ziel der Gesundheitsinformationen des Instituts ist die verständliche Vermittlung von entscheidungsrelevantem Wissen zu gesundheitlichen Fragen. Das Wissen soll Nutzerinnen

und Nutzer in ihrer Autonomie und Kompetenz unterstützen, informiert unter verschiedenen Optionen auszuwählen. Es soll den Nutzerinnen und Nutzern überlassen bleiben, wofür sie sich entscheiden. Die Informationen sind eingebettet in ein Verständnis von Shared Decision-Making, nach dem es keine Vorgaben gibt, auf welchem Weg Nutzerinnen und Nutzer zu ihren Entscheidungen kommen [214]. Ob sie die Entscheidung zum Beispiel an eine Ärztin oder einen Arzt delegieren, diese gemeinsam mit anderen oder allein treffen wollen, bleibt jedem selbst überlassen.

Die Informationen sollen so vermittelt werden, dass sie darüber hinaus die allgemeine Gesundheitskompetenz und die Wissenschaftskenntnisse (Health and Scientific Literacy) fördern.

Zusammenfassend bestehen folgende Ziele:

- verständliche Vermittlung von entscheidungsrelevantem Wissen zu gesundheitlichen Fragen, auch wenn es sich um komplexe wissenschaftliche Sachverhalte handelt,
- Unterstützung einer aktiven und informierten Entscheidungsfindung,
- Förderung der kritischen Nutzung gesundheitsbezogener Dienstleistungen,
- Verbesserung des Wissens um körperliche und seelische Gesundheit,
- Verbesserung des Verständnisses medizinischer und wissenschaftlicher Informationen, darunter auch des Konzepts der evidenzbasierten Medizin,
- Förderung der Unterstützung der Patientinnen und Patienten zum Beispiel durch Angehörige sowie Freundinnen und Freunde,
- Vermittlung von Informationen zur praktischen und emotionalen Bewältigung des Alltags,
- emotionale Unterstützung der Betroffenen und
- Hilfe bei der Orientierung im Gesundheitssystem.

Diese Ziele lassen sich als Empowerment zusammenfassen. Nach der Definition der Weltgesundheitsorganisation (WHO) beinhaltet der Begriff Empowerment im Gesundheitswesen die Befähigung, Entscheidungen zu treffen und Maßnahmen zu ergreifen, die im Einklang mit den eigenen Zielen stehen [546]. Gesundheitskommunikation zur Stärkung der Gesundheitskompetenz spricht an, was Bürgerinnen und Bürger wissen möchten, zeigt Interesse für und Respekt vor ihrer Meinung und erkennt ihre Kompetenz an [184,432,739].

Die besondere Herausforderung evidenzbasierter Gesundheitsinformationen liegt darin, diesen Anforderungen und Zielen gerecht zu werden und dabei für Nutzerinnen und Nutzer attraktiv und verständlich zu sein [176]. Die verschiedenen Anforderungen dürfen also nicht zu einer Überfrachtung einer Gesundheitsinformation führen. Das kann bedeuten, dass sich in der praktischen Umsetzung nicht alle Ziele und Qualitätsdimensionen gleichzeitig innerhalb einer einzelnen Information optimieren lassen.

Das Konzept beinhaltet, einzelne Texte und weitere Formate mit unterschiedlichen Schwerpunkten miteinander zu Informationspaketen angemessener Breite und Tiefe zu verknüpfen. Dem entspricht das Onlineangebot auf www.gesundheitsinformation.de.

7.2 Themenauswahl und Identifizierung der Informationsbedürfnisse

Gemäß § 139a Abs. 3 Nr. 6 SGB V fällt dem Institut folgende Aufgabe zu: „Bereitstellung von für alle Bürgerinnen und Bürger verständlichen allgemeinen Informationen zur Qualität und Effizienz in der Gesundheitsversorgung sowie zu Diagnostik und Therapie von Krankheiten mit erheblicher epidemiologischer Bedeutung“.

7.2.1 Themenkatalog gemäß dem gesetzlichen Auftrag

Eine breit akzeptierte Definition oder eine klar umrissene Zusammenstellung von Krankheiten mit „erheblicher epidemiologischer Bedeutung“ kann in der Literatur nicht ausgemacht werden. Ein grundlegender Aspekt der epidemiologischen Bedeutung ist die Häufigkeit einer Erkrankung. Daher wird angestrebt, Informationen zu den Diagnose- oder Krankheitsgruppen zu erstellen, von denen mindestens 1 Prozent der Bevölkerung dauerhaft (Prävalenz) oder bezogen auf den Zeitraum von 1 Jahr (Inzidenz) betroffen ist. Diese Schwelle wird auf nach Geschlecht und Alter (0 bis 17 Jahre, 18 bis 59 Jahre, 60 Jahre und älter) differenzierte Subpopulationen angewendet, um vorhandene geschlechts- und altersspezifische Besonderheiten dieser Gruppen möglichst zu berücksichtigen [444].

Zur Erstellung des Themenkatalogs wird primär der 2-jährlich aktualisierte Versorgungsreport des Wissenschaftlichen Instituts der Allgemeinen Ortskrankenkasse (WIdO) [282] herangezogen. Der Report beinhaltet auf Basis der Daten von etwa 24 Millionen Allgemeine-Ortskrankenkasse(AOK)-Versicherten Angaben zu Prävalenzen und Hospitalisierungsraten für die 1500 häufigsten Erkrankungsgruppen (3-stellige International-Statistical-Classification-of-Diseases-and-Related-Health-Problems[ICD]-10-Gruppen). Der Themenkatalog des Instituts wird regelmäßig überprüft und, wenn notwendig, angepasst. Dieser Themenkatalog kann erweitert werden, zum Beispiel in Abhängigkeit von Auftragsthemen des IQWiG oder um Erkrankungen, deren Bedeutung und Krankheitslast über die Prävalenz / Inzidenz nur unzureichend erfasst werden.

7.2.2 Identifizierung der Informationsbedürfnisse / Aufbereitung von Krankheitserfahrungen

Evidenzbasierte Gesundheitsinformationen orientieren sich im Idealfall an den Informationsbedürfnissen der Zielgruppe. Neben allgemeinen, für alle Zielgruppen relevanten Aspekten wie Häufigkeit und Krankheitsverlauf kann es bei jedem Thema zum Beispiel spezifische Probleme und Fragen, verbreitete Missverständnisse und Wissenslücken geben.

Patientenzentrierte Gesundheitsinformationen sollen nicht nur medizinisch-fachliche Fragen beantworten und eine informierte Entscheidung ermöglichen, sondern auch Fragen zur

Versorgung und zum Umgang mit einer Erkrankung beantworten sowie emotionale Unterstützung bieten [244]. Dafür ist es einerseits erforderlich zu wissen, auf welche Fragen Nutzerinnen und Nutzer Antworten suchen. Andererseits ist es für eine patientenzentrierte Vermittlung wichtig, dass die Autorinnen und Autoren des Instituts beim Schreiben der Informationen eine Vorstellung von der Situation und den Belastungen haben, die eine Krankheit für Patientinnen und Patienten und ihre Angehörigen mit sich bringen kann. Sie sollen ein Verständnis dafür entwickeln, was es bedeutet, mit einer bestimmten Krankheit zu leben.

Daher wird zur Bearbeitung eines Themas, in der Regel im Rahmen einer fokussierten Informationsbeschaffung (siehe Abschnitt 8.2), nach qualitativer Literatur recherchiert und diese ausgewertet, um Krankheitserfahrungen sowie Fragestellungen und Informationslücken zu identifizieren, die für Nutzerinnen und Nutzer von Interesse und Relevanz sind. Primär werden entsprechende Ergebnisse aus Deutschland ausgewertet. Fehlen diese [572], werden Informationsbedürfnisse aus Studien in vergleichbaren Ländern abgeleitet.

Weiterhin werden kostenlos zugängliche Gesundheitsinformationen anderer Anbieterinnen und Anbieter im Internet, beispielsweise von gesetzlichen Krankenkassen, Behörden, wissenschaftlichen Instituten und kommerziellen Unternehmen, gesichtet. Diese Sichtung liefert einen Überblick über das vorhandene Angebot von Gesundheitsinformationen und deren inhaltliche Schwerpunkte. Zudem werden themenabhängig Selbsthilfeorganisationen zum Informationsbedarf von Betroffenen und zu den Herausforderungen bei der Bewältigung der Erkrankung befragt.

Die Auswertung vermittelt ein Bild davon, welche Stationen Betroffene bei einem bestimmten gesundheitlichen Problem durchlaufen, welche psychologischen und emotionalen Probleme im Zusammenhang mit diesem Thema auftreten können, welche Informationsbedürfnisse bestehen und an welchen Punkten Entscheidungen getroffen werden müssen.

7.2.3 Patientenwege

Insbesondere bei chronischen Erkrankungen wird zur Orientierung über die mit einer Erkrankung verbundenen Fragen und Entscheidungen eine Art Landkarte erstellt, die die möglichen „Wege“ von Patientinnen und Patienten mit einer bestimmten Erkrankung nachzeichnet. Es soll versucht werden, möglichst umfassend nachzuvollziehen, welche Auswirkungen eine Erkrankung auf das Leben von Betroffenen haben kann und zu welchem Zeitpunkt an welcher Stelle wegweisende Entscheidungen anstehen.

Dies wird im Folgenden als Methode der Patientenwege bezeichnet.

Die Entscheidung, ob für ein Thema ein Patientenweg erstellt wird, hängt unter anderem von folgenden Kriterien ab:

- Handelt es sich um eine langwierige oder chronische Erkrankung, die für Patientinnen und Patienten verschiedene Stationen mit sich bringt (zum Beispiel Verarbeiten einer schwerwiegenden Diagnose, belastende Therapie, Bedarf an Nachsorge)?
- Stehen an diesen Stationen teilweise komplexe Entscheidungen an, beispielsweise für verschiedene Behandlungsoptionen mit spezifischem Nutzen- / Schadenprofil?
- Sind mehrere Ärztinnen oder Ärzte und andere Gesundheitsberufe in den Prozess eingebunden? Gibt es verschiedene Versorgungsoptionen (ambulant / [teil-]stationär)?

Patientenwege dienen der kurzen Zusammenfassung und tabellarischen Aufbereitung der verschiedenen sozialen, emotionalen, kognitiven und klinischen Dimensionen, die mit einer Krankheit einhergehen können. So soll erkennbar werden, vor welchen Herausforderungen und Entscheidungen Patientinnen und Patienten im Verlauf einer Erkrankung stehen. Die Methode lehnt sich an die medizinsoziologischen Modelle der sogenannten Illness Trajectory [146] und der Patientenkarriere [280,455] sowie an verschiedene Modelle der sogenannten Patient Journey [461] an.

Patientenwege sollen unter anderem helfen, den inhaltlichen Rahmen der Gesundheitsinformationen abzustecken. Dazu werden folgende Fragen gestellt:

- Wer könnte die Information lesen?
- Welche inhaltlichen Fragen könnten Leserinnen und Leser haben?
- In welchem emotionalen Zustand könnten sich die Leserinnen und Leser befinden?
- Zu welchem Zeitpunkt im Krankheitsverlauf werden welche Informationen möglicherweise benötigt?
- Vor welchen Entscheidungen stehen die Patientinnen und Patienten und wann und wo müssen diese Entscheidungen getroffen werden?
- Was sind die möglichen Auswirkungen von Gesundheitsinformationen zu diesem Thema?

Ziel dieser Methode ist es primär, dass die Autorinnen und Autoren des Instituts ein gutes Verständnis für Patientinnen und Patienten sowie Angehörige und ihre Interaktion mit Information entwickeln; dies wird durch die Orientierung an den in Tabelle 9 angegebenen Dimensionen unterstützt.

Tabelle 9: Verschiedene Dimensionen eines Patientenwegs

Leben / Alltag	Auswirkungen der Erkrankung auf soziale Beziehungen und Rollen: die Familie und Partnerschaft, den Beruf, die Lebensqualität, die „Funktionsfähigkeit“ etc.
Aktivwerden / Bewältigen	jegliche Tätigkeiten im Zusammenhang mit der Erkrankung, zum Beispiel Besuche bei einer Ärztin oder einem Arzt, Einnahme von Medikamenten, Informationssuche, Selbsthilfe
Fühlen	Empfindungen, die im Krankheits- und Behandlungsverlauf auftreten, wie Trauer, Ängste, Sorgen etc.
Wissen	Was wissen Bürgerinnen und Bürger bereits? Wo bestehen potenzielle Informationsbedürfnisse?
Entscheidungen	Welche Entscheidung muss der oder die Betroffene in der jeweiligen Phase treffen?
Klinik	Beschreibung der medizinischen Phasen, zum Beispiel Risikofaktoren, Symptome, Diagnose, Behandlung, Rehabilitation
Kontaktstelle im Gesundheitswesen	Wer ist in der jeweiligen Phase die Ansprechpartnerin oder der Ansprechpartner im Gesundheits- oder Sozialsystem, zum Beispiel Ärztinnen und Ärzte, Pflegekräfte, Physiotherapeutinnen und Physiotherapeuten, Psychotherapeutinnen und Psychotherapeuten, Sozialarbeiterinnen und Sozialarbeiter, Beratungsstellen, Versicherungen?

Als Quellen für den Entwurf eines Patientenwegs dienen unter anderem qualitative Literatur (qualitative Studien und deren Übersichten), evidenzbasierte Leitlinien nationaler und internationaler Fachgesellschaften, Evidenzsynthesen, Literatur zu den Informationsbedürfnissen von Patientinnen und Patienten und zur Versorgungssituation sowie Erfahrungsberichte [450].

7.3 Informationsbeschaffung zur Erstellung von Gesundheitsinformationen

Grundlage der Erstellung evidenzbasierter Gesundheitsinformationen ist eine systematische Recherche in Form einer fokussierten Informationsbeschaffung der für die Fragestellung angemessenen aktuellen Literatur. In der Regel finden zu jedem Thema 2 Recherchen statt:

- Nach qualitativer Literatur (siehe Abschnitt 8.2.2): Die Ergebnisse dienen der Erfassung der Bedürfnisse nach zusätzlichen relevanten Informationen wie Krankheitserfahrungen. Diese Recherche soll den Autorinnen und Autoren ermöglichen, sich so gut wie möglich in die Situation Betroffener zu versetzen.
- Nach systematischen Übersichten (siehe Abschnitt 8.2.1): Die Ergebnisse bilden die Grundlage für Aussagen zum Nutzen und Schaden medizinischer Interventionen.

Ergänzend finden zu diesen Recherchen noch von der Fragestellung abhängige orientierende Recherchen (siehe Abschnitt 8.3) statt, zum Beispiel nach der Erkrankungshäufigkeit.

Die Informationsbeschaffung für systematische Übersichten zielt darauf, solche Übersichten zu allen Fragestellungen oder Teilaspekten einer Gesundheitsinformation zu identifizieren, wie zum Beispiel zu den Ursachen, zum Verlauf und zur Prognose, zur Therapie und Epidemiologie einer Erkrankung. Berücksichtigt werden in erster Linie systematische Übersichten mit Suchen,

die in den letzten 3 Jahren durchgeführt wurden [657,658]. Systematische Übersichten mit einer älteren Suche sind im Hinblick auf die kurze Halbwertszeit medizinischen Wissens möglicherweise veraltet. Themenspezifisch, beispielsweise abhängig von der Forschungsaktivität, kann dieser Zeitraum auch verändert werden.

Die identifizierten Übersichten werden dann den Fragestellungen zugeordnet. Ein ressortinternes redaktionelles und wissenschaftliches Scoping gleicht die Ergebnisse der Informationsbeschaffung mit den identifizierten Informationsbedürfnissen ab.

Sofern ein zentrales Informationsbedürfnis, beispielsweise nach Langzeitdaten, nicht durch eine aktuelle systematische Übersicht abgedeckt wird, kann eine fokussierte Informationsbeschaffung zu Primärstudien in Betracht gezogen werden.

7.4 Auswahl der Evidenz

Für evidenzbasierte Gesundheitsinformationen sind die für die Zielgruppe relevanten vorhandenen wissenschaftlichen Erkenntnisse aus den aktuellen und ausreichend zuverlässigen Studien zu berücksichtigen. Welche Studientypen angemessen sind, hängt von der Fragestellung ab. Aussagen zum Nutzen und Schaden von Interventionen stützen sich in der Regel auf systematische Übersichten von randomisierten kontrollierten Studien (siehe Abschnitt 8.2). Damit eine systematische Übersicht über die Wirkung einer Intervention für Gesundheitsinformationen verwendet werden kann, muss diese gewisse qualitative Mindestvoraussetzungen erfüllen [408,557,559]. Die Bewertung der Qualität erfolgt auf Basis des Oxman-Guyatt-Index [556,557,559]. Bewertet werden 9 Items, unter anderem zur Qualität der Informationsbeschaffung, zur Studienselektion und zur Evidenzsynthese. Um sich als Basis für eine Aussage zum Nutzen und Schaden einer medizinischen Intervention zu eignen, muss eine systematische Übersicht bei diesen Items allenfalls kleinere Mängel („minor flaws“) haben. Das entspricht einem Mindestscore von 5 Punkten. Weiterhin muss das Verzerrungspotenzial der eingeschlossenen Studien regelhaft auf Endpunktebene bewertet worden sein. Zudem wird der Umgang der Reviewautorinnen und -autoren mit einem möglichen Publication Bias beurteilt. Wurde beispielsweise spezifisch nach unpublizierten Daten gesucht oder wurden statistische Tests angewandt, um einen Publication Bias zu identifizieren? Die fehlende Adressierung des Publication Bias oder ein nicht angemessener Umgang kann von einer entsprechenden Darstellung der qualitativen Ergebnisunsicherheit in der resultierenden Gesundheitsinformation bis hin zum Ausschluss eines systematischen Reviews führen.

Eine Rolle bei der Auswahl spielt zudem, ob sich die Ergebnisse auf den deutschen Versorgungskontext anwenden lassen. Dabei werden Aspekte wie Studienpopulation, Zulassungsstatus und Verbreitung der Intervention berücksichtigt. Sofern Aussagen nur für bestimmte Gruppen zutreffen, wird das bei der Darstellung erläutert.

Werden innerhalb einer Gesundheitsinformation Nutzensaussagen auf Basis verschiedener Evidenzsynthesen von unterschiedlicher qualitativer Ergebnissicherheit gemacht, wird dies bei der Darstellung der Ergebnisse adressiert.

Wird ein bestimmtes Thema oder Resultat in mehr als 1 systematischen Übersicht von angemessener methodischer Qualität untersucht, findet eine weitergehende Bewertung der Übersichten statt. So sollen die qualitativ besten Übersichten identifiziert werden. Dazu werden neben den bereits beschriebenen noch folgende Aspekte betrachtet:

- inhaltlicher Schwerpunkt des Reviews, insbesondere auch Relevanz der Fragestellung für eine Patienteninformation,
- Durchführung von Sensitivitätsanalysen und Umgang mit Heterogenität.

Als Quelle für die in den Gesundheitsinformationen dargestellten numerischen Daten dienen die Ergebnisse der Übersicht, die im Hinblick auf die Fragestellung die höchste Qualität aufweist. Wenn Übersichten zu widersprüchlichen Schlussfolgerungen gelangen, wird nach den möglichen Gründen dafür gesucht [409]. Dies können beispielsweise Unterschiede im Studienpool, in der statistischen Auswertung oder in der Interpretation der Ergebnisse sein.

Bei Fragestellungen etwa zur Ätiologie oder Prognose können auch systematische Übersichten auf der Basis von anderen Studientypen als RCTs ausgewertet werden [286]. Bei der Bewertung von solchen systematischen Übersichten werden zur Orientierung die Kriterien des Oxford Centre for Evidence-Based Medicine eingesetzt [127,345]. Die Methoden zur Bewertung qualitativer Studien werden in Abschnitt 9.4 beschrieben.

7.5 Auswahl der dargestellten Ergebnisse (Endpunkte)

Informationen über Behandlungsergebnisse beziehen sich auf Endpunkte, die für Patientinnen und Patienten relevant sind, also insbesondere auf die Sterblichkeit (Mortalität), die Beschwerden und Komplikationen (Morbidität) und die gesundheitsbezogene Lebensqualität. Dafür gelten grundsätzlich die in den Abschnitten 3.1 und 3.2 der Methoden beschriebenen Grundlagen.

Zudem sind die Begleitumstände der Behandlung (z. B. Zeitaufwand, körperliche, seelische, soziale und auch finanzielle Belastungen) für Patientinnen und Patienten oft wichtige Informationen.

7.6 Wahl und Darstellung von Vergleichen

Um Nutzerinnen und Nutzern von gesundheitsinformation.de für die in Abschnitt 7.5 genannten patientenrelevanten Endpunkte eine Abwägung des Nutzens und Schadens zu ermöglichen, erfolgt ein Vergleich zwischen der Maßnahme und dem Verzicht darauf oder einer anderen üblichen Behandlung.

7.7 Umgang mit Zahlen und Risikoangaben

Die ausgewogene Darstellung des möglichen Nutzens und Schadens einer Intervention erfordert in der Regel eine Quantifizierung von Effekten und Risiken. Dabei wird aber darauf

geachtet, dass Zahlen und Risikoangaben so dosiert werden, dass sie eine Information nicht überfrachten und nicht den Lesefluss und das Verständnis behindern.

Da die verbale Umschreibung von Häufigkeiten oft nicht zu einer realistischen Einschätzung führt, werden Zahlen bevorzugt [101].

Bei der Angabe von Zahlen und Wahrscheinlichkeiten werden folgende Grundsätze beachtet:

- Der Effekt einer medizinischen Maßnahme wird durch die Darstellung der absoluten Ereignishäufigkeiten in den zu vergleichenden Gruppen angegeben. Eventuelle Unsicherheiten werden benannt. Die Bezugsgröße wird nach Möglichkeit so gewählt, dass die Größe des Effekts leicht zu erfassen ist.
- Für den Nutzen und den Schaden werden – soweit möglich – einheitliche Bezugsgrößen gewählt. Ausgangspunkt eines Vergleichs ist ein Basisrisiko, zum Beispiel der natürliche Krankheitsverlauf. Dies meint die Wahrscheinlichkeit, mit der sich ein Beschwerdebild auch ohne medizinische Intervention verbessert, verschlechtert oder konstant bleibt. Wenn sich ein Beschwerdebild auch ohne Behandlung bessern kann, wird das den Nutzerinnen und Nutzern vermittelt.
- Wenn es sinnvoll ist, werden zusätzlich zur absoluten Risikoänderung auch relative Änderungen dargestellt.

Ob Effekte als „Gewinn“ oder „Verlust“ dargestellt werden, hängt von der Intervention und der Perspektive der Betroffenen ab. Für den Nutzen und Schaden wird diesbezüglich ein einheitlicher Bezugsrahmen gewählt. Studien, die den Nutzen einer ergänzenden grafischen Aufbereitung numerischer Darstellungen untersuchen, zeigen zu den Endpunkten Wissen, Risikowahrnehmung und Verständlichkeit kein konsistentes Ergebnis. Die evidenzbasierte Leitlinie „Evidenzbasierte Gesundheitsinformation“ kommt in Bezug auf diese Frage auf Basis einer niedrigen Evidenzqualität zu einer Kann-Empfehlung [487].

Bei der Erstellung von Gesundheitsinformationen werden Grafiken vor allem zur ergänzenden Darstellung komplexer Sachverhalte eingesetzt, beispielsweise regelhaft in den Entscheidungshilfen zu Krebsfrüherkennungs-Untersuchungen.

Für die Darstellung in einer Gesundheitsinformation werden die in einer Metaanalyse verwendeten relativen Effektmaße in absolute überführt. Bei der Berechnung der absoluten Effektmaße findet das gleiche Vorgehen Anwendung wie bei der Erstellung der Ergebnistabellen innerhalb eines Cochrane Reviews [353]. Dabei wird die Unsicherheit des Effektmaßes, jedoch nicht die Unsicherheit des Basisrisikos berücksichtigt [161,675].

Grundlage der Berechnungen ist ein gepoolter Effektschätzer aus einer hinreichend homogenen Metaanalyse. Ist das Effektmaß bereits eine Risikodifferenz, so wird diese für die weitere Betrachtung verwendet. Wird ein relatives Effektmaß angegeben, wird für die Herleitung der

absoluten Risikodifferenz zunächst ein plausibles Basisrisiko gewählt. Dieses basiert in der Regel auf dem Median des Risikos der Kontrollgruppen in den eingeschlossenen Einzelstudien.

In begründeten Fällen kann das Basisrisiko auch aus einer geeigneten Einzelstudie (beispielsweise der Studie mit der mit Abstand größten Population oder der höchsten externen Validität) des eingeschlossenen Studienpools oder einer validen externen Quelle (beispielsweise einer deutschen Prävalenzstudie oder aus Registerdaten) hergeleitet werden. Für den Fall, dass das Basisrisiko aus einer externen Quelle kommt und eine hohe Unsicherheit aufweist, wird diese Unsicherheit berücksichtigt. Dabei findet die von Newcombe und Bender beschriebene Methode Anwendung [539].

Auf Grundlage des angenommenen Basisrisikos werden mithilfe des relativen Gesamtschätzers (OR, RR, HR) der Metaanalyse das absolute Risiko in der Interventionsgruppe und die absolute Risikodifferenz errechnet.

Bei heterogenen Basisrisiken der eingeschlossenen Einzelstudien wird jeweils der absolute Effekt für verschiedene angenommene Basisrisiken dargestellt (beispielsweise für ein niedriges und hohes Basisrisiko). Ist dies inhaltlich nicht sinnvoll, wird auf eine Darstellung des absoluten Effektes verzichtet.

7.8 Berücksichtigung von Alters- und Geschlechtsunterschieden

Der natürliche Krankheitsverlauf, die Risiken, Symptome, Morbidität, Mortalität, Wirkungen und unerwünschte Wirkungen von Interventionen, die gesundheitsbezogene Lebensqualität und die Begleitumstände einer Behandlung können je nach Alter oder Geschlecht variieren. Sofern in der identifizierten Literatur zu einer Erkrankung bedeutsame Unterschiede beschrieben werden, werden diese bei der Erstellung einer Information berücksichtigt.

Bei vielen Themen ergeben sich geschlechts- und altersspezifische Aspekte bereits aus der Epidemiologie der Erkrankung. Wenn sich eine Information an die von einer Erkrankung hauptsächlich betroffene Gruppe richtet, wird die Zielgruppe in den Informationen nicht explizit genannt.

Sofern zwischen verschiedenen Gruppen keine relevanten Unterschiede identifiziert werden, wird auf das Fehlen dieser Unterschiede nicht ausdrücklich hingewiesen.

Sofern sich Ergebnisse auf Teilgruppen beziehen, wird das jedoch deutlich gemacht.

Ein neutraler Sprachstil muss gewährleisten, dass Informationen sowohl Frauen als auch Männer erreichen und sich beide Geschlechter gleichermaßen angesprochen fühlen. Die durchgehende Verwendung maskuliner Personenbezeichnungen (generisches Maskulinum) führt zu einer gedanklichen Unterrepräsentation und sprachlichen Benachteiligung von Frauen [402]. In den Texten der Gesundheitsinformationen wird daher ein weitestgehend geschlechtsneutraler Sprachstil verwendet, der möglichst vollständig auf das generische Maskulinum

verzichtet. Stattdessen werden, wo immer möglich, beide Geschlechter explizit genannt, wenn auch beide gemeint sind, oder geschlechtsneutrale Formulierungen gewählt.

7.9 Anpassung an die Zielgruppe

Eine der wichtigsten Herausforderungen im Zusammenhang mit evidenzbasierten Gesundheitsinformationen besteht darin, die Informationen verständlich, dabei jedoch wissenschaftlich präzise und möglichst objektiv zu präsentieren. Zudem sollen die Gesundheitsinformationen des Instituts eine heterogene Zielgruppe ansprechen, deren Mitglieder sich unter anderem durch ihre Kompetenzen, Erkrankungen, ihr Bildungsniveau, ihren persönlichen Hintergrund, ihr Alter und Geschlecht unterscheiden. Auch Besonderheiten wie der muttersprachliche und kulturelle Hintergrund sind für bestimmte Zielgruppen von Bedeutung.

In der Bearbeitung eines Themas wird geprüft, ob sich aus der Epidemiologie der Erkrankung Anforderungen für bestimmte Zielgruppen ergeben.

Um die Informationen an besondere Zielgruppen anzupassen, können folgende Optionen kombiniert werden:

- Surveys, qualitative Primärstudien und Übersichten von qualitativen Studien zum Informationsbedarf der Bevölkerung,
- Erfahrungen von anderen Informationsanbietern, Patientenberatungsstellen und Selbsthilfegruppen,
- Gespräche mit Betroffenen (siehe Abschnitt 9.4.2),
- Erhebung von Erfahrungsberichten als eigenes Format (siehe Abschnitt 7.16.2).

7.9.1 Beteiligung von Betroffenen

Es gibt Hinweise darauf, dass die Beteiligung von Betroffenen an der Erstellung von Gesundheitsinformationen deren Relevanz steigern kann [541]. Die Orientierung an der Sichtweise und an den Informationsbedürfnissen der Bürgerinnen und Bürger zählt zu den Anforderungen an evidenzbasierte Gesundheitsinformationen [175] und ist ein zentraler Bestandteil bei der Erstellung von Gesundheitsinformationen [776]. Dabei kommen verschiedene Maßnahmen zum Einsatz. Im Rahmen der Themenpriorisierung und -fokussierung werden etwa Themenvorschläge der Nutzerinnen und Nutzer und Erfahrungen im Rahmen von Konsultationen mit Selbsthilfegruppen berücksichtigt.

Betroffene werden durch Anfragen an die Patientenvertretung im G-BA und durch Kontakt zu Selbsthilfegruppen identifiziert.

Außerdem können bei geeigneten Themen von Patientinnen und Patienten sowie deren Angehörigen Erfahrungsberichte erhoben werden, um diese individuellen Erfahrungen und Erlebnisse mit Krankheit und Pflegebedürftigkeit in ihren verschiedenen Dimensionen anderen

Betroffenen beziehungsweise Interessierten als Ergänzung zu den Gesundheitsinformationen zur Verfügung zu stellen (siehe Abschnitt 7.16.2).

7.9.2 Nicht öffentliches Stellungnahmeverfahren

Im nicht öffentlichen Stellungnahmeverfahren wird u. a. dem Kuratorium nach Abschluss der internen Qualitätssicherung und Begutachtung durch eine externe Sachverständige oder einen externen Sachverständigen die Möglichkeit zur Kommentierung der Textentwürfe gegeben. Dem Kuratorium gehören auch Vertreterinnen und Vertreter relevanter Institutionen für die Wahrnehmung von Patienteninteressen und von Selbsthilfeorganisationen chronisch kranker und behinderter Menschen an. Nach Ablauf der Stellungnahmefrist sichten die Mitglieder des an der Erstellung beteiligten Teams und ein Mitglied der Ressortleitung die in den Stellungnahmen vorgebrachten Argumente hinsichtlich ihrer Relevanz für die Gesundheitsinformation. Im Rahmen einer Stellungnahmeredaktionskonferenz werden die Kommentare und Argumente der Stellungnehmenden diskutiert und, soweit vorhanden, die entsprechende Evidenz bewertet. Ein möglicher Änderungsbedarf wird konsentiert und dokumentiert. Zeitnah zur Veröffentlichung des finalen Textes werden die Stellungnahmen gewürdigt. Die Würdigung erfolgt regelhaft bei inhaltlichen Änderungsvorschlägen und fristgerechtem Eingang. Stellungnahmen, die nur wenige redaktionelle Anmerkungen enthalten, und verfristete Stellungnahmen werden in der Regel nicht gewürdigt. Stellungnahmen und Würdigungen werden nicht veröffentlicht.

7.9.3 Nutzertestung

Das primäre Mittel, um die Verständlichkeit von Texten zu prüfen, ist die Begutachtung von Entwürfen durch Testleserinnen und Testleser. In der Regel durchlaufen Entwürfe der Gesundheitsinformationen zeitgleich mit dem Stellungnahmeverfahren eine standardisierte externe Nutzertestung. Diese erfolgt durch eine externe Auftragnehmerin oder einen externen Auftragnehmer in Form von Fokusgruppen oder Einzelinterviews. Themenabhängig können auch Patientenorganisationen oder andere Verbände und Institutionen einbezogen werden.

Bei der Nutzertestung kommt ein Methodenmix aus dokumentierter Einzelbewertung und Gruppendiskussion bzw. Einzelinterviews zur Anwendung. Ob eine Gruppendiskussion oder Einzelinterviews durchgeführt werden, wird themenspezifisch entschieden, beispielsweise können sich bei tabuisierten Themen oder bestimmten Zielgruppen Einzelinterviews eher eignen. Anhand eines semistrukturierten Leitfadens testen in der Regel 5 Betroffene oder potenzielle Nutzerinnen und Nutzer die Texte im Hinblick auf ihren Informationsgehalt und ihre Verständlichkeit. Die Testerinnen und Tester werden anhand der vom IQWiG vorgegebenen Kriterien vom externen Auftragnehmer oder von der externen Auftragnehmerin rekrutiert. Die Auswahlkriterien beziehen sich auf die in den zu testenden Informationsmaterialien gesetzten Schwerpunkte und umfassen in der Regel Angaben zu Diagnose, Geschlecht, Alter, Erfahrungen mit bestimmten Behandlungsoptionen und ggf. zum sozialen Status.

Die Ergebnisse der Nutzertestungen fließen in die Überarbeitung der Entwürfe der Gesundheitsinformationen ein.

7.9.4 Kommentare von Nutzerinnen und Nutzern

Darüber hinaus haben die Nutzerinnen und Nutzer der Website www.gesundheitsinformation.de die Möglichkeit, eine Rückmeldung zu geben. Dazu werden auf der Website verschiedene Kanäle angeboten:

- eine Kommentarfunktion zu den einzelnen Texten,
- ein allgemeines Online-Kontaktformular,
- eine zufallsgesteuerte Befragung einzelner Nutzerinnen und Nutzer zur Bewertung der Website.

7.9.5 Informationszugang und Barrierefreiheit

Da die Informationen in erster Linie im Internet zur Verfügung gestellt werden, erfüllt die Website gesundheitsinformation.de die deutsche Barrierefreie-Informationstechnik-Verordnung (BITV) [104].

Die Gesundheitsinformationen werden sowohl in deutscher als auch in englischer Sprache veröffentlicht. Die Verfügbarkeit einer englischsprachigen Version erweitert die Möglichkeiten der Nutzung für Fremdsprachlerinnen und -sprachler und die Übersetzung der Texte in andere Sprachen.

7.10 Sachlich angemessene Darstellung

Informationen im Zusammenhang mit Entscheidungen über diagnostische und therapeutische Maßnahmen sollen ein realistisches, weder in einer direktiven noch in einer wertenden Sprache formuliertes und in einem angemessenen Bezugsrahmen (sog. Framing) dargestelltes Bild des Wissens vermitteln. Tendenziöse und insbesondere unangemessen beunruhigende Formulierungen sind ebenso zu vermeiden wie verharmlosende Darstellungen. Bedeutsame Unsicherheiten sollen inhaltlich und sprachlich in geeigneter Form umgesetzt werden.

Um diese Anforderung an die Autorinnen und Autoren in der täglichen Arbeit zu vermitteln, wird ein Leitfaden (Styleguide) zur Textgestaltung eingesetzt [534]. Dieser wird, basierend auf der Evaluation der Produkte sowie neuer Evidenz aus dem Bereich der evidenzbasierten Kommunikation, kontinuierlich weiterentwickelt.

Um eine sachlich angemessene Darstellung zu erreichen, durchlaufen die Informationen einen mehrstufigen redaktionellen Prozess, inklusive einer IQWiG-internen Qualitätssicherung, eines externen Reviews, einer Nutzertestung und nicht öffentlichen Stellungnahme (siehe Abschnitt 7.9).

7.11 Ableitung von Bewertungen und Empfehlungen

Grundlage für eine individuelle Abwägung des Nutzens und Schadens sind möglichst genaue Informationen über patientenrelevante Ergebnisse. Damit sollen die Nutzerinnen und Nutzer die Möglichkeit erhalten, Entscheidungen zu treffen, die ihren Werten und Präferenzen entsprechen.

Die Evidenz zu erklären und bei der Vermittlung gesundheitsbezogener Informationen neutral zu bleiben, stellt eine besondere Herausforderung dar [218,433,642,716]. In den Gesundheitsinformationen werden in der Regel keine Empfehlungen ausgesprochen. Diesem Anspruch wird durch eine nicht direktive und nicht wertende Sprache Rechnung getragen. Ausnahmen sind möglich, zum Beispiel bei der Darstellung des Umgangs mit Notfällen.

7.12 Vorgehen bei der Erstellung von Entscheidungshilfen

Eine Technik, die Patientinnen und Patienten bei der persönlichen Abwägung des Nutzens und Schadens helfen soll, sind Entscheidungshilfen. Für die Inhalte gelten die allgemeinen Anforderungen an Gesundheitsinformationen. Die Entwicklung von Entscheidungshilfen orientiert sich an den International Patient Decision Aid Standards (IPDAS) [215,364].

7.13 Transparenz bezüglich der Verfasserin / des Verfassers und der Herausgeberin / des Herausgebers

Auf den Websites gesundheitsinformation.de und iqwig.de stellt das Institut seine Grundlagen und Finanzierung als gemeinnütziger und wissenschaftlich unabhängiger Herausgeber von Gesundheitsinformationen dar. Die Selbstangaben gehen über die gesetzlichen Anforderungen hinaus und erfüllen weitergehende Transparenzkriterien.

7.14 Offenlegung von Beziehungen

Das Institut fordert von seinen Mitarbeiterinnen und Mitarbeitern eine sehr weitgehende Offenlegung von Beziehungen.

Sofern Betroffene und externe Gutachterinnen oder Gutachter in die Erstellung von Gesundheitsinformationen einbezogen sind, gilt die generelle Handhabung wie für alle anderen Produkte des Instituts (siehe Abschnitt 2.2).

7.15 Benennung von Wirkstoffen, medizinischen Methoden und Geräten

Bei der Benennung von Wirkstoffen und medizinischen Methoden und Geräten werden generische Namen bevorzugt. Da Menschen aber zum Beispiel Wirkstoffnamen oft nicht kennen und Handelsnamen verwenden, können in den Gesundheitsinformationen des Instituts ergänzend auch Handelsnamen genannt werden.

7.16 Beschreibung der typischen Formate und Inhalte

Der Kern der Website gesundheitsinformation.de ist auf die Darstellung gesundheits- oder krankheitsbezogener Themen ausgerichtet. Ein Thema kann unterschiedliche Text- und Informationsformate beinhalten. Die einzelnen Formate sollen die wesentlichen Aspekte des jeweiligen Themas abdecken und zentrale Fragen der Nutzerinnen und Nutzer beantworten. Diese einzelnen Formate sollen zudem unterschiedlichen Informationsbedürfnissen verschiedener Adressatengruppen gerecht werden.

Zu den Hauptformaten gehören:

- **Überblick:** Der Überblick führt in das Thema ein und bietet die Grundlage für und Anknüpfung an die in der Folge aufgeführten vertiefenden Informationsformate. Der Überblick ist nach einer festen Struktur gegliedert.
- **Mehr Wissen:** Dieses Format informiert ausführlicher über bestimmte Aspekte eines Themas, wie zum Beispiel über medikamentöse und nicht medikamentöse Behandlungsoptionen einer Erkrankung oder bestimmte diagnostische Maßnahmen. „Mehr Wissen“ beschreibt, wenn möglich, auch die Vor- und Nachteile einzelner Behandlungsmöglichkeiten oder bei mangelnder Evidenz auch die resultierenden Unsicherheiten. Ein weiterer Inhalt von „Mehr Wissen“ sollen Ausführungen zum Leben und Umgang mit einer Erkrankung sein. Dabei wird versucht, sowohl die Perspektive der unmittelbar Betroffenen als auch diejenige der Angehörigen zu berücksichtigen. Das Format kann durch Grafiken und Multimediaelemente unterstützt werden.
- **Was Studien sagen:** Diese Texte fassen den aktuellen Wissensstand zu einer im Titel formulierten Frage zusammen. Sie beruhen auf den Ergebnissen qualitativ hochwertiger systematischer Evidenzsynthesen. Sie beschreiben die Studien ausführlicher und erklären, wie die Antwort auf die Forschungsfrage gefunden wurde.

7.16.1 Ergänzende Formate

Die Hauptformate können um verschiedene ergänzende Formate erweitert werden, zum Beispiel um einzelne Aspekte eines Themas zu vertiefen oder andere Zugangswege anzubieten. Beispielsweise kann die Integration von Bildern, Ton und Animationen dazu beitragen, die Attraktivität und Verständlichkeit der Website zu erhöhen, vor allem für Menschen mit eingeschränkter Lesekompetenz.

Als ergänzende Formate sind zu verstehen:

- Erfahrungsberichte von Betroffenen (Näheres findet sich in Abschnitt 7.16.2),
- Grafiken, Fotos und anderes Bildmaterial,
- Animationen mit Ton und Bild,
- Quiz,

- Glossar,
- „Kurz erklärt“: Texte, die themenübergreifend Anatomie, Körperfunktionen, Behandlungs- und Untersuchungsverfahren sowie Prinzipien und Methoden der evidenzbasierten Medizin erklären,
- (Um-)Rechner.

Diese ergänzenden Formate sollen

- das allgemeine Verständnis von gesundheitsbezogenen und medizinischen Fragen fördern;
- Nutzerinnen und Nutzern helfen, die potenziellen Vor- und Nachteile medizinischer Interventionen zu verstehen und gegeneinander abzuwägen;
- Selbstmanagementstrategien unterstützen.

Interaktive Elemente werden in der Regel auch einer externen Nutzertesting unterzogen.

7.16.2 Erfahrungsberichte

Wissenschaftliche Evidenz kann unter anderem anhand von Erfahrungsberichten dargestellt und der Öffentlichkeit zugänglich gemacht werden [287]. Die Bedeutung von Erfahrungsberichten in der medizinischen Praxis und im Gesundheitswesen wird zunehmend wahrgenommen [298,682,773]. Viele Patientinnen und Patienten möchten von den Erfahrungen anderer Betroffener mit derselben Erkrankung hören oder darüber lesen [348,697].

Was Erfahrungsberichte ausmacht, ist die Darstellung des individuellen Erlebens und Umgangs mit einer Situation. Einige Funktionen von Erfahrungsberichten sind folgende [697]:

- Sie bieten die Möglichkeit, eigene Erfahrungen mit den Erfahrungen von anderen zu vergleichen.
- Das Lesen über die Gefühle anderer kann eine „Erlaubnis“ für das Zulassen ähnlicher Emotionen sein.
- Sie können Betroffenen zeigen, dass sie mit ihren Erfahrungen nicht allein sind.

Mit der Erhebung von Erfahrungsberichten von Patientinnen und Patienten sowie deren Angehörigen stellt das Institut individuelle Erfahrungen und Erlebnisse mit Krankheit und Pflegebedürftigkeit in ihren verschiedenen Dimensionen anderen Betroffenen beziehungsweise Interessierten als Ergänzung zu den Gesundheitsinformationen zur Verfügung.

