medizinwissen

HTA-Bericht 102

# Vergleich von Bewertungsinstrumenten für die Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum

**Maren Dreier, Birgit Borutta, Jona Stahmeyer, Christian Krauth, Ulla Walter**

Im Geschäftsbereich des Bundesministerium für Gesundheit

DIMDI deutsche agentur für HTA des Deutschen Instituts für Medizinische Dokumentation und Information

**Schriftenreihe**
**Health Technology Assessment (HTA)**
**In der Bundesrepublik Deutschland**

# Vergleich von Bewertungsinstrumenten für die Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum

**Maren Dreier, Birgit Borutta, Jona Stahmeyer, Christian Krauth, Ulla Walter**

Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung,
Medizinische Hochschule Hannover

## Wir bitten um Beachtung

Dieser HTA-Bericht ist publiziert in der DAHTA-Datenbank des DIMDI (www.dimdi.de – HTA) und in der elektronischen Zeitschrift GMS Health Technology Assessment (www.egms.de).

Die HTA-Berichte des DIMDI durchlaufen ein unabhängiges, grundsätzlich anonymisiertes Gutachterverfahren. Potentielle Interessenkonflikte bezüglich der HTA-Berichte werden dem DIMDI von den Autoren und den Gutachtern offengelegt. Die Literaturauswahl erfolgt nach den Kriterien der evidenzbasierten Medizin. Die durchgeführte Literaturrecherche erhebt keinen Anspruch auf Vollständigkeit. Die Verantwortung für den Inhalt des Berichts obliegt den jeweiligen Autoren.

Die Erstellung des vorliegenden HTA-Berichts des Deutschen Instituts für Medizinische Dokumentation und Information (DIMDI) erfolgte gemäß gesetzlichem Auftrag nach Artikel 19 des GKV-Gesundheitsreformgesetzes 2000. Das Thema stammt aus dem öffentlichen Vorschlagsverfahren beim DIMDI, durch das Kuratorium HTA priorisiert und vom DIMDI beauftragt. Der Bericht wurde mit Mitteln des Bundes finanziert.

---

**Herausgegeben vom**
**Deutschen Institut für Medizinische Dokumentation und Information (DIMDI), Köln**

Das DIMDI ist ein Institut im Geschäftsbereich des Bundesministeriums für Gesundheit (BMG)

**Kontakt**

DAHTA
Deutsche Agentur für Health Technology Assessment des
Deutschen Instituts für Medizinische Dokumentation und Information
Waisenhausgasse 36-38a
50676 Köln

Tel: +49 221 4724-525
Fax: +49 2214724-340

E-Mail: dahta@dimdi.de
www.dimdi.de

Schriftenreihe Health Technology Assessment, Bd. 102
ISSN: 1864-9645
1. Auflage 2010
DOI: 10.3205/hta000085L
URN: urn:nbn:de:0183-hta000085L1

# Inhaltsverzeichnis

# 1 Verzeichnisse

## Tabellenverzeichnis

# Abbildungsverzeichnis



# Abkürzungsverzeichnis

| | |
|---|---|
| AHRQ | Agency for Healthcare Research and Quality |
| AMED | Allied and Complementary Medicine Database |
| AMSTAR | Assessment of multiple systematic reviews |
| ANAES | Agence Nationale d'Accréditation et d'Evaluation en Santé |
| ARIF | Aggressive Research Intelligence Facility |
| BIOSIS | BIOSIS Datenbank |
| BMG | Bundesministerium für Gesundheit |
| BMJ | British Medical Journal |
| BQ | Berichtsqualität |
| CAB | CAB Datenbank |
| CADTH | Canadian Agency for Drugs and Technologies in Health |
| CASP | Critical Appraisal Skills Programme |
| CBA | Kosten-Nutzen-Analyse (engl. Cost-benefit analysis) |
| CCMed | Current Contents Medizin |
| CCOHTA | Canadian Coordinating Office for Health Technology Assessment |
| CDC | Center for Disease Control and Prevention |

**Abkürzungsverzeichnis – Fortsetzung**

| | |
|---|---|
| CDSR | Cochrane Database of Systematic Reviews |
| CEA | Kosten-Effektivitäts-Analyse (engl. cost-effectiveness analysis) |
| CEBM | Centre for Evidence-based Medicine |
| CEBMH | Centre for Evidence-based Mental Health |
| CINAHL | Cumulative Index to Nursing and Allied Health Literature |
| CL | Checkliste |
| CONSORT | Consolidated Standards of Reporting Trials |
| CRD | Centre for Reviews and Dissemination |
| CUA | Kosten-Nutzwert-Analyse (engl. cost-utility analysis) |
| DAHTA | Deutsche Agentur für Health Technology Assessment |
| DALY | Disability-adjusted life years (Behinderungskorrigierte Lebensjahre) |
| DARE | Database of Abstracts of Reviews of Effects |
| DE | Deutsch |
| DIMDI | Deutsches Institut für Medizinische Dokumentation und Information |
| EbM | Evidenzbasierte Medizin |
| ECHTA | European Collaboration for Assessment of Health Interventions |
| EMBASE | Excerpta Medica Database |
| EN | Englisch |
| EPHPP | Effective Public Health Practice Project |
| EQ-5D | EuroQol-Instrument der präferenzbasierten Lebensqualitätsmessung |
| EQUATDUR-2 | Edmonton Quality Assessment Tool for Drug Utilization Reviews |
| ETHMED | Datenbank Ethik in der Medizin |
| EUnetHTA | European network for Health Technology Assessment |
| EV | Externe Validität |
| FKS | Formular für Fall-Kontrollstudien |
| G-BA | Gemeinsamer Bundesausschuss |
| GKV | Gesetzliche Krankenversicherung |
| GRADE | Grading of Recommendations Assessment, Development and Evaluation |
| GSWG-TAHC | German Scientific Working Group Technology Assessment in Health Care |
| HTA | Health Technology Assessment |
| HTAI | Health Technology Assessment International |
| ICC | Intraclass-Correlation |
| IMSIE | Institut für Medizinische Statistik, Informatik und Epidemiologie der Universität zu Köln |
| INAHTA | International Network of Agencies for Health Technology Assessment |
| IPA | International Police Association |
| IQWiG | Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen |
| ISEG | Institut für Sozialmedizin, Epidemiologie und Gesundheitssystemforschung |
| IV | Interne Validität |
| $\kappa$ | Kappa |
| K. A. | Keine Angabe |
| KO | Komponentensystem |
| KS | Formular für Kohortenstudien |
| LBI | Ludwig Boltzmann Institut |
| LQ | Lebensqualität |
| MDK | Medizinischer Dienst der Krankenkassen |
| MEDIKAT | Kataloge der Deutschen Zentralbibliothek für Medizin |
| MEDLINE | Medical Literature Analysis and Retrieval System Online |
| MINORS | Methodological index für non-randomized studies |
| MOOSE | Meta-Analysis of Observational Studies in Epidemiology |

**Abkürzungsverzeichnis – Fortsetzung**

| | |
|---|---|
| MQAS | Methodological Quality Assessment Score |
| NCCHTA | National Coordinating Centre for Health Technology Assessment |
| NHS | National Health Service |
| NHS-CRD-DARE | National Health Service Database of Abstracts of Reviews of Effects |
| NHS-CRD-HTA | International Network of Agencies for Health Technology Assessment |
| NHSEED | National Health Service Economic Evaluation Database |
| NOS | Newcastle Ottawa Scale |
| OPVS | Oxford Pain Validity Scale |
| PEDro | Physiotherapy Evidence Database Scale |
| PHRU | Public Health Research Unit |
| PICO | Patient/Problem; Intervention, Comparison/Control, Outcome/Effects |
| PMA | Pharmaceutical Management Agency |
| PQAQ | Pediatric Quality Appraisal Questionnaire |
| PRISMA | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| PsycLit | Ehemalige Bezeichnung der Datenbank PsycINFO |
| QATSO | Quality assessment tool for systematic reviews of observational studies |
| QB | Qualitätsbewertung |
| QBI | Qualitätsberwertungsinstrument |
| QKB | Qualitative Komponentenbewertung |
| QGB | Qualitative Gesamtbewertung |
| QQAQ | Overview Quality Assessment Questionnaire |
| QUADAS | Quality assessment of diagnostic accuracy studies |
| QALY | Qualitätsadjustierte Lebensjahre |
| QUORUM | The Quality of Reporting of Meta-Analyses |
| RATS | Qualitative Research Review Guidelines |
| RCT | Randomisierte kontrollierte Studie (engl. Randomized controlled/clinical trial) |
| SciSearch | Science Citation Index Expanded |
| SIGN | Scottish Intercollegiate Guidelines Network |
| SK | Skala |
| SOMED | Sozialmedizin |
| SR | Systematischer Review |
| STARD | Standards for Reporting of Diagnostic Accuracy |
| STREGA | Strengthening the Reporting of Genetic Association Studien |
| STROBE | Strengthening the Reporting of Observational Studies in Epidemiology |
| TAPS | Trial Assessment Procedure Scale |
| TREND | Transparent Reporting of Evaluations with Nonrandomized Designs |

# Glossar

| | |
|---|---|
| Beobachtungsstudie | In Beobachtungsstudien wird die Exposition nicht wie in experimentellen Studien dem Studienteilnehmer zugeteilt. Zu den Beobachtungsstudien zählen Fall-Kontroll-, Kohorten-, ökologische und Querschnittstudien. |
| Berichtsqualität (BQ) | Die Vollständigkeit von Informationen zu den Aspekten Design, Durchführung und Auswertung von Studien wird geprüft, ohne dass der Inhalt dieser Informationen im Hinblick auf die Validität beurteilt wird. |
| Bias | Ein systematischer Fehler, der bei der Auswahl der Studienpopulation oder bei der Durchführung einer Studie auftritt und zu einer systematischen Unter- oder Überschätzung des Zusammenhangs von Exposition/Intervention und Outcome führt. |
| Cohens Kappa | Statistisches Maß der Interrater-Reliabilität, der Übereinstimmung von meist zwei Beurteilern nach Jacob Cohen. |

**Glossar – Fortsetzung**

| | |
|---|---|
| Confounding | Durch Confounding wird der tatsächliche Effekt einer Exposition/ Intervention auf das Outcome systematisch unter- oder überschätzt. Confounding entsteht dadurch, dass eine Variable (Confounder) gleichzeitig mit der Exposition/Intervention assoziiert und Prädiktor für das Outcome ist, ohne Teil der Kausalkette zwischen Exposition/Intervention und Outcome zu sein. Dadurch kann die Verteilung der Ausprägung der Variable (Confounder) in den Studiengruppen unterschiedlich sein. |
| Experimentelle Studie | Eine experimentelle Studie ist dadurch definiert, dass die Untersucher im Gegensatz zu einer Beobachtungsstudie die Exposition bzw. Intervention den Studienteilnehmern zuweisen. |
| Fall-Kontrollstudie | In einer Fall-Kontrollstudie werden die Studienteilnehmer nach ihrem Outcome (Fälle bzw. Kontrollen) ausgewählt und die vorausgegangene Exposition erhoben |
| Follow-up | Nachuntersuchung. |
| Friktionskostenansatz | Ansatz für die Bewertung des Produktionsausfalls aus gesellschaftlicher Perspektive. Impliziert, dass bei Erwerbsunfähigkeit und vorzeitigem Tod der Verlust der zukünftigen Arbeitseinkommen bis zur Kompensierung durch Arbeitslose berücksichtigt wird. |
| Humankapitalansatz | Ansatz für die Bewertung des Produktionsausfalls aus gesellschaftlicher Perspektive. Impliziert, dass bei Erwerbsunfähigkeit und vorzeitigem Tod der Verlust der gesamten zukünftigen Arbeitseinkommen bis zum durchschnittlichen Renteneintrittsalter berücksichtigt wird. |
| Inhaltsvalidität | Inhaltliche Analyse dessen, was das Instrument misst, z. B. unter Berücksichtigung von Herstellungsprozess, Definition des Items, Expertenbefragung. |
| Inkrementalanalyse | Bei der Inkrementalanalyse werden die zusätzlichen Kosten und Outcomes erhoben, die mit unterschiedlichen Behandlungen verbunden sind. Der inkrementelle Kosten-Effektivitäts-Quotient wird bestimmt, indem man für beide Therapien die Kostendifferenzen (C2 – C1) durch die Ergebnisdifferenzen (E2 – E1) dividiert. |
| Interrater-Reliabilität | Vergleich der Übereinstimmung der Ergebnisse von zwei Personen bei Anwendung eines gleichen Tests/Verfahrens. Der Grad der Übereinstimmung kann mit der Kappa-Statistik als Kappa-Koeffizient oder als Korrelationskoeffizient angegeben werden. |
| Intention-to-treat-Analyse | Analyse der Studienergebnisse je nach der zu Studienbeginn zugewiesenen Therapieformen. |
| Interventionsstudie | Eine Studie, bei der die Wirksamkeit einer Intervention untersucht wird. |
| Kappa-Wert | Ein Maß für die Übereinstimmung (Interrater-Reliabilität) von Bewertungen oder Einschätzungen durch zwei oder mehr Beurteiler (Rater). S. auch Cohens Kappa. |
| Kaufkraftparität | Die Kaufkraftparität bietet die Möglichkeit eines intervalutarischen Vergleichs verschiedener Länder oder Wirtschaftsräume. |
| Kohortenstudie | In einer Kohortenstudie werden die Studienteilnehmer nach ihrer Exposition in die Studiengruppen eingeteilt und über einen bestimmten Zeitraum (Follow-up) beobachtet. Es wird geprüft, ob das interessierende Outcome eingetreten ist. Es werden prospektive und retrospektive Kohortenstudien unterschieden. |
| Konstruktvalidität | Grad der Übereinstimmung mit einem gleichen Konstrukt, z. B. Vergleich mit einem anderen Instrument. Angabe eines Korrelationskoeffizienten. |
| Kosten, direkte | In der Gesundheitsökonomie wird mit direkten Kosten der Ressourcenverzehr bezeichnet, der unmittelbar mit bestimmten medizinischen Leistungen verbunden ist und direkt zugeordnet werden kann. |
| Kosten, indirekte | Indirekte Kosten bezeichnen den volkswirtschaftlichen Produktivitätsverlust aufgrund von krankheitsbedingter Abwesenheit von Arbeitsplatz, verminderter Leistungsfähigkeit oder vorzeitigem Tod eines Erwerbstätigen. |