Erfahrungsberichte können aber von manchen Menschen als Empfehlung wahrgenommen werden, ähnliche Entscheidungen zu treffen. Diese Wirkung kann dann kollidieren mit dem Anspruch, neutrale Informationen zu erstellen. Damit die auf gesundheitsinformation.de

veröffentlichten Erfahrungsberichte primär einen Zugang zum Erleben einer Erkrankung und zum Umgang mit ihren Folgen schaffen, sind sie redaktionell so bearbeitet, dass sie:

- keine Widersprüche zu Evidenzaussagen in anderen Texten enthalten,
- keine expliziten Empfehlungen enthalten und
- die geschilderten Optionen denen entsprechen, die üblich sind.

Erfahrungsberichte werden folgendermaßen erhoben:

- 1) Suche nach Interviewpartnerinnen und -partnern, insbesondere über Selbsthilfeorganisationen, Patientenuniversitäten und Arztpraxen,
- 2) Einholen der schriftlichen Einverständniserklärung bezüglich der Durchführung und Nutzung des Interviews,
- 3) Durchführung der Interviews (in der Regel telefonisch),
- 4) Dokumentation, redaktionelle Bearbeitung und Einholen der Einverständniserklärung für die Veröffentlichung der endgültigen Version,
- 5) Veröffentlichung auf der Website.

Die redaktionelle Bearbeitung der Erfahrungsberichte umfasst die Überführung des gesprochenen Wortes ins Schriftdeutsch und in der Regel eine Kürzung der Inhalte auf einen im Internet lesbaren Umfang. Die Kürzung und Zusammenfassung der Inhalte orientiert sich an den im Vorfeld festgelegten Schwerpunkten der Gesundheitsinformationen, vor allem in den Bereichen Leben mit der Erkrankung, Erleben der Beschwerden und Umgang mit der Diagnose, dem Krankheitsverlauf und den Folgen einer Erkrankung. Dieser Prozess erfolgt in enger Absprache mit den Interviewpartnerinnen und Interviewpartnern.

Die Methodik zur Erhebung, Bearbeitung und Veröffentlichung von Erfahrungsberichten orientiert sich an den etablierten Vorgehensweisen der Erstellerinnen und Ersteller der Database of Individual Patient Experience (DIPEX) [183].

Die Entscheidung, ob zu einem Thema Erfahrungsberichte erhoben werden, orientiert sich unter anderem an folgenden Kriterien:

- an den möglichen Auswirkungen einer Erkrankung auf das Leben der Patientinnen und Patienten hinsichtlich physischer, psychischer und sozialer Aspekte,
- an der möglichen Dauer und Chronifizierungshäufigkeit einer Erkrankung,
- daran, inwieweit Aspekte einer Erkrankung gesellschaftlich tabuisiert sind und damit womöglich ein Austausch im sozialen Umfeld erschwert ist, beispielsweise bei psychischen Erkrankungen oder Erkrankungen der Geschlechtsorgane.

7.16.3 Website

Die Verbreitung der Gesundheitsinformationen erfolgt in erster Linie über die 2-sprachige Website www.gesundheitsinformation.de bzw. www.informedhealth.org. Ziel ist die Gewährleistung eines hohen Websitestandards im Hinblick auf:

- Benutzerfreundlichkeit und Barrierefreiheit [375,454,540],
- Datenschutz [379],
- Transparenz,
- Sichtbarkeit für Suchmaschinen [711],
- Attraktivität für Nutzerinnen und Nutzer.

Die Website bietet einen kostenlosen elektronischen Newsletter, den man wahlweise 2-wöchentlich oder monatlich abonnieren kann. Er enthält Informationen zu Neuigkeiten auf der Website und zur Aktualisierung von Informationen. Inhalte der Website werden auch im Atom- und RSS-Format bereitgestellt und sind mit üblichen Readern lesbar. Darüber hinaus können die Inhalte der Website über eine Schnittstelle von geeigneten Partnerinnen oder Partnern automatisiert in eigene Websites eingebunden werden.

Die Website gesundheitsinformation.de ist von der Health on the Net Foundation (HON) zertifiziert und erfüllt die 8 Prinzipien des HON Code of Conduct (HONcode) für medizinische und Gesundheitswebsites sowie die Anforderungen der Guten Praxis Gesundheitsinformation [176].

7.17 Aktualisierung der Inhalte

Ein entscheidender Faktor bei der Bereitstellung von evidenzbasierten Gesundheitsinformationen ist die Gewährleistung, dass die Schlussfolgerungen nicht veraltet sind. Eine regelmäßige Aktualisierung ist eines der Qualitätskriterien, die die Europäische Union für gesundheitsbezogene Internetseiten festgelegt hat [142] und die das deutsche Grundsatzpapier Gute Praxis Gesundheitsinformation [175] vorsieht.

Eine Studie zu Leitlinienempfehlungen kam zu dem Schluss, dass nach 3 Jahren mehr als 90 %, nach 6 Jahren jedoch nur noch etwa 50 % der in Leitlinien abgegebenen Empfehlungen aktuell sein dürften [657]. Bei manchen Themen, zu denen beispielsweise sehr starke Evidenz vorliegt, kann die Halbwertszeit der Evidenz deutlich über, in anderen Bereichen dagegen wiederum unter diesen 3 Jahren liegen [658]. Das Institut sieht üblicherweise 3 Jahre als Zeitintervall für eine Aktualitätsprüfung an. Auf Basis dieses Intervalls wird bereits bei der Veröffentlichung eines Themenpaketes das Datum festgelegt, zu dem eine Aktualisierungsrecherche erfolgt.

Unabhängig davon werden im Rahmen eines orientierenden Evidenzscannings regelmäßig die Cochrane Database of Systematic Reviews (Cochrane Reviews) und das McMaster Online

Rating of Evidence (MORE) gesichtet. Zudem werden Warnhinweise deutscher, europäischer und US-amerikanischer Zulassungsbehörden erfasst.

Wenn im Evidenzscanning eine relevante systematische Übersicht, Studie oder Meldung identifiziert wird, werden die Auswirkungen auf den Aktualisierungsbedarf einer Gesundheitsinformation bewertet. Die Konsequenzen hängen davon ab, wie stark sich Aussagen einer Gesundheitsinformation ändern müssten. Sie können von einer vorgezogenen Aktualisierung bis hin zum Zurückziehen der betroffenen Gesundheitsinformation reichen.

7.18 Aktualisierung der Methoden von gesundheitsinformation.de

Die Methoden zur Erstellung von Gesundheitsinformationen werden im Rahmen der generellen Aktualisierung der Methoden des Instituts auf einen Aktualisierungsbedarf geprüft und gegebenenfalls überarbeitet.

8 Informationsbeschaffung

Die Informationsbeschaffung für die Erstellung der Produkte des Instituts erfolgt systematisch. Diese kann dabei ganz unterschiedliche Ziele verfolgen. Danach richtet sich, ob eine umfassende (siehe Abschnitt 8.1) oder fokussierte (siehe Abschnitt 8.2) Informationsbeschaffung durchgeführt wird. Zudem finden orientierende Recherchen (siehe Abschnitt 8.3) statt, um gezielt für bestimmte Fragestellungen nach geeigneten Daten zu suchen. Ausgehend von der detaillierten Beschreibung der Qualitätsstandards einer umfassenden Informationsbeschaffung in Abschnitt 8.1 werden für die Abschnitte 8.2 und 8.3 nur die Änderungen dargestellt. Abschnitt 8.4 erläutert die Recherche nach Leitlinien zur Erstellung von Leitliniensynopsen.

In Abschnitt 8.5 wird das Vorgehen bei der Prüfung einer Informationsbeschaffung beschrieben, wie diese im Rahmen von Dossier- und Potenzialbewertungen sowie Bewertungen gemäß § 137h SGB V durchgeführt wird.

8.1 Umfassende Informationsbeschaffung

Eine umfassende Informationsbeschaffung hat zum Ziel, alle für die jeweilige Fragestellung **relevanten** Studien und zugehörigen Dokumente zu identifizieren. Hierzu wird eine systematische Recherche in mehreren Datenbanken sowie Informationsquellen durchgeführt. In der Regel erfolgt keine Jahreseinschränkung. Jedoch kann eine Beschränkung auf englische und deutsche Publikationen erfolgen, da Auswertungen zeigen, dass der Ausschluss nicht englischer Publikationen das Fazit der meisten systematischen Übersichten nicht ändert [332,416,524]. Dabei ist projektspezifisch zu prüfen, ob eine Sprachbeschränkung Auswirkungen auf das Ergebnis des Berichts erwarten lässt (z. B. Komplementärmedizin) [519]. Die Informationsbeschaffung wird detailliert im Methoden- und Ergebnisteil des Berichts dargestellt.

Vorbemerkung: Reporting Bias (inklusive Publication Bias) begegnen

Bei der Informationsbeschaffung sind übergeordnete Aspekte wie Reporting und Publication Bias zu berücksichtigen, da diese eine grundsätzliche Auswirkung auf die Auswahl der unterschiedlichen Informationsquellen haben. Dabei ist zu bedenken, dass viele Forschungsergebnisse niemals oder nur teilweise publiziert werden [120,512,674] und publizierte Studien dazu neigen, die positiven Effekte von Interventionen zu überschätzen und die negativen Effekte zu unterschätzen [512,674] (siehe Abschnitte 3.2.1 und 9.3.13).

Daher findet bei Nutzenbewertungen standardmäßig auch eine Recherche nach unpublizierten Daten (siehe Abschnitte 8.1.2 und 8.1.3) statt. Über diese Recherche sollen neben unpublizierten Studien auch unpublizierte Daten zu publizierten Studien gefunden werden.

Publizierte Studien werden in der Regel in Fachzeitschriften veröffentlicht und sind somit überwiegend über bibliografische Datenbanken wie MEDLINE und Embase zu recherchieren. Hinweise auf unpublizierte Studien sowie Daten ergeben sich beispielsweise aus Einträgen in Studienregistern sowie aus Studienberichten der Herstellerinnen und Hersteller.

Sofern Studienberichte zur Verfügung stehen, stellen sie die primäre Quelle für die Nutzenbewertung dar, da nur diese nahezu alle Informationen einer Studie beinhalten [445]. Dagegen sind die bereitgestellten Informationen aus anderen Quellen häufig nicht ausreichend für eine zielgerichtete Auswertung gemäß der zugrunde liegenden Fragestellung oder zeigen Diskrepanzen auf [422,576,759]. Jedoch können sich Daten aus Registereinträgen und Publikationen ergänzen [759] oder unpublizierte Daten dazu verwendet werden, die Richtigkeit publizierter Daten zu prüfen [37].

Werden Daten übermittelt, die nach dem Willen der Übermittelnden nicht publiziert werden dürfen, können diese nicht in Bewertungen einfließen, da dies dem Transparenzgebot widerspricht.

Ebenso werden unaufgefordert übermittelte Daten, d. h. außerhalb von Anhörungsverfahren oder außerhalb sonstiger bestehender Regelungen (z. B. Herstelleranfragen), inhaltlich nicht berücksichtigt. Die unaufgeforderte Übermittlung von Studiendaten birgt das Risiko, dass sie selektiv geschieht und es dadurch auch zu einer Verzerrung des Ergebnisses der Nutzenbewertung kommt.

8.1.1 Recherche in bibliografischen Datenbanken

Insbesondere wenn keine Studienberichte zur Verfügung stehen, stellen bibliografische Datenbanken (neben Studienregistern) eine Hauptquelle für die Identifizierung von Studienergebnissen dar. Das detaillierte Vorgehen wird nachfolgend aufgeführt.

A) Durchführung einer Vorabrecherche

Zu Beginn eines Projekts, vor der Entwicklung der eigentlichen Suchstrategie, wird in der Regel eine Vorabrecherche durchgeführt. Die Vorabrecherchen dienen der Vorbereitung des Projekts. Hierfür wird insbesondere nach bereits vorhandenen systematischen Übersichten gesucht [129,259,680], aber auch nach potenziell relevanten Primärstudien zum Thema.

Zur Durchführung einer Vorabrecherche können z. B. die Cochrane Library, die Websites von HTA-Agenturen wie dem NICE oder der AHRQ nach systematischen Übersichten durchsucht werden [234,680,757]. Zudem können ggf. frühere IQWiG-Berichte, veröffentlichte Dossiers der Herstellerinnen und Hersteller sowie Beschlüsse des G-BA gesichtet werden. Um **laufende** HTA-Berichte und systematische Übersichten zu identifizieren, können weitere Datenbanken wie z. B. die Planned and Ongoing Projects (POP) Database [232] und das Prospective Register of Systematic Reviews (PROSPERO) [128] berücksichtigt werden. Des Weiteren können Leitlinien oder klinische Informationssysteme wie Dynamed oder UpToDate als Informationsquelle herangezogen werden.

Werden keine relevanten systematischen Übersichten gefunden, erfolgt ein iterativer Prozess unter Anwendung verschiedener Suchtechniken (siehe Abschnitt 8.1.4) wie etwa das sogenannte Snowballing [145,338] oder die Durchführung der „Similar Articles“-Funktion in PubMed (siehe Abschnitt 8.1.4) [338]. Ausgangspunkt sind dabei mehrere relevante Artikel,

die bereits bekannt sind oder durch eine sehr präzise Recherche gefunden wurden. In mehreren Durchläufen werden dann Artikel identifiziert und auf Relevanz geprüft [584,633].

B) Aufbau einer Suchstrategie

Vor der Entwicklung der Suchstrategie muss die Struktur der Recherche festgelegt werden. Hierzu bedarf es einer klar formulierten PICOS-Fragestellung (PICOS: Population, Intervention, Comparison, Outcome, Study Design) [467]. Es werden zunächst die in der wissenschaftlichen Fragestellung enthaltenen Recherchekomponenten identifiziert und definiert. Dabei werden nur die wichtigsten Recherchekomponenten verwendet, um die Suchstrategie zu entwickeln [634]. Meist enthält eine Suchstrategie Suchbegriffe zur Indikation, Intervention und zum Studientyp [467].

C) Auswahl der Datenbanken

Für die Erstellung von systematischen Übersichten ist eine systematische Recherche in mehreren bibliografischen Datenbanken erforderlich [600,601,619,691].

Neuere Analysen zeigen, dass die meisten der veröffentlichten Studien in einer begrenzten Anzahl von Datenbanken zu finden sind [4,293,318,331]. Das Cochrane Handbook nennt MEDLINE, Embase und CENTRAL als die 3 wichtigsten bibliografischen Datenbanken (für Primärstudien [467]). Dabei handelt es sich auch um die am häufigsten verwendeten Datenbanken bei der Erstellung systematischer Übersichten [65].

Abhängig von der Fragestellung der Nutzenbewertung können regionale oder themenspezifische Datenbanken mit eingebunden werden.

D) Entwicklung von Suchstrategien

Für die Erstellung von Suchstrategien ist eine Kombination aus Schlagwörtern (inkl. Publikationstyp) und Freitextbegriffen notwendig [272,415,429]. Ein objektiver Ansatz, um Suchstrategien zu entwickeln, zeichnet sich dadurch aus, Textanalyseverfahren für die Identifizierung von Freitextbegriffen und Schlagwörtern zu verwenden [564,679]. So stützt sich der „objektive Ansatz“ des IQWiG auf die Analyse bereits bekannter relevanter Artikel [336,339]. Dabei kommen verschiedene Textanalysetools wie Wordstat [578] und AntConc [22] zur Anwendung. In einem nächsten Schritt werden die ausgewählten Suchbegriffe den einzelnen Komponenten der Suchstrategie zugeordnet [623,634].

Falls vorhanden, finden validierte Studienfilter (z. B. für RCTs [206,467,767,769] und systematische Übersichten [769]) oder validierte Klassifikatoren aus dem maschinellen Lernen (z. B. RCT Classifier [501,745]) Anwendung. Bei anderen Studientypen oder Fragestellungen ist im Einzelfall zu prüfen, ob validierte Studienfilter vorliegen und diese zuverlässig anzuwenden sind. Es besteht z. B. eine kontroverse Diskussion zur Nutzung von Suchfiltern bei der Recherche nach Studien zur diagnostischen Güte [56]. Auch bei der Suche nach nicht randomisierten Studien können i. d. R. keine Studienfilter verwendet werden [337,744].

Zudem kann eine zusätzliche Recherche nach nicht verschlagworteten Datensätzen in PubMed / MEDLINE erfolgen. Damit sollen insbesondere sehr aktuelle Referenzen identifiziert werden [202,708]. Die Recherche beruht auf Freitextbegriffen und es erfolgt eine Anpassung der Studienfilter [155,413], da diese normalerweise für eine Kombination aus Schlagwort- und Freitextsuche optimiert sind.

E) Qualitätssicherung von Suchstrategien

Eine qualitativ hochwertige Suchstrategie ist Voraussetzung, um die Vollständigkeit der Evidenzbasis einer Nutzenbewertung sicherzustellen [622,623]. Aufgrund ihrer Komplexität sind Suchstrategien für bibliografische Datenbanken fehleranfällig [620]. Daher wird die Checkliste Peer Review of Electronic Search Strategies (PRESS) [509] verwendet, um den Prozess der Qualitätssicherung zu unterstützen.

Die Qualitätssicherung mit der PRESS-Checkliste stellt zunächst eine formale Prüfung dar und findet immer vor der Durchführung der Recherchen statt. Zusätzlich zur PRESS-Checkliste wird die Suchstrategie gegen ein unabhängiges Set relevanter Referenzen getestet [621]. Somit wird vorab geprüft, ob das Set relevanter Referenzen durch die Suchstrategie zu finden ist.

F) Studienselektion

Die Literaturrecherche in bibliografischen Datenbanken führt aufgrund des primär sensitiven Vorgehens dazu, dass eine große Zahl der identifizierten Zitate nicht für die Bewertung relevant ist. Die Auswahl der relevanten Publikationen erfolgt über mehrere Selektionsschritte:

- Anhand des Titels und, sofern vorhanden, des Abstracts werden zunächst sicher nicht relevante Publikationen (d. h. Publikationen, die nicht die Ein- / Ausschlusskriterien des Berichtsplans bzw. der Projektskizze erfüllen) ausgeschlossen [501].
- Für die verbleibenden, potenziell relevanten Publikationen wird der Volltext beschafft, auf dessen Basis dann die Entscheidung über den Einschluss in die jeweilige Bewertung getroffen wird.

Alle Selektionsschritte werden von 2 Personen unabhängig voneinander durchgeführt. Diskrepanzen werden durch Diskussionen aufgelöst. Beim ersten Selektionsschritt wird im Zweifelsfall der zugehörige Volltext bestellt und bewertet.

Die Dokumentation der Studienselektion erfolgt möglichst transparent und beinhaltet die Entscheidungen zum Ein- und Ausschluss jeder Referenz (nur auf Volltextebene) [129,210]. Die Studienselektion wird in der IQWiG-internen webbasierten Anwendung webTSDB [335] durchgeführt. Die weitere Dokumentation der bewerteten Referenzen im Verlauf der Berichterstellung erfolgt mittels des Literaturverwaltungsprogramms Endnote [136].

G) Dokumentation im Bericht

Eine eindeutige und transparente Darstellung aller Aspekte der Recherche ermöglicht die Beurteilung der Qualität und Vollständigkeit der systematischen Recherche [473,624] sowie die Durchführung späterer Aktualisierungsrecherchen.

- Standardmäßig werden, angelehnt an Mullins et al. [528], dokumentiert:
- genutzte Datenbanken sowie Datenbankanbieter bzw. Suchoberflächen,
- die zeitliche Abdeckung und das letzte Suchdatum,
- die Suchstrategie mit allen Suchlimitationen.

Zudem wird der Selektionsprozess anhand eines Flussdiagramms im Ergebnisteil des Berichts dargestellt [129,210,584] (siehe PRISMA-Vorlage [473,518]) und die Referenzen der ein- und ausgeschlossenen Studien bzw. Dokumente werden in separaten Referenzlisten aufgeführt [653].

8.1.2 Suche in Studienregistern

Studienregister haben in den letzten Jahren stark an Bedeutung gewonnen. So fordert das International Committee of Medical Journal Editors seit 2005 die prospektive Registrierung von klinischen Studien als Bedingung für die Veröffentlichung in Fachzeitschriften [159].

Außerdem gibt es in den USA seit 2007 die gesetzliche Verpflichtung, nahezu alle klinischen Studien zu registrieren, die in den Regulierungsbereich der FDA fallen, sowie deren Ergebnisse zu veröffentlichen [2]. Falls vorhanden, wird mittlerweile in ClinicalTrials.gov auf das Studienprotokoll, den statistischen Analyseplan (SAP) und das Data sharing Statement einer Studie verlinkt [537,699,772].

Auch die Europäische Arzneimittel-Agentur (EMA) veröffentlicht seit 2011 über das EU Clinical Trials Register (EU-CTR) [231] Daten zur Registrierung eines Großteils ihrer Zulassungsstudien für Medikamente. Seit Juli 2014 werden in diesem Register auch Studien-synopsen veröffentlicht [223].

A) Aufbau einer Suchstrategie

Suchen in Studienregistern sollten eine hohe Sensitivität ausweisen, einfach gehalten werden und nach Möglichkeit nur 1 Komponente (in der Regel Indikation oder Intervention) berücksichtigen [283]. Es wird dabei zuerst mit den Begriffen zu dem Konzept gesucht, das sich am besten abbilden lässt und trotz hoher Sensitivität die wenigsten Treffer generiert. Nur bei zu vielen Treffern wird die Suche mit dem 2. Konzept weiter eingeschränkt. Eine weitere Einschränkung (z. B. nach Studienstatus) ist aufgrund der unterschiedlichen Qualität der einzelnen Einträge nur im Ausnahmefall vorzunehmen.

B) Auswahl der Studienregister

In die systematische Suche werden immer mehrere Studienregister eingeschlossen, da kein Studienregister alle Studien beinhaltet [133,283,698]. Die Suche wird mindestens in den Studienregistern ClinicalTrials.gov und International Clinical Trials Registry Platform Search Portal (ICTRP) der WHO durchgeführt [37,283,467]. Beim ICTRP handelt es sich um ein Metaregister, das einen großen Teil der klinischen Studien beinhaltet [133,324]. Jedoch sind die Suchfunktionen sehr eingeschränkt [283,440] und das Studienregister produziert regelmäßig Fehlermeldungen [334]. Somit werden wichtige Studienregister wie ClinicalTrials.gov direkt durchsucht, obwohl sie auch über das ICTRP erfasst werden [283,440].

Für Nutzenbewertungen von Arzneimitteln ist zudem das EU-CTR, das Register der EMA, zu berücksichtigen. Zudem können Studienregister der pharmazeutischen Industrie (Studienregister einzelner Firmen) und das Arzneimittel-Informationssystem (AMIce) des Bundesinstituts für Arzneimittel und Medizinprodukte (BfArM) [103] durchsucht werden. Diese Datenbank soll auch Ergebnisberichte von Studien enthalten, die außerhalb Europas und der USA durchgeführt wurden.

C) Weiteres Vorgehen

Das Vorgehen bezüglich der Qualitätssicherung, Durchführung der Suche, Studienselektion, Dokumentation sowie Aktualisierung der Suche in Studienregistern orientiert sich weitgehend an der Recherche in bibliografischen Datenbanken mit der Ausnahme, dass die Einträge aus Studienregistern in einem 1-schrittigen Vorgehen selektiert werden, da die vollständigen Informationen über die Website des Studienregisters unmittelbar zur Verfügung stehen.

8.1.3 Herstelleranfragen

Für die Nutzenbewertung von Arzneimitteln bzw. Methoden, die maßgeblich auf einem Medizinprodukt beruhen, wird in der Regel bei den Herstellern bezüglich bislang nicht veröffentlichter Informationen angefragt. Ziel dieser Anfrage ist die Identifizierung aller relevanten Informationen zu diesen Studien, unabhängig vom Publikationsstatus.

Basis für die Einbeziehung bislang unveröffentlichter Informationen in die Nutzenbewertung ist der Abschluss einer Vereinbarung zur Übermittlung und Veröffentlichung von Studieninformationen, die vor der Einreichung von Daten zwischen dem Institut und dem beteiligten Herstellerinnen und Hersteller getroffen wird (siehe Mustervertrag zu Arzneimitteln [385]). Diese Vereinbarung spezifiziert den Verfahrensablauf, die Anforderungen an die zu übermittelnden Dokumente sowie die vertraulichen bzw. nicht vertraulichen Anteile der übermittelten Dokumente.

Die Anfrage wird in der Regel in einem 2-stufigen Verfahren durchgeführt. In der 1. Stufe wird bei der Herstellerin oder beim Hersteller eine vollständige Übersicht über alle Studien, die von dieser oder diesem durchgeführt wurden, angefragt. Dabei werden projektspezifische Einschlusskriterien für diese Übersicht definiert. Aus der Übersicht der Studien identifiziert das

Institut die Studien, die für die Nutzenbewertung relevant sind, und fragt zu diesen Studien Detailinformationen an. Dabei kann es sich um eine Anfrage zu unveröffentlichten Studien oder zu ergänzenden, bislang unveröffentlichten Informationen zu veröffentlichten Studien handeln. Auch bisher unveröffentlichte Informationen, die in die Nutzenbewertung einfließen, werden in den Berichten des Instituts publiziert, um die Transparenz zu gewährleisten.

Falls das jeweilige Unternehmen dem Abschluss einer solchen Vereinbarung und damit insbesondere der vollständigen Übermittlung aller angefragten Informationen nicht zustimmt oder der vollständigen Übermittlung der angefragten Informationen trotz des Abschlusses dieser Vereinbarung nicht nachkommt, werden keine weiteren Anfragen an die Herstellerin oder den Hersteller gerichtet. Hierdurch soll eine Verzerrung der Ergebnisse durch die selektive Bereitstellung von Informationen vermieden werden (siehe Abschnitt 3.2.1).

8.1.4 Weitere Informationsquellen und Suchtechniken

Um zusätzliche relevante Studien bzw. Dokumente zu identifizieren, werden je nach Fragestellung weitere Informationsquellen genutzt und weitere Suchtechniken angewendet.

Hinsichtlich der Studienselektion und Dokumentation im Bericht ergeben sich Unterschiede, die zum Teil aus den eingeschränkten Suchmöglichkeiten und zum Teil aus der Art der gesuchten Daten resultieren. So werden bestimmte Informationsquellen nur von 1 Person in Bezug auf Studien gesichtet, die diese dann in Bezug auf ihre Relevanz bewertet; eine 2. Person überprüft den gesamten Prozess inklusive der Bewertungen (z. B. öffentlich zugängliche Dokumente von Zulassungsbehörden, Sichtung von Referenzlisten).

Nachfolgend werden die weiteren Informationsquellen und Suchtechniken dargestellt, die entweder standardmäßig oder optional in der Nutzenbewertung Berücksichtigung finden.

A) Zulassungsbehörden

Wenn Arzneimittel oder Behandlungsmethoden mit einem maßgeblichen Medizinprodukt bewertet werden, sind öffentlich zugängliche Dokumente von Zulassungsbehörden eine mögliche Quelle für die Informationsbeschaffung.

Dokumente der Zulassungsbehörden

Informationen zu zentral zugelassenen Arzneimitteln (wie z. B. die European public Assessment Reports) werden in Europa über die Website der EMA [224] recherchiert. In den USA erhält man Zugang zu den Medical Reviews sowie Statistical Reviews der FDA über Drugs@FDA [723].

Im Gegensatz zu den USA gibt es für Medizinprodukte in Europa keinen zentralen Zulassungsprozess. Öffentlich zugängliche Informationen zu Medizinprodukten sind derzeit lediglich vereinzelt auf Länderebene verfügbar, wie beispielsweise beim NICE die sogenannte List of interventional Procedures [535]. Zukünftig sollen Studienergebnisse zu neuen implantierbaren und Klasse-III-Produkten (teilweise) öffentlich zugänglich in der europäischen

Datenbank für Medizinprodukte (EUDAMED) einsehbar sein [222]. In den USA sind die Informationen zu den von der FDA geprüften Medizinprodukten inklusive der Angaben zur zugrunde liegenden Datenbasis über Devices@FDA zu recherchieren [722].

B) Ausgewählte Fachzeitschriften und Kongressbände

Je nach Fragestellung kann es sinnvoll sein, eine Handsuche in ausgewählten Fachzeitschriften durchzuführen. Dies wird im Einzelfall entschieden. Auf eine Suche nach Kongressbeiträgen wird in der Regel verzichtet, da diese meist wenig Informationen zur Studienmethodik und zu den Ergebnissen beinhalten [203].

C) Durch den G-BA oder das BMG übermittelte Dokumente

Vom jeweiligen Auftraggeber (G-BA, BMG) übermittelte Dokumente werden daraufhin geprüft, ob es sich um Studien handelt, die den Einschlusskriterien der Bewertung entsprechen.

D) G-BA-Website und IQWiG-Website

Je nach Fragestellung kann es sinnvoll sein, die Websites des G-BA und des IQWiG nach früheren IQWiG-Produkten, öffentlich zugänglichen Herstellerunterlagen (z. B. Module 1 bis 4 der Dossiers zu frühen Nutzenbewertungen nach § 35a SGB V) sowie nach Beschlüssen des G-BA zu sichten, ob dort Studien genannt werden, die den Einschlusskriterien der Bewertung entsprechen.

E) Anwendung weiterer Suchtechniken

Im Rahmen von Vorabrecherchen, bei schwer zu recherchierenden Fragestellungen (z. B. zu komplexen Interventionen) und als Prüfung von Suchstrategien in bibliografischen Datenbanken haben sich verschiedene Suchtechniken etabliert [338]. Dazu zählen das Sichten von Referenzlisten, das sogenannte Citation Tracking sowie die Verwendung der „Similar Articles“-Funktion. In der Regel wird für alle Nutzenbewertungen das Sichten von Referenzlisten systematischer Übersichten als zusätzliche Suchtechnik durchgeführt.

F) Anhörung

Im Rahmen der Anhörung zum vorläufigen Berichtsplan oder zum Vorbericht übermittelte Studien oder Studieninformationen werden berücksichtigt.

G) Autorenanfragen

Wenn die zu einer Studie vorhandenen Informationen unvollständig, unklar oder widersprüchlich sind, kann es sinnvoll sein, Kontakt zu den Autorinnen und Autoren aufzunehmen. Autorenanfragen können erfolgen, um die Geeignetheit einer Studie, ihre methodische Qualität oder ihre Ergebnisse besser einschätzen zu können [210,527,584]. Autorenanfragen erfolgen in der Regel nur, wenn sich hierdurch ein relevanter Einfluss auf den Bericht erwarten lässt.

8.2 Fokussierte Informationsbeschaffung

Nicht für alle Fragestellungen ist es erforderlich oder möglich, eine auf Vollständigkeit ausgelegte Informationsbeschaffung durchzuführen. In solchen Fällen wird eine sogenannte fokussierte Informationsbeschaffung durchgeführt, insbesondere wenn weiterhin der Anspruch an ein systematisches und transparentes Vorgehen besteht.

Eine fokussierte Informationsbeschaffung wird beispielsweise 1) bei einer kurzen Bearbeitungszeit (z. B. Dossierbewertungen) durchgeführt, 2) wenn eine Fragestellung nicht auf Vollständigkeit ausgelegt ist (z. B. qualitative Forschung) oder 3) die Bewertung auf systematischen Übersichten beruht. Eine fokussierte Informationsbeschaffung hat das Ziel, eine ausgewogene Relation zwischen Sensitivität (also Vollständigkeit) und Präzision (also Genauigkeit) zu erzielen.

Das Vorgehen bezüglich der Entwicklung der Suchstrategie, Qualitätssicherung, Durchführung der Recherche, Studienselektion sowie Dokumentation orientiert sich an der umfassenden Informationsbeschaffung (siehe Abschnitt 8.1). Jedoch können zum Beispiel in folgenden Bereichen Einschränkungen bzw. Anpassungen erfolgen, wie sie im Bereich der Rapid Reviews häufig Anwendung finden [586,587,717]:

- bei der Auswahl von Datenbanken,
- bei der Auswahl von Studienfiltern,
- bei Jahres- und / oder Spracheinschränkungen,
- bei der Studienselektion (erfolgt durch 1 Person; Qualitätssicherung des Ergebnisses durch 2. Person),
- bei der Darstellung der Methodik und Ergebnisse.

Zudem werden häufig weniger Informationsquellen einbezogen.

8.2.1 Recherche nach systematischen Übersichten

Bei der Recherche nach systematischen Übersichten ist es ausreichend, wenn ein Großteil der hochwertigen und aktuellen systematischen Übersichten zur Fragestellung identifiziert wird. Dabei erfolgt die Recherche mindestens in den Datenbanken

- MEDLINE,
- HTA Database sowie
- Cochrane Database of Systematic Reviews.

Bei der Entwicklung der Suchstrategien werden präzise Studienfilter eingesetzt (z. B. Wong 2006 „High specificity strategy“ [768]). Ggf. erfolgt eine zeitliche Einschränkung der Recherchen. Beispielsweise wird die Recherche nach systematischen Übersichten als Basis für Gesundheitsinformationen regelhaft auf die letzten 3 Jahre beschränkt.

Außerdem können systematische Übersichten als Quelle für Primärstudien dienen, um auf Grundlage derer eine Nutzenbewertung durchzuführen [591]. Hierfür erfolgt vorab eine Prüfung der Ein- und Ausschlusskriterien sowie der Informationsbeschaffung der identifizierten systematischen Übersicht(en) (basierend auf Punkt 3 der AMSTAR-Checkliste [653]). Dann wird eine oder werden ggf. mehrere hochwertige und aktuelle systematische Übersicht(en) ausgesucht, deren Primärstudien extrahiert und dann selektiert werden. Bei diesem Vorgehen wird von der verwendeten systematischen Übersicht allein das Rechercheergebnis, nicht aber die Bewertung der eingeschlossenen Primärstudien oder die Datenextraktion übernommen. Zusätzlich erfolgt eine Aktualisierung der Informationsbeschaffung nach Abschnitt 8.1 für den Zeitraum, der nicht durch die systematische Übersicht(en) abgedeckt wird (siehe Abschnitt 8.2.1).

Wurden im Berichtsplan bzw. in der Projektskizze aufgeführte Informationsquellen in der systematischen Übersicht nicht berücksichtigt oder nicht umfassend durchsucht (z. B. Studienregister), können diese Quellen im Rahmen der Informationsbeschaffung für die Nutzenbewertung durchsucht werden.

8.2.2 Recherche nach qualitativer Forschung

Bei der Recherche nach Literatur zu qualitativer Forschung sollen unter anderem Erfahrungen im Zusammenhang mit einem bestimmten Krankheitsbild erfasst, Probleme im Umgang mit der Erkrankung und deren Behandlung identifiziert sowie potenzielle Informationsbedürfnisse abgeleitet werden.

Es erfolgt eine Recherche mindestens in den bibliografischen Datenbanken wie:

- MEDLINE,
- PsycINFO sowie
- Cinahl.

8.2.3 Recherche für gesundheitsökonomische Fragestellungen

Es gibt sehr unterschiedliche gesundheitsökonomische Fragestellungen, für die eine fokussierte Informationsbeschaffung durchzuführen ist. Dazu zählen die Recherche nach gesundheitsökonomischen Evaluationen, nach entscheidungsanalytischen Modellen, zum Maß des Gesamtnutzens, zur Kostenbestimmung sowie ggf. epidemiologische Daten, wenn Daten aus Deutschland nicht vorhanden sind.

Es erfolgt eine Recherche mindestens in der bibliografischen Datenbank MEDLINE. Zudem kann bei Herstellerinnen und Herstellern bezüglich gesundheitsökonomischer Evaluationen angefragt werden.

8.2.4 Recherchen zu Addenda von §-137e- oder §-137h-Bewertungen

Im Rahmen von Addenda zu §-137e- oder §-137h-Bewertungen wird eine fokussierte Informationsbeschaffung durchgeführt.

Es erfolgt eine systematische Recherche in den bibliografischen Datenbanken

- MEDLINE,
- Embase sowie
- Central.

Zusätzlich wird eine Recherche in den Studienregistern ClinicalTrials.gov und dem ICTRP Search Portal durchgeführt.

8.2.5 Überprüfung der Vollständigkeit eines Studienpools

Im Rahmen der Dossierbewertung erfolgt eine Prüfung der Informationsbeschaffung in den Dossiers (siehe Abschnitt 8.5). Je nach Ergebnis der Prüfung wird eine sogenannte Vollständigkeitsprüfung durchgeführt. Die Vollständigkeitsprüfung hat das Ziel, die Vollständigkeit eines dargelegten Studienpools zu prüfen, und nicht, die vollständige Datenbasis zu identifizieren.

Es erfolgt eine Suche in den Studienregistern

- ClinicalTrials.gov,
- ICTRP Search Portal sowie
- EU Clinical Trials Register.

In Ergänzung dazu wird in bestimmten Fällen (z. B. Recherche nach Arzneimitteln, die vor 2007 zugelassen wurden; Recherche nach anderen Studientypen als RCTs [391]) zusätzlich auch eine bibliografische Recherche mittels Kombination verschiedener Suchtechniken (einfache boolesche Recherche sowie „Similar Articles“-Funktion in PubMed [625,743]) durchgeführt.

8.3 Orientierende Recherchen

Von orientierender Recherche wird immer dann gesprochen, wenn gezielt nach geeigneten Daten gesucht wird. Die Suche endet, sobald die notwendigen Informationen vorliegen.

Die Informationsquellen sind sehr themenabhängig und umfassen häufig klinische Informationssysteme wie Dynamed und UpToDate oder Leitliniendatenbanken. Zudem werden je nach Fragestellung spezielle Datensammlungen wie die des Robert Koch-Instituts, des Statistischen Bundesamts oder des Wissenschaftlichen Instituts der AOK sowie Daten aus regionalen Registern, Gesetzen, Verordnungen oder Richtlinien genutzt.

Anders als bei der umfassenden Informationsbeschaffung findet die Suche und Auswahl der Daten durch 1 Person statt. Die Qualitätssicherung des Ergebnisses erfolgt durch eine 2. Person. Die Dokumentation im Bericht beschränkt sich auf die Darstellung der konkreten Ergebnisse.

Beispiele für orientierende Recherchen sind die Vorabrecherche (siehe Abschnitt 8.1.1) sowie die Suche nach Kostendaten (siehe Abschnitt 4.4.4) und nach epidemiologischen Daten (siehe Abschnitt 4.5.2).

8.4 Recherche nach Leitlinien zur Erstellung von Leitliniensynopsen

Hat eine Recherche zum Ziel, Leitlinien zu identifizieren, erfolgt die Recherche in Leitliniendatenbanken (z. B. der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften [AWMF], Canadian Medical Association [CMA] Infobase, Trip Database) und bei fachübergreifenden bzw. fachspezifischen Leitlinienanbietern.

Bei der Recherche in Leitliniendatenbanken und bei Leitlinienanbietern richtet sich die anzuwendende Suchstrategie nach dem Aufbau und den Möglichkeiten der jeweiligen Internetseiten. Nur wenige Seiten ermöglichen eine Recherche mit Schlagwörtern, sodass in der Regel jeweils die gesamte Liste von veröffentlichten Leitlinien durchgesehen wird. Außerdem ist bei der Recherche in Leitliniendatenbanken und bei Leitlinienanbietern häufig kein standardisierter Export möglich. Aus diesem Grund wird die Dokumentation der Recherche und der Anzahl der Treffer in einem standardisierten Rechercheprotokoll vorgenommen. Die Erfassung der potenziell relevanten Treffer erfolgt in einem Literaturverwaltungsprogramm. Der Ablauf der Selektion von Leitlinien erfolgt wie in Abschnitt 5.1.2 dargestellt. Jedoch wird bei der Recherche in Leitliniendatenbanken und bei Leitlinienanbietern je nach Fragestellung in einem ergänzenden Schritt geprüft, ob bei der Generierung und Formulierung der Leitlinie eine methodische Systematik zur Anwendung kam. In der Regel ist hiermit die Evidenzbasierung der Leitlinie gemeint (siehe Kapitel 5). Im Rahmen der Erstellung des Berichtsplans werden vorab die Ein- und Ausschlusskriterien festgelegt.

Das Titel- / Abstractscreening erfolgt durch 1 Person, die Qualitätssicherung dieses Schritts durch eine 2. Person. Die folgenden Schritte (ab dem Volltextscreening) werden von 2 Personen unabhängig voneinander ausgeführt.

Im Rahmen der Leitlinienbewertung kann es im Einzelfall sinnvoll sein, Kontakt zu Autorinnen und Autoren bzw. Leitlinienerstellerinnen und -erstellern aufzunehmen. Die Anfragen können sich z. B. auf spezielle Detailinformationen zu einzelnen Leitlinien oder auf nicht publizierte Teilaspekte zu Publikationen beziehen.

8.5 Prüfung der Informationsbeschaffung

Für die Erstellung eines Dossiers bzw. Antrags zur Erprobung sind eine Recherche in bibliografischen Datenbanken und eine Suche in öffentlich zugänglichen Studienregistern grundsätzlich durch den Antragsteller durchzuführen; die genauen Vorgaben finden sich in der Verfahrensordnung des G-BA [273,276]. In ähnlicher Weise müssen auch Krankenhäuser, die

Informationen zu einer Bewertung gemäß § 137h SGB V erstellen, systematisch nach Informationen zur angefragten Methode suchen.

Bei der Dossierbewertung, bei der Potenzialbewertung und bei einer Bewertung gemäß § 137h SGB V wird eine Prüfung der in den vorgelegten Unterlagen dokumentierten Informationsbeschaffung durchgeführt. Bei allen 3 Bewertungsverfahren werden die Recherche in bibliografischen Datenbanken, die Suche in Studienregistern sowie die Studienselektion geprüft. Die Prüfung stützt sich auf das in Abschnitt 8.1.1 beschriebene Vorgehen zur Qualitätssicherung von Suchstrategien sowie die Dokumentvorlagen in den Vorgaben der Verfahrensordnung des G-BA [273,276].

Besonderheiten bei der Prüfung von Dossiers

Je nachdem, welche Ergebnisse sich aus der Prüfung der Dossiers ergeben, stehen verschiedene Strategien zur Verfügung, um die Vollständigkeit zu prüfen. Zum Beispiel kann eine stichprobenartige Überprüfung der im Dossier ausgeschlossenen Literaturzitate erfolgen oder eine eigene Recherche und / oder Studienselektion mittels Vollständigkeitsprüfung durchgeführt werden (siehe Abschnitt 8.2.3). Bei hohen Trefferzahlen kann auch ein Abgleich auf Basis von systematischen Übersichten durchgeführt werden (siehe Abschnitt 8.3 oder 8.2.1). Das Ergebnis der Überprüfung der Informationsbeschaffung und die Beschreibung des Vorgehens sind Teil der Dossierbewertung.

9 Informationsbewertung

Grundsätzlich folgt dem Schritt der Informationsbeschaffung (siehe Abschnitt 7.13) ein bewertender Schritt, in dem die beschaffte Information systematisch in Hinblick auf die jeweilige Aussagekraft kritisch eingeordnet wird. In diesem Kapitel werden vor allem die Aspekte erläutert, die bei der Informationsbewertung im Rahmen von Nutzenbewertungen zum Tragen kommen.

In der Forschung bezeichnet der Begriff Bias (Verzerrung) eine systematische Abweichung der Forschungsergebnisse von der „Wahrheit“ [614]. Dabei könnte es sich beispielsweise um eine irrtümlich zu hohe (oder zu niedrige) Einschätzung eines Behandlungseffekts handeln. Ein Hauptziel bei der Nutzenbewertung medizinischer Versorgungsleistungen ist es, die tatsächlichen Wirkungen von Therapien und Interventionen so zuverlässig und unverzerrt wie möglich abzuschätzen. Zur Minimierung von Bias bei der Nutzenbewertung der medizinischen Versorgung werden international verschiedenste Ansätze verfolgt; dazu gehören etwa die Anwendung wissenschaftlich robuster Methoden, die Gewährleistung einer breiten Beteiligung an den jeweiligen Studien sowie die Offenlegung von Beziehungen. Alle diese Methoden bilden auch die gesetzliche Grundlage für die Arbeit des Instituts.

9.1 Qualitätsbewertung von Einzelstudien

9.1.1 Kriterien für den Einschluss von Studien

Häufig ergibt sich das Problem, dass für eine Nutzenbewertung relevante Studien die Einschlusskriterien für die in der systematischen Übersicht definierte Patientenpopulation bzw. die definierten Prüf- und Vergleichsinterventionen nicht vollständig erfüllen. Das Institut verfährt hier in der Regel nach folgenden Kriterien:

Für das Einschlusskriterium bezüglich der Population reicht es aus, wenn bei mindestens 80 % der in der Studie eingeschlossenen Patientinnen und Patienten dieses Kriterium erfüllt ist. Liegen für solche Studien Analysen der relevanten Teilpopulation vor, wird auf diese Analysen zurückgegriffen. Studien, bei denen das Einschlusskriterium bezüglich der Population bei weniger als 80 % der in die Studie eingeschlossenen Patientinnen und Patienten erfüllt ist, werden nur dann eingeschlossen, wenn Analysen der relevanten Teilpopulation vorliegen oder wenn hinreichend sicher plausibel oder nachgewiesen ist, dass die in dieser Studie gewonnenen Ergebnisse auf die Zielpopulation der systematischen Übersicht anwendbar sind (zur Anwendbarkeit siehe Abschnitt 3.3.1).

Ebenfalls eingeschlossen werden Studien, die zu mindestens 80 % das Einschlusskriterium bezüglich der Prüfintervention (Interventionsgruppe der Studie) erfüllen und zu mindestens 80 % das Einschlusskriterium bezüglich der Vergleichsintervention (Vergleichsgruppe der Studie). Ist in einer Studie eines der beiden Kriterien verletzt, wird die Studie aus der Nutzenbewertung ausgeschlossen.

Es ist erforderlich, dass zu einer Studie zumindest die wichtigsten Informationen vorhanden sind, um die Studienergebnisse verwenden zu können. Kongressbeiträge, publizierte Abstracts oder Pressemitteilungen zu Studien enthalten in der Regel deutlich zu wenige und zu unsichere Informationen zu Studienmethodik und -ergebnissen, um diese Quellen für Nutzenbewertungen verwenden zu können [365].

Um die Verwendbarkeit einer Quelle zu beurteilen, orientiert sich das Institut an den diversen Stellungnahmen, die zur Verbesserung der Qualität von Publikationen erarbeitet wurden. Dies sind u. a.

- das Consolidated-Standards-of-Reporting-Trials(CONSORT)-Statement für randomisierte Studien [639] und das dazugehörige Erläuterungsdokument [517],
- ein Vorschlag zur Erweiterung des CONSORT-Statements für randomisierte Studien bei nicht medikamentösen Verfahren [73] und das zugehörige Erläuterungsdokument [72],
- das CONSORT-Statement für clusterrandomisierte Studien [118],
- das CONSORT-Statement zur Dokumentation unerwünschter Ereignisse [399],
- das CONSORT-Statement für Nichtunterlegenheits- und Äquivalenzstudien [571],
- das CONSORT-Statement für pragmatische Versuche [777],
- die CONSORT-PRO-Erweiterung für patientenberichtete Endpunkte [117],
- die CONSORT-Erweiterung für mehrarmige randomisierte Studien [426],
- das Transparent-Reporting-of-Evaluations-with-Non-randomized-Designs(TREND)-Statement für nicht randomisierte Interventionsstudien [166],
- das Strengthening-the-Reporting-of-Observational-Studies-in-Epidemiology(STROBE)-Statement für Beobachtungsstudien in der Epidemiologie [742] und das dazugehörige Erläuterungsdokument [731],
- das Transparent-Reporting-of-a-Multivariable-Prediction-Model-for-Individual-Prognosis-or-Diagnosis(TRIPOD)-Statement für Prognosestudien [141] und das dazugehörige Erläuterungsdokument [523],
- das Standards-for-Reporting-of-Diagnostic-Accuracy(STARD)-Statement für Diagnosestudien [69,70] und das dazugehörige Erläuterungsdokument [71] und
- die International-Society-of-Quality-of-Life-Research(ISOQOL)-Reporting-Standards für patientenberichtete Endpunkte [99].

9.1.2 Zusammenhang zwischen Studientyp / -art und Fragestellung

Es werden hier nur die wichtigsten Designs zusammengefasst, die bei der Nutzenbewertung in der medizinischen Forschung in Abhängigkeit von der Fragestellung eine Rolle spielen.

Für die Nutzenbewertung von Interventionen ist an erster Stelle eine Kontrollgruppe zu fordern. Aus einem reinen Vorher-nachher-Vergleich in einem Design mit abhängigen Stichproben ohne Kontrollgruppe lässt sich in der Regel kein Beleg für einen Effekt einer Intervention ableiten. Ausnahmen bilden Krankheitsbilder mit einem deterministischen (oder nahezu deterministischen) Verlauf (z. B. diabetisches ketoazidotisches Koma oder Kammerflimmern; siehe Abschnitt 3.2.2). Gütekriterien, die die Aussagekraft kontrollierter Studien erhöhen, sind Randomisierung und Verblindung. Häufige Designs bei klinischen Studien bilden parallele Gruppen [574], Cross-over-Studien [421] und clusterrandomisierte Studien [196]. Sind in einer Studie Zwischenauswertungen geplant, so muss dies mithilfe eines adäquaten sequenziellen Designs berücksichtigt werden [421].

Den ersten Informationsgewinn gibt es häufig aus Fallberichten oder Fallserien. Diese sind anfällig für Verzerrungen aller Art, sodass hier, je nach Fragestellung, nur sehr beschränkt eine verlässliche Evidenz ableitbar ist. Bevölkerungsbezogene Querschnittsstudien sind geeignet für die Schätzung von Krankheitsprävalenzen. Weitere grundlegende klassische Studientypen der Epidemiologie sind Fall-Kontroll-Studien [77] zur Untersuchung des Zusammenhangs zwischen Expositionen und seltenen Erkrankungen und Kohortenstudien [78] zur Erforschung des Effekts von Expositionen im Zeitverlauf. Kohortenstudien sind in diesem Sinne prospektiv angelegt; allerdings gibt es auch retrospektive Kohortenstudien, in denen die Exposition aus der Vergangenheit erfasst wird (häufig in der Arbeits- oder auch Pharmakoepidemiologie anzutreffen). Grundsätzlich sind prospektive Studien retrospektiven Designs vorzuziehen. Allerdings sind zum Beispiel Fall-Kontroll-Studien häufig die einzige praktikable Möglichkeit, Informationen über Zusammenhänge zwischen Expositionen und seltenen Erkrankungen zu gewinnen. Neuere Studiendesigns der modernen Epidemiologie enthalten Elemente sowohl von Fall-Kontroll-Studien als auch von Kohortenstudien und sind nicht mehr eindeutig als retrospektiv oder prospektiv zu klassifizieren [428].

Diagnose- und Screeningstudien können sehr unterschiedliche Ziele haben, sodass die Bewertung von der Wahl eines adäquaten Designs abhängt (siehe Abschnitte 3.5 und 3.6).

9.1.3 Rangordnung verschiedener Studienarten / Evidenzgrade

Es gibt verschiedene Ansätze, bestimmten Studientypen bestimmte Evidenzgrade zuzuordnen, um so eine Rangfolge bezüglich der Validität der Evidenz aus verschiedenen Studienarten im Rahmen einer systematischen Übersicht oder für die Entwicklung von Leitlinien zu erstellen [314,322]. Es gibt jedoch kein System der Evidenzbewertung, das allgemein akzeptiert und universell in allen systematischen Übersichten einsetzbar ist [430,750]. Aufgrund der Komplexität der Studienbewertung lässt sich anhand einer Evidenzhierarchie auch kein abschließendes Qualitätsurteil bilden [28,765]. Das Institut orientiert sich gemäß den Standards der evidenzbasierten Medizin an einer groben Hierarchie von Studientypen, die weitgehend mit der Evidenzklassifizierung im 4. Kapitel § 7 Abs. 3 der Verfahrensordnung des G-BA übereinstimmt [276] und Eingang in die Verordnung über die Nutzenbewertung von Arzneimitteln nach § 35a SGB V [106] gefunden hat. Zumindest für die Prüfung von Interventionseffekten

wird der höchste Evidenzgrad RCTs und systematischen Übersichten von RCTs zugeordnet. Einzelne RCTs werden in einigen Einteilungen noch als solche hoher oder weniger hoher Qualität eingestuft (siehe Abschnitt 3.1.4).

Spätestens bei der Einteilung nicht randomisierter Studien bezüglich ihres Verzerrungspotenzials jedoch kann das Studiendesign allein keine hinreichende Orientierung mehr bieten [312,350,740], auch wenn eine grundsätzliche Unterscheidung zwischen vergleichenden und nicht vergleichenden Studien sinnvoll erscheint. Wie in Abschnitt 3.8 dargestellt, wird das Institut bei der Einteilung nicht randomisierter vergleichender Studien neben anderen Designaspekten vor allem die Kontrolle möglicher Confounder prüfen. Diese Graduierung bezieht sich jedoch auf das Verzerrungspotenzial (siehe Abschnitt 9.1.4) und nicht auf den Evidenzgrad einer Studie.

9.1.4 Aspekte der Bewertung des Verzerrungspotenzials

Ein wesentlicher Gesichtspunkt bei der Interpretation von Studienergebnissen ist die Einschätzung von Verzerrungsaspekten (vgl. qualitative Ergebnisunsicherheit, Abschnitt 3.1.4). Dabei spielen die Fragestellung, die Art, das Design und die Durchführung der Studie sowie die Verfügbarkeit der Informationen eine Rolle. Das Verzerrungspotenzial wird von der Qualität der Studiendurchführung zwar maßgeblich beeinflusst, dessen Bewertung ist jedoch nicht gleichbedeutend mit einer Qualitätseinschätzung von Studien. Beispielsweise können einzelne Endpunkte auch in einer qualitativ hochwertigen Studie möglicherweise bedeutsam verzerrt sein. Andererseits können Studien trotz niedriger Qualität im Einzelfall für bestimmte Endpunkte Resultate hoher Ergebnissicherheit liefern. Das Institut wird daher i. d. R. für alle relevanten Ergebnisse das Ausmaß möglicher Verzerrungen problemorientiert sowohl für die Studie als auch endpunktspezifisch einschätzen.

Grundsätzlich soll in einer Studie von der Planung über die Durchführung bis zur Auswertung und Präsentation nach einem anerkannten standardisierten Konzept vorgegangen werden. Hierzu gehört ein Studienprotokoll, in dem alle wichtigen Methoden und Vorgehensweisen beschrieben sind. Die üblichen Standards hierbei bilden bei (randomisierten) klinischen Versuchen die Grundprinzipien der Good clinical Practice (GCP) [396,446] und bei epidemiologischen Studien die Leitlinien und Empfehlungen zur Sicherung Guter Epidemiologischer Praxis (GEP) [169]. Ein zentrales Kriterium für die Vermeidung von Verzerrungen ist dabei, ob die Studie auch so ausgewertet wurde wie in der Planung vorgesehen. Dies ist Publikationen in der Regel nur sehr unzuverlässig zu entnehmen. Ein Abschnitt über die Fallzahlplanung kann hierbei jedoch zumindest Hinweise liefern. Darüber hinaus ist ein Vergleich mit dem eventuell zuvor veröffentlichten Studienprotokoll bzw. einer Publikation zum Studiendesign sinnvoll.

Zentrale Aspekte zur Bewertung des Verzerrungspotenzials der Ergebnisse randomisierter kontrollierter Studien durch das Institut stellen ein adäquates Concealment, d. h. die Unvorhersehbarkeit und Verheimlichung der Gruppenzuteilung (z. B. durch eine externe Randomisierung bei nicht verblindbaren Studien), die verblindete Zielgrößenerhebung bei Studien, in

denen die Ärztin oder der Arzt und die Patientin oder der Patient nicht verblindbar sind, und die adäquate Anwendung des Intention-to-treat-Prinzips dar [685].

Die Interpretation der Ergebnisse von unverblindeten Studien beziehungsweise Studien, in denen es (möglicherweise) zu einer Entblindung kam, muss im Vergleich zu den verblindeten Studien vorsichtiger erfolgen. Wichtige Instrumente zur Vermeidung von Bias in Studien, in denen eine Verblindung der Intervention nicht möglich ist, sind wiederum die Randomisierung und die Wahl geeigneter Zielvariablen. In nicht verblindbaren Studien ist es zentral wichtig, dass eine adäquat verdeckte Zuteilung (Allocation Concealment) der Patientinnen und Patienten zu den zu vergleichenden Gruppen gewährleistet ist. Weiterhin ist es erforderlich, dass die Zielvariable unabhängig von den (unverblindeten) behandelnden Personen ist bzw. unabhängig von den behandelnden Personen verblindet erhoben wird (verblindete Zielgrößen-erhebung). Falls eine verblindete Zielgrößen-erhebung nicht möglich ist, sollte ein möglichst objektiver Endpunkt gewählt werden, der in seiner Ausprägung und in der Stringenz der Erfassung so wenig wie möglich durch diejenige Person, die den Endpunkt (unverblindet) erhebt, beeinflusst werden kann.

Zur Bewertung des Verzerrungspotenzials werden standardisierte Bewertungsbogen verwendet. Für kontrollierte Studien zur Nutzenbewertung von Interventionen werden insbesondere folgende endpunktübergreifende und endpunktspezifische Punkte regelhaft herangezogen:

Endpunktübergreifende Punkte:

- adäquate Erzeugung der Randomisierungssequenz (bei randomisierten Studien)
- Verdeckung der Gruppenzuteilung (Allocation Concealment; bei randomisierten Studien)
- zeitliche Parallelität der Interventionsgruppen (bei nicht randomisierten Studien)
- Vergleichbarkeit der Interventionsgruppen bzw. adäquate Berücksichtigung von prognostisch relevanten Faktoren (bei nicht randomisierten Studien)
- Verblindung der Patientinnen und Patienten sowie der behandelnden Personen bzw. weiterbehandelnden Personen
- ergebnisunabhängige Berichterstattung aller relevanten Endpunkte

Endpunktspezifische Punkte:

- Verblindung der Endpunkterheber
- adäquate Umsetzung des Intention-to-treat-Prinzips
- ergebnisunabhängige Berichterstattung einzelner Endpunkte

Für randomisierte Studien wird anhand dieser Aspekte das Verzerrungspotenzial zusammenfassend als niedrig oder hoch eingestuft. Ein niedriges Verzerrungspotenzial liegt dann

vor, wenn mit großer Wahrscheinlichkeit ausgeschlossen werden kann, dass die Ergebnisse relevant verzerrt sind.

Für die Bewertung eines Endpunkts wird zunächst das Verzerrungspotenzial endpunktübergreifend als niedrig oder hoch eingestuft. Falls diese Einstufung als hoch erfolgt, wird das Verzerrungspotenzial für den Endpunkt auch als hoch bewertet. Ansonsten finden die endpunktspezifischen Punkte Berücksichtigung.

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

Für nicht randomisierte vergleichende Studien wird in der Regel keine zusammenfassende Bewertung der Verzerrungsaspekte durchgeführt, da die Ergebnisse dieser Studien aufgrund der fehlenden Randomisierung generell ein hohes Verzerrungspotenzial besitzen. Bei der Bewertung einzelner Aspekte des Verzerrungspotenzials orientiert sich das Institut an den Kriterien des ROBINS-I-Instruments [684]. Spezifisch von diesem Vorgehen abgewichen wird im Rahmen der Potenzialbewertung (siehe Abschnitt 3.8).

Für die Bewertung von Studien zur diagnostischen Güte werden die QUADAS-2-Kriterien [755,756] berücksichtigt, die ggf. projektspezifisch angepasst werden. Für die methodische Bewertung von Prognosestudien wird primär das PROBAST-Instrument [766] verwendet (siehe Abschnitt 3.5).

Tritt der Fall ein, dass bei einer Projektarbeit des Instituts ältere Studien bewertet werden müssen, die den heute gültigen Qualitätsansprüchen nicht genügen, weil sie in einer Zeit geplant und durchgeführt wurden, zu der diese Ansprüche noch nicht existierten, so zeigt das Institut die Nachteile und Mängel dieser Studien auf und diskutiert mögliche Konsequenzen. Ein unterschiedlicher Umgang im Vergleich zu neueren Studien, die ähnliche Qualitätsmängel haben, ist jedoch nur dann nötig, wenn dies aus der Fragestellung oder anderen Umständen der Bewertung heraus klar begründbar ist.

Die Bewertung formaler Kriterien liefert wesentliche Anhaltspunkte für das Verzerrungspotenzial der Ergebnisse von Studien. Das Institut nimmt jedoch immer eine über rein formale Aspekte hinausgehende Bewertung des Verzerrungspotenzials vor, um beispielsweise Fehler, Widersprüche und Ungereimtheiten in Publikationen darzustellen, und prüft deren Relevanz für die Ergebnisinterpretation.

9.1.5 Interpretation von kombinierten Endpunkten

Ein sogenannter kombinierter Endpunkt umfasst eine von den Untersucherinnen und Untersuchern definierte Gruppe von Ereignissen (z. B. Herzinfarkte, Schlaganfälle, kardiovaskuläre Todesfälle). Dabei unterscheiden sich die einzelnen Ereignisse dieser Gruppe nicht selten in ihrem Schweregrad und in ihrer Bedeutung für Patientinnen und Patienten sowie Ärztinnen und

Ärzte (z. B. Krankenhausaufnahme und kardiovaskuläre Todesfälle). Daher muss man sich bei der Interpretation kombinierter Endpunkte der mit dieser Tatsache verbundenen Konsequenzen bewusst sein [147,248,264]. Die folgenden Ausführungen beschreiben die bei der Ergebnisinterpretation zu berücksichtigenden Aspekte. Sie beziehen sich aber ausdrücklich nicht auf die (ggf. abschließende) Nutzen- / Schadenbewertung mittels kombinierter Endpunkte, wenn beispielsweise der mögliche Schaden einer Intervention (z. B. Häufung schwerer Blutungen) mit dem Nutzen (z. B. Verminderung von Herzinfarkten) in einem Endpunkt erfasst wird.

Eine Voraussetzung für die Berücksichtigung eines kombinierten Endpunkts ist, dass die einzelnen Komponenten des kombinierten Endpunkts sämtliche im Berichtsplan definierten patientenrelevanten Endpunkte darstellen. Hierbei können nur dann Surrogatendpunkte eingeschlossen werden, wenn sie vom Institut dezidiert als valide akzeptiert sind (siehe Abschnitt 3.1.2). Das Resultat für jedes im kombinierten Endpunkt zusammengefasste Einzelereignis soll auch isoliert berichtet werden. Die Komponenten sollen von ähnlicher „Schwere“ sein. D. h. nicht, dass sie von identischer Bedeutung sein müssen. Beispielsweise kann die Mortalität mit dem symptomatischen Herzinfarkt oder Schlaganfall kombiniert werden, jedoch nicht mit dem stillen Herzinfarkt oder der stationären Aufnahme.

Erfüllt ein kombinierter Endpunkt die oben genannten Voraussetzungen, dann sind bei der Interpretation für Aussagen zum Nutzen und Schaden folgende Aspekte zu beachten:

- Ist der Effekt der Intervention auf die einzelnen Komponenten des kombinierten Endpunkts in der Regel gleichgerichtet?
- Wurde ein relevanter, eigentlich in die Kombination passender Endpunkt ohne nachvollziehbare und akzeptable Begründung nicht eingeschlossen oder ausgeschlossen?
- Wurde der kombinierte Endpunkt vorab definiert oder post hoc eingeführt?

Sofern es die vorliegenden Daten und Datenstrukturen erlauben, können Sensitivitätsanalysen unter Ausschluss versus Hinzunahme einzelner Komponenten durchgeführt werden.

Bei entsprechenden Voraussetzungen können im Rahmen der Nutzenbewertung einzelne Endpunkte aus einem kombinierten Endpunkt ermittelt und berechnet werden.

9.1.6 Bewertung der Konsistenz von Daten

Zur Einschätzung der Aussagekraft der Studienergebnisse überprüft das Institut die Daten auf Konsistenz (Plausibilität, Vollständigkeit). Unplausible Daten entstehen zum einen durch eine fehlerhafte Darstellung der Ergebnisse (Tipp-, Formatierungs- oder Rechenfehler), zum anderen durch eine mangel- beziehungsweise fehlerhafte Beschreibung der Methodik oder sogar durch gefälschte oder erfundene Daten [12]. Inkonsistenzen können innerhalb einer Publikation bestehen, aber auch zwischen verschiedenen Publikationen zu derselben Studie.

Ein Problem vieler Publikationen sind unvollständige Angaben im Methodik- und Ergebnisteil. Insbesondere die Darstellung von Lost-to-Follow-up-Patientinnen und -Patienten, Therapieabbrechern etc. sowie deren Art der Berücksichtigung in den Auswertungen sind oft nicht transparent.

Es ist daher angezeigt, mögliche Inkonsistenzen in den Daten aufzudecken. Dazu überprüft das Institut beispielsweise Rechenschritte und vergleicht die Angaben im Text und in den Tabellen und Grafiken. Ein in der Praxis häufiges Problem bei Überlebenszeitanalysen sind Inkonsistenzen zwischen den Angaben zu Lost-to-Follow-up-Patientinnen und -Patienten und denen zu Patientinnen und Patienten unter Risiko bei der grafischen Darstellung von Überlebenskurven. Bei bestimmten Endpunkten, zum Beispiel der Gesamtmortalität, kann die Anzahl der Lost-to-Follow-up-Patientinnen und -Patienten berechnet werden, wenn die Kaplan-Meier-Schätzer mit den Patientinnen und Patienten unter Risiko zu einem Zeitpunkt vor der minimalen Follow-up-Zeit abgeglichen werden. Bei der Aufdeckung erfundener oder gefälschter Daten können statistische Techniken hilfreich sein [12].

Finden sich relevante Inkonsistenzen in der Ergebnisdarstellung, besteht das Ziel des Instituts in der Klärung dieser Inkonsistenzen und / oder Beschaffung fehlender Informationen, beispielsweise durch den Kontakt mit den Autorinnen und Autoren oder die Anforderung von kompletten Studienberichten und weiteren Studienunterlagen. Dabei ist allerdings zu bedenken, dass erstens Autorenanfragen – besonders wenn die Publikation länger zurückliegt – häufig nicht beantwortet werden und dass sie zweitens zu weiteren Inkonsistenzen führen können. Insofern ist im Einzelfall eine Abwägung des Aufwands und Nutzens solcher Anfragen sinnvoll und notwendig. Lassen sich Inkonsistenzen nicht klären, wird der mögliche Einfluss der Inkonsistenzen auf die Effektstärken (Stärke des Bias), die Ergebnisunsicherheit (Erhöhung der Fehlerwahrscheinlichkeiten) und die Präzision (Breite der Konfidenzintervalle) vom Institut abgeschätzt. Dazu können beispielsweise Sensitivitätsanalysen durchgeführt werden. Besteht die Möglichkeit, dass Inkonsistenzen die Ergebnisse relevant beeinflussen, wird dieser Sachverhalt dargestellt und die Ergebnisse werden sehr zurückhaltend interpretiert.

9.2 Berücksichtigung systematischer Übersichten

9.2.1 Einordnung systematischer Übersichten

Auf die Ergebnisse einzelner wissenschaftlicher Studien zu vertrauen, kann irreführend sein. Lediglich eine oder nur manche Studien isoliert von anderen, ähnlichen Studien zu derselben Fragestellung zu betrachten, kann dazu führen, dass Therapien mehr oder weniger nützlich erscheinen, als sie es tatsächlich sind. Das Ziel von hochwertigen systematischen Übersichten ist es, diese Art von Verzerrungen dadurch in den Griff zu bekommen, dass die Evidenz nicht selektiv, sondern systematisch, reproduzierbar und auf transparente Weise identifiziert, bewertet und zusammengefasst wird [197,212,286,567].

In systematischen Übersichten wird die Evidenz aus einem oder mehreren Studientypen, der / die die beste Antwort auf eine spezifische und klar formulierte Frage geben kann / können,

identifiziert, bewertet und zusammengefasst. Zur Identifizierung, Auswahl und kritischen Bewertung der für die interessierende Frage relevanten Studien werden systematische und explizite Methoden angewendet. Hat man Studien identifiziert, werden diese Daten systematisch erhoben und ausgewertet. Systematische Übersichten sind nicht experimentelle Studien, deren Methodik darauf abzielen muss, systematische Fehler (Verzerrungen) auf jeder Ebene des Reviewprozesses zu minimieren [212,353].

Im Falle von systematischen Übersichten über die Wirkungen medizinischer Interventionen geben randomisierte kontrollierte Studien die zuverlässigsten Antworten. Bei anderen Fragestellungen wie etwa zur Ätiologie oder Prognose oder bei der qualitativen Beschreibung von Patientenerfahrungen besteht die geeignete Evidenzbasis für eine systematische Übersicht aus anderen Arten von Primärstudien [286]. Systematische Übersichten über Diagnose- und Screeningtests weisen im Vergleich zu Übersichten über Behandlungsmaßnahmen einige methodische Unterschiede auf [160].

Systematische Übersichten werden für die Arbeit des Instituts zumeist dazu herangezogen, potenziell relevante (Primär-)Studien zu identifizieren. Eine Bewertung kann aber auch teilweise oder sogar ausschließlich auf systematischen Übersichten beruhen (siehe Abschnitt 9.2.2). Die vom Institut für Patientinnen und Patienten sowie Bürgerinnen und Bürger erstellten Gesundheitsinformationen stützen sich zum großen Teil auf systematische Übersichten. Dazu zählen systematische Übersichten über Therapien sowie Übersichten, in denen es um andere Fragen wie Ursachen, unerwünschte Wirkungen und um die Synthese qualitativer Forschungsarbeiten geht (siehe Abschnitt 9.4).

Damit das Institut eine systematische Übersicht über Behandlungseffekte verwenden kann, muss diese gewisse Mindestvoraussetzungen erfüllen, d. h., sie darf nach dem Oxman-Guyatt-Index [410,557,559], anhand des AMSTAR- [652-654], AMSTAR-2- [655] oder ROBIS-Instruments [754] nur geringe methodische Mängel aufweisen. Neben der Stärke der in systematischen Übersichten untersuchten Evidenz berücksichtigt das Institut auch die Relevanz und die Übertragbarkeit der Evidenz. Dazu gehört auch die Untersuchung der Frage, ob die Ergebnisse in verschiedenen Populationen und Untergruppen sowie in verschiedenen Gesundheitsbereichen einheitlich sind. Berücksichtigt werden i. d. R. folgende Faktoren: die Population der Studienteilnehmerinnen und -teilnehmer in den eingeschlossenen Studien (darunter Geschlecht und Erkrankungsrisiko zu Beginn der Studie), der medizinische Kontext (darunter die medizinischen Versorgungsbereiche und die medizinischen Leistungserbringer) sowie die Anwendbarkeit und wahrscheinliche Akzeptanz der Intervention in der Form, in der sie bewertet wurde [63,157].

9.2.2 Nutzenbewertung auf Basis systematischer Übersichten

Eine Nutzenbewertung auf Basis systematischer Übersichten kann – sofern bestimmte Voraussetzungen erfüllt sind – eine ressourcensparende und verlässliche Evidenzbasis für

Empfehlungen an den G-BA bzw. das BMG liefern [148,462]. Für die Verwendung systematischer Übersichten zur Nutzenbewertung ist es notwendig, dass die systematischen Übersichten von ausreichend hoher Qualität sind, d. h.

- nur minimales Verzerrungspotenzial aufweisen sowie
- die Evidenzlage vollständig, transparent und nachvollziehbar darstellen,

und so erlauben, eine klare Schlussfolgerung zu ziehen [27,557,757]. Außerdem ist es eine notwendige Voraussetzung, dass die jeweilige Informationsbeschaffung der Methodik des Instituts nicht widerspricht (siehe Abschnitt 8.2.1) und die Übertragbarkeit der Ergebnisse auf die jeweilige Fragestellung, insbesondere unter Berücksichtigung der genannten Ein- und Ausschlusskriterien, möglich ist.

Die angewendete Methodik muss eine hinreichende Sicherheit dafür bieten, dass eine neue Nutzenbewertung auf Basis der Primärliteratur nicht zu anderen Schlussfolgerungen käme als die Bewertung auf Basis systematischer Übersichten. Dies ist z. B. regelhaft dann nicht der Fall, wenn eine relevante Menge bislang unpublizierter Daten zu erwarten ist.

A) Fragestellungen

Prinzipiell ist diese Methode für alle Fragestellungen geeignet, sofern oben genannte Kriterien erfüllt sind. Bei der Entwicklung der Fragestellung sind insbesondere die folgenden Punkte zu berücksichtigen:

- die Definition der interessierenden Population,
- die Definition der interessierenden Intervention und Vergleichsintervention,
- die Definition aller relevanten Endpunkte,
- ggf. das fokussierte Gesundheitssystem bzw. der geografische Bezug (z. B. Deutschland, Europa).

Die so definierte Frage bildet auch die Grundlage für die Festlegung der für die Nutzenbewertung geltenden Ein- und Ausschlusskriterien und damit für die Feststellung der inhaltlichen und methodischen Relevanz der gefundenen Publikationen. Auf Basis der Fragestellung wird auch entschieden, auf welchem Primärstudientyp die verwendeten systematischen Übersichten beruhen müssen. Je nach Fragestellung ist es möglich, dass bestimmte Teile eines Auftrags auf Basis von systematischen Übersichten beantwortet werden, während für andere Teile auf die Primärliteratur zurückgegriffen wird.

B) Mindestzahl relevanter systematischer Übersichten

Alle qualitativ ausreichenden und thematisch relevanten systematischen Übersichten werden berücksichtigt. In der Regel sollten mindestens 2 Arbeiten von hoher Qualität, die unabhängig voneinander durchgeführt wurden, als Grundlage für die Erstellung eines Berichts auf Basis von Sekundärliteratur vorhanden sein, um die Konsistenz der Ergebnisse überprüfen zu können.

Wenn nur eine Arbeit von hoher Qualität vorliegt und dementsprechend berücksichtigt werden kann, ist eine Begründung für die Bewertung allein anhand dieser einen systematischen Übersicht erforderlich.

C) Qualitätsbewertung der Publikationen einschließlich Mindestanforderungen

Die Bewertung der allgemeinen Qualität systematischer Übersichten erfolgt anhand des validierten Qualitätsindex von Oxman und Guyatt [556,557,559], des AMSTAR- [652-654], des AMSTAR-2- [655] oder des ROBIS-Instruments [754]. Systematische Übersichten gelten nach dem Oxman-Guyatt-Index dann als qualitativ ausreichend, wenn diese von 2 voneinander unabhängigen Gutachterinnen oder Gutachtern in der Gesamtbewertung mit mindestens 5 von 7 möglichen Punkten bewertet wurden. Für AMSTAR, AMSTAR 2 und ROBIS sind keine derartigen Grenzwerte festgelegt, ein solcher muss also ggf. vorab definiert werden. Darüber hinaus werden regelhaft die Sponsorinnen und Sponsoren der systematischen Übersichten sowie die Beziehungen der Autorinnen und Autoren dokumentiert und diskutiert. Die jeweiligen Indexkriterien können je nach Anforderung des Projekts um weitere Items ergänzt werden (beispielsweise Vollständigkeit der Recherche, Recherchen nach unpublizierten Studien z. B. in Registern oder zusätzliche Aspekte bei systematischen Übersichten von Studien zur diagnostischen Güte).

D) Ergebnisse

Die Ergebnisse einer Nutzenbewertung anhand von systematischen Übersichten werden zu jeder Fragestellung, wenn möglich, in tabellarischer Form zusammengefasst. Beim Vorliegen inkonsistenter Ergebnisse aus mehreren Arbeiten zu einem Endpunkt werden mögliche Erklärungen für diese Heterogenität beschrieben [409].

Liefert die Zusammenstellung der systematischen Übersichten Hinweise darauf, dass eine neu erstellte Nutzenbewertung auf Basis von Primärstudien zu anderen Ergebnissen führen könnte, so wird eine Nutzenbewertung auf Basis von Primärstudien durchgeführt.

E) Schlussfolgerung / Empfehlung

Nutzenbewertungen auf Basis systematischer Übersichten fassen die Ergebnisse der zugrunde liegenden systematischen Übersichten zusammen, ergänzen sie ggf. um die Zusammenfassung der aktuelleren Primärstudien oder Primärstudien zu Fragestellungen, die nicht durch die systematischen Übersichten abgedeckt werden, und ziehen daraus ein eigenständiges Fazit.

Die Empfehlungen auf Basis systematischer Übersichten beruhen nicht auf der Zusammenfassung der Empfehlungen bzw. Schlussfolgerungen der zugrunde liegenden systematischen Übersichten. In HTA-Berichten werden diese häufig vor dem Hintergrund spezifischer sozialpolitischer und ökonomischer Rahmenbedingungen eines spezifischen Gesundheitssystems formuliert und sind daher nur selten auf die Versorgungssituation in Deutschland übertragbar.

9.2.3 Berücksichtigung publizierter Metaanalysen

Den Bewertungen des Instituts liegt, dem internationalen Standard der evidenzbasierten Medizin folgend, in der Regel eine fragestellungsbezogene umfassende Informationsbeschaffung zu relevanten Primärstudien zugrunde. Sofern sinnvoll und möglich, werden die Ergebnisse der identifizierten Einzelstudien mittels Metaanalysen zusammenfassend bewertet. Das Institut hat dadurch allerdings in der Regel lediglich Zugriff auf aggregierte Daten der Primärstudien auf Basis der jeweiligen Publikation oder des zur Verfügung gestellten Studienberichts. Es gibt Situationen, in denen Metaanalysen, die auf Basis der individuellen Patientendaten (IPD) der relevanten Studien durchgeführt werden, eine höhere Aussagekraft haben (siehe Abschnitt 9.3.7). Dies ist insbesondere dann der Fall, wenn neben dem alleinigen Interventionseffekt die Evaluation weiterer Faktoren von Interesse ist, die den Interventionseffekt möglicherweise beeinflussen (Interaktion zwischen Interventionseffekt und Kovariablen). Hier liefern Metaanalysen mit individuellen Patientendaten im Vergleich zu Metaregressionen, die auf aggregierten Daten beruhen, im Allgemeinen eine höhere Ergebnissicherheit, d. h. präzisere und nicht durch ökologischen Bias verzerrte Ergebnisse [665]. Solche Analysen können daher im Einzelfall zu genaueren Schlussfolgerungen führen, insbesondere bei Vorliegen heterogener Ergebnisse, die möglicherweise auf unterschiedliche Ausprägungen von Patientenmerkmalen zurückgeführt werden können. Von einer erhöhten Validität von Metaanalysen basierend auf individuellen Patientendaten kann man allerdings nur dann ausgehen, wenn solche Analysen auch tatsächlich auf die Fragestellung der Bewertung des Instituts ausgerichtet sind und zudem eine hohe Ergebnissicherheit aufweisen. Für die Beurteilung der Ergebnissicherheit solcher Analysen ist eine maximale Transparenz Grundvoraussetzung. Dies bezieht sich sowohl auf die Planung als auch auf die Durchführung der Analysen. Diesbezüglich wesentliche, für die Durchführung von Metaanalysen allgemeingültige Aspekte sind z. B. im Preferred-Reporting-Items-for-Systematic-Reviews-Meta-Analyses (PRISMA)-Statement für Metaanalysen randomisierter Studien [518], im zugehörigen Erläuterungsdokument [473], im PRISMA-IPD-Statement für Metaanalysen mit individuellen Patientendaten [686], im PRISMA-P-Statement für Protokolle systematischer Übersichten [520], im zugehörigen Erläuterungsdokument [651], in der PRISMA-Harms-Checkliste [775], im PRISMA-DTA-Statement für Metaanalysen von Studien zur diagnostischen Güte [514] und in einem Dokument der EMA beschrieben [225]. Das Institut berücksichtigt bei der Nutzenbewertung publizierte Metaanalysen auf Basis individueller Patientendaten, sofern sie (Unter-)Fragestellungen des Berichts, die durch Metaanalysen auf Basis aggregierter Daten nicht ausreichend sicher beantwortet werden können, adressieren und von einer hohen Ergebnissicherheit der jeweiligen Analyse ausgegangen werden kann.

9.3 Spezielle biometrische Aspekte

9.3.1 Darstellung von Effekten und Risiken

Die Darstellung von Interventions- oder Expositionseffekten ist zunächst eindeutig an eine klare Zielvariable zu knüpfen. Die Betrachtung einer alternativen Zielvariablen verändert auch

die Darstellung und die Stärke eines möglichen Effekts. Die Wahl eines geeigneten Effektmaßes hängt grundsätzlich vom Messniveau der betrachteten Zielvariablen ab. Bei stetigen Variablen können Effekte in der Regel mithilfe von Mittelwerten sowie Differenzen von Mittelwerten – unter Umständen nach geeigneter Gewichtung – dargestellt werden. Bei kategoriellen Zielvariablen kommen die üblichen Effekt- und Risikomaße aus Vierfeldertafeln infrage [44]. Eine übersichtliche Zusammenfassung der Vor- und Nachteile der üblichen Effektmaße in systematischen Übersichten gibt Kapitel 10 des *Cochrane Handbook for Systematic Reviews of Interventions* [162]. Besondere Aspekte, die beim Vorliegen ordinaler Daten berücksichtigt werden müssen, beschreibt Agresti [9,10].

Zwingend erforderlich ist es, bei jeder Effektschätzung den Grad der statistischen Unsicherheit zu beschreiben. Häufige Methoden hierfür stellen die Berechnung des Standardfehlers sowie die Angabe eines Konfidenzintervalls dar. Wann immer möglich, gibt das Institut adäquate Konfidenzintervalle für Effektschätzungen mit der Information, ob es sich um 1- oder 2-seitige Konfidenzgrenzen handelt, sowie das gewählte Konfidenzniveau an. In der medizinischen Forschung wird üblicherweise das 2-seitige Konfidenzniveau 95 % verwendet, in manchen Situationen aber auch 90 % oder 99 %. Eine Übersicht über die häufigsten Methoden zur Berechnung von Konfidenzintervallen geben Altman et al. [18].

Je nach Datensituation (z. B. bei sehr kleinen Stichproben) und Fragestellung ist zur Einhaltung des Konfidenzniveaus die Anwendung exakter Methoden bei der Intervallschätzung von Effekten und Risiken in Erwägung zu ziehen. Eine aktuelle Diskussion exakter Methoden liefert Agresti [11].

9.3.2 Beurteilung statistischer Signifikanz

Mithilfe statistischer Signifikanztests ist es möglich, a priori formulierte Hypothesen unter Kontrolle der Irrtumswahrscheinlichkeit erster Art zu testen. Die Konvention, von einem statistisch signifikanten Resultat zu sprechen, falls der p-Wert das Signifikanzniveau 0,05 unterschreitet ($p < 0,05$), ist in vielen Fällen durchaus sinnvoll. Je nach Fragestellung und Aussage kann es erforderlich sein, ein möglicherweise kleineres Signifikanzniveau zu fordern. Umgekehrt gibt es Situationen, in denen auch ein höheres Signifikanzniveau akzeptabel sein kann. Solche Ausnahmen begründet das Institut immer explizit.

Bei der Interpretation von p-Werten ist eine Reihe von Aspekten zu betrachten. Es muss im Detail klar sein, auf welche Fragestellung und welche Datensituation sich die Signifikanzangabe bezieht und wie die statistische Hypothesenformulierung lautet. Insbesondere muss klar sein, ob es sich um eine 1- oder 2-seitige Hypothese handelt [60] und ob die betrachtete Hypothese als Teil eines Mehrhypothesenproblems zu sehen ist [719]. Beide Aspekte – ob eine Hypothese 1- oder 2-seitig zu formulieren ist und ob für multiples Testen adjustiert werden muss – werden in der wissenschaftlichen Literatur immer wieder kontrovers diskutiert [243,441].

Bezüglich der Hypothesenformulierung ist traditionell von einem 2-seitigen Testproblem auszugehen. Ausnahmen hiervon bilden zum Beispiel Nichtunterlegenheitsstudien. Die Formulierung eines 1-seitigen Hypothesenproblems ist grundsätzlich immer möglich, bedarf jedoch einer genauen Begründung. Im Fall einer 1-seitigen Hypothesenformulierung sind die Anwendung 1-seitiger Signifikanztests sowie die Berechnung 1-seitiger Konfidenzgrenzen angebracht. Zur besseren Vergleichbarkeit mit 2-seitigen statistischen Verfahren wird in einigen Guidelines für klinische Studien eine Halbierung des üblichen Signifikanzniveaus von 5 % auf 2,5 % gefordert [395]. Das Institut folgt in der Regel diesem Vorgehen. Das Institut folgt weiterhin dem zentralen Grundprinzip, dass die Hypothesenformulierung (1- oder 2-seitig) und das Signifikanzniveau a priori klar festzulegen sind, und begründet Abweichungen von den üblichen Festlegungen (1-seitige anstelle 2-seitiger Hypothesenformulierung, Signifikanzniveau ungleich 5 %) bzw. achtet auf entsprechende Begründungen in der Primärliteratur.

Ist eine betrachtete Hypothese eindeutig Teil eines Mehrhypothesenproblems, so ist eine adäquate Adjustierung für multiples Testen erforderlich, wenn der Fehler 1. Art für das gesamte Mehrhypothesenproblem kontrolliert werden soll [52]. Das Problem der Multiplizität ist in systematischen Übersichten nicht komplett lösbar, sollte aber dennoch zumindest bei der Interpretation der Ergebnisse beachtet werden [47]. Wenn sinnvoll und möglich, wendet das Institut Methoden zur Adjustierung für multiples Testen an. Im Rahmen der Nutzenbewertung (siehe Abschnitt 3.1) versucht das Institut, den Fehler 1. Art bezüglich der Aussagen zu jedem einzelnen Nutzenaspekt getrennt zu kontrollieren. Die zusammenfassende Bewertung wird in der Regel nicht auf quantitative Weise durchgeführt, sodass hier auch keine formalen Methoden zur Adjustierung für multiples Testen angewendet werden können.