**Glossar – Fortsetzung**

| | |
|---|---|
| Kosten-Effektivitäts-Analyse (CEA) | Die Kosten-Effektivitäts-Analyse ist eine Methode zum Vergleich alternativ möglicher Behandlungsweisen, wobei die Behandlungsergebnisse (Outcomes) in der gleichen nicht-monetären (natürlichen) Einheit angegeben werden. |
| Kosten-Nutzen-Analyse (CBA) | Die Kosten-Nutzen-Analyse ist eine Möglichkeit zum Vergleich alternativer Behandlungsmöglichkeiten, wobei sowohl die Kosten, als auch Behandlungsergebnisse (Outcomes) in monetären Einheiten angegeben werden. |
| Kosten-Nutzwert-Analyse (CUA) | Die Kosten-Nutzwert-Analyse ist eine Form der gesundheits-ökonomischen Evaluation, bei der ein Vergleich alternativ möglicher Behandlungsweisen durch die Ermittlung von Nutzengrößen (z. B. qualitätsadjustierte Lebensjahre) erfolgt. |
| Kriteriumsvalidität | Grad der Übereinstimmung mit einem unabhängigen Verfahren, möglichst einem sogenannten Goldstandard. Angabe eines Korrelationskoeffizienten. |
| Lebensqualität (LQ) | Gesundheitsbezogene Lebensqualität ist ein mehrdimensionales, durch die subjektive Sichtweise des Befragten geprägtes Konstrukt, das in medizinischen Interventions-, in epidemiologischen Studien und zunehmend auch bei ökonomischen Evaluationen verwendet wird. |
| Loss-to-follow-up | Studienteilnehmer ohne Daten zu Nachuntersuchungen. |
| Matching | Beim Matching werden die Studienteilnehmer der einen Studiengruppe nach bestimmten Charakteristika der anderen Studiengruppe ausgewählt mit dem Ziel der Gleichverteilung bestimmter Merkmale in den Studiengruppen. |
| Metaanalyse | Eine statistische Methode zur Präzision des Effektschätzers aus den Ergebnissen mehrerer Einzelstudien. |
| Metaepidemiologie | In der Metaepidemiologie werden metaepidemiologische Studien auf der Basis einer Sammlung von Metaanalysen durchgeführt. In den Primärstudien der Metaanalysen wird die Assoziation bestimmter Studiencharakteristika (z. B. verdeckte Zuordnung der Intervention, Ausschluss von Patienten aus der Analyse) mit der Höhe der Effekte untersucht. Unterschiedliche Effekte in Abhängigkeit von den untersuchten Charakteristika weisen auf Bias infolge der entsprechenden Studiencharakteristik hin. |
| Modellierung | Modelle sind vereinfachte Abbilder der Wirklichkeit. Unter Modellieren wird die Vereinfachung der Realität auf eine Stufe verstanden, die die wesentlichen Konsequenzen und Komplikationen verschiedener Optionen für die Entscheidungsfindung beschreibt. |
| Multivariate Analysen | Es wird in einem Datensatz der gleichzeitige Einfluss mehrerer Variablen auf ein bestimmtes Outcome mittels Regressionsanalyse untersucht. Mittel zur Kontrolle von Confounding. |
| Opportunitätskosten | Opportunitätskosten definieren den Wert knapper Ressourcen in der Produktion von Gesundheitsinterventionen und entsprechen dem entgangenen Nutzen der eingesetzten Ressourcen in der nächstbesten Verwendungsalternative. |
| Outcome | Zielgröße einer Studie, auch Endpunkt genannt. |
| Propensity Score | Statistische Methode zur Minimierung des Selektionsbias durch unterschiedliche Patientencharakteristika in den Studiengruppen. |
| Qualitätsadjustierte Lebensjahre (QALY) | Ein QALY ist rechnerisch ein zusätzliches Lebensjahr in optimaler Gesundheit. Analog der Kosten-Effektivitäts-Relation ist aus ökonomischer Perspektive die Alternative mit den geringsten Kosten je QALY zu präferieren. |
| Randomisierte kontrollierte Studie (RCT) | Eine experimentelle Studie mit zufälliger Zuordnung (Randomisierung) der Patienten in die Behandlungsgruppen. |
| Randomisierung | Zuordnung der Patienten in die Behandlungsgruppen nach dem Zufallsprinzip. |

**Glossar – Fortsetzung**

| | |
|---|---|
| Residual Confounding | Rest-Confounding durch unbekannte oder nicht berücksichtigte Störfaktoren. |
| Restriktion | Beschränkung auf eine Ausprägung einer Variablen, z. B. Restriktion auf eine Altersgruppe oder männliches Geschlecht. |
| Sensitivität | Der Anteil der tatsächlich Erkrankten, der auch als krank erkannt wird; die Fähigkeit eines Tests, Kranke zu erkennen. |
| Sensitivitätsanalyse | Sensitivitätsanalysen dienen der Überprüfung von Studienergebnissen auf ihre Abhängigkeit von den getroffenen Annahmen und anderen Studienparametern, die als situationsabhängig, unsicher bzw. veränderlich eingeschätzt werden. |
| Spezifität | Der Anteil der tatsächlich Gesunden, der auch als gesund erkannt wird; die Fähigkeit eines Tests, Gesunde zu erkennen. |
| Standard gamble | Methode zur indirekten Bestimmung von Patientenpräferenzen für Gesundheitszustände zur Bestimmung von qualitätsadjustierten Lebensjahren (QALY). |
| Stratifizieren | Die Daten werden getrennt nach den Kategorien einer Variable betrachtet. |
| Time-trade-off | Verfahren zur indirekten Ermittlung von Präferenzen von Personen für Gesundheitszustände zur Bestimmung von qualitätsadjustierten Lebensjahren (QALY). |
| Triplet | Beim Matching von Fällen und Kontrollen im Verhältnis 1 : 2 entsteht ein Triplet bestehend aus einem Fall und zwei Kontrollen. |
| Validität | Es wird zwischen externer und interner Validität unterschieden.<br>Interne Validität: die Glaubwürdigkeit von Studienergebnissen unter Berücksichtigung von möglichen systematischen Fehlern oder Verzerrungen durch das Design, die Durchführung oder Auswertung einer Studie.<br>Externe Validität: Die Generalisierbarkeit der Studienergebnisse auf eine definierte Population; hängt von der gewählten Population bzw. vom Kontext ab, auf die bzw. den die Ergebnisse übertragen werden sollen. |

# 2 Zusammenfassung

**Gesundheitspolitischer Hintergrund**

Erkenntnisse aus wissenschaftlichen Studien bilden die Grundlage für evidenzbasierte gesundheitspolitische Entscheidungen.

**Wissenschaftlicher Hintergrund**

Zur Einschätzung der Glaubwürdigkeit von Studien sind Qualitätsbewertungen von Studien immanenter Bestandteil von HTA-Berichten (HTA = Health Technology Assessment) und systematischen Übersichtsarbeiten. Diese prüfen, inwieweit die Studienergebnisse systematisch durch Confounding oder Bias verzerrt sein können (interne Validität). Es werden Checklisten, Skalen und Komponentenbewertungen unterschieden.

**Forschungsfragen**

Welche Instrumente zur Qualitätsbewertung von systematischen Reviews, Interventions-, Beobachtungs-, Diagnose- und gesundheitsökonomischen Studien gibt es, wie unterscheiden sich diese und welche Schlussfolgerungen lassen sich daraus für die Qualitätsbewertung ableiten?

**Methodik**

Es wird eine systematische Recherche in einschlägigen Datenbanken ab 1988 durchgeführt, ergänzt um eine Durchsicht der Referenzen, der HTA-Berichte der Deutschen Agentur für Health Technology Assessment (DAHTA) sowie eine Internetrecherche. Die Literaturauswahl, die Datenextraktion und die Qualitätsbewertung werden von zwei unabhängigen Reviewern vorgenommen. Die inhaltlichen Elemente der Qualitätsbewertungsinstrumente (QBI) werden mit modifizierten Kriterienlisten, bestehend aus Items und Domänen spezifisch für randomisierte, Beobachtungs-, Diagnosestudien, systematische Übersichtsarbeiten und gesundheitsökonomische Studien extrahiert. Anhand der Anzahl abgedeckter Items und Domänen werden umfassendere von weniger umfassenden Instrumenten unterschieden. Zwecks Erfahrungsaustausch zu Problemen bei der praktischen Anwendung von Instrumenten wird ein Workshop durchgeführt.

**Ergebnisse**

Es werden insgesamt acht systematische, methodische Reviews und HTA-Berichte sowie 147 Instrumente identifiziert: 15 für systematische Übersichtsarbeiten, 80 für randomisierte Studien, 30 für Beobachtungs-, 17 für Diagnose- und 22 für gesundheitsökonomische Studien. Die Instrumente variieren deutlich hinsichtlich der Inhalte, deren Ausprägung und der Güte der Operationalisierung. Einige Instrumente enthalten neben Items zur internen Validität auch Items zur Berichtsqualität und zur externen Validität. Kein Instrument deckt alle abgefragten Kriterien ab. Designspezifisch werden generische Instrumente dargestellt, die die meisten inhaltlichen Kriterien erfüllen.

**Diskussion**

Die Bewertung von QBI anhand inhaltlicher Kriterien ist schwierig, da kein wissenschaftlicher Konsens über notwendige Elemente der internen Validität bzw. nur für einen Teil der allgemein akzeptierten Elemente ein empirischer Nachweis besteht. Der Vergleich anhand inhaltlicher Parameter vernachlässigt die Operationalisierung der einzelnen Items, deren Güte und Präzision wichtig für Transparenz, Replizierbarkeit, die korrekte Bewertung sowie die Interrater-Reliabilität ist. QBI, die Items zur Berichtsqualität und zur internen Validität vermischen, sind zu vermeiden.

**Schlussfolgerungen**

Es stehen unterschiedliche, designspezifische Instrumente zur Verfügung, die aufgrund ihrer umfassenderen inhaltlichen Abdeckung von Elementen der internen Validität bevorzugt zur Qualitätsbewertung eingesetzt werden können. Zur Minimierung der Subjektivität der Bewertung sind Instrumente mit einer ausführlichen und präzisen Operationalisierung der einzelnen Elemente anzuwenden. Für gesundheitsökonomische Studien sollten Instrumente mit Ausfüllhinweisen entwickelt werden, die die Angemessenheit der Kriterien definieren. Weitere Forschung ist erforderlich, um Studiencharakteristika zu identifizieren, die die interne Validität von Studien beeinflussen.

# 3 Abstract

**Health care policy background**

Findings from scientific studies form the basis for evidence-based health policy decisions.

**Scientific background**

Quality assessments to evaluate the credibility of study results are an essential part of health technology assessment reports and systematic reviews. Quality assessment tools (QAT) for assessing the study quality examine to what extent study results are systematically distorted by confounding or bias (internal validity). The tools can be divided into checklists, scales and component ratings.