Das Institut wertet ein statistisch nicht signifikantes Ergebnis nicht als Evidenz für das Nichtvorhandensein eines Effekts (Abwesenheit bzw. Äquivalenz) [16]. Zum Nachweis einer Gleichheit wendet das Institut adäquate Methoden für Äquivalenzhypothesen an.

Als prinzipielle Alternative zu statistischen Signifikanztests können bayessche Methoden angesehen werden [676,677]. Das Institut wird in Abhängigkeit von der Fragestellung bei Bedarf auch bayessche Methoden anwenden (z. B. bei indirekten Vergleichen, siehe Abschnitt 9.3.8).

9.3.3 Beurteilung klinischer Relevanz

Mit dem Begriff klinische Relevanz werden in der Literatur unterschiedliche Konzepte bezeichnet. Zum einen kann es auf der Gruppenebene um die Frage gehen, ob ein Unterschied zwischen 2 Therapiealternativen bei einem patientenrelevanten Endpunkt (z. B. schwerwiegende Nebenwirkungen) groß genug ist, um eine allgemeine Anwendung der besseren Alternative zu empfehlen. Zum anderen wird unter klinischer Relevanz die Frage verstanden, ob eine Veränderung (z. B. der beobachtete Unterschied von 1 Punkt auf einer Symptomskala) für einzelne Patientinnen und Patienten relevant ist. Soweit das 2. Konzept im Sinne einer Responderdefinition und entsprechender Responderanalysen zur Betrachtung von Gruppenunterschieden führt, sind beide Konzepte für die Bewertungen des Instituts von Bedeutung.

Allgemein spielt im Rahmen von systematischen Übersichten und Metaanalysen die Bewertung der klinischen Relevanz von Gruppenunterschieden eine besondere Rolle, da hier oftmals eine Power erreicht wird, die das statistische Aufdecken kleinster Effekte ermöglicht [730]. Dabei ist die klinische Relevanz eines Effekts oder Risikos grundsätzlich nicht am p-Wert ablesbar. Die statistische Signifikanz ist eine Wahrscheinlichkeitsaussage, in die neben der Stärke eines möglichen Effekts auch die Variabilität der Daten und der Stichprobenumfang eingehen. Bei der Interpretation der Bedeutsamkeit von p-Werten muss insbesondere der Stichprobenumfang der zugrunde liegenden Studie berücksichtigt werden [599]. In einer kleinen Studie ist ein sehr kleiner p-Wert nur bei einem deutlichen Effekt zu erwarten, während in einer großen Studie auch bei einem verschwindend kleinen Effekt eine deutliche Signifikanz nicht ungewöhnlich ist [242,371]. Folglich lässt sich die klinische Relevanz eines Studienergebnisses auf keinen Fall am p-Wert ablesen.

Für die Beurteilung der klinischen Relevanz von Studienergebnissen gibt es noch kein breit akzeptiertes methodisches Vorgehen, unabhängig davon, welches der oben genannten Konzepte adressiert wird. So finden sich beispielsweise nur in wenigen Leitlinien Hinweise auf die Definition von relevanten beziehungsweise irrelevanten Unterschieden zwischen Gruppen [460,707]. Auch Methodenhandbücher zur Erstellung von systematischen Übersichten, wie z. B. das Cochrane Handbook [353], enthalten i. d. R. entweder keine oder keine eindeutige Anleitung zur Bewertung klinischer Relevanz auf System- oder Individualebene. Es gibt aber verschiedene Ansätze zur Beurteilung der klinischen Relevanz von Studienergebnissen. So kann z. B. der beobachtete Unterschied (Effektschätzer und das dazugehörige Konfidenzintervall) ohne prädefinierte Schwellen allein auf Basis medizinischer Sachkenntnis bewertet werden. Als formales Relevanzkriterium kann alternativ gefordert werden, dass das Konfidenzintervall vollständig oberhalb einer bestimmten Irrelevanzschwelle liegen muss, um hinreichend sicher einen klar irrelevanten Effekt ausschließen zu können. Dies entspricht dann der Anwendung eines statistischen Tests mit Verschiebung der Nullhypothese zum statistischen Nachweis relevanter klinischer Effekte [763]. Ein weiterer Vorschlag sieht vor, dass die Relevanz allein auf Basis des Effektschätzers (im Vergleich zu einer Relevanzschwelle) beurteilt wird, vorausgesetzt, es liegt ein statistisch signifikanter Unterschied zwischen den Interventionsgruppen vor [436]. Im Gegensatz zur Anwendung eines statistischen Tests mit Verschiebung der Nullhypothese kann die Wahrscheinlichkeit für den Fehler 1. Art durch die Relevanzbewertung anhand des Effektschätzers nicht kontrolliert werden. Darüber hinaus ist dieser Ansatz unter Umständen weniger effizient. Schließlich liegt eine weitere Möglichkeit zur Relevanzbewertung darin, ein Relevanzkriterium auf individueller Ebene zu formulieren, z. B. im Sinne einer Responderdefinition [437]. Dabei gibt es auch Ansätze, bei denen sich das Responsekriterium innerhalb einer Studie zwischen den untersuchten Probanden unterscheidet, indem vorab individuelle Therapieziele formuliert werden [592].

Patientenrelevante Endpunkte können auch mithilfe von (komplexen) Skalen erhoben werden. Eine Voraussetzung für die Berücksichtigung solcher Endpunkte ist die Verwendung von validierten bzw. etablierten Instrumenten. Bei der Betrachtung patientenrelevanter Endpunkte,

die mithilfe von (komplexen) Skalen operationalisiert werden, ist es in besonderer Weise notwendig, neben der statistischen Signifikanz der Effekte die Relevanz der beobachteten Wirkungen der untersuchten Interventionen zu bewerten, da die Komplexität der Skalen häufig eine sinnvolle Interpretation geringfügiger Unterschiede erschwert. Hier geht es also um die Frage, ob der beobachtete Unterschied zwischen 2 Gruppen überhaupt für die Patientinnen und Patienten spürbar ist. Diese Relevanzbewertung kann auf Basis von Mittelwertdifferenzen und Responderanalysen vorgenommen werden [641]. Ein wesentliches Problem bei der Relevanzbewertung ist die Tatsache, dass Relevanzkriterien skalenspezifisch nicht definiert sind oder aber adäquate Auswertungen auf Basis skalenspezifischer Relevanzkriterien (z. B. Responderanalysen) fehlen [522]. Welches Verfahren bei den Bewertungen des Instituts gewählt werden kann, hängt daher von der Verfügbarkeit der Daten aus den Primärstudien ab.

In den letzten Jahren sind vermehrt Responderanalysen auf Basis eines Responsekriteriums im Sinne einer individuellen Minimal important Difference (MID) durchgeführt worden. Jedoch werden zunehmend methodische Probleme dieses Ansatzes sichtbar.

Grundsätzlich werden für die Bestimmung von MIDs empirische Verfahren präferiert, die eine MID mithilfe patientenberichteter Anker ermitteln; daneben werden auch verteilungsbasierte Verfahren eingesetzt [555,585]. Systematische Zusammenstellungen empirisch ermittelter MIDs ergeben, dass zu einzelnen Instrumenten häufig eine Vielzahl von MIDs publiziert werden, die innerhalb eines Erhebungsinstruments große Spannweiten haben können [122,126,320,544]. Diese Variabilität kann unter anderem dadurch zustande kommen, dass in Studien zur Ermittlung von MIDs unterschiedliche Anker, Beobachtungsperioden oder analytische Methoden eingesetzt werden [180,544,555]. Es gibt derzeit keinen etablierten Standard, mit dem die Qualität dieser Studien bewertet und die Aussagekraft der ermittelten MIDs abgeschätzt werden kann [122,180,419]. Ein erster Vorschlag für ein Instrument zur Qualitätsbewertung von Studien zur Ermittlung von MIDs und zur Aussagekraft der MIDs wurde im Jahr 2020 veröffentlicht [179]. Umfassende Ergebnisse zur Anwendbarkeit stehen noch aus. Jedoch zeigt eine Studie zur Reliabilität dieses neuen Instruments, dass in Publikationen von Studien zur Ermittlung einer MID wesentliche Anteile der Methodik bzw. der Kriterien für eine Bewertung zumeist gar nicht berichtet werden [179]. Eine anhand methodischer Qualitätskriterien begründete Auswahl einer empirisch ermittelten MID für die Nutzenbewertung ist somit derzeit nicht zu treffen [178,180,419].

Neben den methodischen Faktoren, die die Größenordnung ermittelter MIDs beeinflussen, beruht ein anderer Teil der Variabilität von MIDs auf ihrer Abhängigkeit von Charakteristika der Patientenpopulation, in der das Instrument eingesetzt wird, sowie weiteren Kontextfaktoren. So können der Schweregrad der Erkrankung, die Art der eingesetzten Intervention oder die Frage, ob die Patientinnen und Patienten eine Verbesserung oder Verschlechterung ihrer Erkrankung erfahren, Einfluss auf die MID haben [13]. Der Umgang mit diesem Teil der Variabilität von MIDs ist in der internationalen Diskussion weiterhin ungeklärt.

Um vor diesem Hintergrund sicherzustellen, dass in Responderanalysen im Rahmen der Nutzenbewertung geeignete Responseschwellen eingesetzt werden, wird das im Weiteren beschriebene Verfahren angewendet. Dabei soll zum einen erreicht werden, dass eine Responseschwelle hinreichend sicher für die Patientinnen und Patienten spürbare Veränderungen abbildet. Diese Responseschwelle soll eine eher kleine Veränderung darstellen, der Wert kann aber durchaus oberhalb einer minimalen Schwelle liegen. Diese Eigenschaft berücksichtigt auch die empirische Variabilität von MIDs und trägt dazu bei, dass das Responsekriterium in unterschiedlichen Konstellationen nicht zu häufig unter die MIDs fällt. Gleichzeitig soll eine ergebnisgesteuerte Berichterstattung minimiert werden, wie sie z. B. durch eine beliebige Auswahl einer von vielen möglichen MIDs entstehen könnte.

Aus einer Sichtung von systematisch recherchierten systematischen Übersichtsarbeiten zu MIDs [14,190,209,320,414,544,555,678] wurde ein Wert von 15 % der Spannweite der jeweiligen Skalen als plausibler Schwellenwert für eine eher kleine, aber hinreichend sicher spürbare Veränderung identifiziert [393]. Daraus ergibt sich regelhaft folgendes Vorgehen:

- 1) Falls in einer Studie Responderanalysen unter Verwendung einer MID präspezifiziert sind und das Responsekriterium **mindestens** 15 % der Skalenspannweite des verwendeten Erhebungsinstruments entspricht, werden diese Responderanalysen ohne weitere Prüfung des Responsekriteriums für die Bewertung herangezogen.
- 2) Falls präspezifiziert Responsekriterien im Sinne einer MID unterhalb von 15 % der Skalenspannweite liegen, werden diese regelhaft nicht herangezogen. In diesen Fällen und solchen, in denen gar keine Responsekriterien präspezifiziert wurden, aber stattdessen Analysen kontinuierlicher Daten zur Verfügung stehen, bestehen verschiedene Möglichkeiten. Entweder können die Analysen der kontinuierlichen Daten herangezogen werden, für die Relevanzbewertung wird dabei auf ein allgemeines statistisches Maß in Form von standardisierten Mittelwertdifferenzen (SMDs, in Form von Hedges' g) zurückgegriffen. Dabei wird eine Irrelevanzschwelle von 0,2 verwendet: Liegt das zum Effektschätzer korrespondierende Konfidenzintervall vollständig oberhalb dieser Irrelevanzschwelle, wird davon ausgegangen, dass die Effektstärke nicht in einem sicher irrelevanten Bereich liegt. Dies soll gewährleisten, dass der Effekt hinreichend sicher mindestens als klein angesehen werden kann [239]. Alternativ können post hoc spezifizierte Analysen mit einem Responsekriterium von **genau** 15 % der Skalenspannweite berücksichtigt werden.
- 3) Liegen sowohl geeignete Responderanalysen (Responsekriterium präspezifiziert mindestens 15 % der Skalenspannweite oder post hoc genau 15 % der Skalenspannweite) als auch Analysen stetiger Daten vor, werden die Responderanalysen herangezogen.

9.3.4 Nachweis der Verschiedenheit

Beim empirischen Nachweis, dass sich gewisse Gruppen bezüglich eines bestimmten Merkmals unterscheiden, sind verschiedene Aspekte zu beachten. Zunächst ist festzuhalten, dass ein Nachweis hier nicht als Beweis im mathematischen Sinne zu verstehen ist. Mithilfe empirischer

Studiendaten lassen sich Aussagen nur unter Zulassung gewisser Irrtumswahrscheinlichkeiten treffen. Durch die Anwendung statistischer Methoden lassen sich diese Irrtumswahrscheinlichkeiten jedoch gezielt kontrollieren und minimieren, um auf diese Weise eine Aussage statistisch nachzuweisen. Die übliche Methode eines solchen statistischen Nachweises in der medizinischen Forschung ist die Anwendung von Signifikanztests. Diese Argumentationsebene ist zu trennen von der Beurteilung der klinischen Relevanz eines Unterschieds. Die Kombination beider Argumente liefert in der Praxis eine adäquate Beschreibung eines Unterschieds aufgrund empirischer Daten.

Bei der Anwendung eines Signifikanztests zum Nachweis eines Unterschieds sind a priori die wissenschaftliche Fragestellung und darauf basierend die Zielvariable, das Effektmaß und die statistische Hypothesenformulierung festzulegen. Damit eine Studie groß genug ist, um einen Unterschied entdecken zu können, ist es notwendig, den benötigten Stichprobenumfang vor Beginn der Studie zu berechnen. Dafür sind in einfachen Situationen neben obigen Informationen zusätzlich die Angabe des klinisch relevanten Unterschieds sowie ein Schätzwert der Variabilität der Zielgröße erforderlich. Bei komplexeren Designs oder Fragestellungen sind zusätzlich weitere Angaben wie Korrelationsstruktur, Rekrutierungsschema, Drop-out-Schema usw. erforderlich [61,168].

Zur abschließenden Darstellung der Ergebnisse gehören neben der Signifikanzangabe für eine Aussage auch ein mit adäquaten Methoden berechnetes Konfidenzintervall für das gewählte Effektmaß, die deskriptive Angabe weiterer Effektmaße, um unterschiedliche Aspekte des Ergebnisses darzulegen, und eine Diskussion der klinischen Relevanz der Ergebnisse, die auf der Ermittlung patientenrelevanter Zielkriterien aufbauen sollte.

9.3.5 Nachweis der Gleichheit

Einer der häufigen schwerwiegenden Fehler bei der Interpretation medizinischer Daten ist es, ein nicht signifikantes Ergebnis eines traditionellen Signifikanztests als Evidenz für die Richtigkeit der Nullhypothese zu werten [16]. Für den Nachweis einer Gleichheit ist die Anwendung von Methoden zur Untersuchung von Äquivalenzhypothesen erforderlich [420]. Wichtig ist hierbei zunächst das Verständnis, dass eine exakte Gleichheit – also z. B., dass die Differenz der Mittelwerte von 2 Gruppen exakt null beträgt – mithilfe statistischer Methoden nicht nachweisbar ist. In der Praxis ist aber auch nicht der Nachweis einer exakten Gleichheit gefordert, sondern der Nachweis eines höchstens irrelevanten Unterschieds zwischen 2 Gruppen. Um dieses Ziel zu erreichen, muss logischerweise zunächst definiert werden, was ein irrelevanter Unterschied ist, d. h., die Festlegung eines Äquivalenzbereichs ist erforderlich.

Um sinnvoll auf eine Äquivalenz schließen zu können, müssen – ähnlich wie beim Nachweis eines Unterschieds – a priori die wissenschaftliche Fragestellung und darauf basierend die Zielvariable, das Effektmaß und die statistische Hypothesenformulierung festgelegt werden. Zusätzlich ist bei Äquivalenzstudien der Äquivalenzbereich klar zu definieren. Dies kann entweder 2-seitig, was zu einem Äquivalenzintervall führt, oder 1-seitig im Sinne einer

höchstens irrelevanten Unterschiedlichkeit bzw. höchstens irrelevanten Unterlegenheit erfolgen. In letzterem Fall spricht man von einer Nichtunterlegenheitshypothese [152,395,594].

Wie bei Überlegenheitsstudien ist es auch bei Äquivalenzstudien erforderlich, den benötigten Stichprobenumfang vor Beginn der Studie zu berechnen. Welche Methode hierfür infrage kommt, hängt von der genauen Hypothese und von der gewählten Analyse-methode ab [593].

Für die Datenanalyse in Äquivalenzstudien müssen speziell für diesen Zweck entwickelte Methoden verwendet werden. Ein häufig angewendetes Verfahren ist die Konfidenzintervall-Einschluss-Methode. Liegt das berechnete Konfidenzintervall völlig im vorher definierten Äquivalenzbereich, so gilt dies als Nachweis der Äquivalenz. Um das Niveau $\alpha = 0,05$ einzuhalten, genügt hierbei die Berechnung eines Konfidenzintervalls zum Niveau 90 % [420]. In der Regel verwendet das Institut jedoch gemäß internationalem Vorgehen 95 %-Konfidenzintervalle.

Im Vergleich zu Überlegenheitsstudien besitzen Äquivalenzstudien spezielle methodische Probleme. Zum einen ist es häufig schwierig, Äquivalenzbereiche sinnvoll zu definieren [460]. Zum anderen schützen zum Beispiel die üblichen Designkriterien Randomisierung und Verblindung nicht mehr hinreichend sicher vor Verzerrungen [648]. Auch ohne Kenntnis der Therapiegruppe ist es zum Beispiel möglich, die Differenz der Behandlungsunterschiede zur Null und damit zur gewünschten Alternativhypothese hin zu verschieben. Des Weiteren ist mit dem Intention-to-treat-Prinzip sehr vorsichtig umzugehen, da dessen inadäquate Anwendung eine falsche Äquivalenz vortäuschen kann [420]. Somit ist bei der Bewertung von Äquivalenzstudien besondere Vorsicht geboten.

9.3.6 Adjustierung und multifaktorielle Verfahren

Vor allem in nicht randomisierten Studien spielen multifaktorielle Verfahren, die es ermöglichen, den Effekt von Confoundern auszugleichen, eine zentrale Rolle [431]. Ein weiteres wichtiges Anwendungsgebiet für multifaktorielle Methoden sind Studien mit mehreren Interventionen [510]. Die Darstellung von Ergebnissen multifaktorieller Verfahren ist jedoch in der medizinischen Fachliteratur leider oft unzureichend [49,529]. Um die Qualität einer solchen Analyse bewerten zu können, sind eine Darstellung der wesentlichen Aspekte im Rahmen der statistischen Modellbildung notwendig [326,602] sowie Angaben zur Güte des gewählten Modells (Goodness of fit) [366]. Die wichtigsten Informationen hierbei sind in der Regel:

- eine eindeutige Beschreibung und A-priori-Festlegung der Zielvariablen und aller potenziell erklärenden Variablen,
- das Messniveau und die Codierung aller Variablen,
- Angaben zur Selektion von Variablen und Wechselwirkungen,
- eine Beschreibung, wie die Modellannahmen überprüft wurden,

- Angaben zur Modellgüte,
- eine Tabelle mit den wesentlichen Ergebnissen (Parameterschätzung, Standardfehler, Konfidenzintervall) für alle erklärenden Variablen.

Je nach Fragestellung haben diese Informationen eine unterschiedliche Bedeutung. Geht es im Rahmen eines Prognosemodells um eine gute Prädiktion der Zielvariablen, so ist eine hohe Modellgüte wichtiger als bei einem Gruppenvergleich, bei dem man nach wichtigen Confoundern adjustieren muss.

Besonders kritisch ist eine unzureichende Darstellung der Ergebnisse multifaktorieller Verfahren dann, wenn es durch die (nicht ausreichend transparent dargestellte) statistische Modellierung zu einer Verschiebung von Effekten in einen „gewünschten“ Bereich kommt, der bei unifaktorieller Betrachtung so nicht erkennbar ist. Ausführliche Erläuterungen zu den Anforderungen an den Umgang mit multifaktoriellen Verfahren sind in verschiedenen Übersichten und Leitlinien in der Literatur zu finden [33,50,431].

In eigenen regressionsanalytischen Berechnungen setzt das Institut gängige Verfahren ein [325]. Sehr vorsichtig zu interpretieren sind hierbei Ergebnisse multifaktorieller Modelle, die aus einem Variablenselektionsprozess hervorgegangen sind. Falls bei der Modellwahl solche Variablenselektionsverfahren unumgänglich sind, wird eine Form der Backward-Elimination eingesetzt, da diese dem Verfahren der Forward-Selektion vorzuziehen ist [325,692]. Wichtig ist hierbei eine sorgfältige Vorauswahl der potenziellen erklärenden Variablen unter sachwissenschaftlicher Kenntnis [163]. Bei der Modellierung stetiger Einflussgrößen greift das Institut im Bedarfsfall auf flexible Modellansätze wie zum Beispiel das Verfahren der Fractional Polynomials zurück [603,630], um eine adäquate Beschreibung nicht monotoner Zusammenhänge zu ermöglichen.

9.3.7 Metaanalysen

A) Allgemeines

In der Literatur verwendete Begriffe wie Literaturübersicht, systematische Übersicht, Metaanalyse, gepoolte Analyse oder Forschungssynthese sind häufig unterschiedlich definiert und nicht klar voneinander abgegrenzt [212]. 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 Metaanalyse 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 – sofern sinnvoll – ein Gesamteffekt berechnet. Effizientere Auswertungsverfahren sind möglich, wenn individuelle Patientendaten aus den Studien verfügbar sind. Unter einer Metaanalyse 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 Metaanalyse 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 Metaanalyse erfolgt mittels Forest Plots, in denen die Effektschätzer der einzelnen Studien und des Gesamteffekts inklusive der Konfidenzintervalle grafisch aufgetragen werden [471]. 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 [217,649,738]. Liegen Informationen darüber vor, dass die Effekte der einzelnen Studien homogen sind, ist eine Metaanalyse 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 [650]. 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 [84,291]. Wenn das Vorhandensein von Heterogenität nicht ausreichend sicher ausgeschlossen werden kann, so sollte ein Modell mit zufälligen Effekten gewählt werden. Zur Durchführung von Metaanalysen mit zufälligen Effekten existieren zahlreiche Methoden [734]. Gemäß Empfehlungen aus der Literatur sollte zur Durchführung von Metaanalysen mit zufälligen Effekten bei einer ausreichenden Anzahl an Studien regelhaft die Knapp-Hartung-Methode [333,439] verwendet werden, wobei der Heterogenitätsparameter mithilfe der Paule-Mandel-Methode geschätzt wird [398,733,735].

Bei Anwendung der Knapp-Hartung-Methode für Metaanalysen mit zufälligen Effekten ist zu beachten, dass es in homogenen Datensituationen zu irreführend schmalen Konfidenzintervallen kommen kann [598]. Um dies zu vermeiden, kann die von Knapp und Hartung [439] vorgeschlagene Ad-hoc-Varianzkorrektur angewendet werden. Nach Empfehlungen aus der

Literatur sollen bei Anwendung der Knapp-Hartung-Methode immer Sensitivitätsanalysen mithilfe des Modells mit festem Effekt bzw. der DerSimonian-Laird-Methode [165] erfolgen, um zu untersuchen, ob die Anwendung der Ad-hoc-Varianzkorrektur angezeigt ist [405,760]. Ist das Konfidenzintervall nach der Knapp-Hartung-Methode schmäler als das der DerSimonian-Laird-Methode, so sollte die Knapp-Hartung-Methode mit Ad-hoc-Varianzkorrektur angewendet werden.

Die Anwendung von Metaanalysen mit zufälligen Effekten stößt jedoch im Fall sehr weniger Studien (weniger als 5) an ihre Grenzen [406]. Da im Fall sehr weniger Studien die Heterogenität nicht verlässlich schätzbar ist [351], kann die Anwendung von Metaanalysen mit zufälligen Effekten zu sehr breiten Konfidenzintervallen führen, die möglicherweise keine Aussagen zur Evidenzlage mehr zulassen [48]. Gerade im Fall sehr weniger Studien sollte daher die Anwendung des Modells mit festem Effekt oder eine qualitative Zusammenfassung (siehe Abschnitt 3.1.4) in Erwägung gezogen werden [48]. Sprechen keine Gründe gegen die Anwendung eines Modells mit festem Effekt, so sollte vor allem im Fall von nur 2 Studien das Modell mit festem Effekt gewählt werden. Ist die Anwendung des Modells mit festem Effekt nicht vertretbar, so kann eine qualitative Evidenzsynthese erfolgen. Kontextabhängig kommen auch alternative Verfahren in Betracht, wie z. B. bayessche Verfahren [34,265,266,671] oder Methoden aus dem Bereich der generalisierten linearen Modelle [62,457,504,573,664]. Zur Auswahl eines geeigneten Verfahrens sowie für Sensitivitätsanalysen ist insbesondere bei Metaanalysen mit sehr wenigen Studien die Durchführung mehrerer alternativer Methoden notwendig [305]. Des Weiteren wird das Institut, wie im Folgenden beschrieben, nur dann stark heterogene Studienergebnisse metaanalytisch zusammenfassen, wenn plausible Gründe für die Heterogenität ersichtlich sind, die eine Zusammenfassung trotzdem rechtfertigen.

B) Heterogenität

Bevor eine Metaanalyse 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 zusammenfassenden Studien trotz Vergleichbarkeit häufig heterogene Effekte zeigen [355]. In dieser Situation ist es erforderlich, die Heterogenität der Studien bezüglich der Ergebnisse zu untersuchen [285]. Das Vorliegen von Heterogenität kann statistisch getestet werden, wobei diese Verfahren in der Regel eine sehr niedrige Power haben [404,443]. 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 [354]. Für dieses Maß existieren Untersuchungen, die eine grobe Einschätzung der Heterogenität zulassen (z. B. über die Kategorien wahrscheinlich unbedeutend [0 bis 40 %], mittelmäßig [30 bis 60 %], substanziell [50 bis 90 %] und erheblich [75 bis 100 %] [162]). Ist die Heterogenität der Studien zu groß, so ist eine statistische Zusammenfassung der Studienergebnisse unter Umständen nicht sinnvoll [162]. 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 0,05 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 Nutzensaussage führen. In dieser Situation kann ggf. aber auch ohne quantitative Zusammenfassung eine positive Nutzensaussage getroffen werden (siehe Abschnitt 3.1.4). In den übrigen Situationen führt das Institut keine Metaanalyse 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, die eines anderen Maßes jedoch nicht. Bei binären Daten sind relative Effektmaße häufig stabiler als absolute, da sie nicht so stark vom Basisrisiko abhängen [268]. 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 (siehe Abschnitt 7.7).

Bei einer großen Heterogenität der Studien ist es notwendig, mögliche Ursachen hierfür zu untersuchen. Unter Umständen lassen sich mittels Metaregressionen Faktoren finden, die die Heterogenität der Effektstärken erklären können [709,727]. In einer Metaregression 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 Metaregression auf randomisierten Studien basiert, kann aus ihr nur die Evidenz einer Beobachtungsassoziation abgeleitet werden, nicht jedoch ein kausaler Zusammenhang [709]. Besonders schwierig zu interpretieren sind Metaregressionen, 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 [299]. 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 [569,665,709] (siehe Abschnitt 9.2.3).

Zur Darstellung der Heterogenität im Rahmen einer Metaanalyse mit zufälligen Effekten verwendet das Institut Prädiktionsintervalle [304,351,589,733]. Im Gegensatz zu einem Konfidenzintervall, das die Präzision eines geschätzten Effekts quantifiziert, überdeckt ein 95 %-Prädiktionsintervall den wahren Effekt einer einzelnen (neuen) Studie mit der 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. [304], das Prädiktionsintervall deutlich unterscheidbar von einem Konfidenzintervall in Form eines Rechtecks in einen Forest Plot einzufügen. Die Anwendung von Metaanalysen mit zufälligen Effekten und zugehörigen Prädiktionsintervallen im Fall sehr weniger Studien (weniger als 5) wird in der Literatur kritisch gesehen, da eine

mögliche Heterogenität nur sehr unpräzise geschätzt werden kann [351,406]. Das Institut stellt Prädiktionsintervalle in Forest Plots von Metaanalysen mit zufälligen Effekten in der Regel dar, wenn mindestens 5 Studien vorhanden sind und wenn die grafische Darstellung der Heterogenität wichtig ist, z. B. zur Beurteilung, ob und wie stark die Effekte gleichgerichtet sind (siehe Abschnitt 3.1.4).

Wird aufgrund von zu großer Heterogenität gar kein gepoolter Effekt präsentiert, so werden Prädiktionsintervalle auch verwendet um zu prüfen, ob die beobachteten Effekte der vorhandenen Studien mäßig oder deutlich gleichgerichtet sind (siehe Abschnitt 3.1.4). In diesem Fall werden Prädiktionsintervalle regelhaft verwendet, wenn mindestens 4 Studien vorliegen.

C) Geringe Zahl von Ereignissen

Ein häufiges Problem in Metaanalysen bei binären Daten ist das Vorhandensein von sogenannten Nullzellen, also die Beobachtung keines einzigen Ereignisses in einer Interventionsgruppe einer Studie. Das Institut folgt dem üblichen Vorgehen, beim Auftreten von Nullzellen den Korrekturwert von 0,5 zu jeder Zelloberhäufigkeit der entsprechenden Vierfeldertafel zu addieren [162]. 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 Peto-Odds-Ratio-Methode verwendet werden, die keinen Korrekturterm beim Vorliegen von Nullzellen erfordert [75,162]. Die Anwendung dieser Methode ist allerdings nur adäquat, wenn die zu schätzenden Effekte nicht zu groß sind und kein unbalanciertes Design vorliegt [82,83].

Kommen sogar Studien vor, in denen in beiden Studienarmen kein Ereignis beobachtet wird (sogenannte Doppelnulldaten), so werden diese Studien in der Praxis häufig aus der metaanalytischen Berechnung ausgeschlossen. Dieses Verfahren sollte vermieden werden, wenn zu viele Doppelnulldaten auftreten. Es gibt mehrere Methoden, um den Ausschluss von Doppelnulldaten 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. Weitere mögliche Verfahren sind gegeben durch logistische Regressionsmodelle mit zufälligen Effekten [664,721], Betabinomialmodelle [457,504], exakte Methoden [712] oder die Anwendung der Arcus-Sinus-Differenz [605]. Das Institut wird in Abhängigkeit der jeweiligen Datensituation ein geeignetes Verfahren auswählen und ggf. mithilfe von Sensitivitätsanalysen die Robustheit der Ergebnisse untersuchen.

D) Metaanalysen von Studien zur diagnostischen Güte

Auch die Ergebnisse von Studien zur diagnostischen Güte können mithilfe metaanalytischer Techniken statistisch zusammengefasst werden [177,403,491]. 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 Metaanalysen von Studien zur diagnostischen Güte einen in gleicher Weise eingeschränkten Stellenwert haben.

Für eine Metaanalyse von Studien zur diagnostischen Güte gelten die gleichen grundlegenden Prinzipien wie für Metaanalysen von Therapiestudien [177,582]. Dies beinhaltet insbesondere die Notwendigkeit einer systematischen Literaturübersicht, die Bewertung der methodischen Qualität der Primärstudien, die Durchführung von Sensitivitätsanalysen und die Untersuchung des möglichen Einflusses von Publikationsbias.

Bei Metaanalysen von Studien zur diagnostischen Güte 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 [177]. Eine solche metaanalytische Zusammenfassung von Studien zur diagnostischen Güte kann durch getrennte Modelle für Sensitivität und Spezifität erfolgen, wenn die Datenlage zur Anwendung bivariater Modelle unzureichend ist [702]. Bei Interesse an einer summarischen Receiver-Operating-Characteristic(ROC)-Kurve und / oder einem 2-dimensionalen Schätzer für Sensitivität und Spezifität haben jedoch neuere bivariate Metaanalysen mit zufälligen Effekten Vorteile [321,583]. Diese Verfahren ermöglichen auch die Berücksichtigung erklärender Variablen [319]. Auch für komplexe Datensituationen mit mehreren diagnostischen Schwellenwerten auf Studienebene wurden metaanalytische Verfahren entwickelt [367,423,683]. Die grafische Darstellung der Ergebnisse erfolgt entweder über die separate Darstellung der Sensitivitäten und Spezifitäten in Form modifizierter Forest Plots oder über eine 2-dimensionale Abbildung der Schätzer für Sensitivität und Spezifität. Analog zu den Konfidenz- und Prädiktionsintervallen in Metaanalysen von Therapiestudien können bei bivariaten Metaanalysen von diagnostischen Studien Konfidenz- und Prädiktionsregionen im ROC-Raum dargestellt werden.

E) Kumulative Metaanalysen

Es wird seit einiger Zeit verstärkt diskutiert, ob man bei wiederholten Aktualisierungen systematischer Übersichten die darin enthaltenen Metaanalysen als kumulative Metaanalysen mit Korrektur für multiples Testen berechnen und darstellen sollte [66,85,86,545,710,751]. Das Institut verwendet standardmäßig die übliche Form von Metaanalysen und greift in der Regel nicht auf Methoden für kumulative Metaanalysen 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 Metaanalysen mit Korrektur für multiples Testen in Erwägung ziehen.

9.3.8 Indirekte Vergleiche

Unter Methoden für indirekte Vergleiche werden sowohl Verfahren für einen einfachen indirekten Vergleich von 2 Interventionen verstanden als auch Verfahren, in denen direkte und indirekte Evidenz kombiniert wird. Letztere werden Mixed-Treatment-Comparison(MTC)-Metaanalyse [483-485], Multiple-Treatments Meta-Analysis (MTM) [116] oder Netzwerk-Metaanalyse [488,604,617] genannt. Diese Methoden stellen eine wichtige Weiterentwicklung

der üblichen metaanalytischen Verfahren dar [616]. Allerdings gibt es noch zahlreiche ungelöste methodische Probleme, sodass gegenwärtig von einer routinemäßigen Anwendung dieser Methoden im Rahmen der Nutzenbewertung abzuraten ist [32,271,618,673,694]. Aus diesem Grund werden zur Nutzenbewertung von Interventionen primär direkt vergleichende Studien (placebokontrollierte Studien sowie Head-to-Head-Vergleiche) verwendet, d. h., Aussagen für die Nutzenbewertung werden vorzugsweise aus Ergebnissen direkt vergleichender Studien abgeleitet. Um auf Methoden für indirekte Vergleiche zurückzugreifen, ist eine adäquate Begründung erforderlich. Darüber hinaus ist eine wesentliche Voraussetzung für die Berücksichtigung eines indirekten Vergleichs, dass dieser auf die interessierende Fragestellung insgesamt ausgerichtet ist und nicht nur selektiv auf z. B. einzelne Endpunkte.

In bestimmten Situationen wie z. B. bei Bewertungen des Nutzens von Arzneimitteln mit neuen Wirkstoffen [172] sowie bei Kosten-Nutzen-Bewertungen (siehe unten) kann es jedoch erforderlich sein, indirekte Vergleiche einzubeziehen und daraus Aussagen für die Nutzenbewertung unter Berücksichtigung einer geringeren Ergebnissicherheit abzuleiten.

Bei der Kosten-Nutzen-Bewertung von Interventionen sind i. d. R. gemeinsame quantitative Vergleiche multipler (d. h. von mehr als 2) Interventionen notwendig. Die Einschränkung auf direkte Head-to-Head-Vergleiche würde bedeuten, dass sich die Kosten-Nutzen-Bewertung auf einen einzigen paarweisen Vergleich beschränken würde oder sogar ganz unmöglich wäre. Um eine Kosten-Nutzen-Bewertung multipler Interventionen zu ermöglichen, kann es regelhaft erforderlich sein, unter Inkaufnahme einer – im Vergleich zum Ansatz der reinen Nutzenbewertung – geringeren Ergebnissicherheit auch indirekte Vergleiche zur Bewertung von Kosten-Nutzen-Verhältnissen heranzuziehen (siehe Kapitel 4).

Allerdings müssen adäquate Methoden für indirekte Vergleiche verwendet werden. Die Anwendung nicht adjustierter indirekter Vergleiche (d. h. die Verwendung einzelner Arme aus verschiedenen Studien) wird abgelehnt [53]. Dies gilt auch für Methoden für indirekte Vergleiche, in denen über Modellierungen mit starken Annahmen über die unbekannten Effekte [121] oder mithilfe von Verfahren aus dem Bereich der Kausalmodelle für Beobachtungsstudien mit untestbaren Annahmen [662] versucht wird, Effektschätzungen trotz fehlender Brückenkompaktoren zu ermöglichen. Es werden ausschließlich adjustierte indirekte Vergleiche über adäquate Brückenkompaktoren akzeptiert. Hierzu zählen insbesondere das Verfahren von Bucher et al. [100] sowie die oben erwähnten Methoden der Netzwerk-Metaanalysen.

Neben den Annahmen einer ausreichenden Ähnlichkeit und Homogenität der paarweisen Metaanalysen, die auch hier erfüllt sein müssen, muss bei Netzwerk-Metaanalysen zusätzlich eine ausreichende Konsistenz der geschätzten Effekte aus direkter und indirekter Evidenz gegeben sein. Letzteres ist ein kritischer Punkt, da Netzwerk-Metaanalysen nur valide Ergebnisse liefern, wenn die Konsistenzannahme erfüllt ist. Zur Untersuchung der Konsistenzannahme gibt es mehrere Methoden [182,195,484]. Allerdings sind diese noch nicht ausreichend untersucht und es konnte sich hier noch kein methodischer Standard etablieren [690]. Zudem

lässt sich die Konsistenz nicht immer untersuchen, da ein Vergleich von direkter und indirekter Evidenz nicht möglich ist (z. B. bei der Methode nach Bucher et al. [100]). Daher ist insbesondere in diesen Fällen eine sehr sorgfältige Überprüfung der Ähnlichkeit und Homogenität notwendig [435]. Bestehen große Zweifel, dass eine oder mehrere der grundlegenden Annahmen in ausreichendem Maß erfüllt sind, so sollte von der Anwendung indirekter Vergleiche abgesehen werden [434]. In der Praxis ist eine vollständige Beschreibung des verwendeten Modells zusammen mit verbleibenden Unklarheiten notwendig [694]. Letztere sollen in Sensitivitätsanalysen sorgfältig untersucht werden. Die in der Literatur vorhandenen Leitlinien zur Durchführung und Bewertung indirekter Vergleiche sind zu beachten [5,358,373,411,412,434].

9.3.9 Subgruppenanalysen

Mit Subgruppenanalysen wird untersucht, ob sich die Ergebnisse 1 oder mehrerer Studien zwischen verschiedenen in diese Studien eingeschlossenen Subgruppen (beispielsweise Patientinnen und Patienten mit versus ohne Nierenfunktionsstörung) unterscheiden. Dieser Unterschied kann qualitativ (Effektumkehr oder Effekt in einer Subgruppe, nicht jedoch in einer anderen) oder quantitativ (unterschiedliche Effektstärke) sein.

Für den zielgerichteten Einsatz medizinischer Interventionen sind solche Subgruppenanalysen sinnvoll, da mit ihnen gegebenenfalls Patientengruppen definiert werden können, für die eine Intervention einen Nutzen hat, und solche, denen dieselbe Intervention eher schadet als nutzt. Diese Informationen können auch zur Einschränkung des Anwendungsbereichs einer Intervention führen, beispielsweise bei der Zulassung von Arzneimitteln. Auch wenn Subgruppenanalysen für die Optimierung der Behandlung sinnvoll sind, werden sie in der methodischen Literatur teilweise kritisch diskutiert [26,558]:

- Geringe Power: Oft erreichen Subgruppen nicht die Stichprobengröße, die für das (inferenzstatistische) Aufdecken moderater Unterschiede nötig ist, sodass auch beim Vorliegen tatsächlicher Effekte das Ergebnis innerhalb einer Subgruppe nicht statistisch signifikant sein muss [292]. Anders ist die Situation, wenn schon bei der Fallzahlkalkulation eine der Subgruppenanalyse angemessene Power berücksichtigt und ein entsprechend erhöhter Stichprobenumfang geplant wurde [87].
- Multiples Testen: Bei der Analyse zahlreicher Subgruppen besteht eine manchmal recht hohe Wahrscheinlichkeit, dass die Ergebnisse irgendeiner Subgruppe statistische Signifikanz erreichen, obwohl es sich in Wahrheit um ein zufälliges Ergebnis handelt.
- Vergleichbarkeit zwischen Behandlungsgruppen innerhalb der Subgruppen: Erfolgte die Randomisierung nicht nach dem Subgruppenmerkmal stratifiziert, so könnten sich die Behandlungsgruppen bei kleinen Fallzahlen innerhalb der Subgruppen in prognostischen Faktoren unterscheiden [151,693]. In diesem Fall ist die Vergleichbarkeit der Behandlungsgruppen innerhalb der Subgruppen gefährdet, sodass (nicht) vorhandene Unterschiede zwischen Subgruppen allein durch diese Imbalance bedingt sein können.

- Effektmodifikation durch mehr als 1 Subgruppenmerkmal (Wechselwirkung höherer Ordnung): Liegt für einen Endpunkt beispielsweise ein Unterschied sowohl zwischen 2 Altersgruppen als auch zwischen Männern und Frauen vor, so ist zur Interpretation eine Auswertung separat für jede Altersgruppe und getrennt für Männer und Frauen nötig (also eine Analyse zu 4 Subgruppen). Eine solche Auswertung liegt jedoch selten vor.

Darüber hinaus wird diskutiert, dass Subgruppenanalysen generell keinen Beweischarakter haben, insbesondere wenn sie nicht a priori geplant wurden. Wenn Subgruppenanalysen bezüglich mehr oder weniger arbiträrer subgruppenbildender Merkmale post hoc stattfinden, könnten ihre Ergebnisse nicht als methodisch korrekte Prüfung einer Hypothese betrachtet werden. Während im Allgemeinen post hoc durchgeführte Subgruppenanalysen auf Studienebene kritisch zu interpretieren sind (auch in Anbetracht der oben genannten methodischen Limitationen), 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. Solche Subgruppenanalysen sind dann im Sinne der systematischen Übersicht auch nicht als post hoc zu bezeichnen, sondern entsprechen einer in dieser Übersicht zu überprüfenden Hypothese. Insofern ist die Analyse der Heterogenität zwischen den einzelnen Studien und damit gegebenenfalls der Subgruppenanalysen eine wissenschaftliche Notwendigkeit. Subgruppenanalysen zu Merkmalen, die nicht vor der Randomisierung, sondern im Studienverlauf erhoben wurden (z. B. Patientinnen und Patienten mit vs. ohne Herzinfarkt unter der untersuchten Therapie), sind grundsätzlich nicht verlässlich.

Einerseits erfordern die oben aufgeführten Aspekte die Bewertung der Verlässlichkeit von Subgruppenanalysen; Sun et al. [693] identifizieren hierzu Kriterien. Andererseits können trotz dieser Einschränkungen Subgruppenanalysen für manche Fragestellungen die auf absehbare Zeit beste wissenschaftliche Evidenz für die Bewertung von Effekten in ebendiesen Subgruppen darstellen [262], da beispielsweise ethische Überlegungen gegen die Verifizierung beobachteter Effekte in weiteren Studien sprechen können.

Eine wichtige Indikation für Subgruppenanalysen stellen erwartete Effektunterschiede zwischen verschiedenen, klar voneinander abgrenzbaren Patientenpopulationen dar [449,597]. Gibt es a priori Kenntnis von einem möglichen wichtigen Effektmodifikator (beispielsweise Alter, Pathologie), so ist es sogar erforderlich, eine mögliche Heterogenität bezüglich des Effekts in den verschiedenen Patientengruppen zu untersuchen.

Subgruppenanalysen können auch sozialrechtlich impliziert sein: Nach § 139a Abs. 2 SGB V ist es beispielsweise notwendig, alters-, geschlechts- und lebenslagenspezifische Besonderheiten zu berücksichtigen. Darüber hinaus soll auch erarbeitet werden, für welche Patientengruppen ein neues Arzneimittel eine maßgebliche Verbesserung des Behandlungserfolgs erwarten lässt, mit dem Ziel, dass diese Patientinnen und Patienten das neue Arzneimittel erhalten sollen [170]. Eine entsprechende Zielsetzung findet sich auch in § 35a SGB V zur Bewertung des Nutzens von Arzneimitteln mit neuen Wirkstoffen [172]. Bei

dieser Bewertung sollen Patientengruppen abgegrenzt werden, für die ein therapeutisch bedeutsamer Zusatznutzen besteht.

Bei der Interpretation von Subgruppenanalysen ist zu beachten, dass ein statistisch signifikanter Effekt in einer Subgruppe, aber kein Effekt oder ein entgegengesetzter Effekt in einer anderen Subgruppe, für sich allein (inferenzstatistisch) nicht als das Vorliegen von zwischen den Subgruppen unterschiedlichen Effekten interpretiert werden kann. Stattdessen muss zunächst der statistische Nachweis unterschiedlicher Effekte zwischen verschiedenen Subgruppen anhand eines adäquaten Homogenitäts- beziehungsweise Interaktionstests vorgenommen werden. Damit überhaupt eine gewisse Wahrscheinlichkeit für einen solchen Nachweis vorliegt, werden Subgruppenanalysen regelhaft nur dann durchgeführt, falls jede Subgruppe mindestens 10 Personen umfasst und bei binären Daten und Überlebenszeiten mindestens 10 Ereignisse in einer der Subgruppen aufgetreten sind. Mit einem Ereignis ist hier gemeint, dass dieses im Verlauf einer Beobachtung auftritt und zu Baseline noch nicht eingetreten ist (z. B. Erlangen einer Virusfreiheit bei infizierten Personen).

Ist das Ergebnis eines Heterogenitäts- oder Interaktionstests zwischen wichtigen Subgruppen zum Niveau $\alpha = 0,05$ signifikant, liegt eine Effektmodifikation, das heißt zwischen den Subgruppen unterschiedliche Effekte, vor. In diesem Fall werden die Ergebnisse der Subgruppen nicht zu einem gemeinsamen Effektschätzer gepoolt. Bei mehr als 2 Subgruppen werden – wenn sinnvoll – die paarweisen statistischen Tests auf Interaktionen durchgeführt. Paare, die zum Niveau $\alpha = 0,05$ nicht statistisch signifikant sind (bei gleichzeitiger Signifikanz der restlichen Paare), werden zu einer Gruppe zusammengefasst. Die Ergebnisse der verbleibenden Gruppen werden dann getrennt berichtet und es werden getrennte Nutzensaussagen für diese Gruppen abgeleitet. Finden sich keine Paare, die zum Niveau $\alpha = 0,05$ nicht statistisch signifikant sind, werden keine Paare gebildet, sondern es wird eine separate Aussage für jede Subgruppe getroffen. In Situationen, die auf Basis der paarweisen Interaktionstests nicht eindeutig sind (z. B. wenn nur ein einziger paarweiser Interaktionstest statistisch signifikant ist), wird im Einzelfall entschieden, ob und, wenn ja, welche Subgruppenpaare zu einer Gruppe zusammengefasst werden.

Eine Ausnahme von der Notwendigkeit eines adäquaten Homogenitäts- beziehungsweise Interaktionstests liegt vor, wenn sich durch den Zulassungsstatus von Arzneimitteln eine sozialrechtliche Notwendigkeit für Subgruppenanalysen ergibt. Zum einen kann es eine Folge der Abwägungsentscheidung in Bezug auf die Wirksamkeit und Risiken durch die Zulassungsbehörden sein, dass das Arzneimittel nur für einen Teil der in den Zulassungsstudien untersuchten Patientenpopulationen zugelassen wird. Den Abwägungen können dabei auch post hoc durchgeführte Subgruppenanalysen zugrunde liegen. Zum anderen können auch Studien, die nach der Zulassung durchgeführt wurden, Patientengruppen einschließen, für die das Arzneimittel in Deutschland nicht zugelassen ist. Dies gilt umso mehr, je stärker sich die Zulassungen international voneinander unterscheiden. In solchen Fällen muss gegebenenfalls auf Subgruppenanalysen zurückgegriffen werden, die den Zulassungsstatus des Arzneimittels

abbilden, und zwar unabhängig davon, ob diese Analysen innerhalb der Studie a priori geplant waren oder nicht.

9.3.10 Umgang mit nicht oder nicht vollständig publizierten Daten

In der Praxis ergibt sich häufig das Problem, dass für die Qualitätsbewertung von Publikationen notwendige Daten oder Informationen ganz oder teilweise fehlen (siehe Abschnitt 8.1). Darüber hinaus ist es möglich, dass Studien zum Zeitpunkt der Bewertung einer Technologie durch das Institut (noch) nicht publiziert sind.

Das Ziel des Instituts ist es, die Bewertung auf einer möglichst vollständigen Datenbasis vorzunehmen. Im Fall fehlender relevanter Informationen versucht das Institut deshalb, diese fehlenden Informationen zu vervollständigen. Dies geschieht unter anderem über den Kontakt mit den Autorinnen und Autoren von Publikationen oder den Sponsorinnen und Sponsoren von Studien (siehe Abschnitte 3.2.1 und 8.1.3). Je nach Art des zu erarbeitenden Produkts kann es allerdings aufgrund zeitlicher Vorgaben nur eingeschränkt möglich sein, nicht publizierte Informationen zu erfragen.

Ein häufig auftretendes Problem besteht darin, dass wichtige Angaben zur Durchführung einer Metaanalyse (z. B. Varianzen der Effektschätzer) fehlen. In vielen Fällen lassen sich jedoch die fehlenden Daten aus den vorhandenen Angaben berechnen oder zumindest schätzen [181,368,561]. Das Institut wendet nach Möglichkeit solche Verfahren an.

Bei nur teilweise vorhandenen Informationen bzw. bei der Verwendung geschätzter Werte wird die Robustheit der Ergebnisse ggf. mithilfe von Sensitivitätsanalysen untersucht und diskutiert. Dies kann beispielsweise im Sinne von Best-Case- und Worst-Case-Szenarien durchgeführt werden. Ein Worst-Case-Szenario kann jedoch hier nur als Beleg der Robustheit eines gefundenen Effekts verwendet werden. Aus einem Worst-Case-Szenario, in dem ein vorher gefundener Effekt nicht bestätigt wird, kann nicht ohne Weiteres abgeleitet werden, dass ein solcher Effekt nicht nachgewiesen ist. Beim weitgehenden oder vollständigen Fehlen relevanter Informationen kann es vorkommen, dass eine Publikation nicht bewertet bzw. eine Studie nicht für die Analyse herangezogen werden kann. In diesen Fällen wird dargestellt, dass es zu einem bestimmten Thema weitere Daten gibt, die aber für eine Qualitätsbewertung bzw. für die Analyse nicht zur Verfügung stehen; in der studienübergreifenden Bewertung wird dies entsprechend berücksichtigt (z. B. durch eine Herabstufung der Aussagesicherheit, siehe Abschnitt 3.2.1).

9.3.11 Umgang mit unvollständigen Daten

Bei jeder Nachbeobachtung innerhalb einer Studie kann es unvermeidlich passieren, dass Studienteilnehmer für die Datenerhebung zu Endpunkten nicht mehr zur Verfügung stehen. Für diese sogenannten Follow-up-Verluste kommen verschiedenste Gründe in Betracht, z. B. Tod, Abbruch oder Wechsel der Intervention (arzt- oder patientenseitig), Widerruf der Einwilligungserklärung aufgrund von Nebenwirkungen oder ausbleibender Wirksamkeit, neu

auftretende Komorbidität oder der belastende Charakter der Nachuntersuchung selbst. Die Standards zum Berichten klinischer Studien sehen vor, dass die Gründe für Follow-up-Verluste – soweit bekannt – für jede Interventionsgruppe getrennt anzugeben sind (siehe Abschnitt 9.1.1). Von den Follow-up-Verlusten, also Studienteilnehmer, die mit unvollständigen Daten in die Analysen eingehen, sind die Studienteilnehmer abzugrenzen, die zwar in die Studie eingeschlossen wurden, aber in den Analysen vollständig unberücksichtigt bleiben (unabhängig davon, ob die Nachbeobachtung erfolgte, vollständig oder unvollständig ist). Für eine Intention-to-treat-Analyse sind möglichst alle randomisierten Studienteilnehmer auch in der Analyse zu berücksichtigen [246].

Sowohl Follow-up-Verluste als auch Studienteilnehmer, die in den Analysen vollständig unberücksichtigt bleiben, führen zu fehlenden Daten (partiell bzw. komplett). Fehlende Daten erhöhen nicht nur die statistische Unsicherheit der Effektschätzung, sondern führen auch zu einem Verzerrungspotenzial, wenn nicht anhand der Gründe für Follow-up-Verlust oder Ausschluss aus den Analysen klar ersichtlich ist, dass die Verluste (Drop-outs) zufällig erfolgten (Missing completely at random) [480]. Das Verzerrungspotenzial steigt mit steigendem Anteil fehlender Werte.

Neben Follow-up-Verlusten können fehlende Werte auch vereinzelt oder temporär – sogar zu Baseline – auftreten, selbst wenn ein Studienteilnehmer zum Untersuchungstermin erscheint (z. B. Verlust von Proben, technisches Versagen oder Protokollieren von Messungen). Fehlende Werte aus solchen Gründen haben jedoch in der Praxis eine untergeordnete Bedeutung.

In einem 2-stufigen Prozess werden die fehlenden Werte bewertet. Im 1. Schritt werden die Studienteilnehmer betrachtet, die vollständig aus den Analysen ausgeschlossen wurden. Ergebnisse fließen in der Regel nicht in die Nutzenbewertung ein, wenn diese auf weniger als 70 % der in die Auswertung einzuschließenden Studienteilnehmer basieren, das heißt, wenn der Anteil der Studienteilnehmer, die gar nicht in der Auswertung berücksichtigt werden, größer als 30 % ist. In der Literatur werden zum Teil bereits Auswertungen, in denen 20 % der Studienteilnehmer nicht berücksichtigt werden, als nicht mehr aussagekräftig betrachtet [640]. Die Ergebnisse werden ferner auch dann nicht in die Nutzenbewertung einbezogen, wenn der Unterschied der Anteile nicht berücksichtigter Studienteilnehmer zwischen den Gruppen größer als 15 Prozentpunkte ist, da dies einen nicht zufälligen Ausschluss andeutet. Diese Regeln sind je nach Kontext nicht starr, sondern als Orientierung anzusehen. Ausnahmen von diesen Regeln werden dann gemacht, wenn die Ausschlüsse nachweislich zufällig erfolgten.

Im 2. Schritt werden die Follow-up-Verluste betrachtet, also die Studienteilnehmer, die zwar in die Analysen eingehen, aber unvollständig beobachtet wurden. Für die Bewertung des Verzerrungspotenzials durch diese Verluste werden keine festen Grenzen angelegt. Die Stärke des Verzerrungspotenzials richtet sich kontextabhängig nach folgenden Faktoren: Anzahl und Zeitpunkt der Follow-up-Verluste sowie Gründe für die Follow-up-Verluste. Insbesondere Unterschiede zwischen den Gruppen bei diesen Faktoren erhöhen das Verzerrungspotenzial.

Falls der Umfang fehlender Werte zu hoch ist oder ungeeignete Ersetzungsstrategien angewendet wurden, können die Ergebnisse ggf. nicht berücksichtigt werden.

Die Ergebnisse werden als robust angesehen, wenn die möglichen Probleme durch Follow-up-Verluste durch adäquate Ersetzungsstrategien oder statistische Auswertungsmethoden ausgeglichen werden [43,253]. Hierzu ist es erforderlich, den Mechanismus, der zu den fehlenden Daten geführt hat, im Detail zu kennen und beschreiben zu können. Dies wird in der Praxis unmöglich sein, wenn der Anteil fehlender Daten zu hoch ist.

9.3.12 Umgang mit variablen Beobachtungszeiten

Wenn in klinischen Studien der Eintritt der Patientinnen und Patienten zeitversetzt ist (Rekrutierungszeitraum), der Nachbeobachtungszeitpunkt jedoch identisch, ergibt sich, dass die Beobachtungszeiten für die einzelnen Patientinnen und Patienten unterschiedlich sind. Um dies adäquat zu berücksichtigen, stehen die Methoden der Überlebenszeitanalyse zur Verfügung [438,465]. Häufig angewendete Standardmethoden sind die Kaplan-Meier-Kurve zur Schätzung der Überlebensfunktion, der Log-Rank-Test zum Vergleich mehrerer Überlebensfunktionen und das Cox-Modell zur Schätzung von (adjustierten) Hazard Ratios. Bei der Betrachtung von Endpunkten ohne Todesfolge (wie z. B. nicht tödlicher Herzinfarkt) muss beachtet werden, dass der Tod immer ein konkurrierendes Ereignis für die nicht tödlichen Ereignisse darstellt. In diesem Fall müssen spezielle Methoden der Überlebenszeitanalyse für konkurrierende Risiken angewendet werden [30,427]. Die Anwendung der gewöhnlichen Kaplan-Meier-Kurve führt hierbei zu einer Überschätzung des absoluten Risikos [459] und sollte daher nicht angewendet werden. Stattdessen sollte der Aalen-Johansen-Schätzer für die kumulative Inzidenzfunktion verwendet werden [636]. Die Anwendung des Cox-Modells für die ereignisspezifische Hazardfunktion stellt dagegen eine korrekte Auswertung dar. Für eine komplette Analyse aller Daten benötigt man aber auch noch das Cox-Modell für das konkurrierende Ereignis [636]. Eine weitere Möglichkeit stellt die Anwendung des Fine-Gray-Modells dar [250]; Erläuterungen hierzu finden sich in der Literatur [31,580].

Insbesondere bei der Untersuchung unerwünschter Ereignisse sowie patientenberichteter Endpunkte (PROs) in onkologischen Studien tritt häufig das zusätzliche Problem auf, dass diese nur während der Behandlung erhoben und bei Behandlungsabbruch oder Behandlungswechsel die entsprechenden Beobachtungen zensiert werden. Dadurch entstehen häufig unterschiedliche mittlere Beobachtungsdauern in den beiden zu vergleichenden Gruppen. Obwohl bereits in der Arzneimittelzulassung schon lange die Berücksichtigung der Beobachtungsdauer bei der Analyse unerwünschter Ereignisse durch Anwendung adäquater Verfahren für Überlebenszeiten gefordert wurde [548,549], dominieren in diesem Bereich nach wie vor einfache Methoden basierend auf relativen Häufigkeiten oder Inzidenzdichten [46,725].

In der Nutzenbewertung werden jedoch für alle patientenrelevanten Endpunkte Auswertungen benötigt, die Schlussfolgerungen auf Basis einer adäquaten Inferenzstatistik zulassen. Dies bedeutet, dass insbesondere bei Gruppenvergleichen mit unterschiedlichen mittleren Beobachtungsdauern adäquate Verfahren für Überlebenszeiten benötigt werden [45,46]. Ist die Ursache

für die unterschiedlichen mittleren Beobachtungsdauern eine unvollständige Datenerfassung (Zensierung bei Behandlungsabbruch oder -wechsel), so entsteht allerdings auch bei Anwendung von Verfahren für Überlebenszeiten das Problem informativer Zensierungen, woraus ggf. ein hohes Verzerrungspotenzial folgt. Um solche Probleme zu vermeiden, sollte eine vollständige Datenerfassung der unerwünschten Ereignisse erfolgen, und zwar auch nach einem Behandlungsabbruch oder -wechsel [46]. Die häufig angewendete Praxis, unerwünschte Ereignisse nur bis maximal 30 Tage nach Behandlungsabbruch oder -wechsel zu erfassen, ist für die Nutzenbewertung unzureichend, da Auswertungen auf Basis des Treatment Policy Estimand (d. h. einer Schätzung des Effekts für die gesamte Behandlungsstrategie unabhängig von einem Behandlungsabbruch oder -wechsel) und des Intention-to-treat-Prinzips benötigt werden [46,771]. Für PROs werden bei unterschiedlichen Beobachtungszeiten zwar teilweise Überlebenszeitanalysen durchgeführt (z. B. Responderanalysen für die Zeit bis zur Verschlechterung), es fehlen jedoch in diesen Auswertungen die medizinisch notwendigen Informationen zu PROs nach Behandlungsabbruch oder -wechsel.

Liegen bei den zu vergleichenden Gruppen unterschiedlich lange Beobachtungsdauern vor, ohne dass dies in der Datenanalyse adäquat berücksichtigt worden ist, so kann dies dazu führen, dass die entsprechenden Ergebnisse für die Nutzenbewertung nicht verwertbar sind.

9.3.13 Darstellung von Verzerrungsarten

Unter Verzerrung (Bias) versteht man eine systematische Abweichung der Schätzung eines Effekts aus Studiendaten vom wahren Effekt. Es gibt vielfältige mögliche Ursachen, die eine Verzerrung hervorrufen können [131]. Nur die wichtigsten Verzerrungsarten werden im Folgenden beschrieben; eine ausführliche Übersicht über verschiedene Verzerrungsarten in unterschiedlichsten Situationen gibt Feinstein [241].

Ein Selection Bias entsteht durch eine Verletzung der Zufallsprinzipien bei Stichprobenziehungen, d. h. bei der Zuteilung der Patientinnen und Patienten zu den Interventionsgruppen. Speziell bei Gruppenvergleichen kann ein Selection Bias zu systematischen Unterschieden zwischen den Gruppen führen. Sind dadurch wichtige Confounder in den Gruppen ungleich verteilt, so sind die Ergebnisse eines Vergleichs in aller Regel nicht mehr interpretierbar. Beim Vergleich von Gruppen ist die Randomisierung die beste Methode zur Vermeidung eines Selection Bias [352], da die entstehenden Gruppen sich sowohl bezüglich bekannter als auch unbekannter Confounder nicht systematisch unterscheiden. Die Gewährleistung einer Strukturgleichheit ist allerdings nur bei ausreichend großen Stichproben gegeben. In kleinen Studien kann es trotz Randomisierung zufällig zu bedeutsamen Gruppenunterschieden kommen. Beim Vergleich von Gruppen mit bestehender Strukturungleichheit kann der Effekt bekannter Confounder mithilfe multifaktorieller Methoden berücksichtigt werden, es bleibt aber das Problem eines systematischen Unterschieds zwischen den Gruppen aufgrund von nicht bekannten oder nicht ausreichend erhobenen Confoundern.

Neben der Vergleichbarkeit der Gruppen bezüglich potenzieller prognostischer Faktoren spielen die Behandlungsgleichheit und die Beobachtungsgleichheit aller Probandinnen und

Probanden eine entscheidende Rolle. Eine Verzerrung durch unterschiedliche Behandlungen (mit Ausnahme der zu untersuchenden Intervention) wird als Performance Bias bezeichnet. Eine Verletzung der Beobachtungsgleichheit kann zu einem Detection Bias führen. Die Verblindung ist ein wirksamer Schutz vor beiden Biasarten [425], die in der Epidemiologie als Information Bias zusammengefasst werden.

Protokollverletzungen und Studienaustritte können bei Nichtberücksichtigung in der Auswertung das Studienergebnis systematisch verzerren, was als Attrition Bias bezeichnet wird. Zur Verminderung von Attrition Bias kann in Studien, die eine Überlegenheit zeigen wollen, das Intention-to-treat-Prinzip eingesetzt werden, das besagt, dass alle randomisierten Probandinnen und Probanden in der Analyse berücksichtigt werden, und zwar in der durch die Randomisierung zugeordneten Gruppe, unabhängig von Protokollverletzungen [425,453].

Ein ähnliches Problem stellen fehlende Werte auch anderer Ursache dar. Fehlende Werte, die nicht durch einen Zufallsmechanismus zustande kommen, können ebenso eine Verzerrung im Ergebnis verursachen [480]. Die möglichen Ursachen und Auswirkungen fehlender Werte sind daher im Einzelfall zu diskutieren. Bei Bedarf sind statistische Verfahren erforderlich, die eine Verzerrung berücksichtigen bzw. ausgleichen. Ersetzungsverfahren (Imputationsverfahren) für fehlende Werte sind dabei nur eine Klasse von verschiedenen Verfahren, von denen keines als allgemein akzeptiert gilt. Beispielsweise empfiehlt die EMA, in Sensitivitätsanalysen unterschiedliche Verfahren zum Umgang mit fehlenden Werten gegenüberzustellen [230].

Bei der Untersuchung von Screeningprogrammen besteht die Gefahr, dass durch eine Vorverlagerung der Diagnose eine nur scheinbare Verlängerung der Überlebenszeit beobachtet wird, die jedoch in Wahrheit auf nicht vergleichbare Anfangszeitpunkte zurückzuführen ist (Lead Time Bias). Eine verlängerte Überlebenszeit kann ebenso vorgetäuscht werden, wenn eine Screeningmaßnahme bevorzugt milder oder langsamer verlaufende Frühformen einer Krankheit erkennen kann (Length Bias). Einen Schutz vor diesen Verzerrungsmechanismen bietet die Durchführung einer randomisierten Studie zur Beurteilung der Wirksamkeit einer Screeningmaßnahme [255].

Ein Reporting Bias wird verursacht durch die selektive Berichterstattung lediglich einer Teilmenge aller relevanten Daten und kann zur Überschätzung des Nutzens einer Intervention in systematischen Übersichten führen. Werden innerhalb einer Publikation in Abhängigkeit der Ergebnisse Teilmengen der Analysen oder Zielgrößen nicht bzw. weniger detailliert oder in einer von der Planung abweichenden Weise berichtet, dann liegt ein Bias durch ergebnisgesteuerte Berichterstattung (Selective Reporting Bias bzw. Outcome Reporting Bias) vor [130,204,289,290,560]. Publication Bias dagegen beschreibt die Tatsache, dass Studien, die einen statistisch signifikanten negativen oder keinen statistisch signifikanten Unterschied zwischen der Interventions- und der Kontrollgruppe finden, gar nicht oder ggf. später publiziert werden als Studien mit positiven und statistisch signifikanten Resultaten [74]. Bei der Zusammenfassung der publizierten Ergebnisse kann es deshalb zu einer systematischen

Verzerrung des gemeinsamen Effektschätzers kommen. Für die Entdeckung und Berücksichtigung von Publication Bias können grafische Methoden wie der Funnel Plot [213] und statistische Methoden wie die Metaregression eingesetzt werden. Diese können das Vorliegen eines Publication Bias weder sicher nachweisen noch ausschließen, was die Bedeutung der Suche auch nach unpublizierten Daten unterstreicht.

In Studien zur Ermittlung der Güte einer diagnostischen Strategie (Indextest) kann es zu verzerrten Ergebnissen kommen, wenn der Referenztest die Probandinnen und Probanden nicht korrekt in gesund und krank unterscheidet (Misclassification Bias). Wenn nur eine nicht zufällige Stichprobe von Probandinnen oder Probanden, die den Indextest erhalten haben, dem Referenztest unterzogen wird (Partial Verification Bias) oder wenn der verwendete Referenztest vom Ergebnis des Indextests abhängt (Differential Verification Bias), besteht die Gefahr, verzerrte Schätzungen der diagnostischen Güte zu erhalten. In Fällen, in denen der Indextest selbst Bestandteil des Referenztests ist, kann es zu Überschätzungen der diagnostischen Güte kommen (Incorporation Bias) [466].

Ein weiterer in der internationalen Literatur aufgeführter Bias ist der sogenannte Spectrum Bias. Dieser spielt eine Rolle in Studien, in denen die Stichprobe zur Validierung eines diagnostischen Tests aus bereits als erkrankt bekannten Personen und gesunden Freiwilligen als Kontrollgruppe zusammengesetzt wird [477]. Die Validierung eines Testes in solchen Studien führt häufig zu Schätzungen für die Sensitivität und Spezifität, die höher sind, als sie es in einer klinischen Situation wären, bei der Patientinnen und Patienten mit Verdacht auf eine Erkrankung untersucht werden [753]. Der Begriff Bias (im Sinne einer systematischen Beeinträchtigung der internen Validität) ist allerdings in diesem Zusammenhang nicht günstig, da die Ergebnisse solcher Studien bei sachgerechter Durchführung durchaus intern valide sind [753]. Trotzdem können Studien des oben beschriebenen Designs Merkmale (insbesondere bezüglich der Stichprobenzusammensetzung) aufweisen, aufgrund derer sie für klinische Fragestellungen im Sinne der externen Validität nicht aussagekräftig sind.

Wie auch bei Interventionsstudien ist bei diagnostischen Studien die vollständige Berücksichtigung aller Teilnehmerinnen und Teilnehmer der Studie (auch derjenigen mit nicht eindeutigen Testergebnissen) notwendig, um eine systematische Verzerrung des Ergebnisses zu vermeiden [466]. Während es zur Bedeutung des sowie zum Umgang mit Publication Bias im Zusammenhang mit Interventionsstudien bereits zahlreiche Untersuchungen gibt, ist dieses Problem für Studien zur diagnostischen Genauigkeit noch deutlich weniger erforscht [466].

Ein allgemeines Problem bei der Schätzung von Effekten ist eine Verzerrung durch Messfehler in den erhobenen Studiendaten [123,132]. In der Praxis sind Messfehler kaum zu vermeiden und es ist bekannt, dass auch nicht differenzielle Messfehler zu einer Verzerrung in der Effektschätzung führen können. Im Fall eines einfachen linearen Regressionsmodells mit klassischem Messfehler in der erklärenden Variablen kommt es zum Dilution Bias, d. h. zu einer Verzerrung des Schätzers in Richtung des Nulleffekts. In anderen Modellen und komplexeren Situationen sind jedoch Verzerrungen in alle Richtungen möglich. Je nach

Fragestellung sollte die Stärke möglicher Messfehler diskutiert werden und bei Bedarf sollten Methoden zur Adjustierung von Messfehlerverzerrungen angewendet werden.

9.3.14 Auswertung abhängiger Daten

Die einfachen Standardverfahren der medizinischen Statistik gehen von unabhängigen Beobachtungseinheiten innerhalb eines Behandlungsarms aus. In bestimmten Situationen (z. B. bei clusterrandomisierten Studien oder Untersuchungen an Augen oder Zähnen), bei denen mehrere Beobachtungseinheiten eines Clusters oder einer Patientin / eines Patienten innerhalb eines Behandlungsarms in die Analyse eingehen, ist diese Unabhängigkeit nicht gegeben. Die Anwendung der einfachen Standardverfahren für unabhängige Beobachtungseinheiten innerhalb eines Behandlungsarms führt dann zu einer Unterschätzung der Varianz und möglicherweise zu falschen Signifikanzaussagen, da die errechneten p-Werte zu klein sind [17]. Die Evidenz aus Studien, die trotz korrelierter Daten einfache Standardverfahren verwendet haben, kann unter Umständen dennoch berücksichtigt werden, falls sich der damit verbundene Fehler gut einordnen lässt. Ist dies nicht der Fall, so lassen sich die Ergebnisse aus Studien mit korrelierten Daten nicht adäquat interpretieren. Liegen abhängige Daten innerhalb eines Behandlungsarms vor, so muss die damit verbundene Korrelationsstruktur durch Anwendung geeigneter Methoden für abhängige Daten berücksichtigt werden. In der Praxis häufig angewendete Methoden sind hierbei die Generalized Estimating Equations [110] sowie Verfahren aus der Klasse der gemischten Modelle [95].

Auch zwischen den zu vergleichenden Behandlungsarmen können Abhängigkeiten vorhanden sein (z. B. durch Matching oder bei Cross-over-Studien), die in der Datenanalyse durch die Anwendung geeigneter statistischer Verfahren für verbundene Stichproben bzw. Cross-over-Studien berücksichtigt werden müssen [421,647].

9.4 Qualitative Methoden

Qualitative Forschungsmethoden werden eingesetzt, um subjektives Erleben, subjektive Erfahrungen und individuelles Handeln sowie die soziale Realität zu erforschen und zu verstehen [185,323,494,531]. Sie können unter anderem einen Zugang zu den Sichtweisen und Erfahrungen von Patientinnen und Patienten und deren Angehörigen mit einem bestimmten Krankheitsbild oder einer Intervention schaffen.

Instrumente der qualitativen Forschung sind z. B. Fokusgruppen etwa mit Teilnehmenden einer randomisierten kontrollierten Studie. Qualitative Daten können auch über Interviews, Beobachtungen und schriftliche Dokumente wie Tagebücher erhoben werden.

An die Erhebung schließt sich eine Analyse an, die zumeist das Ziel hat, übergreifende Themen und Konzepte in den erhobenen Daten zu identifizieren und zu analysieren. Qualitative Methoden können unter anderem als eigenständige Forschungsmethode, als Vorbereitung von oder in Ergänzung zu quantitativen Studien, im Rahmen einer Triangulation bzw. eines Mixed-Method-Ansatzes oder nach der Durchführung quantitativer Studien zur Erklärung von

Prozessen oder Ergebnissen eingesetzt werden. Qualitative Forschung wird als Mittel angesehen, um die Verbindung zwischen Evidenz und Praxis zu fördern [187].

9.4.1 Qualitative Forschung bei der Erstellung von Gesundheitsinformationen

Im Rahmen der Entwicklung von Gesundheitsinformationen werden Forschungsergebnisse aus qualitativen Primärstudien und aus Übersichten qualitativer Studien zur Ermittlung von (potenziellen) Informationsbedürfnissen und zur Eruierung von Erfahrungen mit einem bestimmten Krankheitsbild bzw. mit einer Intervention sowie zum Umgang mit einer Erkrankung genutzt. Insbesondere sollen folgende Fragen beantwortet werden:

- Welche Fragen bestehen zu dieser Erkrankung / zu diesem Thema?
- Wie erleben Menschen diese Erkrankung / diese Intervention?
- Wann suchen Menschen eine Ärztin oder einen Arzt auf und wie erleben sie die Symptome?
- Welche Erfahrungen, Probleme, Herausforderungen und Fragen bestehen zu den diagnostischen Verfahren und zur Diagnose?
- Welche Erfahrungen wurden mit der Behandlung bzw. der Behandlungsentscheidung gemacht?
- Welche Erfahrungen, Probleme, Herausforderungen und Fragen bestehen bei der Bewältigung des Alltags?
- Welche Sachinformationen werden benötigt?
- Wie können die Informationen unterstützen (z. B. durch die Aufbereitung und das Format)?

Hierfür wird eine fokussierte Informationsbeschaffung in bibliografischen Datenbanken durchgeführt (siehe Abschnitt 8.2.2).

Die Abschätzung der Studienqualität erfolgt anhand vorher definierter Kriterien. In den letzten Jahren wurden verschiedene Instrumente zur Qualitätsbewertung qualitativer Studien entwickelt [154]. Die Bewertung qualitativer Studien zielt darauf ab festzustellen, ob das Forschungsdesign, die Studienqualität und die Reliabilität für die untersuchte Fragestellung angemessen sind. Bezüglich der Kriterien für die Durchführung, die Bewertung und die Synthese qualitativer Studien gibt es im Vergleich zu anderen Forschungsrichtungen noch keinen allgemeinen Konsens [185,188,323,531]. Das Institut wird die methodischen Entwicklungen beobachten und orientiert sich derzeit am Instrument The Confidence in the Evidence from Reviews of Qualitative research (CERQual).

Die identifizierten qualitativen Studien werden anhand folgender Aspekte, angelehnt an die Checkliste des Critical Appraisal Skills Programme (CASP), auf ihre Qualität geprüft [150]:

- Sind die Forschungsfrage und / oder Ziele der Studie beschrieben?
- Ist die Samplingstrategie beschrieben?
- Ist das Sample beschrieben und für die Fragestellung geeignet?
- Sind die Methoden der Datenerhebung beschrieben und für das Thema geeignet?
- Sind die Methoden der Datenauswertung beschrieben?
- Waren mindestens 2 Wissenschaftlerinnen oder Wissenschaftler an der Auswertung der Daten beteiligt?
- Ist der Prozess der Datenanalyse transparent und nachvollziehbar beschrieben?
- Sind die Ergebnisse der Studie klar ausgewiesen?

Nach der Extraktion der eingeschlossenen Studien erfolgen eine übergreifende Analyse und eine deskriptive Zusammenfassung der Ergebnisse. Weiterhin werden potenzielle Informationsbedürfnisse aus den Ergebnissen abgeleitet.

9.4.2 Qualitative Studien bei der Erstellung anderer IQWiG-Produkte

Verschiedene Informationsquellen können die Erarbeitung systematischer Übersichten unterstützen [186,472,706]. Eine dieser Quellen können Forschungsergebnisse aus qualitativen Studien sein [323,472,533,706]. Qualitative Forschung kann unter anderem Informationen zur Akzeptanz, Eignung und Implementation von Interventionen in der Praxis liefern [29,185,470,530]. Die Ergebnisse qualitativer Forschung können bei der Interpretation einer systematischen Übersicht hilfreich sein [706] und im Rahmen von Primärstudien oder systematischen Übersichten zur Ermittlung patientenrelevanter Endpunkte eingesetzt werden [185,187,452,531,533].

Das Institut kann qualitative Forschungsergebnisse zur Ermittlung patientenrelevanter Endpunkte, im Rahmen der Darstellung von Hintergrundinformationen zu den Erfahrungen von Patientinnen und Patienten und der Bedeutung des zu prüfenden Behandlungsverfahrens für Patientinnen und Patienten sowie für die Diskussion und Interpretation der Ergebnisse einer systematischen Übersicht nutzen. In den HTA-Berichten (siehe Abschnitte 6.5.3 bis 6.5.6) bilden zudem Ergebnisse aus eigenen qualitativen Erhebungen und Analysen (Einzel- oder Fokusgruppeninterviews) sowie aus qualitativen Studien und Übersichten die Basis zur Bearbeitung der Domänen Ethik, Soziales und Organisatorisches.

Anhang A – Rationale der Methodik zur Feststellung des Ausmaßes des Zusatznutzens

In diesem Anhang 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 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 10 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 bis 3 macht der Verordnungsgeber deutlich, dass die den Kategorien zugeordneten Kriterien nicht abschließend zu verstehen sind. Es ist nicht davon auszugehen, dass der Verordnungsgeber bei 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 10: Feststellung des Ausmaßes des Zusatznutzens – Kriterien gemäß AM-NutzenV

Ausmaßkategorie	Erheblich nachhaltige und gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte große Verbesserung des therapielevanten 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 therapielevanten 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 therapielevanten Nutzens			Verringerung von nicht schwerwiegenden Symptomen	relevante Vermeidung von Nebenwirkungen
AM-NutzenV: Arzneimittel-Nutzenbewertungsverordnung					

In einem ersten Schritt ist es also sinnvoll, die Kriterienliste anzupassen und durch qualitativ und quantitativ gleichwertige Kriterien zu ergänzen. In Tabelle 11 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 gleichkommt. Die Verkürzung der Symptombdauer, 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 12):

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

		Zielgrößenkategorie			
		<i>Gesamtmortalität</i>	<i>Symptome (Morbidität)</i>	<i>Gesundheitsbezogene Lebensqualität</i>	<i>Nebenwirkungen</i>
Ausmaßkategorie	Erheblich nachhaltig 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)	<i>erhebliche Verbesserung der Lebensqualität</i>	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) <i>Bedeutsame Verringerung von nicht schwerwiegenden (bzw. schweren) Symptomen (bzw. Folgekomplikationen)</i>	<i>bedeutsame Verbesserung der Lebensqualität</i>	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	<i>jegliche Verlängerung der Überlebensdauer</i>	jegliche Verringerung schwerwiegender (bzw. schwerer) Symptome (bzw. Folgekomplikationen) Verringerung von nicht schwerwiegenden (bzw. schweren) Symptomen (bzw. Folgekomplikationen)	<i>relevante Verbesserung der Lebensqualität</i>	<i>jegliche (statistisch signifikante) Verringerung schwerwiegender (bzw. schwerer) Nebenwirkungen</i> relevante Vermeidung von (anderen, nicht schwerwiegenden bzw. schweren) Nebenwirkungen
* Ergänzungen gegenüber AM-NutzenV <i>kursiv</i> gesetzt AM-NutzenV: Arzneimittel-Nutzenbewertungsverordnung					

Tabelle 12: Feststellung des Ausmaßes des Zusatznutzens – hierarchisierte Kriterien gemäß AM-NutzenV mit Ergänzungen*

		Zielgrößenkategorie			
		Gesamtmortalität	Schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen	Gesundheitsbezogene Lebensqualität	Nicht schwerwiegende (bzw. nicht schwere) Symptome (bzw. Folgekomplikationen) und Nebenwirkungen
Ausmaßkategorie	Erheblich nachhaltig und gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte große Verbesserung des therapielevanten Nutzens	erhebliche Verlängerung der Überlebensdauer	langfristige Freiheit bzw. weitgehende Vermeidung	<i>erhebliche Verbesserung</i>	<i>nicht besetzt</i>
	Beträchtlich gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte deutliche Verbesserung des therapielevanten Nutzens	moderate Verlängerung der Überlebensdauer	Abschwächung bzw. relevante Vermeidung	<i>bedeutsame Verbesserung</i>	bedeutsame Vermeidung
	Gering gegenüber der zweckmäßigen Vergleichstherapie bisher nicht erreichte moderate und nicht nur geringfügige Verbesserung des therapielevanten Nutzens	<i>jegliche Verlängerung der Überlebensdauer</i>	<i>jegliche Verringerung</i>	<i>relevante Verbesserung</i>	relevante Vermeidung
* Ergänzungen gegenüber AM-NutzenV <i>kursiv</i> gesetzt AM-NutzenV: Arzneimittel-Nutzenbewertungsverordnung					

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 Metaanalysen [670]. Diese hohe Sensitivität für Rahmenbedingungen stellt die Übertragbarkeit von absoluten Effektmaßen aus Studien auf die Versorgung infrage. Daher ist es übliche Praxis, Effekte in klinischen Studien vorzugsweise als relatives Risiko, Odds Ratio oder Hazard (oder auch Incidence) Ratio auszudrücken [161].
- 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 größer als 80 % grundsätzlich kein erheblicher Zusatznutzen (für den entsprechenden Endpunkt) mehr darstellbar.

- Ein weiterer Nachteil der Verwendung von absoluten Risikoreduktionen als Effektmaß für die Operationalisierung zur Feststellung des Ausmaßes des Zusatznutzens ist zudem, dass ein genauer Zeitpunkt definiert werden muss, zu 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 eine 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 2 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 12 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 aus der AM-NutzenV nach einer Berücksichtigung der Krankheitsschwere Rechnung getragen. Demgegenüber 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 [390]. 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. [189] als Anforderung für einen „Durchbruch“ postuliert – als ein Effekt erheblichen Ausmaßes für die Zielgröße Gesamtmortalität verankert [390].

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 [390], werden aber auch am Ende dieses Anhangs noch einmal aufgegriffen. Sie führten dazu, dass für einen Schwellenwert von 0,85 die gleichzeitigen Anforderungen bezüglich Realisierbarkeit und Stringenz als erfüllt angesehen werden können.

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 auf die Gesamtmortalität – die Unterschreitung eines Schwellenwerts von 0,90. Voraussetzung für die Einstufung eines Effekts auf 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 für die Feststellung des Ausmaßes des Effekts auf Endpunkte zur gesundheitsbezogenen Lebensqualität (wie für alle patientenberichteten Endpunkte) 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 (das 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 auf 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 begründet, dass es sich um eine moderate und nicht nur geringfügige Verbesserung handeln muss. Das Verfahren impliziert somit, dass (auch statistisch signifikante) Effekte, die aber nur als geringfügig bewertet werden, zur Einstufung in die Kategorie kein Zusatznutzen führen.

In der folgenden Tabelle 13 sind die jeweiligen Schwellenwerte für alle Ausmaßkategorien und Zielgrößenkategorien abgebildet.

Tabelle 13: 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
Ausmaß-kategorie	erheblich	0,85	0,75 und Risiko $\geq 5\%$ ^b	nicht besetzt
	beträchtlich	0,95	0,90	0,80
	gering	1,00	1,00	0,90
a: Voraussetzung ist, wie für alle patientenberichteten Endpunkte, die Verwendung eines validierten bzw. etablierten Instruments sowie eines validierten bzw. etablierten Responsekriteriums.				
b: Das Risiko muss für mindestens 1 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

$$H_0: RR \geq RR_0 \text{ vs. } H_1: RR < RR_0$$

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 für die Testung der üblichen (nicht verschobenen) Hypothesen festgelegt wurde, muss die Fallzahl erhöht werden, und zwar entweder innerhalb der Studie oder durch die Kombination mehrerer Studien. Vom Regelfall des Vorliegens von 2 (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 2 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 2-seitigen 95 %-Konfidenzintervalls für das relative Risiko. Bei Vorgabe eines Signifikanzniveaus von 5 % (2-seitig) und einer Power von 90 % (sowohl für die übliche als auch für die verschobene Hypothesengrenze) sowie 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 in Anhang A der Nutzenbewertung zu Ticagrelor [390] aufgeführte Formel für den Zusammenhang zwischen dem tatsächlichen Effekt und dem Schwellenwert 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 [629] wird auf die Arbeit von Fleiss et al. [251] verwiesen. Ein Austausch mit Herrn Joachim 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 folgendermaßen ermittelt:

- 1) Das Signifikanzniveau für die oben genannte Hypothese beträgt 2,5 %, die Power beträgt 90 %. Der Parameter RR_1 durchläuft alle Werte zwischen 0,2 und 0,95 mit einer Schrittweite von 0,01. Das Risiko in der Kontrollgruppe p_C durchläuft alle Werte

zwischen 0,05 und 0,95 mit einer Schrittweite von 0,05. Für jedes dieser Tupel (RR_1, p_c) wird die benötigte Fallzahl n nach der Formel von Farrington und Manning [238] mit $RR_0 = 1$ berechnet und anschließend verdoppelt ($m := 2n$).