**Research questions**

What QAT are available to assess the quality of interventional studies or studies in the field of health economics, how do they differ from each other and what conclusions can be drawn from these results for quality assessments?

**Methods**

A systematic search of relevant databases from 1988 onwards is done, supplemented by screening of the references, of the HTA reports of the German Agency for Health Technology Assessment (DAHTA) and an internet search. The selection of relevant literature, the data extraction and the quality assessment are carried out by two independent reviewers. The substantive elements of the QAT are extracted using a modified criteria list consisting of items and domains specific to randomized trials, observational studies, diagnostic studies, systematic reviews and health economic studies. Based on the number of covered items and domains, more and less comprehensive QAT are distinguished. In order to exchange experiences regarding problems in the practical application of tools, a workshop is hosted.

**Results**

A total of eight systematic methodological reviews is identified as well as 147 QAT: 15 for systematic reviews, 80 for randomized trials, 30 for observational studies, 17 for diagnostic studies and 22 for health economic studies. The tools vary considerably with regard to the content, the performance and quality of operationalisation. Some tools do not only include the items of internal validity but also the items of quality of reporting and external validity. No tool covers all elements or domains. Design-specific generic tools are presented, which cover most of the content criteria.

**Discussion**

The evaluation of QAT by using content criteria is difficult, because there is no scientific consensus on the necessary elements of internal validity, and not all of the generally accepted elements are based on empirical evidence. Comparing QAT with regard to contents neglects the operationalisation of the respective parameters, for which the quality and precision are important for transparency, replicability, the correct assessment and interrater reliability. QAT, which mix items on the quality of reporting and internal validity, should be avoided.

**Conclusions**

There are different, design-specific tools available which can be preferred for quality assessment, because of its wider coverage of substantive elements of internal validity. To minimise the subjectivity of the assessment, tools with a detailed and precise operationalisation of the individual elements should be applied. For health economic studies, tools should be developed and complemented with instructions, which define the appropriateness of the criteria. Further research is needed to identify study characteristics that influence the internal validity of studies.

# 4 Kurzfassung

## 4.1 Gesundheitspolitischer Hintergrund

Gesundheitspolitische Entscheidungen sollen evidenzbasiert auf der Grundlage von wissenschaftlichen Erkenntnissen getroffen werden. Evidenz basiert auf der Synthese von Studienergebnissen, die möglichst unverzerrt sind und damit eine hohe Glaubwürdigkeit aufweisen.

## 4.2 Wissenschaftlicher Hintergrund

Zur Einschätzung der Glaubwürdigkeit von Studien sind Qualitätsbewertungen immanenter Bestandteil von HTA-Berichten (HTA = Health Technology Assessment) und systematischen Übersichtsarbeiten. Diese prüfen, inwieweit die Studienergebnisse systematisch durch Confounding oder Bias verzerrt sein können (interne Validität).

Es gibt keinen Goldstandard für die Bewertung der Studienqualität, da die wahren Zusammenhänge von Exposition/Intervention und Outcome unbekannt sind. Die eingesetzten Instrumente können als Skalen, Checklisten und Komponentenbewertungen klassifiziert werden. Bei einer Skala erhält jedes Item eine numerische Bewertung, die zu einem Summenscore addiert wird. Skalen werden nicht mehr empfohlen, da sie die Höhe der Validität nicht korrekt abbilden. Eine Checkliste besteht aus mindestens zwei Items ohne numerisches Bewertungssystem. Die Komponentenbewertung enthält als Items Komponenten wie „Randomisierung“ und „Verblindung“, die ebenfalls nicht numerisch, sondern qualitativ bewertet werden. Von der methodischen Qualität, die in diesem Bericht synonym zum Begriff Studienqualität verwendet wird, muss die Berichtsqualität abgegrenzt werden, die nicht Bestandteil dieses Berichts ist.

Die Qualität gesundheitsökonomischer Studien wird bestimmt durch (a) die Validität der Studienergebnisse, (b) die Einhaltung methodischer Standards der gesundheitsökonomischen Evaluation und (c) den Zugang zu belastbaren Kosten- und Outcomedaten. Die methodischen Standards der gesundheitsökonomischen Evaluation sind in Standardlehrbüchern und gesundheitsökonomischen Leitlinien beschrieben. Gesundheitsökonomische Evaluation basiert auf den theoretischen Konzepten der Wohlfahrtsökonomik und Entscheidungsanalyse. Bei den Standards der gesundheitsökonomischen Evaluation hat sich ein Konsens über konstitutive Elemente der gesundheitsökonomischen Evaluation und über zulässige Ansätze der Kostenanalyse und Outcomebestimmung herausgebildet. Teilweise wird in Leitlinien explizit gefordert, alternative Ansätze zu kalkulieren. Die Elemente der gesundheitsökonomischen Evaluation umfassen (1) die begründete Auswahl der Studienform, (2) die Identifizierung und Festlegung der Vergleichsalternativen, (3) die Perspektive der Evaluation, (4) die Bestimmung von Ressourcenkonsum und Kosten, (5) die Identifizierung und Bestimmung der relevanten Effekte und Nutzen, (6) die Festlegung des Zeithorizonts, (7) die Modellierung, (8) die Diskontierung, (9) die Inkremental- und (10) die Unsicherheitsanalyse.

## 4.3 Fragestellung

Welche Instrumente zur Qualitätsbewertung von systematischen Übersichtsarbeiten, Interventions-, Beobachtungs-, Diagnose- und gesundheitsökonomischen Studien gibt es, wie unterscheiden sich diese und welche Schlussfolgerungen lassen sich daraus für die Qualitätsbewertung ableiten?

## 4.4 Methodik

Zur Identifikation von Instrumenten wird eine systematische Recherche in einschlägigen Datenbanken ab 1988 durchgeführt, ergänzt um eine Durchsicht der Referenzen, der HTA-Berichte der Deutschen Agentur für Health Technology Assessment (DAHTA) sowie eine Internetrecherche. Es werden formale Charakteristika und inhaltliche Elemente der Instrumente extrahiert. Die inhaltliche Datenextraktion wird spezifisch für Interventions-, Beobachtungs-, Diagnosestudien, systematische Übersichtsarbeiten und gesundheitsökonomische Studien durchgeführt. Die Literaturauswahl, die Datenextraktion und die Qualitätsbewertung werden jeweils von zwei unabhängigen Reviewern vorgenommen, bei Diskrepanzen erfolgt eine Konsensentscheidung.

Die Inhalte von Instrumenten zur Bewertung von randomisierten Interventions-, Beobachtungs-, Diagnosestudien und systematischen Übersichtsarbeiten werden anhand von modifizierten Kriterienlisten extrahiert. Die Elemente der Listen setzen sich aus Studiencharakteristika zusammen, für die entweder empirisch ein Einfluss auf die Höhe der Studienergebnisse nachgewiesen oder deren Einfluss allgemein akzeptiert bzw. theoretisch fundiert ist. Die Elemente für Studiencharakteristika von Interventions-, Beobachtungsstudien und systematischen Übersichtsarbeiten werden in mehrere Domänen zusammengefasst. Von den Elementen werden diejenigen als relevant definiert, für die empirische Evidenz als potenzielle Biasquelle besteht bzw. die von anderen Autoren als essenziell eingestuft werden.

Als Basis für die Auswahl eines Instruments zur Qualitätsbewertung werden designspezifisch nur generische Instrumente und ihre Elemente der internen Validität betrachtet. Außerdem wird das Vorhandensein von Ausfüllhinweisen berücksichtigt. Anhand der Anzahl abgedeckter Elemente insgesamt, abgedeckter relevanter Elemente sowie abgedeckter Domänen werden umfassendere von weniger umfassenden Instrumenten unterschieden.

Für die Datenextraktion der inhaltlichen Elemente für gesundheitsökonomische Studien wird ein Formular entwickelt, da keine Übersichtsarbeiten vorliegen, die als Referenz dienen können. Im ersten Schritt des Entwicklungsprozesses werden Standardlehrbücher sowie aktuelle nationale und internationale Leitlinien zur Erstellung gesundheits- und pharmakoökonomischer Studien gesichtet. Inhaltlich sprechen die Lehrbücher und Leitlinien weitgehend identische Themenschwerpunkte an (Elemente der gesundheitsökonomischen Evaluation). In einem zweiten Schritt werden die herausgearbeiteten Themenschwerpunkte auf den Bezug zur Studienqualität (interne Validität) gesundheitsökonomischer Studien untersucht. Es werden Domänen und Items entwickelt, die auf den Themenschwerpunkten der Lehrbücher und Leitlinien basieren. Sie werden in ein Formular zur Extraktion von gesundheitsökonomischen Qualitätsbewertungsinstrumenten (QBI) überführt, mit dessen Hilfe die verschiedenen Bewertungsinstrumente extrahiert werden. Bei der Entwicklung der Domänen und Items wird darauf geachtet, dass sich diese primär auf die interne Validität beziehen.

Im gesundheitsökonomischen Extraktionsformular wird für die Bewertung der Items der berücksichtigten QBI eine Abstufung vorgenommen: „angemessen“, „begründet“, „berichtet“ und „fehlend“. Eine Bewertung „berichtet“ wird vergeben, wenn ein QBI lediglich abfragt, ob ein Item in einer gesundheitsökonomischen Studie berichtet wird (z. B. Perspektive der Analyse, einbezogene Outcomeparameter oder Diskontierungsrate). Die Beurteilung „begründet“ bedeutet, dass das QBI explizit nach Begründungen für die Ausprägung des Items fragt. Die Bewertung „angemessen“ heißt, dass ein Instrument eine Überprüfung der Angemessenheit des Items fordert. Die Überprüfung der Angemessenheit sollte an den Standards der gesundheitsökonomischen Evaluation orientiert sein.

Zwecks Erfahrungsaustausch zu Problemen bei der praktischen Anwendung von Instrumenten wird ein Workshop durchgeführt. Ziele des Workshops sind der Austausch und die Diskussion der Erfahrungen sowie des Umgangs mit Bewertungsinstrumenten zur Qualität von randomisierten und nichtrandomisierten Interventionsstudien, Anforderungen sowie Inhalte an/von Bewertungsinstrumente/n zur Qualität von Interventionsstudien. Der Austausch dient zur Ergänzung von wissenschaftlichen Untersuchungen um praktische Aspekte, deren Stellenwert in Publikationen oft nicht thematisiert wird. Eine Konsensbildung zu einzelnen Aspekten wird nicht angestrebt. Zielgruppe des Workshops sind Autoren von deutschsprachigen HTA-Berichten oder systematischen Reviews des Deutschen Instituts für Medizinische Dokumentation und Information (DIMDI) und des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Experten auf dem Gebiet der Methodik, Wissenschaftler (aus den Disziplinen Medizin, Public Health, Epidemiologie, Prävention, Gesundheitsökonomie), die mit gesundheitspolitisch relevanten Evaluationen befasst sind, sowie Institute/Verbände, die systematische Reviews mit Qualitätsbewertung durchführen. Referenten werden mit ihren Vorträgen die entsprechenden Themen einleiten. Im Anschluss an die Vorträge sind jeweils 20 bis 30 Minuten für eine moderierte Diskussion vorgesehen. Zur Dokumentation wird u. a. eine Audio-Aufzeichnung mit anschließender Transkription durchgeführt.

## 4.5 Ergebnisse

Die umfassende Recherche ergibt insgesamt 147 Instrumente zur Bewertung der Studienqualität: 15 für systematische Reviews/HTA-Berichte/Metaanalysen, 80 für Interventions-, 30 für Beobachtungs-, 17 für Diagnose- und 22 für gesundheitsökonomische Studien. Unter den QBI sind 16 Instrumente, die sowohl für Interventions- als auch für Beobachtungsstudien eingesetzt werden können.

Ein initiales Screening von HTA-Berichten in der DAHTA-Datenbank zeigt, dass in 87 % der Berichte die Durchführung einer Qualitätsbewertung angegeben wird. Von diesen wird ist jedoch nur bei der Hälfte das verwendete QBI dokumentiert.

Die identifizierten Instrumente weisen eine große Variation hinsichtlich der formalen und inhaltlichen Charakteristika auf. Einige Instrumente enthalten neben Items zur internen Validität auch welche zur Berichtsqualität und zur externen Validität. Designspezifisch werden generische Instrumente für die Bewertung von systematischen Reviews/HTA-Berichten/Metaanalysen, Interventions-, Beobachtungs- und Diagnosestudien ermittelt, die die meisten Elemente zur internen Validität, die meisten Domänen mit mindestens einem bzw. 50 % der enthaltenen Elemente sowie die meisten als relevant definierten Elemente abdecken. Es können umfassendere von weniger umfassenden Instrumenten unterschieden werden.