- 2) Für jedes Tripel (RR_1, p_c, m) durchläuft ein Schwellenwert SW absteigend alle Werte zwischen 1 und 0 mit einer Schrittweite von $-0,005$. Für jedes SW wird die Power für die oben genannte Hypothese mit $RR_0 = SW$ approximiert. Das Signifikanzniveau beträgt 2,5 %. Dazu werden per Zufallsgenerator 50 000 Vierfeldertafeln simuliert, die obere Konfidenzintervallgrenze für das relative Risiko mittels Normalverteilungsapproximation und Deltamethode zur Varianzschätzung berechnet und anschließend der Anteil der Simulationsdurchläufe bestimmt, für die die obere Konfidenzintervallgrenze kleiner als SW ist. Der Durchlauf von SW wird gestoppt, sobald eine approximierte Power kleiner als 90 % ist. Das dazugehörige Tripel (RR_1, p_c, SW) wird in eine Liste geschrieben.
- 3) Nach dem Durchlauf aller Parameter in Schritt 1 und 2 werden alle Tripel der Liste selektiert, für die der Schwellenwert SW um weniger als 0,01 von einem der Werte 0,75, 0,8, 0,85, 0,9 und 0,95 abweicht.

In Abbildung 17 sind die resultierenden (genaueren) tatsächlichen Effekte in Abhängigkeit des Risikos in der Kontrollgruppe für alle oben festgelegten Schwellenwerte aufgetragen (Punkte durch geglättete Kurven approximiert).

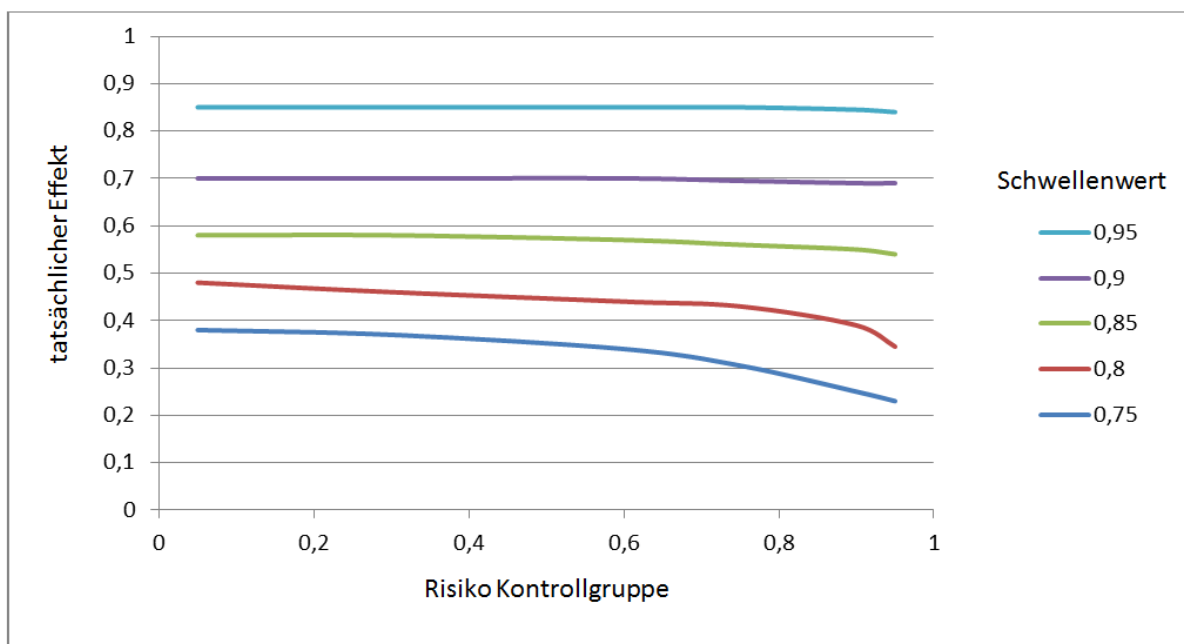


Abbildung 17: Tatsächliche Effekte in Abhängigkeit des Basisrisikos

In Tabelle 14 sind die Bereiche, in denen sich die tatsächlichen Effekte (in Abhängigkeit des Risikos der Kontrollgruppe) realisieren, pro Zielgrößen- und Ausmaßkategorie noch einmal eingetragen.

Tabelle 14: Tatsächliche Effekte für das relative Risiko

		Zielgrößenkategorie		
		Gesamt-mortalität	schwerwiegende (bzw. schwere) Symptome (bzw. Folgekomplika-tionen) und Nebenwirkungen und gesundheitsbezogene Lebensqualität	nicht schwerwiegende (bzw. nicht schwere) Symptome (bzw. Folgekomplika-tionen) und Nebenwirkungen
Ausmaß-kategorie	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 [390] 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 13 veranschlagte Rasterung der Schwellenwerte erscheint vernünftig und praktikabel.

Literaturverzeichnis

1. Memorandum for the evaluation of diagnostic measures. Journal of Clinical Chemistry and Clinical Biochemistry 1990; 28(12): 873-879.
2. Food and Drug Administration Amendments Act of 2007: public law 110–85 [online]. 27.09.2007 [Zugriff: 20.11.2019]. URL: <http://www.gpo.gov/fdsys/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf>.
3. SGB V Handbuch Sozialgesetzbuch V: Krankenversicherung. Altötting: KKF-Verlag; 2011.
4. Aagaard T, Lund H, Juhl C. Optimizing literature search in systematic reviews: are MEDLINE, EMBASE and CENTRAL enough for identifying effect studies within the area of musculoskeletal disorders? BMC Med Res Methodol 2016; 16: 161.
5. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7: a reviewer's checklist. Med Decis Making 2013; 33(5): 679-691.
6. Ades AE, Claxton K, Sculpher MJ. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. Health Econ 2006; 15(4): 373-381.
7. AGREE Collaboration. Appraisal of guidelines for research & evaluation: AGREE instrument. London: St. George's Hospital Medical School; 2001.
8. AGREE Next Steps Consortium. Appraisal of guidelines for research & evaluation II, AGREE II instrument [online]. 12.2017 [Zugriff: 05.11.2019]. URL: <https://www.agreetrust.org/wp-content/uploads/2017/12/AGREE-II-Users-Manual-and-23-item-Instrument-2009-Update-2017.pdf>.
9. Agresti A. Modelling ordered categorical data: recent advances and future challenges. Stat Med 1999; 18(18): 2191-2207.
10. Agresti A (Ed). Categorical data analysis. Hoboken: Wiley; 2002.
11. Agresti A. Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. Stat Methods Med Res 2003; 12(1): 3-21.
12. Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005; 331(7511): 267-270.
13. Alma H, De Jong C, Jelusic D, Wittmann M, Schuler M, Kollen B et al. Baseline health status and setting impacted minimal clinically important differences in COPD: an exploratory study. J Clin Epidemiol 2019; 116: 49-61.
14. Alma H, De Jong C, Tsiligianni I, Sanderman R, Kocks J, Van der Molen T. Clinically relevant differences in COPD health status: systematic review and triangulation. Eur Respir J 2018; 52(3): 1800412.

15. Altman DG. Systematic reviews of evaluations of prognostic variables. In: Egger M, Davey Smith G, Altman DG (Ed). Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group; 2001. S. 228-247.
16. Altman DG, Bland JM. Statistic notes: absence of evidence is not evidence of absence. BMJ 1995; 311(7003): 485.
17. Altman DG, Bland JM. Statistics notes: units of analysis. BMJ 1997; 314(7098): 1874.
18. Altman DG, Machin D, Bryant TN, Gardner MJ (Ed). Statistics with confidence: confidence intervals and statistical guidelines. London: BMJ Books; 2005.
19. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. PLoS Med 2012; 9(5): e1001216.
20. American Society of Clinical Oncology. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol 1996; 14(2): 671-679.
21. Amir E, Seruga B, Martinez-Lopez J, Kwong R, Pandiella A, Tannock IF et al. Oncogenic targets, magnitude of benefit, and market pricing of antineoplastic drugs. J Clin Oncol 2011; 29(18): 2543-2549.
22. Anthony L. AntConc: a freeware corpus analysis toolkit for concordancing and text analysis [online]. [Zugriff: 04.11.2019]. URL: <https://www.laurenceanthony.net/software/antconc/>.
23. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. JAMA 1992; 268(2): 240-248.
24. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten, Deutsche Gesellschaft für Sozialmedizin und Prävention, Deutsche Gesellschaft für Epidemiologie. Gute Praxis Sekundärdatenanalyse (GPS): Leitlinien und Empfehlungen [online]. 2014 [Zugriff: 05.11.2019]. URL: https://www.dgepi.de/assets/Leitlinien-und-Empfehlungen/GPS_revision2-final_august2014.pdf.
25. Assasi N, Schwartz L, Tarride JE, Campbell K, Goeree R. Methodological guidance documents for evaluation of ethical considerations in health technology assessment: a systematic review. Expert Review of Pharmacoeconomics & Outcomes Research 2014; 14(2): 203-220.
26. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355(9209): 1064-1069.
27. 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.

28. Atkins D, Eccles MP, Flottorp S, Guyatt GH, Henry D, Hill S et al. Systems for grading the quality of evidence and the strength of recommendations; I: critical appraisal of existing approaches. *BMC Health Serv Res* 2004; 4: 38.
29. Atkins S, Lewin S, Smith H, Engel M, Fretheim A, Volmink J. Conducting a meta-ethnography of qualitative literature: lessons learnt. *BMC Med Res Methodol* 2008; 8: 21.
30. Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. *Stat Med* 2017; 36(8): 1203-1209.
31. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat Med* 2017; 36(27): 4391-4400.
32. Bafeta A, Trinquart L, Seror R, Ravaud P. Reporting of results from network meta-analyses: methodological systematic review. *BMJ* 2014; 348: g1741.
33. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol* 2001; 54(10): 979-985.
34. Bai O, Chen M, Wang X. Bayesian estimation and testing in random effects meta-analysis of rare binary adverse events. *Stat Biopharm Res* 2016; 8(1): 49-59.
35. Baio G. Bayesian methods in health economics. Boca Raton: CRC Press; 2013. (Chapman & Hall/CRC Biostatistics Series; Band 53).
36. Baker SG. Surrogate endpoints: wishful thinking or reality? *J Natl Cancer Inst* 2006; 98(8): 502-503.
37. Balshem H, Stevens A, Ansari M, Norris S, Kansagara D, Shamliyan T et al. Finding grey literature evidence and assessing for outcome and analysis reporting biases when comparing medical interventions: AHRQ and the effective health care program [online]. 11.2013 [Zugriff: 21.11.2019]. URL: https://www.ncbi.nlm.nih.gov/books/NBK174882/pdf/Bookshelf_NBK174882.pdf.
38. Banta D. The development of health technology assessment. *Health Policy* 2003; 63(2): 121-132.
39. Barro RJ, Sala-i-Martin X. World real interest rates. In: Blanchard OJ, Fischer S (Ed). *NBER Macroeconomics Annual 1990*. Cambridge: MIT Press; 1990. S. 15-61.
40. Barron BA, Bukantz SC. The evaluation of new drugs: current Food and Drug Administration regulations and statistical aspects of clinical trials. *Arch Intern Med* 1967; 119(6): 547-556.
41. Beauchamp TL. Methods and principles in biomedical ethics. *J Med Ethics* 2003; 29(5): 269-274.
42. Beauchamp TL, Childress JF. Principles of biomedical ethics. New York: Oxford University Press; 2013.

43. Bell ML, Kenward MG, Fairclough DL, Horton NJ. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. *BMJ* 2013; 346: e8668.
44. Bender R. Interpretation von Effizienzmaßen der Vierfeldertafel für Diagnostik und Behandlung. *Medizinische Klinik* 2001; 96(2): 116-121.
45. Bender R, Beckmann L. Limitations of the incidence density ratio as approximation of the hazard ratio. *Trials* 2019; 20: 485.
46. Bender R, Beckmann L, Lange S. Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharm Stat* 2016; 15(4): 292-296.
47. Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL et al. Attention should be given to multiplicity issues in systematic reviews. *J Clin Epidemiol* 2008; 61(9): 857-865.
48. Bender R, Friede T, Koch A, Kuss O, Schlattmann P, Schwarzer G et al. Methods for evidence synthesis in the case of very few studies. *Res Syn Meth* 2018; 9(3): 382–392.
49. Bender R, Grouven U. Logistic regression models used in medical research are poorly presented. *BMJ* 1996; 313(7057): 628.
50. Bender R, Grouven U. Ordinal logistic regression in medical research. *J R Coll Physicians Lond* 1997; 31(5): 546-551.
51. Bender R, Grouven U. Möglichkeiten und Grenzen statistischer Regressionsmodelle zur Berechnung von Schwellenwerten für Mindestmengen. *Z Evid Fortbild Qual Gesundheitswes* 2006; 100(2): 93-98.
52. Bender R, Lange S. Adjusting for multiple testing: when and how? *J Clin Epidemiol* 2001; 54(4): 343-349.
53. Bender R, Schwenke C, Schmoor C, Hauschke D. Stellenwert von Ergebnissen aus indirekten Vergleichen: gemeinsame Stellungnahme von IQWiG, GMDS und IBS-DR [online]. 02.02.2012 [Zugriff: 21.11.2019]. URL: https://www.iqwig.de/download/12-03-07_Gemeinsame_Stellungnahme_IQWiG_GMDS_IBS-DR_zum_indirekten_Vergleich.pdf.
54. Bennett K, Duda S, Brouwers M, Szatmari P, Newton A, McLennan J et al. Towards high-quality, useful practice guidelines for child and youth mental health disorders: protocol for a systematic review and consensus exercise. *BMJ Open* 2018; 8(2): e018053.
55. Bent S, Padula A, Avins AL. Brief communication: better ways to question patients about adverse medical events; a randomized, controlled trial. *Ann Intern Med* 2006; 144(4): 257-261.
56. Beynon R, Leeftang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database Syst Rev* 2013; (9): MR000022.
57. Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. *Ann Epidemiol* 2006; 16(7): 540-544.

58. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. *Prev Sci* 2000; 1(1): 31-49.
59. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics* 2001; 69(3): 89-95.
60. Bland JM, Altman DG. One and two sided tests of significance. *BMJ* 1994; 309(6949): 248.
61. Bock J, Toutenburg H. Sample size determination in clinical research. In: Rao CR, Chakraborty R (Ed). *Statistical methods in biological and medical sciences*. Amsterdam: Elsevier; 1991. S. 515-538. (Handbook of Statistics; Band 8).
62. Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. *Biom J* 2015; 57(4): 633-648.
63. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *BMJ* 2006; 333(7563): 346-349.
64. Bonhoeffer J, Zimbrun B, Heininger U. Reporting of vaccine safety data in publications: systematic review. *Pharmacoepidemiology and Drug Safety* 2005; 14(2): 101-106.
65. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017; 7(2): e012545.
66. 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.
67. Bossuyt PM, Irwig LM, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006; 332(7549): 1089-1092.
68. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet* 2000; 356(9244): 1844-1847.
69. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; 351: h5527.
70. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003; 138(1): 40-44.
71. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003; 138(1): W1-W12.

72. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Ann Intern Med* 2008; 148(4): 295-309.
73. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Methods and processes of the CONSORT group: example of an extension for trials assessing nonpharmacologic treatments. *Ann Intern Med* 2008; 148(4): W60-W66.
74. Boutron I, Page MJ, Higgins JPT, Altman DG. Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al (Ed). *Cochrane handbook for systematic reviews of interventions*. Hoboken: Wiley-Blackwell; 2019. S. 177-204.
75. 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.
76. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ* 2004; 13(9): 873-884.
77. Breslow NE, Day NE. *Statistical methods in cancer research; volume I: the analysis of case-control studies*. Lyon: International Agency for Research on Cancer; 1980. (IARC Scientific Publications; Band 32). URL: <https://publications.iarc.fr/publications/media/download/3468/200a06ebe55fa3d17ed551c5fcd83f5cf0f3752.pdf>.
78. Breslow NE, Day NE. *Statistical methods in cancer research; volume II: the design and analysis of cohort studies*. Lyon: International Agency for Research on Cancer; 1987. (IARC Scientific Publications; Band 82). URL: <http://www.iarc.fr/en/publications/pdfs-online/stat/sp82/SP82.pdf>.
79. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA et al. Conjoint analysis applications in health: a checklist; a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value Health* 2011; 14(4): 403-413.
80. 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.
81. Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. *Med Decis Making* 2012; 32(5): 722-732.
82. Brockhaus AC, Bender R, Skipka G. The Peto odds ratio viewed as a new effect measure. *Stat Med* 2014; 33(28): 4861-4874.
83. Brockhaus AC, Grouven U, Bender R. Performance of the Peto odds ratio compared to the usual odds ratio estimator in the case of rare events. *Biom J* 2016; 58(6): 1428-1444.

84. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2001; 20(6): 825-840.
85. 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.
86. 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.
87. Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol* 2004; 57(3): 229-236.
88. Brouwer W, Rutten F, Koopmanschap M. Costing in economic evaluations. In: Drummond M, McGuire A (Ed). *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press; 2001. S. 68-93.
89. Brouwer WB, Culyer AJ, Van Exel NJ, Rutten FF. Welfarism vs. extra-welfarism. *J Health Econ* 2008; 27(2): 325-338.
90. Brouwer WBF, Koopmanschap MA, Rutten FFH. Productivity costs in cost-effectiveness analysis: numerator or denominator; a further discussion. *Health Econ* 1997; 6(5): 511-514.
91. Brouwer WBF, Koopmanschap MA, Rutten FFH. Productivity costs measurement through quality of life: a response to the recommendation of the Washington Panel. *Health Econ* 1997; 6(3): 253-259.
92. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G et al. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ* 2010; 182(18): E839-E842.
93. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G et al. Development of the AGREE II; part 1: performance, usefulness and areas for improvement. *CMAJ* 2010; 182(10): 1045-1052.
94. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G et al. Development of the AGREE II; part 2: assessment of validity of items and tools to support application. *CMAJ* 2010; 182(10): E472-E478.
95. Brown H, Prescott R. *Applied mixed models in medicine*. Chichester: Wiley; 2006.
96. Brozek JL, Akl EA, Alonso-Coello P, Lang D, Jaeschke R, Williams JW et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines; part 1 of 3: an overview of the GRADE approach and grading quality of evidence about interventions. *Allergy* 2009; 64(5): 669-677.

97. Brozek JL, Akl EA, Compalati E, Kreis J, Terracciano L, Fiocchi A et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines; part 3 of 3: the GRADE approach to developing recommendations. *Allergy* 2011; 66(5): 588-595.
98. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines; part 2 of 3: the GRADE approach to grading quality of evidence about diagnostic tests and strategies. *Allergy* 2009; 64(8): 1109-1116.
99. Brundage M, Blazeby J, Revicki D, Bass B, De Vet H, Duffy H et al. Patient-reported outcomes in randomized clinical trials: development of ISOQOL reporting standards. *Qual Life Res* 2013; 22(6): 1161-1175.
100. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* 1997; 50(6): 683-691.
101. Büchter RB, Fechtelpeter D, Knelangen M, Ehrlich M, Waltering A. Words or numbers? Communicating risk of adverse effects in written consumer health information: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making* 2014; 14: 76.
102. Bundesärztekammer, Kassenärztliche Bundesvereinigung, Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. Programm für Nationale VersorgungsLeitlinien: Methoden-Report [online]. 30.07.2010 [Zugriff: 20.11.2019]. URL: <http://www.leitlinien.de/mdb/downloads/nvl/methodik/mr-aufl-4-version-1.pdf>.
103. Bundesinstitut für Arzneimittel und Medizinprodukte. PharmNet.Bund-Arzneimittel-Informationssystem [online]. 31.08.2020 [Zugriff: 30.09.2020]. URL: <https://www.pharmnet-bund.de/dynamic/de/arzneimittel-informationssystem/index.html>.
104. Bundesministerium der Justiz. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung - BITV 2.0) [online]. 21.05.2019 [Zugriff: 25.11.2019]. URL: http://www.gesetze-im-internet.de/bitv_2_0/BJNR184300011.html.
105. Bundesministerium für Gesundheit. Verordnung über die Verfahrensgrundsätze der Bewertung von Untersuchungs- und Behandlungsmethoden in der vertragsärztlichen Versorgung und im Krankenhaus (Methodenbewertungsverfahrensverordnung - MBVerfV) [online]. 23.06.2020 [Zugriff: 01.09.2020]. URL: <http://www.gesetze-im-internet.de/mbverfv/MBVerfV.pdf>.
106. 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) [online]. 09.08.2019 [Zugriff: 25.11.2019]. URL: <http://www.gesetze-im-internet.de/bundesrecht/am-nutzenv/gesamt.pdf>.

107. Bundesministerium für Gesundheit, Bundesministerium für Bildung und Forschung. Verordnung über die Voraussetzungen für die Bewertung neuer Untersuchungs- und Behandlungsmethoden mit Medizinprodukten hoher Risikoklasse nach § 137h des Fünften Buches Sozialgesetzbuch (Medizinprodukteverfahrenbewertungsverordnung - MeMBV) [online]. 15.12.2015 [Zugriff: 21.11.2019]. URL: <http://www.gesetze-im-internet.de/membv/MeMBV.pdf>.
108. Bundessozialgericht. Urteil: Aktenzeichen B 6 A 1/08 R [online]. 06.05.2009 [Zugriff: 25.11.2019]. URL: <https://sozialgerichtsbarkeit.de/sgb/esgb/export.php?modul=esgb&id=121138&exportformat=PDF>.
109. Bundesverfassungsgericht. Leitsatz zum Beschluss des Ersten Senats: Aktenzeichen 1 BvR 347/98 [online]. 06.12.2005 [Zugriff: 20.11.2019]. URL: http://www.bverfg.de/entscheidungen/rs20051206_1bvr034798.html.
110. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998; 17(11): 1261-1291.
111. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat* 2006; 5(3): 173-186.
112. Burzykowski T, Molenberghs G, Buyse M (Ed). The evaluation of surrogate endpoints. New York: Springer; 2005.
113. Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T et al. Modelling in economic evaluation: an unavoidable fact of life. *Health Econ* 1997; 6(3): 217-227.
114. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000; 1(1): 49-67.
115. Cairns J. Discounting in economic evaluation. In: Drummond MF, McGuire A (Ed). *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press; 2001. S. 236-255.
116. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005; 331(7521): 897-900.
117. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 2013; 309(8): 814-822.
118. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012; 345: e5661.

119. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada; CADTH methods and guidelines [online]. 03.2017 [Zugriff: 05.11.2019]. URL: https://www.cadth.ca/sites/default/files/pdf/guidelines_for_the_economic_evaluation_of_health_technologies_canada_4th_ed.pdf.
120. Canestaro WJ, Hendrix N, Bansal A, Sullivan SD, Devine EB, Carlson JJ. Favorable and publicly funded studies are more likely to be published: a systematic review and meta-analysis. *J Clin Epidemiol* 2017; 92: 58-68.
121. Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. *Pharmacoeconomics* 2010; 28(10): 957-967.
122. Carrasco-Labra A, Devji T, Qasim A, Phillips M, Devasenapathy N, Zeraatkar D et al. Interpretation of patient-reported outcome measures: an inventory of over 3000 minimally important difference estimates and an assessment of their credibility. *Cochrane Database Syst Rev* 2018; (9 Suppl 1): 135-136.
123. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. London: Chapman & Hall; 2006.
124. Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. *BMJ* 2015; 350: g7773.
125. Carter SM, Rogers W, Heath I, Degeling C, Doust J, Barratt A. The challenge of overdiagnosis begins with its definition. *BMJ* 2015; 350: h869.
126. Çelik D, Çoban Ö, Kılıçoğlu Ö. Minimal clinically important difference of commonly used hip-, knee-, foot-, and ankle-specific questionnaires: a systematic review. *J Clin Epidemiol* 2019; 113: 44-57.
127. Centre for Evidence-based Medicine. Levels of evidence (March 2009) [online]. 03.2009 [Zugriff: 21.11.2019]. URL: <https://www.cebm.net/2009/06/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/>.
128. Centre for Reviews and Dissemination. PROSPERO: international prospective register of systematic reviews [online]. [Zugriff: 20.11.2019]. URL: <http://www.crd.york.ac.uk/PROSPERO/>.
129. Centre for Reviews and Dissemination (Ed). Systematic reviews: CRD's guidance for undertaking reviews in health care. York: CRD; 2009.
130. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; 291(20): 2457-2465.
131. Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. *J Clin Epidemiol* 2010; 63(11): 1205-1215.

132. Cheng CL, Van Ness JW. Statistical regression with measurement error. London: Arnold; 1999.
133. Chi C. Shall we search all trial registers? A comparative study of the sensitivity of five trial registers used by the Cochrane Skin Group [online]. In: 20th Cochrane Colloquium; 30.09.-03.10.2012; Auckland, Neuseeland. [Zugriff: 22.11.2019]. URL: <http://abstracts.cochrane.org/2012-auckland/shall-we-search-all-trial-registers-comparative-study-sensitivity-five-trial-registers>.
134. Chiou CF, Hay JW, Wallace JF, Bloom BS, Neumann PJ, Sullivan SD et al. Development and validation of a grading system for the quality of cost-effectiveness studies. Med Care 2003; 41(1): 32-44.
135. Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. BMJ 2013; 346: f457.
136. Clarivate Analytics. EndNote [online]. [Zugriff: 04.11.2019]. URL: <https://endnote.com/>.
137. Clark MD, Determann D, Petrou S, Moro D, De Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics 2014; 32(9): 883-902.
138. Claxton K, Paulden M, Gravelle H, Brouwer W, Culyer AJ. Discounting and decision making in the economic evaluation of health-care technologies. Health Econ 2011; 20(1): 2-15.
139. Cochrane Effective Practice and Organisation of Care Review Group. The data collection checklist [online]. 06.2002 [Zugriff: 20.11.2019]. URL: <http://epoc.cochrane.org/sites/epoc.cochrane.org/files/uploads/datacollectionchecklist.pdf>.
140. Collège des Économistes de la Santé. French guidelines for the economic evaluation of health care technologies [online]. 09.2004 [Zugriff: 05.11.2019]. URL: http://www.ces-asso.org/docs/France_Guidelines_HE_Evaluation.PDF.
141. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015; 350: g7594.
142. Commission of the European Communities. eEurope 2002: quality criteria for health related websites [online]. 29.11.2002 [Zugriff: 20.11.2019]. URL: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2002:0667:FIN:EN:PDF>.
143. Committee of Experts on Management of Safety and Quality in Health Care, Expert Group on Safe Medication Practices. Glossary of terms related to patient and medication safety [online]. 20.10.2005 [Zugriff: 20.11.2019]. URL: http://www.who.int/patientsafety/highlights/COE_patient_and_medication_safety_gl.pdf.

144. Concannon TW, Fuster M, Saunders T, Patel K, Wong JB, Leslie LK et al. A systematic review of stakeholder engagement in comparative effectiveness and patient-centered outcomes research. *J Gen Intern Med* 2014; 29(12): 1692-1701.
145. Cooper C, Booth A, Britten N, Garside R. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. *Syst Rev* 2017; 6: 234.
146. Corbin JM, Strauss AL. *Weiterleben lernen: Verlauf und Bewältigung chronischer Krankheiten*. Bern: Huber; 2010.
147. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ* 2010; 341: c3920.
148. Cornell JE, Laine C. The science and art of deduction: complex systematic overviews. *Ann Intern Med* 2008; 148(10): 786-788.
149. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008; 337: a1655.
150. Critical Appraisal Skills Programme. 10 questions to help you make sense of qualitative research [online]. 31.05.2013 [Zugriff: 20.11.2019]. URL: http://media.wix.com/ugd/dded87_29c5b002d99342f788c6ac670e49f274.pdf.
151. Cui L, Hung HMJ, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. *J Biopharm Stat* 2002; 12(3): 347-358.
152. D'Agostino RBS, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues; the encounters of academic consultants in statistics. *Stat Med* 2003; 22(2): 169-186.
153. 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.
154. Daly J, Willis K, Small R, Green J, Welch N, Kealy M et al. A hierarchy of evidence for assessing qualitative health research. *J Clin Epidemiol* 2007; 60(1): 43-49.
155. Damarell RA, Tieman JJ, Sladek RM. OvidSP Medline-to-PubMed search filter translation: a methodology for extending search filter range to include PubMed's unique content. *BMC Med Res Methodol* 2013; 13: 86.
156. Danner M, Hummel JM, Volz F, Van Manen JG, Wiegard B, Dintsios CM et al. Integrating patients' views into health technology assessment: Analytic Hierarchy Process (AHP) as a method to elicit patient preferences. *Int J Technol Assess Health Care* 2011; 27(4): 369-375.

157. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature; XIV: how to decide on the applicability of clinical trial results to your patient. *JAMA* 1998; 279(7): 545-549.
158. Dans LF, Silvestre MA, Dans AL. Trade-off between benefit and harm is crucial in health screening recommendations; part I: general principles. *J Clin Epidemiol* 2011; 64(3): 231-239.
159. De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R et al. Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. *Ann Intern Med* 2005; 143(2): 146-148.
160. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001; 323(7305): 157-162.
161. 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.
162. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al (Ed). *Cochrane handbook for systematic reviews of interventions*. Hoboken: Wiley-Blackwell; 2019. S. 241-284.
163. Derksen S, Keselman HJ. Backward, forward, and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 1992; 45(2): 265-282.
164. Derry S, Loke YK, Aronson JK. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. *BMC Med Res Methodol* 2001; 1: 7.
165. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; 7(3): 177-188.
166. Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *Am J Public Health* 2004; 94(3): 361-366.
167. Desroches B, Francis M. World real interest rates: a global savings and investment perspective [online]. 03.2007 [Zugriff: 05.11.2019]. (Bank of Canada Working Papers; Band 2007-16). URL: <http://www.bankofcanada.ca/wp-content/uploads/2010/03/wp07-16.pdf>.
168. Desu MM, Raghavarao D. Sample size methodology. Boston: Academic Press; 1990.
169. Deutsche Gesellschaft für Epidemiologie. Leitlinien und Empfehlungen zur Sicherung von Guter Epidemiologischer Praxis (GEP): Langversion [online]. 09.2018 [Zugriff: 06.11.2019]. URL: https://www.dgepi.de/assets/Leitlinien-und-Empfehlungen/Leitlinien_fuer_Gute_Epidemiologische_Praxis_GEP_vom_September_2018.pdf.

170. Deutscher Bundestag. Gesetzentwurf der Fraktionen SPD, CDU/CSU und BÜNDNIS 90/DIE GRÜNEN: Entwurf eines Gesetzes zur Modernisierung der gesetzlichen Krankenversicherung (GKV-Modernisierungsgesetz - GMG); BT Drucksache 15/1525 [online]. 08.09.2003 [Zugriff: 20.11.2019]. URL: <http://dipbt.bundestag.de/doc/btd/15/015/1501525.pdf>.
171. Deutscher Bundestag. Gesetz zur Modernisierung der gesetzlichen Krankenversicherung (GKV-Modernisierungsgesetz - GMG). Bundesgesetzblatt Teil 1 2003; (55): 2190-2258.
172. Deutscher Bundestag. Gesetz zur Neuordnung des Arzneimittelmarktes in der gesetzlichen Krankenversicherung (Arzneimittelmarktneuordnungsgesetz – AMNOG) vom 22. Dezember 2010. Bundesgesetzblatt Teil 1 2010; (67): 2262-2277.
173. 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>
https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/DER_StnAllo-Aufl2_Online.pdf.
174. Deutsches Institut für Normung. Klinische Prüfung von Medizinprodukten an Menschen: gute klinische Praxis (ISO 14155:2011 + Cor. 1:2011); deutsche Fassung EN ISO 14155:2011 + AC:2011. Berlin: Beuth; 2012.
175. Deutsches Netzwerk Evidenzbasierte Medizin. Die "Gute Praxis Gesundheitsinformation". Z Evid Fortbild Qual Gesundheitswes 2010; 104(1): 66-68.
176. Deutsches Netzwerk Evidenzbasierte Medizin. Gute Praxis Gesundheitsinformation: ein Positionspapier des Deutschen Netzwerks Evidenzbasierte Medizin e.V.; Version 2.0 [online]. 21.07.2016 [Zugriff: 21.11.2019]. URL: https://www.ebm-netzwerk.de/de/medien/pdf/gpgi_2_20160721.pdf.
177. 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.
178. Devji T, Carrasco-Labra A, Lytvyn L, Johnston B, Ebrahim S, Furukawa T et al. A new tool to measure credibility of studies determining minimally important difference estimates. Cochrane Database Syst Rev 2017; (9 Suppl 1): 58.
179. Devji T, Carrasco-Labra A, Qasim A, Phillips M, Johnston BC, Devasenapathy N et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020; 369: m1714.

180. Devji T, Guyatt GH, Lytvyn L, Brignardello-Petersen R, Foroutan F, Sadeghirad B et al. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. *BMJ Open* 2017; 7(5): e015587.
181. Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. *Stat Med* 2006; 25(13): 2299-2322.
182. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med* 2010; 29(7-8): 932-944.
183. DIPEX. Healthtalk.org [online]. [Zugriff: 20.11.2019]. URL: <http://www.healthtalk.org/>.
184. Dixon-Woods M. Writing wrongs? An analysis of published discourses about the use of patient information leaflets. *Social Science and Medicine* 2001; 52(9): 1417-1432.
185. Dixon-Woods M, Agarwal S, Young B, Jones D, Sutton A. Integrative approaches to qualitative and quantitative evidence. London: Health Development Agency; 2004. URL: https://www.webarchive.org.uk/wayback/archive/20140616160402mp_/http://nice.org.uk/nice/media/documents/integrative_approaches.pdf.
186. Dixon-Woods M, Fitzpatrick R. Qualitative research in systematic reviews: has established a place for itself. *BMJ* 2001; 323(7316): 765-766.
187. Dixon-Woods M, Fitzpatrick R, Roberts K. Including qualitative research in systematic reviews: opportunities and problems. *J Eval Clin Pract* 2001; 7(2): 125-133.
188. Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B et al. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. *Journal of Health Services Research and Policy* 2007; 12(1): 42-47.
189. Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepko 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.
190. Doganay Erdogan B, Leung YY, Pohl C, Tennant A, Conaghan PG. Minimal clinically important difference as applied in rheumatology: an OMERACT Rasch Working Group systematic review and critique. *J Rheumatol* 2016; 43(1): 194-202.
191. Dolan JG. Shared decision-making: transferring research into practice; the Analytic Hierarchy Process (AHP). *Patient Educ Couns* 2008; 73(3): 418-425.
192. 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.
193. Dolan P, Edlin R, Tsuchiya A. The relative societal value of health gains to different beneficiaries [online]. 04.06.2014 [Zugriff: 05.11.2019]. (HEDS Discussion Papers; Band 08/12). URL: http://eprints.whiterose.ac.uk/10902/1/HEDS_DP_08-12.pdf.

194. 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.
195. Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. Assessing key assumptions of network meta-analysis: a review of methods. *Res Syn Meth* 2013; 4(4): 291-323.
196. Donner A, Klar J. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
197. Draborg E, Gyrd-Hansen D, Poulsen PB, Horder M. International comparison of the definition and the practical application of health technology assessment. *Int J Technol Assess Health Care* 2005; 21(1): 89-95.
198. Drazen JM. COX-2 inhibitors: a lesson in unexpected problems. *N Engl J Med* 2005; 352(11): 1131-1132.
199. Droste S, Herrmann-Frank A, Scheibler F, Krones T. Ethical issues in autologous stem cell transplantation (ASCT) in advanced breast cancer: a systematic literature review. *BMC Med Ethics* 2011; 12: 6.
200. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. *BMJ* 1996; 313(7052): 275-283.
201. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press; 2015.
202. Duffy S, De Kock S, Misso K, Noake C, Ross J, Stirk L. Supplementary searches of PubMed to improve currency of MEDLINE and MEDLINE In-Process searches via Ovid. *J Med Libr Assoc* 2016; 104(4): 309-312.
203. Dunder Y, Dodd S, Dickson R, Walley T, Haycox A, Williamson PR. Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies. *Health Technol Assess* 2006; 10(5): iii-iv, ix-145.
204. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 2008; 3(8): e3081.
205. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias: an updated review. *PLoS One* 2013; 8(7): e66844.
206. Eady AM, Wilczynski NL, Haynes RB. PsycINFO search strategies identified methodologically sound therapy studies and review articles for use by clinicians and researchers. *J Clin Epidemiol* 2008; 61(1): 34-40.
207. Ebbesen M, Jensen TG, Andersen S, Pedersen FS. Ethical perspectives on RNA interference therapeutics. *Int J Med Sci* 2008; 5(3): 159-168.

208. Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, Ewigman B et al. Strength of Recommendation Taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician* 2004; 69(3): 548-556.
209. Ebrahim S, Vercammen K, Sivanand A, Guyatt GH, Carrasco-Labra A, Fernandes RM et al. Minimally important differences in patient or proxy-reported outcome studies relevant to children: a systematic review. *Pediatrics* 2017; 139(3): e20160833.
210. Eden J, Levit L, Berg A, Morton S. Finding what works in health care: standards for systematic reviews. Washington: National Academies Press; 2011.
211. Edwards JE, McQuay HJ, Moore RA, Collins SL. Reporting of adverse effects in clinical trials should be improved: lessons from acute postoperative pain. *Journal of Pain and Symptom Management* 1999; 18(6): 427-437.
212. Egger M, Davey Smith G, Altman DG (Ed). Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group; 2001.
213. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315(7109): 629-634.
214. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P et al. Shared decision making: a model for clinical practice. *J Gen Intern Med* 2012; 27(10): 1361-1367.
215. Elwyn GJ, O'Connor A, Stacey D, Volk R, Edwards AGK, Coulter A et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. *BMJ* 2006; 333(7565): 417-424.
216. Engelke K, Droste S. Bewertungen der rechtlichen Aspekte von Technologien. In: Perleth M, Busse R, Gerhardus A, Gibis B, Lühmann D, Zentner A (Ed). *Health Technology Assessment: Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung*. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. S. 280-296.
217. 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.
218. Epstein RM, Alper BS, Quill TE. Communicating evidence for participatory decision making. *JAMA* 2004; 291(19): 2359-2366.
219. Etzioni R, Gulati R, Mallinger L, Mandelblatt J. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Ann Intern Med* 2013; 158(11): 831-838.
220. Europäisches Parlament, Rat der Europäischen Union. Verordnung (EG) Nr. 141/2000 des Europäischen Parlaments und des Rates vom 16. Dezember 1999 über Arzneimittel für seltene Leiden. *Amtsblatt der Europäischen Gemeinschaften* 2000; 43(L18): 1-5.

221. Europäisches Parlament, Rat der Europäischen Union. Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte, zur Änderung der Richtlinie 2001/83/EG, der Verordnung (EG) Nr. 178/2002 und der Verordnung (EG) Nr. 1223/2009 und zur Aufhebung der Richtlinien 90/385/EWG und 93/42/EWG des Rates [online]. 05.04.2017 [Zugriff: 25.10.2019]. URL: <https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32017R0745>.
222. European Commission. European database on medical devices (EUDAMED) [online]. [Zugriff: 04.11.2019]. URL: https://ec.europa.eu/growth/sectors/medical-devices/new-regulations/eudamed_en.
223. European Medicines Agency. Posting of clinical trial summary results in European Clinical Trials Database (EudraCT) to become mandatory for sponsors as of 21 July 2014 [online]. 19.06.2014 [Zugriff: 20.11.2019]. URL: http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2014/06/news_detail_002127.jsp&mid=WC0b01ac058004d5c1.
224. European Medicines Agency. Webauftritt [online]. [Zugriff: 24.06.2014]. URL: <http://www.ema.europa.eu>.
225. European Medicines Agency. Points to consider on application with: 1. meta-analyses; 2. one pivotal study [online]. 31.05.2001 [Zugriff: 20.11.2019]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.
226. European Medicines Agency. Guideline on the choice of the non-inferiority margin [online]. 27.07.2005 [Zugriff: 20.11.2019]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf.
227. European Medicines Agency. Reflection paper on the regulatory guidance for the use of Health Related Quality of Life (HRQL) measures in the evaluation of medicinal products [online]. 27.07.2005 [Zugriff: 20.11.2019]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf.
228. European Medicines Agency. Guideline on clinical trials in small populations [online]. 27.07.2006 [Zugriff: 20.11.2019]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003615.pdf.
229. European Medicines Agency. Guideline on clinical investigation of medicinal products in the treatment of diabetes mellitus: draft [online]. 20.01.2010 [Zugriff: 20.11.2019]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/02/WC500073570.pdf.

230. European Medicines Agency. Guideline on missing data in confirmatory clinical trials [online]. 02.07.2010 [Zugriff: 20.11.2019]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf.
231. European Medicines Agency. EU Clinical Trials Register [online]. [Zugriff: 20.11.2019]. URL: <https://www.clinicaltrialsregister.eu/ctr-search/search>.
232. European Network for Health Technology Assessment. EUnetHTA POP database [online]. [Zugriff: 20.11.2019]. URL: <https://eunetha.dimdi.de/PopDB/>.
233. European Network for Health Technology Assessment. Joint Action on HTA 2012-2015: HTA core model; version 3.0 [online]. 25.01.2016 [Zugriff: 20.11.2019]. URL: <https://eunetha.eu/wp-content/uploads/2018/03/HTACoreModel3.0-1.pdf>.
234. European Network for Health Technology Assessment. Process of information retrieval for systematic reviews and health technology assessments on clinical effectiveness [online]. 12.2017 [Zugriff: 04.11.2019]. URL: https://www.eunetha.eu/wp-content/uploads/2018/01/Guideline_Information_Retrieval_V1-2_2017.pdf.
235. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. JAMA 1992; 268(17): 2420-2425.
236. Eyding D, Lelgemann M, Grouven U, Harter M, Kromp M, Kaiser T et al. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. BMJ 2010; 341: c4737.
237. Facey K, Boivin A, Gracia J, Hansen HP, Lo Scalzo A, Mossman J et al. Patients' perspectives in health technology assessment: a route to robust evidence and fair deliberation. Int J Technol Assess Health Care 2010; 26(3): 334-340.
238. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990; 9(12): 1447-1454.
239. Fayers P, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. Chichester: Wiley; 2007.
240. Feeny D. As good as it gets but good enough for which applications? Med Decis Making 2006; 26(4): 307-309.
241. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: Saunders; 1985.
242. Feinstein AR. Invidious comparisons and unmet clinical challenges. Am J Med 1992; 92(2): 117-120.
243. Feise RJ. Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol 2002; 2: 8.