Die Instrumente, die die Qualität gesundheitsökonomischer Studien untersuchen, weisen ebenfalls erhebliche Unterschiede auf sowohl in der Betrachtung der verschiedenen Themenbereiche, als auch in der Bewertung der Qualität. Zudem bestehen beträchtliche Differenzen in den Operationalisierungen. Über alle Studiendesigns hinweg erfüllt keines der eingeschlossenen Instrumente alle Bereiche.

Am Workshop nehmen insgesamt 27 Personen aus HTA- und EbM-assoziierten (EbM = Evidenzbasierte Medizin) Institutionen teil. Folgende Diskussionspunkte werden von den Teilnehmern vorgeschlagen: externe Validität als Bestandteil von Bewertungsinstrumenten, Subjektivität der Bewertung, Umgang mit geringer Berichtsqualität, endpunkt- statt studienbezogene Qualitätsbewertung und Integration der Ergebnisse der Bewertung. Eine Konsensbildung ist im Rahmen des Workshops nicht vorgesehen, es werden daher Einzelmeinungen wiedergegeben. Externe und interne Validität sollten getrennt voneinander bewertet werden. Items, die einen großen Spielraum für subjektive Bewertungen lassen, führen zu mangelnder Übereinstimmung der Bewertung und hohem Diskussionsbedarf. Dies kann durch eine präzise Operationalisierung der Items vermieden werden.

## 4.6 Diskussion

Studienqualität kann unterschiedlich operationalisiert werden. Es überwiegt die Auffassung, dass eine Bewertung der Studienqualität die Höhe der internen Validität bzw. das Verzerrungspotenzial abbilden sollte. Die Bestandsaufnahme der zahlreichen identifizierten Instrumente zeigt jedoch, dass viele Instrumente auch Items der Berichtsqualität enthalten. Diese Vermischung von Berichtsqualität und interner Validität kann zu einer Fehleinschätzung der Studienqualität führen, wenn Elemente der Berichtsqualität als Surrogatparameter für die Einschätzung der methodischen Qualität herangezogen werden.

Anhand der tabellarischen Darstellung abgedeckter inhaltlicher Items können die identifizierten QBI verglichen werden. Dieses Vorgehen ist jedoch mit Einschränkungen verbunden, da kein Konsens über geeignete Kriterien existiert und nicht für alle Elemente Evidenz vorliegt, dass sie die Höhe der internen Validität einer Studie beeinflussen. Daher ist eine hohe Zahl an abgedeckten Elementen nicht notwendigerweise ein Hinweis auf ein gutes Instrument.

Zur weiteren Differenzierung der QBI wird die Anzahl der als relevant definierten Elemente dargestellt. Während für die relevanten Elemente in Interventions- und Diagnosestudien nur evidenzbasierte Biasquellen ausgewählt werden, trifft dies nur für einige der relevanten Elemente in Beobachtungsstudien und systematischen Übersichtsarbeiten zu. Insgesamt kann die Erfüllung von relevanten Elementen nur als erste Einschätzung dienen, um Instrumente zu identifizieren, die mehr oder weniger umfassend sind. Je nach Themenbereich sollte jeweils geprüft werden, ob alle Items des Instruments relevant sind bzw. ob für das jeweilige Thema zusätzliche Items einbezogen werden sollten.

Einige inhaltliche Elemente von QBI waren nicht eindeutig der Berichtsqualität, der internen oder externen Validität zuzuordnen. Beispielsweise ist die Berechnung der erforderlichen Stichprobengröße zunächst nur mit der Präzision der Ergebnisse assoziiert ohne dass die Höhe des Effektschätzers

beeinflusst wird. Die Präzision der Effektschätzer kann jedoch Einfluss auf die Signifikanz der Ergebnisse haben.

Sicher werden nicht alle jemals eingesetzten Instrumente gefunden. Gleichwohl wird die Möglichkeit, bedeutsame und häufig eingesetzte Instrumente übersehen zu haben, als gering eingeschätzt, u. a. auch durch die Nutzung mehrerer Datenquellen einschließlich Internet.

Generell gilt, je höher der Spielraum für subjektive Bewertungen ist, desto geringer ist die Übereinstimmung der Reviewer. Die einzelnen Items der Instrumente sollten daher möglichst präzise und ausführlich operationalisiert sein. Ggf. sind die Ausfüllhinweise anzupassen, um eine eindeutige Bewertungsgrundlage für alle Reviewer sicherzustellen. Etwa 40 % der eingeschlossenen Instrumente geben eine ausführlichere Anleitung zur Durchführung der Qualitätsbewertung.

Die Bewertung der Qualität gesundheitsökonomischer Studien ist ein zwingend erforderlicher Bestandteil bei der Erstellung von HTA-Berichten. Insgesamt werden 22 gesundheitsökonomische QBI identifiziert. Zwischen den untersuchten Instrumenten gibt es deutliche Unterschiede bezüglich:

- Anzahl der untersuchten Items aus dem Extraktionsformular (Themenschwerpunkte)
- Bewertungsqualität: angemessen – begründet – berichtet
- Differenziertheit der Qualitätsabfragen.

Keines der untersuchten Bewertungsinstrumente deckt die gesamte Bandbreite der Themenschwerpunkte (Elemente der gesundheitsökonomischen Evaluation) ab. Nur wenige Instrumente berücksichtigen fast alle Bereiche des Extraktionsbogens. Nur drei Instrumente überprüfen überwiegend die Angemessenheit der methodischen Verfahren. In vielen Instrumenten wird zumindest bei einigen Items nach der Angemessenheit der Verfahren gefragt. Für keines der Instrumente wird jedoch erläutert, was unter „angemessen“ zu verstehen ist. Die Mehrzahl der Instrumente fordert Begründungen für konkrete Ausprägungen der Items ein oder untersucht lediglich, ob und welche Items berichtet werden.

Deutliche Unterschiede bestehen auch in der Differenziertheit der Qualitätsabfragen. Wie differenziert ein Bewertungsinstrument die Themenschwerpunkte erfragt, wird über die Anzahl der Items abgebildet. Wenn sich die Qualitätsbewertung auf wenige Items stützt, müssen die Fragen global gestellt werden. Reviewern bleiben dann größere Spielräume bei der Interpretation von Items. Bei umfangreicheren Instrumenten mit großer Itemanzahl lassen sich Items stärker operationalisieren, sodass die Interpretationsspielräume deutlich eingeschränkt werden und objektivere Bewertungen unterstützt werden.

## 4.7 Schlussfolgerungen

Die Qualitätsbewertung von Studien ist ein obligatorischer Arbeitsschritt bei der Erstellung von systematischen Übersichtsarbeiten, der transparent darzustellen ist. Es stehen unterschiedliche designspezifische Instrumente zur Verfügung, die entsprechend ihrer inhaltlichen Abdeckung von Elementen der internen Validität für die Qualitätsbewertung ausgewählt werden können.

Für die Auswahl eines QBI gilt, dass Skalen nicht bzw. ohne quantitative Gesamtbewertung eingesetzt werden sollten. Zur Minimierung der Subjektivität der Bewertung sind Instrumente mit einer ausführlichen und präzisen Operationalisierung der einzelnen Elemente vorteilhaft. Wenn möglich, sollten die ausgewählten Instrumente zuvor an ausgewählten Studien getestet und bei Bedarf die Operationalisierung der Items ergänzt bzw. präzisiert werden, um die Subjektivität der Bewertung zu minimieren und eine hohe Übereinstimmung der Bewertungen sicherzustellen.

Weitere Forschung ist erforderlich, um Studiencharakteristika zu identifizieren, die die interne Validität von Studien beeinflussen. Dies gilt insbesondere für Beobachtungsstudien. Offen ist auch, inwieweit die Validität von Studien durch eine qualitative Gesamtbewertung korrekt gemessen wird.

Für die gesundheitsökonomische Qualitätsbewertung sollten Instrumente entwickelt werden, die (1) die gesamten Themenschwerpunkte abbilden, (2) die angemessene Umsetzung von Items in gesundheitsökonomischen Studien überprüfen und (3) die Themenschwerpunkte hinreichend differenziert abfragen. Die Angemessenheit sollte sich an den Standards der gesundheitsökonomischen Evaluation orientieren (definiert durch Standardlehrbücher und internationale Guidelines). Es sollten Erläuterungen und Ausfüllhinweise zu den Bewertungsinstrumenten entwickelt werden, in denen beschrieben wird, wie Angemessenheit definiert ist.

# 5 Summary

## 5.1 Health political background

Healthcare policy decisions should be based on the best available scientific evidence. Scientific evidence is based on the synthesis of study results, which are if possible unbiased and thus have a high credibility.

## 5.2 Scientific background

Quality assessments to evaluate the credibility of studies is an inherent component of HTA reports (HTA = Health Technology Assessment) and systematic reviews. There are various quality assessment tools (QAT) that rate the extent of systematic distortion in study results by confounding or bias (internal validity).

There is no gold standard for assessing the study quality, since the true associations of exposures/ interventions and outcomes are unknown. The existing tools for assessing study quality can be classified into scales, checklists and component ratings. In a scale, each item receives a numerical rating that will be added to a sum score. Scales are no longer recommended, because they do not reflect the correct extent of validity. A checklist consists of at least two items without a numerical rating system. The component rating includes components like “randomization” and “blinding”, which are also not evaluated numerically, but qualitatively.

In this report methodological quality that is used synonymously with the expression study quality and must be distinguished from the reporting quality, which is not part of this report.

The quality of health economic studies is determined by (a) the validity of study results, (b) the compliance with methodological standards of health economic evaluation and (c) the access to appropriate cost data. The methodological standards of health economic evaluations are described in health economic literature and international guidelines for providing health economic evaluations. Health economic evaluations are based on the theoretical concepts of welfare economics and decision analysis. The standards of economic evaluation have reached a broad consensus regarding the constitutive elements of health economic evaluation and approaches to cost analysis and outcome determination. Nevertheless, some guidelines recommend different approaches to be used. The elements of health economic evaluation contain (1) the justification and the choice of the evaluation type, (2) the identification and the selection of comparators, (3) the perspective, (4) the identification of resource use and costs, (5) the identification of all relevant effects and benefits, (6) the declaration of the time horizon, (7) modelling, (8) discounting, (9) incremental analysis, (10) uncertainty analysis.

## 5.3 Research questions

What QAT are available to assess the quality of systematic reviews/HTA reports, intervention studies, observational studies, diagnostic studies and health economic studies, how do they differ among each other and what conclusions can be drawn from these results for quality assessments?

## 5.4 Methods

A systematic search of relevant electronic databases from 1988 onwards is done to identify QAT, supplemented by screening of the references of the HTA reports of the German Agency for Health Technology Assessment (DAHTA) and in addition an internet search. Formal characteristics and substantive elements of the tools are extracted. The substantive elements of the QAT are extracted specific to systematic reviews, intervention studies, observational studies, diagnostic studies, and health economic studies. The literature search, the data extraction and the quality assessment are carried out independently by two reviewers. Different ratings of the reviewers are solved by consensus.

The content of tools for the quality assessment of systematic reviews, intervention studies, observational studies, and diagnostic studies is extracted by using modified criteria lists. The elements of the lists are made up of study characteristics, which have either empirically demonstrated evidence of

an effect on the level of the study results or its distorting effect on study results is generally accepted. The elements for study characteristics of systematic reviews, intervention studies, and observational studies are summarised in several domains. Out of all elements, those elements with empirical evidence as a potential source of bias or elements being classified on a theoretical basis as essential for internal validity are defined as relevant elements.

In order to provide a basis for the selection of a tool, only generic tools and their elements of internal validity are considered. Furthermore, the presence of sufficient operationalisation is required. The tools are distinguished by the total number of covered elements, covered relevant elements, and covered domains. Tabular summaries of the results are prepared for each study design and the results across the QAT are assessed qualitatively to identify more and less comprehensive tools.

For the data extraction of the basic elements of health economic studies, a form is developed, because there are no systematic reviews that can provide a basis for the data extraction. In the first step of the development process health economic literature and current national and international guidelines for creating health and pharmacoeconomic studies are screened. Literature and guidelines address mainly similar topics (elements of health economic evaluation). In the second step, the key elements are worked out to investigate the relation to study quality (internal validity) of health economic studies. Domains and items are developed based on the elements of health economic evaluations adapted from literature and guidelines. Domains and items are transferred into a form for analysing the quality assessment tools for health economic evaluation studies. This form helps to extract the various tools. In the development of domains and items, effort is made to ensure that items relate primarily to the internal validity.

In the health economic extraction form a gradation for rating the different items is made as such: “appropriate”, “justified”, “reported” and “missing”. If a quality assessment tool asks for a special item addressed in a study, a rating with “reported” is made (e. g. perspective of analysis, outcome parameter or discount rate). An item is rated with “justified”, the quality assessment tool asks for the rationale for choosing a special specification. The rating “appropriate” is assigned when the quality assessment tool asks for the adequacy of used methodology in an item.

In order to find out about problems in the practical application of tools, a workshop is conducted. Objectives of the workshop are to exchange and discuss user experiences with quality assessment tools for intervention studies, requirements, and content of tools on the quality of intervention studies. These discussions will examine practical issues that are rarely discussed in the literature. A consensus on individual aspects is not pursued. The target audience include authors of the German HTA reports and systematic reviews of the German Institute for Medical Documentation and Information (DIMDI) and the Institute for Quality and Efficiency in Health Care (IQWiG), experts in the field of methodology, researchers (from the disciplines of medicine conducting public health, epidemiology, prevention, health economics), involved in healthcare policy-relevant evaluations, as well as institutes/associations conducting systematic reviews. Topics are introduced by presentations of invited experts followed by moderated discussions. Presentations and discussions are documented by audio recordings and transcriptions.

## 5.5 Results

The extensive literature search yields a total of 147 tools to assess the study quality: 15 for systematic reviews/HTA reports/meta-analysis, 80 for intervention studies, 30 for observational studies, 17 for diagnostic studies and 22 for health economic studies. Among the QAT are 16 tools that can be used both for intervention and observational studies.

An initial screening of HTA reports in the DAHTA database indicates that a quality assessment was reported in 87 % of the identified documents. However, in only half of these reports the chosen QAT was mentioned.

The tools show a wide variation of the formal and content characteristics. Some tools contain not only items of internal validity, but also of reporting quality and external validity. Design-specific generic tools for the assessment of systematic reviews/HTA reports/meta-analysis, intervention studies, observational studies and diagnostic studies are identified, which cover most elements for internal validity,

most of the domains with at least one, or 50 % of the contained elements as well as the most relevant elements. More and less comprehensive tools can be distinguished.

The tools that examine the quality of health economic studies also reveal significant differences both in the consideration of various topics, as well as in the assessment of quality. In addition, substantial differences exist in the operationalisation of the items. Across all study designs, none of the included tools meet all elements.

A total of 27 people from HTA and EBM-associated (EBM = evidence-based medicine) institutions take part in the workshop. The following discussion points are suggested by the participants: the external validity as a part of assessment tools, the subjectivity of the assessment process, dealing with low reporting quality, endpoint versus study related quality assessment and incorporation of the results of the quality assessment. As consensus at the workshop is not intended, individual opinions are presented. External and internal validity should be assessed separately from each other. Items, which leave much room for subjective ratings, lead to a lack of interrater reliability and result in a high need for discussions. This can be avoided by a precise and detailed operationalisation of the items.

## 5.6 Discussion

The quality of studies can be defined in various ways. It is a dominating view that an assessment of study quality can either express the level of internal validity or the possibility of distortion. However, the inventory of the numerous identified tools shows that many of them include the assessment of reporting quality. Mixing the reporting quality and the internal validity can lead to a misinterpretation of the study quality, if the elements of the reporting quality are used as a surrogate for assessing the methodological quality.

Based on the tabular presentation of covered content items, the identified QAT can be compared. However, this approach has limitations, since there is no scientific consensus on the necessary elements of the internal validity, and not all of the generally accepted elements are based on empirical evidence. Therefore, the highest possible number of covered elements is not necessarily an indication of an appropriate tool.

For further differentiation of the QAT, the number of covered relevant elements is presented. While for relevant elements of intervention and diagnostic studies only evidence based elements affecting the internal validity are selected, this is true for only some of the relevant elements of observational studies and systematic reviews. Overall, the performance of relevant elements should be used cautiously to identify tools that are more or less comprehensive. Depending on the topic, it should be examined, whether all items of a chosen tool are relevant, and whether additional quality items should be assessed as a part of the assessment.

Some elements of QAT cannot be clearly assigned to the reporting quality, the internal or external validity. For example, the calculation of the required sample size is only associated with the precision of the results without affecting the size of the effect estimator. However, the precision of the effect estimates may affect the significance of the results.

Not all the tools ever used have been found. However, the possibility of having missed important and frequently applied tools is low, since different data sources including the internet were screened.

In general, the higher the scope for subjective assessments, the lower the agreement between the reviewers is. Therefore, every item of a tool should be operationalised as detailed and precisely as possible. Where necessary, the instructions can be adjusted to ensure that all reviewers are clear on how to score study quality. About 40 % of the included tools provide more detailed guidance for assessment.

The quality assessment of health economic studies is an essential part of creating HTA reports. A total of 22 health economic QAT is identified. There are considerable differences regarding:

- the number of included items of the health economic extraction form (elements of health economic evaluation)
- the assessment quality: appropriate – justified – reported
- the diversity of quality sampling

None of the analysed QAT covers the whole range of relevant themes (elements of health economic evaluation). Only few consider most domains of the extraction form. Only three tools check the adequacy of the methodological procedures. Many tools ask for the methodological adequacy in few items. None of the QAT defines what is meant with adequacy. Most tools demand a justification for the methodological procedures or analyse, which items are reported.

Significant differences also exist in the sophistication of the quality assessment. The question how differentiated an assessment tool discusses the different elements of health economic evaluation can be answered by the number of items in a QAT. Because a tool is based only on few items, questions have to be more generally introduced. Reviewers will have a considerable scope for interpretation. For extensive tools with a great number of items, they can be operationalised to be more specific, so the scope for interpretation will be significantly reduced and more objective assessments are supported.

## 5.7 Conclusions

The quality assessment of studies is a mandatory part of systematic reviews, and has to be documented transparently. There are different, design-specific QAT available that can be selected according to their substantive coverage of the elements of internal validity.

There is consensus that scales should not be used for quality assessments or should be used without quantitative assessment. To minimise the subjectivity of the evaluation, tools with a detailed and precise operationalisation of the items are preferable. If possible, the chosen tool should be tested in a few studies in advance to check if the operationalisation of the items needs to be supplemented or clarified to minimise the subjectivity of the evaluation and to ensure uniform scoring of all reviewers.

Further research is needed to identify study characteristics that influence the internal validity of studies, especially for observational studies. So far, there is no evidence that qualitative overall assessment of study quality is correctly associated with the internal validity.

For assessing the quality of health economic studies, tools should be developed, which (1) cover all relevant elements of health economic evaluation, (2) assess the appropriate use of methodological procedures and (3) differentiate the various topics sufficiently. The adequacy should be based on the standards of health economic evaluation (defined by standard literature and international guidelines). Advice for filling in and operationalisations should be part of the assessment tools and, in addition, adequacy should be accurately described and defined.

# 6 Hauptdokument

## 6.1 Gesundheitspolitischer Hintergrund

Gesundheitspolitische Entscheidungen sollen auf der Grundlage von wissenschaftlichen Erkenntnissen getroffen werden, sie müssen somit evidenzbasiert sein. Evidenz gründet auf der Synthese von Studienergebnissen, die möglichst unverzerrt sind und damit eine hohe Validität aufweisen. In Deutschland dienen dem Gemeinsamen Bundesausschuss (G-BA) als evidenzbasierte Entscheidungsgrundlage unter anderem HTA-Berichte (HTA = Health Technology Assessment), die vom Deutschen Institut für Medizinische Dokumentation und Information (DIMDI) in Auftrag gegeben werden. Seit dem Inkrafttreten des GKV-Modernisierungsgesetzes im Jahr 2004 können der G-BA sowie das Bundesministerium für Gesundheit (BMG) auch das neu eingerichtete Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) mit der Bearbeitung wissenschaftlicher Fragestellungen bzw. das IQWiG das DIMDI mit der Erstellung von HTA-Berichten beauftragen. HTA ist damit Teil eines zeitgemäßen Wissens- und Informationsmanagements für Entscheidungsträger im Gesundheitswesen.

Um die Glaubwürdigkeit (Validität) von Studien einzuschätzen, werden Qualitätsbewertungen von Studien durchgeführt (methodische Qualität). Diese prüfen, ob die angegebenen Studienergebnisse systematisch durch Confounding, Selektions- oder Informationsbias verzerrt sein können. Je höher die Wahrscheinlichkeit für systematische Verzerrungen ist, desto geringer wird die Qualität einer Studie eingeordnet. Dies ist wichtig, da die Studienqualität sich auf die Abschätzung der Höhe des Nutzens einer Intervention auswirkt. Studien mit höherer Qualität schätzen die Wirkung einer Intervention oftmals geringer ein als Studien geringerer Qualität[141].

Die bestehenden Bewertungsinstrumente variieren stark: Es existieren Instrumente, die den jeweiligen Studientyp, wie beispielsweise randomisierte kontrollierte Studie (RCT), Kohorten- oder Fall-Kontrollstudie, berücksichtigen; es werden qualitative und quantitative Instrumente, Checklisten und Skalen unterschieden[141, 166]. Wissenschaftlern ist die Relevanz von Qualitätsbewertungen der einzuschließenden Dokumente für systematische Berichte, Metaanalysen oder HTA-Berichte bewusst. Dennoch werden nicht durchgängig Qualitätsbewertungen durchgeführt[56, 141]. Ferner fließen für systematische Berichte bei Einsatz eines Bewertungsinstruments dessen Ergebnisse nicht per se auch in die Datensynthese und -analyse ein.

Bezüglich der Kriterien, die erfüllt sein müssen, damit Studien als Evidenzbasis herangezogen werden, gibt es kein einheitliches Vorgehen. Einige Autoren von systematischen Übersichtsarbeiten schließen ausschließlich RCT ein, während andere auch kontrollierte Beobachtungsstudien akzeptieren. Inwieweit dieses Vorgehen tatsächlich die bestmögliche Evidenz liefert und ob nicht gut durchgeführte nicht-randomisierte Studien einbezogen werden sollten, wird kontrovers diskutiert[68, 141, 165].

Der vorliegende HTA-Bericht hat das Ziel, einen Überblick über Qualitätsbewertungsinstrumente (QBI) für Primär- und Sekundärstudien zu geben, diese Instrumente zu vergleichen und ggf. aus den Ergebnissen Schlussfolgerungen für die Durchführung von Qualitätsbewertungen abzuleiten.

## 6.2 Wissenschaftlicher Hintergrund

HTA-Berichte haben das Ziel der systematischen Bewertung medizinischer Technologien, für die ein gesundheitspolitischer Handlungs- bzw. Entscheidungsbedarf besteht. Der Begriff Technologie kann sowohl Arzneimittel, Instrumente, Geräte, präventive, therapeutische und diagnostische Verfahren als auch Organisations- und Managementsysteme umfassen. Diese Vielzahl inhaltlicher Schwerpunkte kann auf unterschiedliche Settings ausgedehnt werden, wie ambulante und stationäre Versorgung, sowie sich auf einzelne Bevölkerungsgruppen oder die gesamte Gesellschaft beziehen.

Neben der Bewertung der Wirksamkeit bestimmter Verfahren ist die gesundheitsökonomische Untersuchung der Effizienz dieser Verfahren ein wichtiger Bestandteil von HTA-Berichten. Ergänzt wird die Einschätzung des Nutzens und der Kosten-Effektivität eines Verfahrens um assoziierte soziale, ethische und juristische Aspekte.

Die Erstellung von HTA-Berichten ist durch eine systematische und wissenschaftlich-methodische Vorgehensweise bestimmt, um eine größtmögliche Sicherheit der Validität von Informationen zu ge-

währleisten. Neben dem systematischen Ansatz ist die absolute Transparenz dieses Bewertungsprozesses ein wesentliches Element.

Die Vorgehensweise bei der Erstellung von HTA-Berichten und systematischen Übersichtsarbeiten folgt einem strukturierten Ablauf, beginnend mit der Formulierung einer präzisen Fragestellung, an die sich eine systematische Literaturrecherche und -selektion anschließen. Die ausgewählte Literatur sollte einer Qualitätsbewertung unterzogen werden. Anschließend wird die Datenextraktion durchgeführt. Die extrahierten Daten werden in einer qualitativen und/oder quantitativen Informationssynthese zusammengefasst und entsprechende Schlussfolgerungen daraus abgeleitet.