244. Feldman-Stewart D, Brennenstuhl S, Brundage MD. A purpose-based evaluation of information for patients: an approach to measuring effectiveness. *Patient Education and Counseling* 2007; 65(3): 311-319.
245. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves-facts, fallacies and frequently asked questions. *Health Econ* 2004; 13(5): 405-415.
246. Fergusson D, Aaron SD, Guyatt G, Hebert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. *BMJ* 2002; 325(7365): 652-654.
247. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. *BMJ* 2012; 344: e686.
248. Ferreira-González I, Busse JW, Heels-Ansdell D, Montori VM, Alk EA, Byrant DM et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomized controlled trials. *BMJ* 2007; 334(7597): 786-792.
249. Field MJ, Lohr KN (Ed). *Clinical practice guidelines: directions for a new program*. Washington: National Academy Press; 1990.
250. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999; 94(446): 496-509.
251. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* 1980; 36(2): 343-346.
252. Fleming TR. Surrogate endpoints and FDA's accelerated approval process. *Health Aff (Millwood)* 2005; 24(1): 67-78.
253. Fleming TR. Addressing missing data in clinical trials. *Ann Intern Med* 2011; 154(2): 113-117.
254. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996; 125(7): 605-613.
255. Fletcher RH, Fletcher SW. *Klinische Epidemiologie: Grundlagen und Anwendung*. Bern: Huber; 2007.
256. Flynn TN. Using conjoint analysis and choice experiments to estimate QALY values: issues to consider. *Pharmacoeconomics* 2010; 28(9): 711-722.
257. Food and Drug Administration. Guidance for industry: developing medical imaging drug and biological products; part 2: clinical indications [online]. 06.2004 [Zugriff: 20.11.2019]. URL: <https://www.fda.gov/media/71226/download>.
258. Food and Drug Administration. Guidance for industry: patient-reported outcome measures; use in medical product development to support labeling claims [online]. 12.2009 [Zugriff: 20.11.2019]. URL: <https://www.fda.gov/media/77832/download>.

259. Forsetlund L, Kirkehei I, Harboe I, Odgaard-Jensen J. A comparison of two search methods for determining the scope of systematic reviews and health technology assessments. *Int J Technol Assess Health Care* 2012; 28(1): 59-64.
260. Franken M, Heintz E, Gerber-Grote A, Raftery J. Health economics as rhetoric: the limited impact of health economics on funding decisions in four European countries. *Value Health* 2016; 19(8): 951-956.
261. Fransen GA, Van Marrewijk CJ, Mujakovic S, Muris JW, Laheij RJ, Numans ME et al. Pragmatic trials in primary care: methodological challenges and solutions demonstrated by the DIAMOND-study. *BMC Med Res Methodol* 2007; 7: 16.
262. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? *BMJ* 2001; 322(7292): 989-991.
263. Freemantle N, Blonde L, Bolinder B, Gerber RA, Hobbs FD, Martinez L et al. Real-world trials to answer real-world questions. *Pharmacoeconomics* 2005; 23(8): 747-754.
264. Freemantle N, Calvert M. Weighing the pros and cons for composite outcomes in clinical trials. *J Clin Epidemiol* 2007; 60(7): 658-659.
265. Friede T, Rover C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Syn Meth* 2017; 8: 79-91.
266. Friede T, Rover C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biom J* 2017; 59(4): 658-671.
267. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991; 11(2): 88-94.
268. 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.
269. Gafni A, Birch S, Mehrez A. Economics, health and health economics: HYE's versus QALYs. *J Health Econ* 1993; 12(3): 325-339.
270. Garber AM, Weinstein MC, Torrance GW, Kamlet MS. Theoretical foundations of cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996. S. 25-53.
271. Gartlehner G, Moore CG. Direct versus indirect comparisons: a summary of the evidence. *Int J Technol Assess Health Care* 2008; 24(2): 170-177.
272. Gehanno JF, Rollin L, Le Jean T, Louvel A, Darmoni S, Shaw W. Precision and recall of search strategies for identifying studies on return-to-work in Medline. *J Occup Rehabil* 2009; 19(3): 223-230.

273. Gemeinsamer Bundesausschuss. Anlage I zum 2. Kapitel der Verfahrensordnung: Antrag zur Erprobung von Untersuchungs- und Behandlungsmethoden nach § 137e des Fünften Buches Sozialgesetzbuch (SGB V) [online]. [Zugriff: 20.11.2019]. URL: https://www.g-ba.de/downloads/17-98-3627/Anlage%20I_2-Kapitel-VerfO_Erprobungsantrag_Formular_2017-12-21_iK-2018-03-28.pdf.
274. Gemeinsamer Bundesausschuss. Mindestmengenregelungen gemäß § 136b Abs. 1 Satz 1 Nr. 2 SGB V [online]. URL: <https://www.g-ba.de/richtlinien/5/>.
275. Gemeinsamer Bundesausschuss. Richtlinie des Gemeinsamen Bundesausschusses zur Zusammenführung der Anforderungen an strukturierte Behandlungsprogramme nach § 137f Absatz 2 SGB V(DMP-Anforderungen-Richtlinie/DMP-A-RL) [online]. 18.04.2019 [Zugriff: 05.11.2019]. URL: https://www.g-ba.de/downloads/62-492-1938/DMP-A-RL_2019-04-18_iK-2019-10-01.pdf.
276. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses [online]. URL: <https://www.g-ba.de/informationen/richtlinien/42/>.
277. Gemeinsamer Bundesausschuss. Beauftragung des Instituts für Qualität imd Wirtschaftlichkeit im Gesundheitswesen [online]. 21.12.2004 [Zugriff: 08.09.2020]. URL: <https://www.g-ba.de/downloads/39-261-216/2004-12-21-Generalauftrag-IQWiG.pdf>.
278. Gemeinsamer Bundesausschuss. Beschluss des Gemeinsamen Bundesausschusses über die Anpassung der Beauftragung des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen zur Erstellung von Patienteninformationen [online]. 13.03.2008 [Zugriff: 20.11.2019]. URL: <https://www.g-ba.de/downloads/39-261-650/2008-03-13-IQWiG-Anpassung-Generalauftrag.pdf>.
279. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses: Anlage II zum 5. Kapitel; Format und Gliederung des Dossiers, einzureichende Unterlagen, Vorgaben für technische Standards [online]. 16.03.2018 [Zugriff: 25.10.2019]. URL: <https://www.g-ba.de/richtlinien/anlage/167/>.
280. Gerhardt U. Patientenkarrieren. Frankfurt am Main: Suhrkamp; 1986.
281. Gerhardus A, Stich AK. Die Bewertung sozio-kultureller Aspekte im HTA. In: Perleth M, Busse R, Gerhardus A, Gibis B, Lühmann D, Zentner A (Ed). Health Technology Assessment: Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. S. 312-320.
282. Gerste B, Drohan D, Günster C. Diagnosehäufigkeit und Inanspruchnahme von Gesundheitsleistungen. In: Klauber J, Günster C, Gerste B, Robra BP, Schmacke N (Ed). Versorgungs-Report 2015/2016. Stuttgart: Schattauer; 2016. S. 391-444.
283. Glanville JM, Duffy S, McCool R, Varley D. Searching ClinicalTrials.gov and the International Clinical Trials Registry Platform to inform systematic reviews: what are the optimal search approaches? J Med Libr Assoc 2014; 102(3): 177-183.

284. Glasziou PP, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007; 334(7589): 349-351.
285. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. *Stat Med* 2002; 21(11): 1503-1511.
286. Glasziou PP, Vandenbroucke JP, Chalmers I. Assessing the quality of research. *BMJ* 2004; 328(7430): 39-41.
287. Glenton C, Nilsen ES, Carlsen B. Lay perceptions of evidence-based information: a qualitative evaluation of a website for back pain sufferers. *BMC Health Serv Res* 2006; 6: 34.
288. Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Med Res Methodol* 2003; 3: 28.
289. Goldacre B, Drysdale H, Dale A, Milosevic I, Slade E, Hartley P et al. COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials* 2019; 20(1): 118.
290. Goldacre B, Drysdale H, Marston C, Mahtani KR, Dale A, Milosevic I et al. COMPare: qualitative analysis of researchers' responses to critical correspondence on a cohort of 58 misreported trials. *Trials* 2019; 20(1): 124.
291. Gonnermann A, Framke T, Grohennig A, Koch A. No solution yet for combining two independent studies in the presence of heterogeneity. *Stat Med* 2015; 34(16): 2476-2480.
292. Gonnermann A, Kottas M, Koch A. Biometrische Entscheidungsunterstützung in Zulassung und Nutzenbewertung am Beispiel der Implikationen von heterogenen Ergebnissen in Untergruppen der Studienpopulation. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2015; 58(3): 274-282.
293. Goossen K, Tenckhoff S, Probst P, Grummich K, Mihaljevic AL, Buchler MW et al. Optimal literature search for systematic reviews in surgery. *Langenbecks Archives of Surgery* 2018; 403(1): 119-129.
294. Gøtzsche PC, Liberati A, Torri V, Rossetti L. Beware of surrogate outcome measures. *Int J Technol Assess Health Care* 1996; 12(2): 238-246.
295. Graf von der Schulenburg JM, Greiner W, Jost F, Klusen N, Kubin M, Leidl R et al. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation: dritte und aktualisierte Fassung des Hannoveraner Konsens. *Gesundheitsökonomie & Qualitätsmanagement* 2007; 12(5): 285-290.
296. Graham RM, Mancher M, Miller-Wolman D, Greenfield S, Steinberg E. Clinical practice guidelines we can trust. Washington: National Academies Press; 2011. URL: http://www.awmf.org/fileadmin/user_upload/Leitlinien/International/IOM_CPG_lang_2011.pdf.
297. Gray JAM. How to get better value healthcare. Oxford: Offox Press; 2007.

298. Greenhalgh T, Hurwitz B. Narrative based medicine: why study narrative? *BMJ* 1999; 318(7175): 48-50.
299. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 1989; 18(1): 269-274.
300. Greiner W, Damm O. Die Berechnung von Kosten und Nutzen. In: Schöffski O, Graf von der Schulenburg JM (Ed). *Gesundheitsökonomische Evaluationen*. Berlin: Springer; 2012. S. 23-42.
301. Grimes DA, Schulz K. An overview of clinical research: the lay of the land. *Lancet* 2002; 359(9300): 57-61.
302. Grimes DA, Schulz KF. Surrogate end points in clinical research: hazardous to your health. *Obstetrics and Gynecology* 2005; 105(5 Pt 1): 1114-1118.
303. Gu Y, Norman R, Viney R. Estimating health state utility values from discrete choice experiments: a QALY space model approach. *Health Econ* 2014; 23(9): 1098-1114.
304. 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.
305. Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. *Stat Methods Med Res* 2017; 26(3): 1500-1518.
306. Guyatt G, Rennie D, Meade MO, Cook DJ (Ed). *Users' guides to the medical literature: a manual for evidence-based clinical practice*. New York: McGraw-Hill Education; 2015.
307. Guyatt GH. Evidence-based medicine. *ACP J Club* 1991; 114(2): A16.
308. Guyatt GH, Jaeschke R, Roberts R. N-of-1 randomized clinical trials in pharmacoepidemiology. In: Strom BL (Ed). *Pharmacoepidemiology*. Chichester: Wiley; 2005. S. 665-680.
309. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A et al. Going from evidence to recommendations. *BMJ* 2008; 336(7652): 1049-1051.
310. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008; 336(7651): 995-998.
311. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P et al. GRADE guidelines; 9: rating up the quality of evidence. *J Clin Epidemiol* 2011; 64(12): 1311-1316.
312. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P et al. GRADE guidelines; 4: rating the quality of evidence; study limitations (risk of bias). *J Clin Epidemiol* 2011; 64(4): 407-415.
313. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336(7650): 924-926.

314. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature; IX: a method for grading health care recommendations. *JAMA* 1995; 274(22): 1800-1804.
315. Guyatt GH, Sackett DL, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy: randomized trials in individual patients. *N Engl J Med* 1986; 314(14): 889-892.
316. Guyatt GH, Tugwell P, Feeny DH, Drummond MF, Haynes RB. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. *J Chronic Dis* 1986; 39(4): 295-304.
317. Hall AE, Chowdhury S, Hallowell N, Pashayan N, Dent T, Pharoah P et al. Implementing risk-stratified screening for common cancers: a review of potential ethical, legal and social issues. *J Public Health (Oxf)* 2014; 36(2): 285-291.
318. Halladay CW, Trikalinos TA, Schmid IT, Schmid CH, Dahabreh IJ. Using data sources beyond PubMed has a modest impact on the results of systematic reviews of therapeutic interventions. *J Clin Epidemiol* 2015; 68(9): 1076-1084.
319. Hamza TH, Van Houwelingen HC, Heijnenbroek-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.
320. Hao Q, Devji T, Zeraatkar D, Wang Y, Qasim A, Siemieniuk RAC et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. *BMJ Open* 2019; 9(2): e028777.
321. 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.
322. Harbour RT, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ* 2001; 323(7308): 334-336.
323. Harden A, Garcia J, Oliver S, Rees R, Shepherd J, Brunton G et al. Applying systematic review methods to studies of people's views: an example from public health research. *Journal of Epidemiology and Community Health* 2004; 58(9): 794-800.
324. Hardt JL, Metzendorf MI, Meerpohl JJ. Surgical trials and trial registers: a cross-sectional study of randomized controlled trials published in journals requiring trial registration in the author instructions. *Trials* 2013; 14: 407.
325. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
326. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15(4): 361-387.

327. Harris J. QALYfying the value of life. *J Med Ethics* 1987; 13(3): 117-123.
328. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM et al. Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med* 2001; 20(3 Suppl): 21-35.
329. Hart D. Der regulatorische Rahmen der Nutzenbewertung: vom Arzneimittelrecht zum HTA-Recht. *Medizinrecht* 2004; 22(9): 469-481.
330. Hart D. Leitlinien und Haftungsrecht: Inkorporation, Rezeption und Wissensbasis in Wissenschaft und Praxis. In: Hart D (Ed). *Klinische Leitlinien und Recht*. Baden-Baden: Nomos; 2005. S. 81-103. (Gesundheitsrecht und Gesundheitswissenschaften; Band 8).
331. Hartling L, Featherstone R, Nuspl M, Shave K, Dryden DM, Vandermeer B. The contribution of databases to the results of systematic reviews: a cross-sectional study. *BMC Med Res Methodol* 2016; 16: 127.
332. Hartling L, Featherstone R, Nuspl M, Shave K, Dryden DM, Vandermeer B. Grey literature in systematic reviews: a cross-sectional study of the contribution of non-English reports, unpublished studies and dissertations to the results of meta-analyses in child-relevant reviews. *BMC Med Res Methodol* 2017; 17: 64.
333. Hartung J. An alternative method for meta-analysis. *Biom J* 1999; 41(8): 901-916.
334. Hausner E. Problems encountered with ICTRP Search Portal (comment on: "Van Enst WA et al. Identification of additional trials in prospective trial registers for Cochrane systematic reviews. *PLoS One* 2012; 7(8): e42812") [online]. 23.07.2014 [Zugriff: 20.11.2019]. URL: <https://journals.plos.org/plosone/article/comment?id=info:doi/10.1371/annotation/5aac3b1b-56ed-43bf-a07e-cb38afd31478>.
335. Hausner E, Ebrahim S, Herrmann-Frank A, Janzen T, Kerekes MF, Pischedda M et al. Study selection by means of a web-based Trial Selection DataBase (webTSDB) [online]. In: 19th Cochrane Colloquium; 19.-22.10.2011; Madrid, Spanien. [Zugriff: 20.11.2019]. URL: <https://abstracts.cochrane.org/2011-madrid/study-selection-means-web-based-trial-selection-database-webtsdb>.
336. Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. *J Clin Epidemiol* 2015; 68(2): 191-199.
337. Hausner E, Metzendorf MI, Richter B, Lotz F, Waffenschmidt S. Study filters for non-randomized studies of interventions consistently lacked sensitivity upon external validation. *BMC Med Res Methodol* 2018; 18: 171.
338. Hausner E, Waffenschmidt S. Value of using different search approaches [online]. 28.05.2019 [Zugriff: 04.11.2019]. URL: <http://vortal.htai.org/?q=node/993>.

339. Hausner E, Waffenschmidt S, Kaiser T, Simon M. Routine development of objectively derived search strategies. *Syst Rev* 2012; 1: 19.
340. Haute Autorité de Santé. Choices in methods for economic evaluation [online]. 10.2012 [Zugriff: 20.11.2019]. URL: http://www.has-sante.fr/portail/upload/docs/application/pdf/2012-10/choices_in_methods_for_economic_evaluation.pdf.
341. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. *Ann Intern Med* 2006; 144(6): 427-437.
342. Hayden JA, Van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013; 158(4): 280-286.
343. Hayes MJ, Kaestner V, Mailankody S, Prasad V. Most medical practices are not parachutes: a citation analysis of practices felt by biomedical authors to be analogous to parachutes. *CMAJ Open* 2018; 6(1): E31-E38.
344. Haynes RB. Forming research questions. *J Clin Epidemiol* 2006; 59(9): 881-886.
345. Haynes RB, Cotoi C, Holland J, Walters L, Wilczynski N, Jedraszewski D et al. Second-order peer review of the medical literature for clinical practitioners. *JAMA* 2006; 295(15): 1801-1808.
346. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *Evid Based Med* 2002; 7(2): 36-38.
347. Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. *Am J Psychiatry* 2006; 163(2): 185-194.
348. Herxheimer A, McPherson A, Miller R, Shepperd S, Yaphe J, Ziebland S. Database of Patients' Experiences (DIPEx): a multi-media approach to sharing experiences and information. *Lancet* 2000; 355(9214): 1540-1543.
349. Hessel F, Kohlmann T, Krauth C, Nowy R, Seitz R, Siebert U et al. Gesundheitsökonomische Evaluation in der Rehabilitation; Teil 1: Prinzipien und Empfehlungen für die Leistungserfassung. In: Verband Deutscher Rentenversicherungsträger (Ed). Förderschwerpunkt "Rehabilitationswissenschaften": Empfehlungen der Arbeitsgruppen "Generische Methoden", "Routinedaten" und "Reha-Ökonomie". Frankfurt: VDR; 1999. S. 103-193. (DRV-Schriften; Band 16).
350. Higgins JP, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC et al. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Res Syn Meth* 2013; 4(1): 12-25.
351. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc A* 2009; 172(1): 137-159.

352. Higgins JPT, Savović J, Page MJ, Elbers RG, Sterne JAC. Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al (Ed). *Cochrane handbook for systematic reviews of interventions*. Hoboken: Wiley-Blackwell; 2019. S. 205-228.
353. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al (Ed). *Cochrane handbook for systematic reviews of interventions*. Hoboken: Wiley-Blackwell; 2019.
354. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; 21(11): 1539-1558.
355. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327(7414): 557-560.
356. Hill AB (Ed). *Controlled clinical trials*. Oxford: Blackwell; 1960.
357. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ* 2013; 346: e5793.
358. Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC et al. Conducting indirect-treatment comparison and network-meta-analysis studies: report of the ISPOR task force on indirect treatment comparisons good research practices; part 2. *Value Health* 2011; 14(4): 429-437.
359. Hoffmann-Eßer W, Siering U, Neugebauer EAM, Lampert U, Eikermann M. Systematic review of current guideline appraisals performed with the appraisal of guidelines for research & evaluation II instrument: a third of AGREE II users apply a cut-off for guideline quality. *J Clin Epidemiol* 2018; 95: 120-127.
360. Hofmann B. Toward a procedure for integrating moral issues in health technology assessment. *Int J Technol Assess Health Care* 2005; 21(3): 312-318.
361. Hofmann B, Droste S, Oortwijn W, Cleemput I, Sacchini D. Harmonization of ethics in health technology assessment: a revision of the Socratic approach. *Int J Technol Assess Health Care* 2014; 30(1): 3-9.
362. Hofmann B, Haustein D, Landeweerd L. Smart-glasses: exposing and elucidating the ethical issues. *Sci Eng Ethics* 18.07.2016 [Epub ahead of print].
363. Hofmann B, Lysdahl KB, Droste S. Evaluation of ethical aspects in health technology assessment: more methods than applications? *Expert Review of Pharmacoeconomics & Outcomes Research* 2015; 15(1): 5-7.
364. Holmes-Rovner M. International Patient Decision Aid Standards (IPDAS): beyond decision aids to usual design of patient education materials. *Health Expect* 2007; 10(2): 103-107.
365. Hopewell S, Clarke M, Askie L. Reporting of trials presented in conference abstracts needs to be improved. *J Clin Epidemiol* 2006; 59(7): 681-684.

366. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health* 1991; 81(12): 1630-1635.
367. Hoyer A, Hirt S, Kuss O. Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. *Res Syn Meth* 2018; 9(1): 62-72.
368. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med Res Methodol* 2005; 5: 13.
369. Hummel M, IJzerman M. The past and future of the AHP in health care decision making [online]. In: XI International Symposium on the Analytic Hierarchy Process (ISAHP); 15.-18.06.2011; Sorrent, Italien. [Zugriff: 05.11.2019]. URL: https://www.isahp.org/uploads/71_0111_hummel.pdf.
370. Hummel MJM, Steuten LMG, Groothuis-Oudshoorn KGM, IJzerman MJ. How the Analytic Hierarchy Process may fill missing gaps in early decision modeling. *ISPOR Connections* 2011; 17(3): 10-11.
371. Hung HM, O'Neill RT, Bauer P, Köhne K. The behavior of the P-value when the alternative hypothesis is true. *Biometrics* 1997; 53(1): 11-22.
372. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *Pharmacoeconomics* 2013; 31(5): 361-367.
373. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med* 2015; 162(11): 777-784.
374. ICH Expert Working Group. ICH harmonised tripartite guideline: the extent of population exposure to assess clinical safety for drugs intended for long-term treatment of non-life-threatening conditions; E1; current step 4 version [online]. 27.10.1994 [Zugriff: 18.03.2015]. URL: https://database.ich.org/sites/default/files/E1_Guideline.pdf.
375. Inan H. Measuring the success of your website: a customer-centric approach to website management. Frenchs Forest: Pearson Education Australia; 2002.
376. Institut für Arbeitsmarkt- und Berufsforschung. Webauftritt [online]. [Zugriff: 05.11.2019]. URL: <http://www.iab.de/de/befragungen/stellenangebot/aktuelle-ergebnisse.aspx>.
377. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anhang: Dokumentation der Stellungnahmen zur „Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung Version 1.0“ [online]. 30.09.2008 [Zugriff: 20.11.2019]. URL: https://www.iqwig.de/download/08-10-14_Dokumentation_der_Stellungnahmen_KNB_Version_1_0.pdf.

378. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Aussagekraft von Surrogatendpunkten in der Onkologie: Rapid Report; Auftrag A10-05; Version 1.1 [online]. 21.11.2011 [Zugriff: 20.11.2019]. (IQWiG-Berichte; Band 80). URL: https://www.iqwig.de/download/A10-05_Rapid_Report_Version_1-1_Surrogatendpunkte_in_der_Onkologie.pdf.

379. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Datenschutzbestimmungen für gesundheitsinformation.de [online]. 12.11.2019 [Zugriff: 25.11.2019]. URL: <http://www.gesundheitsinformation.de/hinweise-zur-nutzung.2010.de.html>.

380. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Dokumentation und Würdigung der Stellungnahmen zur „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ [online]. 28.11.2013 [Zugriff: 20.11.2019]. URL: https://www.iqwig.de/download/Dokumentation_und_Wuerdigung_der_Stellungnahmen_IQWiG_Methoden_4-1.pdf.

381. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Entwurf einer Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung: Version 2.0 [online]. 16.03.2009 [Zugriff: 20.11.2019]. URL: https://www.iqwig.de/download/09-03-18_Entwurf_Methoden_Kosten-Nutzen-Bewertung_Version_2_0.pdf.

382. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung: Version 1.1 [online]. 09.10.2008 [Zugriff: 20.11.2019]. URL: https://www.iqwig.de/download/08-10-14_Entwurf_Methoden_Kosten-Nutzen-Bewertung_Version_1_1.pdf.

383. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Technischer Anhang: Modellierung [online]. 09.10.2008 [Zugriff: 20.11.2019]. URL: https://www.iqwig.de/download/TA_KNB_Modellierung_v_1_0.pdf.

384. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Technischer Anhang: Unsicherheit [online]. 09.10.2008 [Zugriff: 20.11.2019]. URL: https://www.iqwig.de/download/TA_KNB_Unsicherheit_v_1_0.pdf.

385. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Vereinbarung über die vertrauliche Behandlung von Unterlagen [online]. 19.08.2005 [Zugriff: 20.11.2019]. URL: <https://www.iqwig.de/download/IQWiG-VFA-Mustervertrag.pdf>.

386. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Würdigung der Stellungnahmen zur „Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung Version 1.0“ [online]. 09.10.2008 [Zugriff: 20.11.2019]. URL: https://www.iqwig.de/download/08-10-14_Wuerdigung_der_Stellungnahmen_KNB_Version_1_0.pdf.

387. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Entwicklung und Anwendung von Modellen zur Berechnung von Schwellenwerten bei Mindestmengen für die Koronarchirurgie: Vorbericht; Auftrag B05/01B [online]. 24.04.2006 [Zugriff: 04.11.2019]. URL: http://www.iqwig.de/download/B05-01B_Vorbericht_Entwicklung_und_Anwendung_von_Modellen_zur_Berechnung_von_Schwellenwerten_bei_Mindestmengen_fuer_die_Koronarchirurgie.pdf.
388. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Zusammenhang zwischen Menge der erbrachten Leistungen und der Ergebnisqualität für die „Perkutane Transluminale Coronare Angieplastie (PTCA)“: Abschlussbericht; Auftrag Q05-01B [online]. 06.06.2006 [Zugriff: 04.11.2019]. (IQWiG-Berichte; Band 8). URL: https://www.iqwig.de/download/Q05-01B_Abschlussbericht_Zusammenhang_Menge_erbrachter_Leistung_und_Ergebnisqualitaet_bei_PTCA..pdf.
389. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Bupropion, Mirtazapin und Reboxetin bei der Behandlung von Depression: Abschlussbericht; Auftrag A05-20C [online]. 09.11.2009 [Zugriff: 20.11.2019]. (IQWiG-Berichte; Band 68). URL: https://www.iqwig.de/download/A05-20C_Abschlussbericht_Bupropion_Mirtazapin_und_Reboxetin_bei_Depressionen.pdf.
390. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Ticagrelor: Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung; Auftrag A11-02 [online]. 29.09.2011 [Zugriff: 20.11.2019]. (IQWiG-Berichte; Band 96). URL: https://www.iqwig.de/download/A11-02_Ticagrelor_Nutzenbewertung_35a_SGB_V_.pdf.
391. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Suchen in Studienregistern nach Studien zu neu zugelassenen Arzneimitteln: Arbeitspapier; Auftrag GA14-01 [online]. 03.02.2016 [Zugriff: 20.11.2019]. (IQWiG-Berichte; Band 361). URL: https://www.iqwig.de/download/GA14-01_Arbeitspapier_Suchen-in-Studienregistern-nach-Studien-zu-neu-zugelassenen-Arzneimitteln.pdf.
392. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Prostatakrebs [online]. 07.02.2018 [Zugriff: 25.10.2019]. URL: <https://www.gesundheitsinformation.de/prostatakrebs.2066.de.html>.
393. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Dokumentation und Würdigung der Anhörung zum Entwurf der Allgemeinen Methoden 6.0. Köln: IQWiG; 2020.
394. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzepte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V: Rapid Report; Auftrag A19-43; Version 1.1 [online]. 13.05.2020 [Zugriff: 02.06.2020]. (IQWiG-Berichte; Band 863). URL: https://www.iqwig.de/download/A19-43_Versorgungsnahe-Daten-zum-Zwecke-der-Nutzenbewertung_Rapid-Report_V1-1.pdf.

395. International Conference on Harmonisation Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials; an introductory note on an international guideline. *Stat Med* 1999; 18(15): 1905-1942.
396. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Webauftritt [online]. [Zugriff: 18.03.2015]. URL: <http://www.ich.org>.
397. International Society for Pharmacoeconomics and Outcomes Research. ISPOR good practices for outcomes research [online]. [Zugriff: 05.11.2019]. URL: <https://www.ispor.org/heor-resources/good-practices-for-outcomes-research>.
398. Int'Hout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014; 14: 25.
399. Ioannidis JPA, Evans S, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004; 141(10): 781-788.
400. Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA* 2001; 285(4): 437-443.
401. Ioannidis JPA, Mulrow CD, Goodman SN. Adverse events: the more you search, the more you find. *Ann Intern Med* 2006; 144(4): 298-300.
402. Irmen L, Linner U. Die Repräsentation generisch maskuliner Personenbezeichnungen: eine theoretische Integration bisheriger Befunde. *Z Psychol* 2005; 213(3): 167-175.
403. 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.
404. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. *Stat Med* 2006; 25(15): 2688-2699.
405. Jackson D, Law M, Rücker G, Schwarzer G. The Hartung-Knapp modification for random-effects meta-analysis: a useful refinement but are there any residual concerns? *Stat Med* 2017; 36(25): 3923-3934.
406. Jackson D, Turner R. Power analysis for random-effects meta-analysis. *Res Syn Meth* 2017; 8(3): 290-302.
407. Jackson N, Waters E. Criteria for the systematic review of health promotion and public health interventions. *Health Promot Int* 2005; 20(4): 367-374.
408. Jadad AR. Randomised controlled trials: a user's guide. London: BMJ Books; 1998.
409. Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *CMAJ* 1997; 156(10): 1411-1416.

410. Jadad AR, Enkin MW. Randomized controlled trials: questions, answers and musings. Malden: Blackwell Publishing; 2007.
411. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices; part 1. *Value Health* 2011; 14(4): 417-428.
412. Jansen JP, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R et al. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health* 2014; 17(2): 157-173.
413. Janzen T, Hausner E, Waffenschmidt S. Entwicklung und Evaluation von RCT- und SR-Filtern für die Suche nach nicht verschlagworteten Datensätzen in PubMed [online]. In: Entscheiden trotz Unsicherheit: 14. Jahrestagung des Deutschen Netzwerks Evidenzbasierte Medizin; 15.-16.03.2013; Berlin, Deutschland. 11.03.2013 [Zugriff: 21.11.2019]. URL: <http://www.egms.de/static/de/meetings/ebm2013/13ebm059.shtml>.
414. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. *J Clin Epidemiol* 2017; 89: 188-198.
415. Jenuwine ES, Floyd JA. Comparison of medical subject headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. *J Med Libr Assoc* 2004; 92(3): 349-353.
416. Jiao S, Tsutani K, Haga N. Review of Cochrane reviews on acupuncture: how Chinese resources contribute to Cochrane reviews. *Journal of Alternative and Complementary Medicine* 2013; 19(7): 613-621.
417. Johannesson M. Avoiding double-counting in pharmacoeconomic studies. *Pharmacoeconomics* 1997; 11(5): 385-388.
418. Johnson RF. Sample size issues for conjoint analysis. In: Orme BK (Ed). *Getting started with conjoint analysis: strategies for product design and pricing research*. Madison: Research Publishers LLC; 2010. S. 57-66.
419. Johnston BC, Ebrahim S, Carrasco-Labra A, Furukawa TA, Patrick DL, Crawford MW et al. Minimally important difference estimates and methods: a protocol. *BMJ Open* 2015; 5(10): e007953.
420. Jones B, Jarvis P, Lewis J, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996; 313(7048): 36-39.
421. Jones B, Kenward MG. *Design and analysis of cross-over trials*. Boca Raton: CRC Press; 2015. (Monographs on Statistics and Applied Probability; Band 138).

422. Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. *BMC Med* 2015; 13: 282.
423. Jones HE, Gatsonsis CA, Trikalinos TA, Welton NJ, Ades AE. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. *Stat Med* 2019; 38(24): 4789-4803.
424. Jull A, Bennett D. Do n-of-1 trials really tailor treatment? *Lancet* 2005; 365(9476): 1992-1994.
425. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001; 323(7303): 42-46.
426. Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of multi-arm parallel-group randomized trials: extension of the CONSORT 2010 statement. *JAMA* 2019; 321(16): 1610-1620.
427. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. Hoboken: Wiley-Interscience; 2002.
428. Kass PH, Gold EB. Modern epidemiologic study designs. In: Ahrens W, Pigeot I (Ed). *Handbook of epidemiology*. Berlin: Springer; 2005. S. 321-344.
429. Kastner M, Wilczynski NL, Walker-Dilks C, McKibbin KA, Haynes B. Age-specific search strategies for Medline. *J Med Internet Res* 2006; 8(4): e25.
430. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. *BMC Med Res Methodol* 2004; 4: 22.
431. Katz MH. Multivariable analysis: a primer for readers of medical research. *Ann Intern Med* 2003; 138(8): 644-650.
432. Kettunen T, Liimatainen L, Villberg J, Perko U. Developing empowering health counseling measurement: preliminary results. *Patient Education and Counseling* 2006; 64(1-3): 159-166.
433. Kickbusch IS. Health literacy: addressing the health and education divide. *Health Promot Int* 2001; 16(3): 289-297.
434. Kiefer C, Sturtz S, Bender R. Indirekte Vergleiche und Netzwerk-Metaanalysen: Schätzung von Effekten bei fehlenden direkten Vergleichsstudien; Teil 22 der Serie zur Bewertung wissenschaftlicher Publikationen. *Deutsches Ärzteblatt* 2015; 112(47): 803-808.
435. Kiefer C, Sturtz S, Bender R. A simulation study to compare different estimation approaches for network meta-analysis and corresponding methods to evaluate the consistency assumption. *BMC Med Res Methodol* 2020; 20(1): 36.
436. Kieser M. Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharm Stat* 2005; 4(2): 101-107.

437. Kieser M, Röhmel J, Friede T. Power and sample size determination when assessing the clinical relevance of trial results by 'responder analyses'. *Stat Med* 2004; 23(21): 3287-3305.
438. Kleinbaum DG, Klein M. *Survival analysis: a self-learning text*. New York: Springer; 2005.
439. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med* 2003; 22(17): 2693-2710.
440. Knelangen M, Hausner E, Metzendorf MI, Sturtz S, Waffenschmidt S. Trial registry searches for randomized controlled trials of new drugs required registry-specific adaptation to achieve adequate sensitivity. *J Clin Epidemiol* 2018; 94: 69-75.
441. Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. *J Clin Epidemiol* 2001; 54(2): 109-110.
442. Köbberling J. Der Zweifel als Triebkraft des Erkenntnisgewinns in der Medizin. In: Kunz R, Ollenschläger G, Raspe H, Jonitz G, Donner-Banzhoff N (Ed). *Lehrbuch evidenzbasierte Medizin in Klinik und Praxis*. Köln: Deutscher Ärzteverlag; 2007. S. 3-14.
443. Koch A, Ziegler S. Metaanalyse als Werkzeug zum Erkenntnisgewinn. *Medizinische Klinik* 2000; 95(2): 109-116.
444. Koch K, Waltering A. IQWiG-Gesundheitsinformation: pragmatischer Weg zum Themenkatalog. *Deutsches Ärzteblatt* 2016; 113(11): A489-A493.
445. Köhler M, Haag S, Biester K, Brockhaus AC, McGauran N, Grouven U et al. Information on new drugs at market entry: retrospective analysis of health technology assessment reports versus regulatory reports, journal publications, and registry reports. *BMJ* 2015; 350: h796.
446. Kolman J, Meng P, Scott G. *Good clinical practice: standard operating procedures for clinical researchers*. Chichester: Wiley; 1998.
447. Kommission der Europäischen Gemeinschaften. Richtlinie 2003/63/EG der Kommission vom 25. Juni 2003 zur Änderung der Richtlinie 2001/83/EG des Europäischen Parlaments und des Rates zur Schaffung eines Gemeinschaftskodexes für Humanarzneimittel. *Amtsblatt der Europäischen Gemeinschaften* 2003; 46(L159): 46-94.
448. Koopmanschap MA, Rutten FFH, Van Ineveld BM, Van Roijen L. The friction cost method for measuring indirect costs of disease. *J Health Econ* 1995; 14(2): 171-189.
449. Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. *JAMA* 2006; 296(10): 1286-1289.
450. Krankheitserfahrungen.de. Webauftritt [online]. [Zugriff: 21.11.2019]. URL: www.krankheitserfahrungen.de.
451. Kreis J, Puhan MA, Schünemann HJ, Dickersin K. Consumer involvement in systematic reviews of comparative effectiveness research. *Health Expect* 2013; 16(4): 323-337.

452. Kristensen FB, Sigmund H. Health technology assessment handbook. Kopenhagen: Danish Centre for Health Technology Assessment; 2007. URL: <https://www.sst.dk/~media/ECAAC5AA1D6943BEAC96907E03023E22.ashx>.
453. Kristman V, Manno M, Cote P. Loss to follow-up in cohort studies: how much is too much? *Eur J Epidemiol* 2004; 19(8): 751-760.
454. Krug S. Don't make me think! Web Usability; das intuitive Web. Heidelberg: mitp; 2006.
455. Kulbe A. Grundwissen Psychologie, Soziologie und Pädagogik: Lehrbuch für Pflegeberufe. Stuttgart: Kohlhammer; 2009.
456. Kunz R, Djulbegovic B, Schünemann HJ, Stanulla M, Muti P, Guyatt G. Misconceptions, challenges, uncertainty, and progress in guideline recommendations. *Semin Hematol* 2008; 45(3): 167-175.
457. Kuss O. Statistical methods for meta-analyses including information from studies without any events: add nothing to nothing and succeed nevertheless. *Stat Med* 2015; 34(7): 1097-1116.
458. Laaser U, Hurrelmann K. Gesundheitsförderung und Krankheitsprävention. In: Hurrelmann K, Laaser U (Ed). *Handbuch Gesundheitswissenschaften*. Weinheim: Juventa Verlag; 1998. S. 395-424.
459. Lacny S, Wilson T, Clement F, Roberts DJ, Faris P, Ghali WA et al. Kaplan-Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. *J Clin Epidemiol* 2018; 93: 25-35.
460. Lange S, Freitag G. Choice of delta: requirements and reality; results of a systematic review. *Biom J* 2005; 47(1): 12-27.
461. Lapsley P. The patient's journey: travelling through life with a chronic illness. *BMJ* 2004; 329(7466): 582-583.
462. Lavis JN. How can we support the use of systematic reviews in policymaking? *PLoS Med* 2009; 6(11): e1000141.
463. Law AM. How to build valid and credible simulation models. In: Mason SJ, Hill RR, Mönch L, Rose O, Jefferson T, Fowler JW (Ed). *Proceedings of the 2008 Winter Simulation Conference*. Piscataway: IEE; 2008. S. 39-47.
464. Law AM. *Simulation modeling and analysis*. New York: McGraw-Hill Education; 2015.
465. Lee ET, Wang JW. *Statistical methods for survival data analysis*. Hoboken: Wiley-Interscience; 2003.
466. Leeftang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008; 149(12): 889-897.

467. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf M et al. Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 67-107.
468. Leidl R, Graf von der Schulenburg JM, Wasem J (Ed). Ansätze und Methoden der ökonomischen Evaluation: eine internationale Perspektive. Baden-Baden: Nomos Verlagsgesellschaft; 1999. (Health Technology Assessments; Band 9).
469. Lelgemann M. Kritische Bewertung medizinischer Leitlinien: eine Analyse und Diskussion der Ergebnisse des Deutschen Leitlinien-Clearingverfahrens [Dissertation]. Köln: Universität; 2009. URL: <http://www.aezq.de/mdb/edocs/pdf/literatur/diss-ml-2009.pdf>.
470. Lewin S, Glenton C, Munthe-Kaas H, Carlsen B, Colvin CJ, Gülmezoglu M et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). PLoS Med 2015; 12(10): e1001895.
471. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001; 322(7300): 1479-1480.
472. Leys M. Health care policy: qualitative evidence and health technology assessment. Health Policy 2003; 65(3): 217-226.
473. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009; 339: b2700.
474. Liberati A, Sheldon TA, Banta HD. EUR-ASSESS project subgroup report on methodology: methodological guidance for the conduct of health technology assessment. Int J Technol Assess Health Care 1997; 13(2): 186-219.
475. Lieb K, Klemperer D, Koch K, Baethge C, Ollenschläger G, Ludwig WD. Interessenskonflikt in der Medizin: mit Transparenz Vertrauen stärken. Deutsches Ärzteblatt 2011; 108(6): A256-A260.
476. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol 2009; 62(4): 364-373.
477. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, Van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11): 1061-1066.
478. Lipscomb J, Drummond M, Fryback D, Gold M, Revicki D. Retaining, and enhancing, the QALY. Value Health 2009; 12(Suppl 1): S18-S26.
479. Lipscomb J, Weinstein MC, Torrance GW. Time preference. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. S. 214-246.

480. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken: Wiley; 2002.
481. Lo B, Field MJ (Ed). Conflict of interest in medical research, education, and practice. Washington: National Academies Press; 2009.
482. Lord SJ, Irwig LM, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006; 144(11): 850-855.
483. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004; 23(20): 3105-3124.
484. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc* 2006; 101(474): 447-459.
485. Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med* 2007; 26(20): 3681-3699.
486. Luce BR, Manning WG, Siegel JE, Lipscomb J. Estimating costs in cost-effectiveness analysis. In: Gold MR, Russell LB, Siegel JE, Weinstein MC (Ed). *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996. S. 176-213.
487. Lühnen J, Albrecht M, Mühlhauser I, Steckelberg A. Leitlinie evidenzbasierte Gesundheitsinformation [online]. 20.02.2017 [Zugriff: 26.04.2017]. URL: <http://www.leitlinie-gesundheitsinformation.de/wp-content/uploads/2014/05/Leitlinie-evidenzbasierte-Gesundheitsinformation.pdf>.
488. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002; 21(16): 2313-2324.
489. Lysdahl KB, Mozygemba K, Burns L, Chilcott JB, Brönneke JB, Hofmann B. Guidance for assessing effectiveness, economic aspects, ethical aspects, socio-cultural aspects and legal aspects in complex technologies. INTEGRATE-HTA; 2016. URL: http://www.integrate-hta.eu/wp-content/uploads/2016/08/IPP_Guidance-INTEGRATE-HTA_Nr.3_FINAL.pdf.
490. Lysdahl KB, Oortwijn W, Van der Wilt GJ, Refolo P, Sacchini D, Mozygemba K et al. Ethical analysis in HTA of complex health interventions. *BMC Med Ethics* 2016; 17: 16.
491. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: analysing and presenting results [online]. In: Deeks JJ, Bossuyt PM, Gatsonis C (Ed.). *Handbook for systematic reviews of diagnostic test accuracy version 1.0*. 23.12.2010 [Zugriff: 07.09.2020]. URL: <https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter%2010%20-%20Version%201.0.pdf>.
492. MacDermid JC, Brooks D, Solway S, Switzer-McIntyre S, Brosseau L, Graham ID. Reliability and validity of the AGREE instrument used by physical therapists in assessment of clinical practice guidelines. *BMC Health Serv Res* 2005; 5: 18.