### 6.2.1 Studienqualität

Die Qualitätsbewertung als Teil des systematischen Bewertungsprozesses ist Grundlage der Einschätzung der Validität der Studien, die als Evidenzbasis für Schlussfolgerungen dienen. Mögliche Definitionen von Studienqualität lauten:

*„Qualität einer Studie (methodische Qualität): bezieht sich auf das Bemühen in einer Studie, Bias zu minimieren. Um die Qualität zu bewerten können Merkmale des Designs, der Durchführung und der statistischen Analyse einer Studie herangezogen werden. Diese determinieren die Glaubwürdigkeit (Validität) der Studienergebnisse“.*[117]

*„(…) the confidence that the trial design, conduct and analysis has minimised or avoided biases in its treatment comparisons“.*[143]

*„Quality is a set of parameters in the design and conduct of a study that reflects the validity of the outcome, related to the external and internal validity and the statistical model used“.*[227]

*„The extent to which all aspects of a study‘s design and conduct can be shown to protect against systematic bias, non-systematic bias, and inferential error“.*[132]

Ziel der Bewertung der Studienqualität ist die Einschätzung, inwieweit systematische Fehler (Bias) und Confounding die Studienergebnisse verzerren und damit die interne Validität beeinträchtigen können. Der Zusammenhang von Studienergebnissen und dem Grad der Validität beruht sowohl auf Lehrmeinungen als auch auf empirischen Erkenntnissen. So ist vielfach gezeigt worden, dass das Gesamtergebnis von Metaanalysen von der Qualität der eingeschlossenen Studien abhängt. Studien mit höherer Qualität schätzen die Wirkung einer Intervention oftmals geringer ein als Studien geringerer Qualität[141]. Andere Autoren finden, dass Studien mit hoher Studienqualität sowohl größere, gleich hohe oder auch geringere Effekte im Vergleich zu Studien mit niedriger Qualität gemessen haben[118].

Auch der Einfluss einzelner Studienaspekte auf die Studienergebnisse ist untersucht. Bei Vergleichen von randomisierten und nicht-randomisierten Studien derselben Intervention, werden in den meisten Fällen stärkere Effekte in den nicht-randomisierten Studien festgestellt[119]. Bei einigen Studien ist es umgekehrt und in einem Fall haben die Effekte eine unterschiedliche Richtung: Randomisierte Studien finden einen schädigenden Effekt, während nicht-randomisierte Studien mit historischen Kontrollen einen Nutzen der Intervention anzeigen[56].

Untersuchungen zum Einfluss der verdeckten Interventionszuteilung (allocation concealment) ergeben, dass Studien mit fehlender oder inadäquater Verdeckung der Zuteilung größere Effekte erzielen als Studien mit verdeckter Zuteilung[68, 119, 141, 193]. Das gleiche Muster kann bei offenen im Vergleich mit doppelblinden Studien beobachtet werden[68, 108, 141].

Trotz der allgemein anerkannten Relevanz der Qualitätsbewertung wird diese nicht standardmäßig durchgeführt[141]. Eine Analyse ergibt, dass fast 50 % aller systematischen Reviews zu diagnostischen Tests keine Qualitätsbewertung enthalten[236]. Für systematische Reviews mit Beobachtungsstudien kann von vergleichbaren Ergebnissen ausgegangen werden[56, 137].

### 6.2.2 Berichtsqualität

Von der methodischen Qualität, die in diesem Bericht synonym zum Begriff Studienqualität verwendet wird, muss die Berichtsqualität abgegrenzt werden. Zur Erfassung der Berichtsqualität von Publi-

kationen wird das Vorhandensein bzw. die Vollständigkeit von Informationen zu den Aspekten Design, Durchführung und Auswertung von Studien geprüft[143], ohne dass der Inhalt dieser Informationen im Hinblick auf die Validität beurteilt wird.

Seit Mitte der 90-er Jahre existieren mehrere Instrumente zur systematischen Bewertung der Berichtsqualität von systematischen Reviews[124, 140, 209], RCT[147], Beobachtungs-[58, 232] und diagnostischen Studien[21] (Tabelle 1: Übersicht über Checklisten für die Bewertung der Berichtsqualität), da methodische Details oft unzureichend in Publikationen dargestellt werden[15, 175, 192]. Diese Instrumente zur Erfassung der Berichtsqualität sind Checklisten zur systematischen Abfrage von relevanten inhaltlichen Komponenten. Zielgruppe dieser Instrumente sind Autoren von Studien, Reviewer, Editoren und Herausgeber wissenschaftlicher Zeitschriften.

Da die Beurteilung der methodischen Qualität in der Regel auf publizierten Berichten beruht, ist eine adäquate Berichtsqualität Voraussetzung für die Bewertung der Studienqualität. Werden relevante methodische Aspekte nicht berichtet, können diese nicht zur Bewertung der Studienqualität herangezogen werden. Es ist daher denkbar, dass eine lückenhaft berichtete, methodisch jedoch gut durchgeführte Studie zu einer systematischen Unterschätzung der tatsächlichen Studienqualität führt[143, 228]. Dabei kann die Summe der fehlenden Informationen einen Hinweis auf die Möglichkeit einer Fehleinschätzung der Studienqualität geben. Weitere Untersuchungen zeigen, dass Studien gleicher Berichtsqualität große methodische Unterschiede aufweisen können[99]. Daher sollte zwischen Berichts- und methodischer Qualität klar unterschieden sowie die Berichtsqualität nicht als Surrogatparameter für die methodische Studienqualität herangezogen werden. Aus diesen Gründen ist die Berichtsqualität explizit nicht Gegenstand dieses Berichts.

**Tabelle 1: Übersicht über Checklisten für die Bewertung der Berichtsqualität**

| **Kurzname** | **Name** | **Studientyp**(en) | **Referenzen** |
|---|---|---|---|
| CONSORT | Consolidated Standards of Reporting Trials | RCT | Begg et al.[16] |
| CONSORT (Revision) | Consolidated Standards of Reporting Trials | RCT | Moher et al.[147, 142]<br>Altman et al.[6]<br>www.consort-statement.org |
| CONSORT: extension to cluster randomised trials | Consolidated Standards of Reporting Trials statement: extension to cluster randomised trials | Cluster-RCT | Campbell et al.[28] |
| STROBE | Strengthening the Reporting of Observational Studies in Epidemiology | Kohortenstudien<br>Fall-Kontrollstudien<br>Querschnittstudien | Elm et al.[232]<br>Vandenbroucke et al.[225] |
| STREGA | Strengthening the Reporting of Genetic Association Studien (STREGA) – An Extension of the STROBE Statement | Genetische Assoziationsstudien | Little et al.[129] |
| MOOSE | Meta-Analysis of Observational Studies in Epidemiology | Metaanalyse von Beobachtungsstudien | Stroup et al.[209] |
| QUOROM | The Quality of Reporting of Meta-Analyses | Metaanalyse von RCT | Moher et al.[140] |
| PRISMA<br>((Weiter-) Entwicklung der QUORUM-Empfehlungen) | Preferred Reporting Items for Systematic Reviews and Meta-Analyses | Metaanalyse von Studien, die Interventionen im Gesundheitswesen bewerten | Little et al.[129]<br>www.prisma-statement.org |

**Tabelle 1: Übersicht über Checklisten für die Bewertung der Berichtsqualität – Fortsetzung**

| Kurzname | Name | Studientyp(en) | Referenzen |
|---|---|---|---|
| RATS (nicht eindeutig Berichtsqualität) | Qualitative Research Review Guidelines | | Clark et al.[46] |
| STARD | Standards for Reporting of Diagnostic Accuracy | Diagnosestudien | Bossuyt et al.[21] |
| TREND | Transparent Reporting of Evaluations with Nonrandomized Designs | Nicht-randomisierte Interventionsstudien im Bereich Public Health | Des Jarlais et al.[58] |

RCT = Randomisierte kontrollierte Studie.

### 6.2.3 Validität

Die interne Validität beschreibt die Wahrscheinlichkeit, inwieweit Ergebnisse einer Studie die „wahren" Effekte einer Exposition/Intervention darstellen und somit glaubwürdig sind. Es gibt prinzipiell folgende Erklärungen für das Zustandekommen von Studienergebnissen. Das Ergebnis spiegelt tatsächlich den „wahren" Effekt wieder oder das Ergebnis ist durch Zufallsfehler, Confounding oder Bias zu erklären. Die Höhe der internen Validität kann abgeschätzt werden, indem das Ausmaß von Zufallsfehlern, Confounding und Bias beurteilt wird.

Zufallsfehler werden durch Berechnung von p-Werten und Konfidenzintervallen berücksichtigt. Confounding und Bias können durch Maßnahmen beim Studiendesign und bei der Studiendurchführung vermieden oder minimiert werden. Zusätzlich kann Confounding bei der Datenanalyse durch Auswertungsverfahren kontrolliert werden. Bias und Confounding werden im Folgenden detailliert erläutert.

Die externe Validität beschreibt die Generalisierbarkeit bzw. Übertragbarkeit der Studienergebnisse auf Personen, die nicht an der Studie teilgenommen haben, beispielsweise Patienten in der Routineversorgung. Voraussetzung für die externe Validität ist z. B. die Richtigkeit der Studienergebnisse, also die interne Validität.

### 6.2.4 Faktoren, die die interne Validität beeinflussen

**Confounding**

Confounding führt zu einer systematischen Über- oder Unterschätzung des „wahren" Effekts, indem der Zusammenhang zwischen Exposition/Intervention mit dem Outcome durch eine Störgröße (Confounder) verzerrt wird. Als Confounder gelten Variablen oder Merkmale von Studienteilnehmern, die sowohl mit der Exposition/Intervention als auch mit dem Outcome assoziiert sind, ohne Teil der Kausalkette von Exposition/Intervention und Outcome zu sein. Das bedeutet, dass dieses Merkmal in den Studiengruppen unterschiedlich verteilt und gleichzeitig ein Risikofaktor oder protektiver Faktor für das Outcome ist. Daher kann der Zusammenhang von Exposition und Outcome nicht nur ausschließlich durch die Exposition bedingt sein, sondern einen Mischeffekt von Exposition und Confounder darstellen. Typische Confounder können Alter, Geschlecht, sozioökonomischer Status sowie das Rauchverhalten sein.

Zu den Maßnahmen auf der Ebene des Studiendesigns zur Vermeidung von Confounding zählen Randomisierung, Restriktion und Matching. Sie haben zum Ziel, die Studiengruppen hinsichtlich der Häufigkeit bestimmter Merkmale anzugleichen (Strukturgleichheit). Eine erfolgreiche Randomisierung gilt als die effektivste Maßnahme, da eine Strukturgleichheit der Studiengruppen nicht nur für erfasste, sondern auch für alle nicht-erfassten Merkmale erzielt wird[112, 145, 193]. In diesem Zusammenhang ist die verdeckte Zuordnung der Gruppenzuteilung (allocation concealment) ein wichtiges Element zur korrekten Umsetzung der Randomisierung[174, 242].

In nicht-randomisierten Studien und Beobachtungsstudien kann Confounding durch Beschränkung auf eine bestimmte Ausprägung eines potenziellen Confounders vermieden werden (Restriktion). Wenn beispielsweise Geschlecht ein Confounder sein könnte, würden nur Männer bzw. Frauen in die Studie aufgenommen. Allerdings ist dadurch die Generalisierbarkeit der Ergebnisse in der Regel auf Männer bzw. Frauen limitiert.

Beim Matching werden entsprechend der definierten Matchingfaktoren Teilnehmern der einen Studiengruppe Teilnehmer mit gleicher Häufigkeitsverteilung der Matchingfaktoren in der anderen Studiengruppe zugeordnet (Häufigkeitsmatching). Werden als Matchingfaktoren Alter und Geschlecht gewählt, so sind diese Merkmale folglich in den Studiengruppen gleich verteilt. Beim individuellen Matching werden einem Studienteilnehmer der einen Gruppe ein oder mehrere Studienteilnehmer der anderen Studiengruppe mit gleicher Ausprägung der Matchingfaktoren zugeordnet, sodass Paare oder Triplets etc. entstehen. Wird ein individuelles Matching durchgeführt, sind spezielle Auswertungsverfahren (z. B. konditionale logistische Regressionsanalyse) erforderlich.

Auf der Ebene der Datenauswertung kann Confounding kontrolliert werden, indem nach Kategorien des Confounders stratifizierte Analysen durchgeführt oder mit multivariaten Analyseverfahren für die Confounder adjustiert werden. Es ist jedoch schwer abzuschätzen, inwieweit unbekannte oder nicht erhobene Störgrößen das Ergebnis weiterhin verzerren (residual confounding). Bei erfolgreich durchgeführter Randomisierung in RCT ist eine Kontrolle für Confounding bei der Datenauswertung nicht erforderlich.

**Bias**

Unter Bias werden systematische Fehler verstanden, die bei der Auswahl der Studienteilnehmer (Selektionsbias) oder während der Durchführung der Studie (Informationsbias) entstehen.

Ein Selektionsbias kann auftreten durch die Selbstselektion der Studienteilnehmer infolge der Freiwilligkeit der Studienteilnahme und ist daher prinzipiell bei jeder Studie, bei der die Teilnehmer ihre Einwilligung für die Teilnahme geben müssen, möglich. Studienteilnehmer können sich entsprechend von der Zielpopulation unterscheiden, z. B. durch bessere prognostische Faktoren. In Fall-Kontrollstudien kann durch den Einschluss prävalenter Fälle ein Bias entstehen (prognostischer Bias), wenn die Dauer der Erkrankung mit einer besseren Prognose assoziiert ist. In diesem Fall sollten nur inzidente Fälle eingeschlossen werden. In Interventions- und Kohortenstudien kann ein Loss-to-follow-up zu einem Selektionsbias führen, wenn sich die „verlorenen“ Studienteilnehmer oder Studienabbrecher von den in der Studie verbliebenen Personen systematisch unterscheiden. Dieser Verlustbias (attrition bias) erfordert in randomisierten Studien eine Intention-to-treat-Analyse, um den Effekt der Randomisierung (die Strukturgleichheit der Studiengruppen) aufrecht zu erhalten. Ein Ausschluss von Studienteilnehmern aus der Analyse führt oft zu einer Verzerrung der Effektschätzer[164]. Der Healthy-worker-Effekt beschreibt eine andere Form des Selektionsbias in Kohortenstudien, wenn die Morbidität oder die Mortalität der arbeitenden Bevölkerung mit der der Gesamtbevölkerung verglichen wird, da eine arbeitende Population durchschnittlich gesünder ist als die Gesamtbevölkerung, da diese auch Frührentner etc. umfasst.

Als Maßnahmen zur Vermeidung von Selektionsbias gelten gleiche Ein- und Ausschlusskriterien für alle Studienteilnehmer, hohe Teilnahmeraten, Beachtung und Einheitlichkeit der Rekrutierungswege sowie falls möglich Vermeidung der Selbstrekrutierung.

Informationsbias entsteht bei der Datenerhebung. Prinzipiell können in Interventionsstudien ein Durchführungs- (performance bias) und ein Messungsbias (measurement bias) unterschieden werden. Ein Durchführungsbias liegt vor, wenn systematische Unterschiede in den Studiengruppen hinsichtlich der geleisteten Versorgung außerhalb der untersuchten Intervention bestehen, z. B. wenn andere, nicht zur Forschungsfrage gehörende Behandlungen in einer Gruppe häufiger eingesetzt werden (Kointerventionen). Ziel ist daher die Behandlungsgleichheit, die durch ein standardisiertes therapeutisches Vorgehen und die Verblindung von Teilnehmern und Ärzten erreicht werden kann. Auch eine unterschiedliche Compliance in den Studiengruppen führt zu einer systematischen Verzerrung der Studienergebnisse. Wenn möglich, sollte daher die Compliance der Studienteilnehmer gemessen werden.

Ein Messungsbias kann durch systematische Unterschiede bei der Erhebung des Outcomes in den Studiengruppen entstehen, vor allem, wenn subjektive Endpunkte erhoben werden. Neben standardisierten Messmethoden ist die Verblindung von Teilnehmern und Untersuchern eine wichtige Maßnahme zur Sicherstellung der Beobachtungsgleichheit der Gruppen[174, 242]. Empfehlenswert sind ferner möglichst objektive Datenquellen und die Instruktion der Untersucher bzw. eine Interviewerschulung sowie die verblindete Erhebung der Zielgrößen.

**Bias durch selektives Berichten**

Es kann empirisch nachgewiesen werden, dass innerhalb einer Studie eher die signifikanten und nicht die nicht-signifikanten Ergebnisse einer Analyse berichtet werden[40-42]. Es handelt sich quasi um einen Publikationsbias innerhalb einer Studie.

**Finanzielle Förderung**

Konkurrierende Interessen der Autoren können zu einer systematischen Verzerrung der Studienergebnisse führen. Eine Studie zeigt, dass Reviews, die keine Assoziation von Passivrauchen und gesundheitlichen Schäden finden, überwiegend von Autoren mit Verbindungen zur Tabakindustrie stammen[13]. Eine andere Studie weist hinsichtlich der Frage der Sicherheit eines Medikaments (Kalziumkanalantagonist) nach, dass die befürwortende, neutrale oder kritische Haltung gegenüber diesem Medikament mit finanziellen Verflechtungen mit den entsprechenden pharmazeutischen Unternehmen assoziiert ist[207].

**Bias bei der Auswertung einer Studie**

Systematische Fehler bei der Auswertung können ebenfalls zu einer Verzerrung der Studienergebnisse führen. Hierzu gehören unter anderem die Verwendung parametrischer statt nicht-parametrischer Verfahren unter falschen Voraussetzungen, die Nichtbeachtung der Analyseeinheiten (Cluster vs. Individuen) oder Voraussetzungen, die für bestimmte Modellbildungen erfüllt sein müssen[5]. Außerdem kann ein Bias induziert werden, wenn nicht für alle Studienteilnehmer Daten zum Outcome vorliegen bzw. berücksichtigt werden.

**Bias in Diagnosestudien**

Diagnosestudien untersuchen die Validität von diagnostischen Tests, indem sie deren Sensitivität und Spezifität bestimmen. Ein diagnostischer Test kann eine Laboruntersuchung sein, die körperliche Untersuchung, die gezielte Befragung (Anamnese), ein Fragebogen, eine Röntgenuntersuchung, also prinzipiell jede Maßnahme, die durchgeführt wird, um die Wahrscheinlichkeit für eine bestimmte Diagnose zu erhöhen.

Auch für diesen Studientyp ist empirisch nachgewiesen, dass der Grad der internen Validität mit der Höhe der beobachteten Ergebnisse assoziiert ist[126]. In Diagnosestudien können neben den bereits aufgeführten Biasformen weitere mögliche, für diese Studien spezifische Biasformen auftreten, die im Folgenden erläutert werden. Whiting et al.[237] haben in einem umfangreichen HTA-Bericht systematisch die empirische Evidenz für potenzielle Biasquellen in diagnostischen Studien untersucht.

Als Referenzstandard, mit dem der Indextest verglichen wird, sollte die Goldstandardmethode gewählt werden, d. h. die Methode, die die Zielerkrankung nach aktuellem Kenntnisstand mit der größtmöglichen Sicherheit nachweist. Idealerweise hat der Referenzstandard daher eine Sensitivität und Spezifität von jeweils 100 %. Ist dies nicht gegeben, kann ein Bias durch einen nicht geeigneten Referenzstandard resultieren.

Verifikationsbias (auch Work-up-Bias genannt) bezeichnet den Sachverhalt, dass nicht bei allen Teilnehmern der Zielparameter mit dem (gleichen) Referenztest überprüft wird. Handelt es sich z. B. um einen invasiven und damit risikoreichen Referenzstandard, so wird dieser u. U. nur bei Teilnehmern mit positivem Indextest durchgeführt. Dieses Vorgehen führt zu einer Unterschätzung der Sensitivität und Überschätzung der Spezifität des Indextests.

Eine weitere Biasquelle entsteht, wenn der Indextest Teil des Referenzstandards ist, beispielsweise, wenn der Referenzstandard aus einer Kombination von mehreren Tests besteht, von denen einer der Indextest ist. Indextest und Referenzstandard sind dann nicht unabhängig voneinander, die Validität des Indextests könnte überschätzt werden (Inkorporationsbias).

Ein Reviewbias kann auftreten, wenn Index- und Referenztest nicht jeweils für das Ergebnis des anderen Tests verblindet ausgewertet werden. Wenn klinische Informationen bei der Testauswertung vorliegen, handelt es sich um einen klinischen Reviewbias.

Eine Änderung des Krankheitsstatus zwischen Durchführung des Index- und Referenztests kann zu einem Bias durch Krankheitsprogression führen.

Wird basierend auf den Ergebnissen des Indextests eine Behandlung eingeleitet, wirkt sich dies auf die zu erwartenden Ergebnisse des Referenztests aus. Dies verzerrt somit die Übereinstimmung von Index- und Referenztest (Behandlungsparadox).

Um einen sogenannten Spektrumbias zu vermeiden, sollte die Studienpopulation die Krankheitsprävalenz und das Spektrum der Krankheitsschwere der Population abbilden, in der der Test später eingesetzt wird.

Im Hinblick auf die Auswertung ist der Umgang mit nicht bewertbaren Testergebnissen wichtig. Werden diese nicht angemessen berücksichtigt, kann eine Verzerrung der Ergebnisse eintreten.

**Bias in systematischen Reviews**

Als für diese Studienart typische Biasformen können ein Publikations-, ein Sprachbias und ein Bias durch die Subjektivität der Reviewer bei der Literaturauswahl, der Datenextraktion und der Qualitätsbewertung auftreten.

Der Subjektivität der Reviewer bei der Qualitätsbewertung wird in der Regel durch mindestens zwei unabhängige Reviewer sowie Konsensentscheidungen bei Differenzen begegnet (s. auch Kapitel Qualitätsbewertung). Weitere Faktoren, die die Subjektivität der Bewertung beeinflussen können, sind die Güte der Operationalisierung der einzelnen Items bzw. Komponenten sowie die methodischen Kenntnisse der Reviewer.

Ein Publikationsbias liegt vor, wenn ein Zusammenhang zwischen der statistischen Signifikanz eines Studienergebnisses und der Publikationswahrscheinlichkeit besteht. Es spiegelt die Beobachtung wider, dass vor allem Studien veröffentlicht und daher gefunden werden, die den (gewünschten) in der Regel signifikanten Effekt zeigen. Mit einer grafischen Darstellung, dem sogenannten Funnel-Plot, kann orientierend das Vorhandensein eines Publikationsbias geprüft werden. In einem Funnel-Plot werden die Effektschätzer der Einzelstudien (x-Achse) den jeweiligen Stichprobenumfängen oder dem Kehrwert der Varianz (y-Achse) in einer Punktwolke gegenüber gestellt. Wenn kein Publikationsbias vorliegt, streuen die kleineren Studien/Studien mit größerer Varianz stärker als die größeren Studien bzw. Studien mit geringerer Varianz um den wahren Effektschätzer, sodass eine symmetrische, spitz nach oben zulaufende Verteilung (umgekehrter Trichter) entsteht. Jede asymmetrische Verteilung ist verdächtig für einen Publikationsbias.

Mit einer umfassenden Literaturrecherche, die möglichst viele Datenbanken einschließt, ergänzt durch die Durchsicht der Referenzen, die Befragung von Experten, die Handsuche in einschlägigen Zeitschriften sowie der Suche nach unveröffentlichten Studienergebnissen kann ein Publikationsbias minimiert werden. Durch die Einführung von Studienregistern wird außerdem die Identifizierung nicht publizierter Studien erleichtert.

Die Beschränkung auf englischsprachige Publikationen (Sprachbias) scheint dagegen die Studienergebnisse nicht wesentlich zu verzerren[144, 146].

### 6.2.5 Gesundheitsökonomische Studien

Die Qualität gesundheitsökonomischer Studien wird bestimmt durch (a) die Validität der Studienergebnisse, (b) die Einhaltung methodischer Standards der gesundheitsökonomischen Evaluation und (c) den Zugang zu belastbaren Kosten- und Outcomedaten.

Die Einschätzung der Validität erfolgt wie bei Studien zur Wirksamkeit von Interventionen. Die methodischen Standards der gesundheitsökonomischen Evaluation werden in Standardlehrbüchern[61, 64, 191] und gesundheitsökonomischen Guidelines[1, 29, 74, 86, 87, 101, 172, 173, 212, 231] beschrieben. Die gesundheitsökonomische Evaluation basiert auf den theoretischen Konzepten der Wohlfahrtsökonomik und Entscheidungsanalyse. Die Standards der gesundheitsökonomischen Evaluation sind keine eineindeutigen Festlegungen. Es hat sich jedoch ein Konsens über konstitutive Elemente der gesundheitsökonomischen Evaluation und über zulässige Ansätze der Kostenanalyse und Outcomebestimmung herausgebildet. Teilweise wird in Guidelines explizit gefordert, alternative Ansätze zu kalkulieren.