493. Maetzel A. Der Gebrauch von Nutzwerten im gesundheitsökonomischen Vergleich von Interventionen bei verschiedenen Krankheitsbildern: eine Einführung. *Z Rheumatol* 2004; 63(5): 380-384.
494. Malterud K. The art and science of clinical knowledge: evidence beyond measures and numbers. *Lancet* 2001; 358(9279): 397-400.
495. Mandelblatt JS, Fryback DG, Weinstein MC, Russell LB, Gold MR, Hadorn DC. Assessing the effectiveness of health interventions. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996. S. 135-175.
496. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol* 2009; 27(24): 4027-4034.
497. Mandrekar SJ, Sargent DJ. All-comers versus enrichment design strategy in phase II trials. *J Thorac Oncol* 2011; 6(4): 658-660.
498. Mangiapane S, Velasco Garrido M. Surrogatendpunkte als Parameter der Nutzenbewertung [online]. 2009 [Zugriff: 21.11.2019]. (Schriftenreihe Health Technology Assessment; Band 91). URL: https://portal.dimdi.de/de/hta/hta_berichte/hta250_bericht_de.pdf.
499. March JS, Silva SG, Compton S, Shapiro M, Califf R, Krishnan R. The case for practical clinical trials in psychiatry. *Am J Psychiatry* 2005; 162(5): 836-846.
500. Marsh K, Lanitis T, Neasham D, Orfanos P, Caro J. Assessing the value of healthcare interventions using multi-criteria decision analysis: a review of the literature. *Pharmacoeconomics* 2014; 32(4): 345-365.
501. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Res Synth Methods* 2018; 9(4): 602-614.
502. Martin LP, Arias-Gallo J, Perez-Chrzanowska H, Seco PR, Moro JG, Burgueno-Garcia M. Transfusion requirements in microsurgical reconstruction in maxillofacial surgery: ethical and legal problems of patients who are Jehovah's Witnesses. *Craniofac Trauma Reconstr* 2013; 6(1): 31-36.
503. Martini P. *Methodenlehre der therapeutischen Untersuchung*. Berlin: Springer; 1932.
504. Mathes T, Kuss O. A comparison of methods for meta-analysis of a small number of studies with binary outcomes. *Res Syn Meth* 2018; 9(3): 366-381.
505. Matthias K, Gruber S, Pietsch B. Evidenz von Volume-Outcome-Beziehungen und Mindestmengen: Diskussion in der aktuellen Literatur. *Gesundheits- und Sozialpolitik* 2014; 68(3): 23-30.

506. Mattke S, Kelley E, Scherer P, Hurst J, Lapetra MLG. Health Care Quality Indicators Project: initial indicators report [online]. 09.03.2006 [Zugriff: 21.11.2019]. (OECD Health Working Papers; Band 22). URL: <http://dx.doi.org/10.1787/481685177056>.
507. Mauskopf JA, Earnshaw S, Mullins CD. Budget impact analysis: review of the state of the art. *Expert Rev Pharmacoecon Outcomes Res* 2005; 5(1): 65-79.
508. Mauskopf JA, Sullivan SD, Annemans L, Caro J, Mullins CD, Nuijten M et al. Principles of good practice for budget impact analysis: report of the ISPOR Task Force on Good Research Practices; budget impact analysis. *Value Health* 2007; 10(5): 336-347.
509. Mc Gowan J, Sampson M, Salzwedel DM, Cogo E, Foester V, Lefebvre C. PRESS: peer review electronic search strategies; 2015 guideline explanation and elaboration (PRESS E&E) [online]. 01.2016 [Zugriff: 21.11.2019]. URL: https://www.cadth.ca/sites/default/files/pdf/CP0015_PRESS_Update_Report_2016.pdf.
510. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA* 2003; 289(19): 2545-2553.
511. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. *BMJ* 2002; 324(7351): 1448-1451.
512. McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research: a narrative review. *Trials* 2010; 11: 37.
513. McGregor M, Caro JJ. QALYs: are they helpful to decision makers? *Pharmacoeconomics* 2006; 24(10): 947-952.
514. McInnes MDF, Moher D, Thoms BD, McGrath TA, Bossuyt PM, Clifford T et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* 2018; 319(4): 388-396.
515. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 2005; 97(16): 1180-1184.
516. Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. *Int J Technol Assess Health Care* 2013; 29(3): 343-350.
517. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340: c869.
518. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; 339: b2535.
519. Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. *Health Technol Assess* 2003; 7(41): 1-90.

520. Moher D, Shamseer L, Clarke M, Gherzi D, Liberati A, Petticrew M et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015; 4: 1.
521. Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M. A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. *Stat Methods Med Res* 2010; 19(3): 205-236.
522. Molnar FJ, Man-Son-Hing M, Fergusson D. Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia. *J Am Geriatr Soc* 2009; 57(3): 536-546.
523. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162(1): W1-W73.
524. Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. *Int J Technol Assess Health Care* 2012; 28(2): 138-144.
525. Mozygemba K, Hofmann B, Lysdal KB, Pfadenhauer L, Van der Wilt G, Gerhardus A. Guidance to assess socio-cultural aspects. In: Lysdahl KB, Mozygemba K, Burns L, Chilcott JB, Brönneke JB, Hofmann B (Ed). *Guidance for assessing effectiveness, economic aspects, ethical aspects, socio-cultural aspects and legal aspects in complex technologies. INTEGRATE-HTA*; 2016. S. 76-100. URL: http://www.integrate-hta.eu/wp-content/uploads/2016/08/IPP_Guidance-INTEGRATE-HTA_Nr.3_FINAL.pdf.
526. Mühlbacher AC, Bethge S, Tockhorn A. Präferenzmessung im Gesundheitswesen: Grundlage von Discrete-Choice-Experimenten. *Gesundheitsökonomie & Qualitätsmanagement* 2013; 18(4): 159-172.
527. Mullan RJ, Flynn DN, Carlberg B, Tleyjeh IM, Kamath CC, LaBella ML et al. Systematic reviewers commonly contact study authors but do so with limited rigor. *J Clin Epidemiol* 2009; 62(2): 138-142.
528. Mullins MM, DeLuca JB, Crepaz N, Lyles CM. Reporting quality of search methods in systematic reviews of HIV behavioral interventions (2000-2010): are the searches clearly explained, systematic and reproducible? *Res Syn Meth* 2013; 5(2): 116-130.
529. Müllner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. *Ann Intern Med* 2002; 136(2): 122-126.
530. Munn Z, Porritt K, Lockwood C, Aromataris E, Pearson A. Establishing confidence in the output of qualitative research synthesis: the ConQual approach. *BMC Med Res Methodol* 2014; 14: 108.
531. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. *Health Technol Assess* 1998; 2(16): 1-274.

532. Muth C, Gensichen J, Beyer M, Hutchinson A, Gerlach FM. The systematic guideline review: method, rationale, and test on chronic heart failure. BMC Health Serv Res 2009; 9: 74.
533. National Advisory Committee on Health and Disability. Screening to improve health in New Zealand: criteria to assess screening. Wellington: National Health Committee; 2003. URL: https://www.nsu.govt.nz/system/files/resources/screening_to_improve_health.pdf.
534. National Health and Medical Research Council. Cultural competency in health: a guide for policy, partnerships and participation. Canberra: Commonwealth of Australia; 2006. URL: <https://www.nhmrc.gov.au/file/2771/download?token=2dNegUjD>.
535. National Institute for Health and Care Excellence. Guidance and advice list: interventional procedures guidance [online]. [Zugriff: 22.11.2019]. URL: <https://www.nice.org.uk/guidance/indevelopment?type=ipg>.
536. National Institute for Health and Care Excellence. Guide to the processes of technology appraisal. London: NICE; 2014. (NICE Process and Methods; Band 19). URL: <https://www.nice.org.uk/process/pmg19/resources/guide-to-the-processes-of-technology-appraisal-pdf-72286663351237>.
537. National Institutes of Health, Department of Health and Human Services. Clinical trials registration and results information submission: final rule. Fed Regist 2016; 81(183): 64981-65157.
538. Neidhardt K, Wasmuth T, Schmid A. Die Gewichtung multipler patientenrelevanter Endpunkte: ein methodischer Vergleich von Conjoint Analyse und Analytic Hierarchy Process unter Berücksichtigung des Effizienzgrenzenkonzepts des IQWiG; Diskussionspapier [online]. 02.2012 [Zugriff: 05.11.2019]. (Wirtschaftswissenschaftliche Diskussionspapiere; Band 02-12). URL: http://www.fwi.uni-bayreuth.de/de/download/WP_02-12.pdf.
539. Newcombe RG, Bender R. Implementing GRADE: calculating the risk difference from the baseline risk and the relative risk. Evid Based Med 2014; 19(1): 6-8.
540. Nielsen J, Loranger H. Web Usability. München: Addison-Wesley; 2008.
541. Nilsen ES, Myrhaug HT, Johansen M, Oliver S, Oxman AD. Methods of consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. Cochrane Database Syst Rev 2006; (3): CD004563.
542. Nord E. An alternative to QALYs: the saved young life equivalent (SAVE). BMJ 1992; 305(6858): 875-877.
543. Nord E. Cost-value analysis in health care: making sense out of QALYs. Cambridge: Cambridge University Press; 1999.
544. Nordin A, Taft C, Lundgren-Nilsson A, Dencker A. Minimal important differences for fatigue patient reported outcome measures: a systematic review. BMC Med Res Methodol 2016; 16: 62.

545. Nüesch E, Jüni P. Commentary: which meta-analyses are conclusive? *Int J Epidemiol* 2009; 38(1): 298-303.
546. Nutbeam D. Health promotion glossary. *Health Promot Int* 1998; 13(4): 349-364.
547. O'Mahony JF, Paulden M. NICE's selective application of differential discounting: ambiguous, inconsistent, and unjustified. *Value Health* 2014; 17(5): 493-496.
548. O'Neill RT. Statistical analyses of adverse event data from clinical trials: special emphasis on serious events. *Drug Inf J* 1987; 21(1): 9-20.
549. O'Neill RT. Assessment of safety. In: Peace KE (Ed). *Biopharmaceutical statistics for drug development*. New York: Dekker; 1988. S. 543-604. (Statistics; Band 86).
550. Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schünemann H et al. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev* 2011; (4): MR000012.
551. Oliver A. A normative perspective on discounting health outcomes. *J Health Serv Res Policy* 2013; 18(3): 186-189.
552. Oostenbrink JB, Koopmanschap MA, Rutten FF. Standardisation of costs: the Dutch Manual for Costing in economic evaluations. *Pharmacoeconomics* 2002; 20(7): 443-454.
553. Organisation for Economic Co-operation and Development. Secretary-General's report to ministers 2016. Paris: OECD Publishing; 2016. URL: http://dx.doi.org/10.1787/sg_report-2016-en.
554. Orlewska E, Mierzejewski P. Proposal of Polish guidelines for conducting financial analysis and their comparison to existing guidance on budget impact in other countries. *Value Health* 2004; 7(1): 1-10.
555. Ousmen A, Touraine C, Deliu N, Cottone F, Bonnetain F, Efficace F et al. Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. *Health Qual Life Outcomes* 2018; 16(1): 228.
556. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ* 1988; 138(8): 697-703.
557. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991; 44(11): 1271-1278.
558. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992; 116(1): 78-84.
559. Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA et al. Agreement among reviewers of review articles. *J Clin Epidemiol* 1991; 44(1): 91-98.

560. Page MJ, Higgins JPT, Sterne JAC. Assessing risk of bias due to missing results in a synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 349-374.
561. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998; 17(24): 2815-2834.
562. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980; 302(20): 1109-1117.
563. Paulden M, Claxton K. Budget allocation and the revealed social rate of time preference for health. *Health Econ* 2012; 21(5): 612-618.
564. Paynter R, Banez LL, Berliner E, Erinoff E, Lege-Matsuura J, Potter S et al. EPC methods: an exploration of the use of text-mining software in systematic reviews; AHRQ publication 16-EHC023-EF [online]. 04.2016 [Zugriff: 18.11.2019]. URL: https://www.ncbi.nlm.nih.gov/books/NBK362044/pdf/Bookshelf_NBK362044.pdf.
565. Perleth M, Busse R, Gerhardus A, Gibis B, Lühmann D, Zentner A (Ed). *Health Technology Assessment: Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung*. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014.
566. Perleth M, Gibis B, Velasco Garrido M, Zentner A. Organisationsstrukturen und Qualität. In: Perleth M, Busse R, Gerhardus A, Gibis B, Zentner A (Ed). *Health Technology Assessment: Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung*. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. S. 265-280.
567. Perleth M, Jakubowski E, Busse R. What is 'best practice' in health care? State of the art and perspectives in improving the effectiveness and efficiency of the European health care systems. *Health Policy* 2001; 56(3): 235-250.
568. Petitti DB, Teutsch SM, Barton MB, Sawaya GF, Ockene JK, DeWitt T. Update on the methods of the U.S. Preventive Services Task Force: insufficient evidence. *Ann Intern Med* 2009; 150(3): 199-205.
569. Petkova E, Tarpey T, Huang L, Deng L. Interpreting meta-regression: application to recent controversies in antidepressants' efficacy. *Stat Med* 2013; 32(17): 2875-2892.
570. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* 2004; 8(36): iii-iv, ix-xi, 1-158.
571. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA* 2012; 308(24): 2594-2604.

572. Pieper D, Jülich F, Antoine SL, Bächle C, Chernyak N, Genz J et al. Studies analysing the need for health-related information in Germany: a systematic review. *BMC Health Serv Res* 2015; 15: 407.
573. Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. *Stat Med* 1999; 18(6): 643-654.
574. Pocock SJ. *Clinical trials: a practical approach*. Chichester: Wiley; 1983.
575. Poltawski L, Abraham C, Forster A, Goodwin VA, Kilbride C, Taylor RS et al. Synthesising practice guidelines for the development of community-based exercise programmes after stroke. *Implement Sci* 2013; 8: 115.
576. Pranic S, Marusic A. Changes to registration elements and results in a cohort of Clinicaltrials.gov trials were not reflected in published articles. *J Clin Epidemiol* 2016; 70: 26-37.
577. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989; 8(4): 431-440.
578. Provalis Research. WordStat: content analysis and text mining software [online]. [Zugriff: 04.11.2019]. URL: <https://provalisresearch.com/products/content-analysis-software/>.
579. Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol* 2012; 12: 173.
580. Putter H, Schumacher M, Van Houwelingen HC. On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: the Fine-Gray model revisited. *Biom J* 2020; 62(3): 790-807.
581. Raftery JP. How should we value future health? Was NICE right to change? *Value Health* 2013; 16(5): 699-700.
582. Raum E, Perleth M. *Methoden der Metaanalyse von diagnostischen Genauigkeitsstudien*. Köln: Deutsches Institut für Medizinische Dokumentation und Information; 2003. (Schriftenreihe Health Technology Assessment; Band 2). URL: http://portal.dimdi.de/de/hta/hta_berichte/hta025_bericht_de.pdf.
583. 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.
584. Relevo R, Balshem H. Finding evidence for comparing medical interventions [online]. 01.2011 [Zugriff: 21.11.2019]. URL: http://www.effectivehealthcare.ahrq.gov/ehc/products/276/605/MethodsGuide_Relevo_FindingEvidenceforComparing_20120430.pdf.

585. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008; 61(2): 102-109.
586. Reynen E, Robson R, Ivory J, Hwee J, Straus SE, Pham B et al. A retrospective comparison of systematic reviews with same-topic rapid reviews. *J Clin Epidemiol* 2018; 96: 23-34.
587. Rice M, Ali MU, Fitzpatrick-Lewis D, Kenny M, Raina P, Sherifali D. Testing the effectiveness of simplified search strategies for updating systematic reviews. *J Clin Epidemiol* 2017; 88: 148-153.
588. Richardson J, Khan MA, Iezzi A, Maxwell A. Cross-national comparison of twelve quality of life instruments: MIC paper 5 Canada. Clayton: Centre for Health Economics; 2012. (Research Papers; Band 82). URL: <https://www.aqol.com.au/papers/researchpaper82.pdf>.
589. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011; 342: 964-967.
590. Ringbaek T, Brøndum E, Martinez G, Lange P. EuroQoL in assessment of the effect of pulmonary rehabilitation COPD patients. *Respir Med* 2008; 102(11): 1563-1567.
591. Robinson KA, Whitlock EP, O'Neil ME, Anderson JK, Hartling L, Dryden DM et al. Integration of existing systematic reviews: AHRQ publication no. 14-EHC016-EF [online]. 06.2014 [Zugriff: 21.11.2019]. URL: https://www.ncbi.nlm.nih.gov/sites/books/NBK216379/pdf/Bookshelf_NBK216379.pdf.
592. Rockwood K, Fay S, Song X, MacKnight C, Gorman M. Attainment of treatment goals by people with Alzheimer's disease receiving galantamine: a randomized controlled trial. *CMAJ* 2006; 174(8): 1099-1105.
593. Roebruck P, Elze M, Hauschke D, Leverkus F, Kieser M. Literaturübersicht zur Fallzahlplanung für Äquivalenzprobleme. *Inform Biom Epidemiol Med Biol* 1997; 28(2): 51-63.
594. Röhm J, Hauschke D, Koch A, Pigeot I. Biometrische Verfahren zum Wirksamkeitsnachweis im Zulassungsverfahren: Nicht-Unterlegenheit in klinischen Studien. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2005; 48(5): 562-571.
595. Ross SM. Simulation. San Diego: Academic Press; 2013.
596. Rossouw JE. Estrogens for prevention of coronary heart disease: putting the brakes on the bandwagon. *Circulation* 1996; 94(11): 2982-2985.
597. Rothwell PM. Treating individuals 2: subgroup analysis in randomised controlled trials; importance, indications, and interpretation. *Lancet* 2005; 365(9454): 176-186.

598. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol* 2015; 15: 99.
599. Royall RM. The effect of sample size on the meaning of significance tests. *Am Stat* 1986; 40(4): 313-315.
600. Royle P, Bain L, Waugh N. Systematic reviews of epidemiology in diabetes: finding the evidence. *BMC Med Res Methodol* 2005; 5: 2.
601. Royle P, Waugh N. Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system. *Health Technol Assess* 2003; 7(34): 1-51.
602. Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Stat Med* 2000; 19(14): 1831-1847.
603. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Appl Stat* 1994; 43(3): 429-467.
604. Rücker G. Network meta-analysis, electrical networks and graph theory. *Res Syn Meth* 2012; 3(4): 312-324.
605. 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.
606. Russell LB, Siegen JE, Daniels N, Gold MR, Luce BR, Mandelblatt JS. Cost-effectiveness analysis as a guide to resource allocation in health: roles and limitations. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996. S. 3-24.
607. Ryan M, Gerard K, Amaya-Amaya M (Ed). *Using discrete choice experiments to value health and health care*. Dordrecht: Springer; 2008. (The Economics of Non-Market Goods and Resources; Band 11).
608. Saaty T, Vargas LG. *Decision making with the analytic network process: economic, political, social and technological applications with benefits, opportunities, costs and risks*. New York: Springer; 2013. (International Series in Operations Research & Management Science; Band 195).
609. Saaty TL. A scaling method for priorities in hierarchical structures. *J Math Psychol* 1977; 15(3): 234-281.
610. Saaty TL. Decision making with the Analytic Hierarchy Process. *International Journal of Services Sciences* 2008; 1(1): 83-98.
611. Saaty TL. *Theory and applications of the analytic network process: decision making with benefits, opportunities, costs, and risks*. Pittsburgh: RWS Publications; 2009.

612. 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.
613. Sachverständigenrat für die Konzertierte Aktion im Gesundheitswesen. Bedarfsgerechtigkeit und Wirtschaftlichkeit; Band III: Über- Unter- und Fehlversorgung; Gutachten 2000/2001; ausführliche Zusammenfassung [online]. 08.2001 [Zugriff: 21.11.2019]. URL: http://www.svr-gesundheit.de/fileadmin/user_upload/Gutachten/2000-2001/Kurzf-de-01.pdf.
614. Sackett DL. Bias in analytic research. *J Chronic Dis* 1979; 32(1-2): 51-63.
615. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; 312(7023): 71-72.
616. Salanti G. Indirect and mixed-treatment comparison, network, or multipletreatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Syn Meth* 2012; 3(2): 80-97.
617. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008; 17(3): 279-301.
618. Salanti G, Marinho V, Higgins JPT. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *J Clin Epidemiol* 2009; 62(8): 857-864.
619. Sampson M, Barrowman NJ, Moher D, Klassen TP, Pham B, Platt R et al. Should meta-analysts search Embase in addition to Medline? *J Clin Epidemiol* 2003; 56(10): 943-955.
620. Sampson M, McGowan J. Errors in search strategies were identified by type and frequency. *J Clin Epidemiol* 2006; 59(10): 1057-1063.
621. Sampson M, McGowan J. Inquisitio validus index medicus: a simple method of validating MEDLINE systematic review searches. *Res Syn Meth* 2011; 2(2): 103-109.
622. Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. *J Clin Epidemiol* 2009; 62(9): 944-952.
623. Sampson M, McGowan J, Lefebvre C, Moher D, Grimshaw J. PRESS: Peer Review of Electronic Search Strategies. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2008. URL: http://www.cadth.ca/media/pdf/477_PRESS-Peer-Review-Electronic-Search-Strategies_tr_e.pdf.
624. Sampson M, McGowan J, Tetzlaff J, Cogo E, Moher D. No consensus exists on search reporting methods for systematic reviews. *J Clin Epidemiol* 2008; 61(8): 748-754.
625. Sampson MJ. Updating searches for systematic reviews [Dissertation]. Aberystwyth: University; 2009. URL: <http://cadair.aber.ac.uk/dspace/bitstream/2160/3760/1/Sampson%20-%20Updating%20Searches%20for%20Systematic%20Reviews%20-%20PhD.pdf>.
626. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 2005; 23(9): 2020-2027.

627. Sargent DJ, Mandrekar SJ. Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. *Clin Trials* 2013; 10(5): 647-652.
628. Sargent RG. Verification and validation of simulation models. In: Johansson B, Jain S, Montoya-Torres J, Hagan J, Yücesan E (Ed). *Proceedings of the 2010 Winter Simulation Conference*. Piscataway: IEEE; 2010. S. 166-183.
629. SAS Institute. SAS/STAT 9.2 user's guide: second edition [online]. 2009 [Zugriff: 21.11.2019]. URL: <http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf>.
630. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J R Stat Soc Ser A* 1999; 162(1): 71-94.
631. Sauerland S, Fujita-Rohwerder N, Zens Y, Molnar S. Premarket evaluation of medical devices: a cross-sectional analysis of clinical studies submitted to a German ethics committee. *BMJ Open* 2019; 9(2): e027041.
632. Sawaya GF, Guirguis-Blake J, LeFevre M, Harris R, Petitti D. Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. *Ann Intern Med* 2007; 147(12): 871-875.
633. Sayers A. Tips and tricks in performing a systematic review. *Br J Gen Pract* 2007; 57(545): 999.
634. Schlosser RW, Wendt O, Bhavnani S, Nail-Chiwetalu B. Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive pearl growing; a review. *International Journal of Language and Communication Disorders* 2006; 41(5): 567-582.
635. Schluter PJ, Ware RS. Single patient (n-of-1) trials with binary treatment preference. *Stat Med* 2005; 24(17): 2625-2636.
636. Schmoor C, Schumacher M, Finke J, Beyersmann J. Competing risks and multistate models. *Clin Cancer Res* 2013; 19(1): 12-21.
637. Schöffski O. Grundformen gesundheitsökonomischer Evaluationen. In: Schöffski O, Graf von der Schulenburg JM (Ed). *Gesundheitsökonomische Evaluationen*. Berlin: Springer; 2012. S. 43-70.
638. Schöffski O, Graf von der Schulenburg JM (Ed). *Gesundheitsökonomische Evaluationen*. Berlin: Springer; 2012.
639. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340: c332.
640. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet* 2002; 359(9308): 781-785.

641. Schünemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. *Health Qual Life Outcomes* 2006; 4: 62.
642. Schünemann HJ, Best D, Vist GE, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. *CMAJ* 2003; 169(7): 677-680.
643. Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development; 9: grading evidence and recommendations. *Health Res Policy Syst* 2006; 4: 21.
644. Schwarzer R, Rochau U, Saverno K, Jahn B, Bornschein B, Muehlberger N et al. Systematic overview of cost-effectiveness thresholds in ten countries across four continents. *J Comp Eff Res* 2015; 4(5): 485-504.
645. Sculpher M. The role and estimation of productivity costs in economic evaluation. In: Drummond MF, McGuire A (Ed). *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press; 2001. S. 94-112.
646. Sculpher MJ, O'Brien BJ. Income effects of reduced health and health effects of reduced income: implications for health-state valuation. *Med Decis Making* 2000; 20(2): 207-215.
647. Senn S. *Cross-over trials in clinical research*. Chichester: Wiley; 2002.
648. Senn SJ. Inherent difficulties with active control equivalence studies. *Stat Med* 1993; 12(24): 2367-2375.
649. Senn SJ. The many modes of meta. *Drug Inf J* 2000; 34(2): 535-549.
650. Senn SJ. Trying to be precise about vagueness. *Stat Med* 2007; 26(7): 1417-1430.
651. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015; 349: g7647.
652. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). *PLoS One* 2007; 2(12): e1350.
653. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007; 7: 10.
654. Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009; 62(10): 1013-1020.
655. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* 2017; 358: j4008.

656. Shechter SM, Schaefer AJ, Braithwaite RS, Roberts MS. Increasing the efficiency of Monte Carlo cohort simulations with variance reduction techniques. *Med Decis Making* 2006; 26(5): 550-553.
657. Shekelle PG, Ortiz E, Rhodes S, Morton SC, Eccles MP, Grimshaw JM et al. Validity of the agency for healthcare research and quality clinical practice guidelines: how quickly do guidelines become outdated? *JAMA* 2001; 286(12): 1461-1467.
658. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med* 2007; 147(4): 224-233.
659. Siebert U. Entscheidungsanalytische Modelle zur Sicherung der Übertragbarkeit internationaler Evidenz von HTA auf den Kontext des deutschen Gesundheitssystems: ein Methodenbeitrag zu HTA [online]. 2005 [Zugriff: 05.11.2019]. (Schriftenreihe Health Technology Assessment; Band 16). URL: https://portal.dimdi.de/de/hta/hta_berichte/hta099_bericht_de.pdf.
660. Siebert U. When should decision-analytic modeling be used in the economic evaluation of health care? *Eur J Health Econ* 2003; 4(3): 143-150.
661. Siering U, Eikermann M, Hausner E, Hoffmann-Esser W, Neugebauer EA. Appraisal tools for clinical practice guidelines: a systematic review. *PLoS One* 2013; 8(12): e82915.
662. Signorovitch JE, Wu EQ, Yu AP, Gerrits CM, Kantor E, Bao Y et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics* 2010; 28(10): 935-945.
663. Silvestre MAA, Dans LF, Dans AL. Trade-off between benefit and harm is crucial in health screening recommendations; part II: evidence summaries. *J Clin Epidemiol* 2011; 64(3): 240-249.
664. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. *Stat Methods Med Res* 2016; 25(6): 2858-2877.
665. 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.
666. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* 1994; 69(6): 979-985.
667. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst* 2009; 101(21): 1446-1452.
668. Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, Ioannidis JPA. Diagnostic tests often fail to lead to changes in patient outcomes. *J Clin Epidemiol* 2014; 67(6): 612-621.

669. Skipka G, Wieseler B, Kaiser T, Thomas S, Bender R, Windeler J et al. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biom J* 2016; 58(1): 43-58.
670. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses: sometimes informative, usually misleading. *BMJ* 1999; 318(7197): 1548-1551.
671. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995; 14(24): 2685-2699.
672. Sölétormos G, Duffy MJ, Hayes DF, Sturgeon CM, Barak V, Bossuyt PM et al. Design of tumor biomarker-monitoring trials: a proposal by the European Group on Tumor Markers. *Clin Chem* 2013; 59(1): 52-59.
673. Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ* 2009; 338: b1147.
674. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ et al. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess* 2010; 14(8): 1-193.
675. Spencer FA, Iorio A, You J, Murad MH, Schünemann HJ, Vandvik PO et al. Uncertainties in baseline risk estimates and confidence in treatment effects. *BMJ* 2012; 345: e7401.
676. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. *J R Stat Soc Ser A* 1994; 157(3): 357-416.
677. Spiegelhalter DJ, Myles JP, Jones D, Abrams KR. Methods in health service research: an introduction to Bayesian methods in health technology assessment. *BMJ* 1999; 319(7208): 508-512.
678. St-Pierre C, Desmeules F, Dionne CE, Fremont P, MacDermid JC, Roy JS. Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review. *Disability and Rehabilitation* 2016; 38(2): 103-122.
679. Stansfield C, O'Mara-Eves A, Thomas J. Text mining for search term development in systematic reviewing: a discussion of some methods and challenges. *Res Synth Methods* 2017; 8(3): 355-365.
680. Statens Beredning för Medicinsk Utvärdering. Utvärdering av metoder i hälso- och sjukvården: en handbok. Stockholm: SBU; 2012. URL: <http://www.sbu.se/globalassets/ebm/metodbok/sbushandbok.pdf>.
681. Statistisches Bundesamt. Verbraucherpreisindizes [online]. [Zugriff: 14.11.2019]. URL: <https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/Publikationen/publikationen-verbraucherpreisindex.html>.

682. Steiner JF. The use of stories in clinical research and health policy. *JAMA* 2005; 294(22): 2901-2904.
683. Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol* 2016; 16(1): 97.
684. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016; 355: i4919.
685. Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; 366: l4898.
686. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G et al. Preferred reporting items for systematic review and meta-analyses of individual participant data: the PRISMA-IPD statement. *JAMA* 2015; 313(16): 1657-1665.
687. Steyerberg EW, Moons KG, Van der Windt DA, Hayden JA, Perel P, Schroter S et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; 10(2): e1001381.
688. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med Decis Making* 1998; 18(2 Suppl): S68-S80.
689. Stollenwerk B, Lhachimi SK, Briggs A, Fenwick E, Caro JJ, Siebert U et al. Communicating the parameter uncertainty in the IQWiG efficiency frontier to decision-makers. *Health Econ* 2015; 24(4): 481-490.
690. Sturtz S, Bender R. Unsolved issues of mixed treatment comparison meta-analysis: network size and inconsistency. *Res Syn Meth* 2012; 3(4): 300-311.
691. Suarez-Almazor ME, Belseck E, Homik J, Dorgan M, Ramos-Remus C. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. *Control Clin Trials* 2000; 21(5): 476-487.
692. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996; 49(8): 907-916.
693. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010; 340: c117.
694. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* 2008; 26(9): 753-767.
695. Swart E, Bitzer EM, Gothe H, Harling M, Hoffmann F, Horenkamp-Sonntag D et al. Standardisierte Berichtsroutine für Sekundärdaten Analysen (STROSA): ein konsentierter Berichtsstandard für Deutschland; Version 2. *Gesundheitswesen* 2016; 78(Suppl 1): e145-e160.

696. Swart E, Ihle P, Gothe H, Matusiewicz D (Ed). Routinedaten im Gesundheitswesen: Handbuch Sekundärdatenanalyse; Grundlagen, Methoden und Perspektiven. Bern: Huber; 2014.
697. Swift TL, Dieppe PA. Using expert patients' narratives as an educational resource. *Patient Education and Counseling* 2005; 57(1): 115-121.
698. Tai FM, Willson ML, Ghera D. Implications of searching multiple trial registries: how should we search ClinicalTrials.gov and WHO ICTRP? [online]. In: 20th Cochrane Colloquium; 30.09.-03.10.2012; Auckland, Neuseeland. [Zugriff: 21.11.2019]. URL: <http://abstracts.cochrane.org/2012-auckland/implications-searching-multiple-trial-registries-how-should-we-search>.
699. Taichman D, Sahni P, Pinborg A, Peiperl L, Laine C, James A. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. *PLoS Med* 2017; 14(6): e1002315.
700. Tainio M, Tuomisto JT, Hänninen O, Ruuskanen J, Jantunen MJ, Pekkanen J. Parameter and model uncertainty in a life-table model for fine particles (PM_{2.5}): a statistical modeling study. *Environ Health* 2007; 6: 24.
701. Tajik P, Zwiderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. *Clin Cancer Res* 2013; 19(17): 4578-4588.
702. Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Stat Methods Med Res* 2017; 26(4): 1896-1911.
703. Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med* 2013; 158(7): 544-554.
704. Tan SS. Microcosting in economic evaluations: issues of accuracy, feasibility, consistency and generalisability [Dissertation]. Rotterdam: Erasmus Universität; 2009. URL: http://repub.eur.nl/res/pub/17354/091127_Tan,%20Siok%20Swan.pdf.
705. Tan SS, Bouwmans CAM, Rutten FFH, Hakkaart-van Roijen L. Update of the Dutch manual for costing in economic evaluations. *Int J Technol Assess Health Care* 2012; 28(2): 152-158.
706. Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R et al. Integrating qualitative research with trials in systematic reviews. *BMJ* 2004; 328(7446): 1010-1012.
707. Thomas S. Klinische Relevanz von Therapieeffekten: systematische Sichtung, Klassifizierung und Bewertung methodischer Konzepte [Dissertation]. Duisburg/Essen: Universität; 2009.
708. Thompson JC, Quigley JM, Halfpenny NJ, Scott DA, Hawkins NS. Importance and methods of searching for E-publications ahead of print in systematic reviews. *Evid Based Med* 2016; 21(2): 55-59.

709. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002; 21(11): 1559-1573.
710. 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.
711. Thurow S. Search engine visibility. Berkley: New Riders; 2008.
712. Tian L, Cai T, Pfeiffer MA, Piankov N, Cremieux PY, Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2×2 tables with all available data but without artificial continuity correction. *Biostatistics* 2009; 10(2): 275-281.
713. Tierney JF, Stewart LA, Gherzi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007; 8: 16.
714. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *BMJ* 2001; 322(7282): 355-357.
715. Torrance GW, Siegel JE, Luce BR, Gold MR, Russell LB, Weinstein MC. Framing and designing the cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). *Cost-effectiveness in health and medicine*. New York: Oxford University Press; 1996. S. 54-81.
716. Trevena LJ, Davey HM, Barratt A, Butow P, Caldwell P. A systematic review on communicating with patients about evidence. *J Eval Clin Pract* 2006; 12(1): 13-23.
717. Tricco AC, Langlois EV, Straus SE (Ed). *Rapid reviews to strengthen health policy and systems: a practical guide*. Genf: World Health Organization; 2017. URL: <https://apps.who.int/iris/bitstream/handle/10665/258698/9789241512763-eng.pdf;jsessionid=78C22F1C3E46FC746BEC184B66A79AF5?sequence=1>.
718. Trueman P, Drummond M, Hutton J. Developing guidance for budget impact analysis. *Pharmacoeconomics* 2001; 19(6): 609-621.
719. Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977; 198(4318): 679-684.
720. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003; 290(12): 1624-1632.
721. 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.
722. U.S. Food and Drug Administration. Devices@FDA [online]. [Zugriff: 21.11.2019]. URL: <http://www.accessdata.fda.gov/scripts/cdrh/devicesatfda/index.cfm>.
723. U.S. Food and Drug Administration. Drugs@FDA: FDA approved drug products [online]. [Zugriff: 21.11.2019]. URL: <https://www.accessdata.fda.gov/scripts/cder/daf/>.

724. UK National Screening Committee. Programme appraisal criteria: criteria for appraising the viability, effectiveness and appropriateness of a screening programme [online].

23.10.2015 [Zugriff: 21.11.2019]. URL:

<https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes/criteria-for-appraising-the-viability-effectiveness-and-appropriateness-of-a-screening-programme>.

725. Unkel S, Amiri M, Benda N, Beyersmann J, Knoerzer D, Kupas K et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. *Pharm Stat* 2019; 18(2): 166-183.

726. Van Aert RCM, Wicherts JM, Van Assen M. Publication bias examined in meta-analyses from psychology and medicine: a meta-meta-analysis. *PLoS One* 2019; 14(4): e0215052.

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

728. Van Tinteren H, Hoekstra OS, Boers M. Do we need randomised trials to evaluate diagnostic procedures? *European Journal of Nuclear Medicine and Molecular Imaging* 2004; 31(1): 129-131.

729. Van Tinteren H, Hoekstra OS, Smit EF, Van den Bergh JH, Schreurs AJ, Stallaert RA et al. Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: the PLUS multicentre randomised trial. *Lancet* 2002; 359(9315): 1388-1393.

730. Van Tulder M, Furlan A, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine (Phila Pa 1976)* 2003; 28(12): 1290-1299.

731. Vandenbroucke JP, Von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Ann Intern Med* 2007; 147(8): W163-W194.

732. Veerman JL, Mackenbach JP, Barendregt JJ. Validity of predictions in health impact assessment. *J Epidemiol Community Health* 2007; 61(4): 362-366.

733. Veroniki AA, Jackson D, Bender R, Kuss O, Langan D, Higgins JPT et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Res Syn Meth* 2019; 10(1): 23-43.

734. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Syn Meth* 2016; 7(1): 55-79.

735. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Knapp G, Kuss O et al. Recommendations for quantifying the uncertainty in the summary intervention effect and estimating the between-study heterogeneity variance in random-effects meta-analysis. *Cochrane Database Syst Rev* 2015; (Suppl 1: Cochrane Methods): 25-27.
736. Vidanapathirana J, Abramson MJ, Forbes A, Fairley C. Mass media interventions for promoting HIV testing. *Cochrane Database Syst Rev* 2005; (3): CD004775.
737. Vijan S. Should we abandon QALYs as a resource allocation tool? *Pharmacoeconomics* 2006; 24(10): 953-954.
738. 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.
739. Virtanen H, Leino-Kilpi H, Salantera S. Empowering discourse in patient education. *Patient Education and Counseling* 2007; 66(2): 140-146.
740. Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters M et al. Assessing the risk of bias of individual studies in systematic reviews of health care interventions: Agency for Healthcare Research and Quality methods guide for comparative effectiveness reviews [online]. 03.2012 [Zugriff: 18.03.2015]. URL: https://www.ncbi.nlm.nih.gov/books/NBK91433/pdf/Bookshelf_NBK91433.pdf.
741. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. *Int J Qual Health Care* 2005; 17(3): 235-242.
742. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007; 335(7624): 806-808.
743. Waffenschmidt S, Janzen T, Hausner E, Kaiser T. Simple search techniques in PubMed are potentially suitable for evaluating the completeness of systematic reviews. *J Clin Epidemiol* 2013; 66(6): 660-665.
744. Waffenschmidt S, Navarro-Ruan T, Hobson N, Hausner E, Sauerland S, Haynes RB. Development and validation of study filters for identifying controlled non-randomized studies in PubMed and Ovid MEDLINE. *Res Syn Meth* 2020; 11(5): 617-626.
745. Wallace BC, Noel-Storr A, Marshall IJ, Cohen AM, Smalheiser NR, Thomas J. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. *J Am Med Inform Assoc* 2017; 24(6): 1165-1168.
746. Walter U, Schwartz FW. Prävention. In: Schwartz FW, Badura B, Busse R, Leidl R, Raspe H, Siegrist J et al (Ed). *Das Public Health Buch: Gesundheit und Gesundheitswesen*. München: Urban und Fischer; 2003. S. 189-214.

747. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C et al. Principles of good practice for decision analytic modeling in health care evaluation: report of the ISPOR Task Force on Good Research Practices; modeling studies. *Value Health* 2003; 6(1): 9-17.
748. Weinstein MC, Siegel JE, Garber AM, Lipscomb J, Luce BR, Manning WG et al. Productivity costs, time costs and health-related quality of life: a response to the Erasmus Group. *Health Econ* 1997; 6(5): 505-510.
749. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat Med* 2006; 25(2): 183-203.
750. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF et al. Systems to rate the strength of scientific evidence: AHRQ publication no. 02-E016 [online]. 03.2002 [Zugriff: 21.11.2019]. (Evidence Report/Technology Assessment (Summaries); Band 47). URL: <https://archive.ahrq.gov/clinic/epcsums/strengthsum.pdf>.
751. 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.
752. Wetzel H. Mindestmengen zur Qualitätssicherung: Konzeptionelle und methodische Überlegungen zur Festlegung und Evaluation von Fallzahlgrenzwerten für die klinische Versorgung. *Z ärztl Fortbild Qual Gesundheitswes* 2006; 100(2): 99-106.
753. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Ann Intern Med* 2004; 140(3): 189-202.
754. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. *J Clin Epidemiol* 2016; 69: 225-234.
755. Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J Clin Epidemiol* 2013; 66(10): 1093-1104.
756. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 155(8): 529-536.
757. Whitlock EP, Lin JS, Chou R, Shekelle P, Robinson KA. Using existing systematic reviews in complex systematic reviews. *Ann Intern Med* 2008; 148(10): 776-782.
758. Widrig D, Tag B. HTA and its legal issues: a framework for identifying legal issues in health technology assessment. *Int J Technol Assess Health Care* 2014; 30(6): 587-594.
759. Wieseler B, Kerekes MF, Vervoelgyi V, McGauran N, Kaiser T. Impact of document type on reporting quality of clinical drug trials: a comparison of registry reports, clinical study reports, and journal publications. *BMJ* 2012; 344: d8141.

760. Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Stat Med* 2016; 35: 2503-2515.
761. Windeler J. Bedeutung randomisierter klinischer Studien mit relevanten Endpunkten für die Nutzenbewertung. In: Begriffsdefinitionen und Einführung: Dokumentation des ersten gemeinsamen Workshops von GFR und IQWiG am 4. September 2007 in Berlin. Bonn: Gesundheitsforschungsrat (GFR) des Bundesministeriums für Bildung und Forschung; 2007. S. 26-31. URL: https://www.iqwig.de/download/DLR_Nutzenbewert_07-11-22_Druckversion.pdf.
762. Windeler J. Externe Validität. *Z Evid Fortbild Qual Gesundhwes* 2008; 102(4): 253-259.
763. Windeler J, Conradt C. Wie können "Signifikanz" und "Relevanz" verbunden werden? *Medizinische Klinik* 1999; 94(11): 648-651.
764. Windeler J, Lange S. Nutzenbewertung in besonderen Situationen: seltene Erkrankungen. *Z Evid Fortbild Qual Gesundhwes* 2008; 102(1): 25-30.
765. Windeler J, Ziegler S. Evidenzklassifizierungen. *Z Arztl Fortbild Qualitatssich* 2003; 97(6): 513-514.
766. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; 170(1): 51-58.
767. Wong SS, Wilczynski NL, Haynes RB. Optimal CINAHL search strategies for identifying therapy studies and review articles. *J Nurs Scholarsh* 2006; 38(2): 194-199.
768. Wong SSL, Wilczynski NL, Haynes RB. Comparison of top-performing search strategies for detecting clinically sound treatment studies and systematic reviews in MEDLINE and EMBASE. *J Med Libr Assoc* 2006; 94(4): 451-455.
769. Wong SSL, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. *J Med Libr Assoc* 2006; 94(1): 41-47.
770. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008; 336(7644): 601-605.
771. Yang F, Wittes J, Pitt B. Beware of on-treatment safety analyses. *Clin Trials* 2019; 16(1): 63-70.
772. Zarin DA, Tse T, Williams RJ, Carr S. Trial reporting in ClinicalTrials.gov: the final rule. *N Engl J Med* 2016; 375(20): 1998-2004.
773. Ziebland S, McPherson A. Making sense of qualitative data analysis: an introduction with illustrations from DIPEX (personal experiences of health and illness). *Med Educ* 2006; 40(5): 405-414.

774. Ziegler DK, Mosier MC, Buenaver M, Okuyemi K. How much information about adverse effects of medication do patients want from physicians? Arch Intern Med 2001; 161(5): 706-713.

775. Zorzela L, Loke YK, Ioannidis JP, Golder S, Santaguida P, Altman DG et al. PRISMA harms checklist: improving harms reporting in systematic reviews. BMJ 2016; 352: i157.

776. Zschorlich B, Knelangen M, Bastian H. Die Entwicklung von Gesundheitsinformationen unter Beteiligung von Bürgerinnen und Bürgern am Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Gesundheitswesen 2011; 73(7): 423-429.

777. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008; 337: a2390.