In gesundheitsökonomischen Evaluationen sind die folgenden Elemente festzulegen[29, 64]:

- Studienform
- Vergleichsalternativen
- Perspektive der gesundheitsökonomischen Analyse
- Ressourcenkonsum und Kosten

- Effektivität und Nutzen
- Zeithorizont
- Modellierung
- Diskontierung
- Inkrementalanalyse
- Variabilität und Unsicherheit

Studienformen der gesundheitsökonomischen Evaluation sind (1) Kosten-Effektivitäts-Analysen (cost-effectiveness analysis [CEA]), (2) Kosten-Nutzwert-Analysen (cost-utility analysis [CUA]) und (3) Kosten-Nutzen-Analysen (cost-benefit-analysis [CBA]). Die Studienformen unterscheiden sich nach den einbezogenen Outcomeparametern:

- CEA: gesundheitliche Outcomes (wie Vermeidung eines Herzinfarkts) und Lebenserwartung
- CUA: qualitätsadjustierte Lebensjahre (QALY)
- CBA: Zahlungsbereitschaften und (sonstige) monetäre Outcomes.

In Guidelines zur gesundheitsökonomischen Evaluation wird empfohlen, die zu evaluierende – im Allgemeinen neue – Gesundheitstechnologie mit (1) der üblichen Versorgung (eventuell auch einem nach Marktanteilen gewichtetem Interventionsmix), (2) der effektivsten Intervention und (3) der bisher kosteneffektivsten Intervention zu vergleichen (Vergleichsalternativen). Es sollte auch (4) ein Vergleich mit der kostenminimalen Intervention (soweit möglich der Nichtintervention) erfolgen[2, 29, 231].

Die gesundheitsökonomische Evaluation kann aus unterschiedlichen Perspektiven durchgeführt werden, z. B. dem Blickwinkel der Gesellschaft, dem der Kostenträger oder dem der Patienten. Generell wird in Standardlehrbüchern auf die gesellschaftliche Perspektive abgestellt (um bei Allokationsentscheidungen ein gesellschaftliches Optimum zu realisieren)[64, 191]. Guidelines von HTA-Agenturen empfehlen teilweise die Perspektive der Kostenträger (für die der HTA-Bericht erstellt wird)[29].

Bei der Kostenbestimmung werden die gesamten Kosten (Einsparungen) berücksichtigt, die durch die Intervention induziert werden, soweit sie aus der eingenommenen Perspektive relevant sind. Es werden unterschieden (1) direkte medizinische Kosten (Kosten der Leistungserstellung durch das Gesundheitssystem), (2) direkte nicht-medizinische Kosten (Ressourcenkonsum von Patienten und Angehörigen, der die Erstellung medizinischer Leistungen im Gesundheitssektor unterstützt) und (3) indirekte Kosten (Produktivitätsverluste wegen krankheitsbedingter Arbeitsausfälle). Die Kostenbestimmung erfolgt in einem dreistufigen Prozess: (1) Identifikation der relevanten Kostenparameter, (2) Mengenerfassung des Ressourcenkonsums und (3) Bewertung der Ressourcen. Die Bewertung erfolgt zu Opportunitätskosten. Häufig sind nur staatlich administrierte Preise und/oder kollektiv-vertraglich vereinbarte Preise bekannt wie z. B. Gebührenordnungsziffern im ambulanten oder Fallpauschalen im stationären Sektor, die als Näherung genutzt werden – eventuell korrigiert um begründete Zu- und Abschläge, wie z. B. einen Zuschlag zu den akutstationären Tagespflegesätzen für die Investitionskosten. Bei der Bewertung indirekter Kosten existieren zwei alternative Bewertungsansätze: Humankapital- und Friktionskostenansatz[25, 64].

Bei der Bestimmung der Outcomes sind die relevanten Effekte und Nutzen zu identifizieren, um die Fragestellung der Evaluation zu beantworten. Zunehmend wird in gesundheitsökonomischen Studien auf gesundheitsbezogene Lebensqualität abgestellt. Die Lebensqualitätsbestimmung erfolgt durch direkte Nutzenbewertung (Standard gamble, Time-trade-off und Rating scale) oder indirekte Nutzenbewertung (standardisierte gesundheitsökonomische Lebensqualitätsinstrumente wie z. B. der EQ-5D [= EuroQol-Instrument der präferenzbasierten Lebensqualitätsmessung]). Für die Messung der Lebensqualität stehen verschiedene Nutzenkonzepte wie z. B. QALY oder DALY (behinderungskorrigierte Lebensjahre) zur Verfügung. Das QALY-Konzept ist am weitesten verbreitet. Es berücksichtigt sowohl die Lebensqualität, als auch die Lebenserwartung einer Person.

Der Zeithorizont einer gesundheitsökonomischen Evaluation sollte hinreichend lang sein, um alle relevanten Kosten- und Outcomeunterschiede zwischen den Vergleichsalternativen abzubilden. Bei chronischen Erkrankungen ist häufig ein lebenslanger Zeithorizont erforderlich, insbesondere wenn Lebenszeitgewinne erwartet werden[25, 29, 64, 138, 215].

Im Allgemeinen können in prospektiven Studien aber keine hinreichend langen Zeiträume überblickt werden, um die gesundheitsökonomisch relevanten Effektivitätsparameter – wie Auswirkungen auf die Lebenserwartung – abzubilden. Es ist dann erforderlich, die primäre Datenerhebung in der prospektiven Studie um Modellanalysen – basierend etwa auf epidemiologischen Studien – zu ergänzen. Formen der Modellierung sind z. B. Entscheidungsbaumanalysen, Markov-Modelle und Simulationsmodelle auf Patientenebene.

In der gesundheitsökonomischen Methodenliteratur besteht Konsens darüber, dass Kosten und Nutzen auf die aktuelle Periode zu diskontieren sind, um die zu unterschiedlichen Zeitpunkten anfallenden Kosten und den entstehenden Nutzen vergleichbar zu machen[64, 128]. Die Diskontierung wird begründet mit (1) der positiven Zeitpräferenz von Individuen und (2) den Opportunitätskosten des Kapitals. Eine positive Zeitpräferenz bedeutet, dass Individuen heutigen gegenüber zukünftigem Konsum bevorzugen. Begründet wird die Bevorzugung des heutigen Konsums mit der Unsicherheit der Individuen, ob sie zukünftige Zeitperioden erleben und damit zukünftigen Konsum realisieren werden. Gleichzeitig wird davon ausgegangen, dass in einer wachsenden Ökonomie das Konsumpotenzial in zukünftigen Zeitperioden ansteigt, was bei dem in der mikroökonomischen Theorie unterstellten abnehmenden Grenznutzen im Konsum zu einer Präferierung des heutigen Konsums führt[128]. Mit den Opportunitätskosten des Kapitals wird berücksichtigt, dass durch Investition von Ressourcen ein Konsumpotenzial in späteren Zeitperioden erzeugt wird, das den Ressourceneinsatz (Verzicht auf heutigen Konsum) übersteigt. Individuen werden aber nur bereit sein auf heutigen Konsum zu verzichten, wenn sie einen Ausgleich erhalten – dies sind die Zinsen auf das Kapital, die aus dem Zusatzkonsumpotenzial beglichen werden können. Umstritten ist, ob Nutzen analog den Kosten zu diskontieren sind. Teils wird in der Methodenliteratur argumentiert, dass Kosten stärker diskontiert werden sollten als Nutzen. Gesundheitsökonomische Guidelines empfehlen jedoch überwiegend, beide Parameter – Kosten und Nutzen – mit derselben Rate zu diskontieren. Auch die angemessene Diskontierungsrate ist umstritten. Häufig schlagen internationale Guidelines 3 % oder 5 % für Kosten und Nutzen vor[246].

Gesundheitsökonomische Analysen basieren auf einem Vergleich von Interventionsalternativen. Bei dem Alternativenvergleich wird nicht primär auf die gesamten Kosten und Outcomes (gesundheitliche Effekte, QALY oder monetärer Nutzen) der neuen Gesundheitstechnologie abgestellt, sondern auf die Kosten- und Outcomedifferenzen gegenüber den Vergleichsinterventionen (inkrementelle Kosten-Effektivitäts-Relation).

In empirischen Evaluationsstudien ist das Problem unsicherer Datengrundlagen der Kosten- und Outcomeparameter zu berücksichtigen. Die Unsicherheit in gesundheitsökonomischen Evaluationsstudien leitet sich ab aus[24, 29]

- Stichprobenvariationen bei patientenbasierten Daten in prospektiven Erhebungen
- Punktschätzungen bei Leistungs- und Ressourcenpreisen. Die Punktschätzungen basieren auf Unternehmensdaten, Sekundärdatenquellen und/oder Experteneinschätzungen
- Extrapolation von intermediären auf finale Outcomes (wie zum Beispiel von Bluthochdruck auf Lebenserwartung) in Modellanalysen.

In gesundheitsökonomischen Evaluationsstudien ist die Unsicherheit der Kosten- und Outcomeparameter auszuweisen. Die Auswirkungen von Parameterunsicherheiten, die auf Stichprobenerhebungen basieren, auf die inkrementelle Kosten-Effektivitäts-Relation lassen sich in statistischen Analysen überprüfen (z. B. Bootstrapping und Kosten-Effektivitäts-Akzeptanzkurven). Die sonstigen unsicheren Annahmen wie Punktschätzungen bei Leistungs- und Ressourcenpreisen sowie Extrapolationen werden in deterministischen und stochastischen Sensitivitätsanalysen untersucht[64], indem Annahmen über zentrale Parameter variiert und Auswirkungen auf das Evaluationsergebnis bestimmt werden.

### 6.2.6 Qualitätsbewertungsinstrumente

Die Qualitätsbewertung dient der Einschätzung der Glaubwürdigkeit von Studienergebnissen. Es existiert kein Goldstandard für die Bewertung der Studienqualität, da die wahren Zusammenhänge von Exposition/Intervention und Outcome unbekannt sind. Die Durchführung einer Qualitätsbewertung ist ein anspruchsvoller Prozess, der profunde methodische Kenntnisse erfordert.

Es gibt eine Vielzahl von QBI, die eine ausgeprägte Variabilität aufweisen, die überwiegend auf Expertenansicht und weniger auf empirischen Erkenntnissen und deren Konzeption, häufig nicht auf stringenten Methoden zur Instrumentenentwicklung basiert[132, 235].

**Klassifikation von QBI**

Zur Bewertung der Studienqualität werden prinzipiell drei Arten von Instrumenten verwendet: Skalen, Checklisten und die Komponentenbewertungen[56, 69, 143, 189].

Bei einer Skala erhält jedes Item einen numerischen Score, der zu einem Summenscore addiert wird. Neben dieser impliziten Gewichtung der Items durch einfache Addition werden in einigen Skalen mehr oder weniger komplexe Gewichtungen der einzelnen Items durchgeführt, abgeleitet von der jeweils zugeschriebenen Bedeutsamkeit des Items für die Validität. Als Beispiel für eine Skala sind im Anhang die von Downs & Black[60] sowie die von Jadad et al.[106] entwickelten Instrumente abgebildet (Kapitel 8.16.4).

Eine Checkliste ist eine Liste bestehend aus mindestens zwei Items ohne numerisches Bewertungssystem. Es gibt Checklisten, die einzelne Komponenten, die aus mehreren Items bestehen, qualitativ bewerten sowie Checklisten, die zu einer qualitativen Gesamtbewertung kommen. Als Beispiele für Checklisten ohne qualitative Bewertung sind die Checklisten der German scientific working group (GSWG)[69] und als Beispiel für Checklisten mit einer qualitativen Gesamtbewertung sind die Instrumente des Ludwig Boltzmann Instituts (LBI) im Anhang (Kapitel 8.16.1 und 8.16.2) dargestellt.

Die Komponentenbewertung enthält Komponenten wie „Randomisierung“ und „Verblindung“, die ebenfalls nicht numerisch, sondern qualitativ beurteilt werden. Die einzelnen Komponenten setzen sich oft aus mehreren Studienaspekten zusammen und benötigen daher in der Regel eine sehr ausführliche Operationalisierung, um zu einer eindeutigen Bewertung zu gelangen. Sowohl bei Checklisten als auch bei der Komponentenbewertung gibt es Instrumente, die anhand definierter Kriterien zu einer qualitativen Gesamtbewertung kommen und z. B. ein hohes, mittleres oder niedriges Risiko für systematische Verzerrungen angeben. Als Beispiel für eine Komponentenbewertung ist im Anhang das von der Cochrane Collaboration empfohlene QBI für RCT[92] abgebildet (s. 8.16.3 Beispiel für ein Komponentensystem).

Obgleich die Bildung eines einzigen numerischen Werts (Skalenbewertung) zur Abbildung der Studienqualität verlockend scheint, wird als Argument gegen Skalen eingewendet, dass durch die Gesamtbewertung die Ergebnisse einzelner Items nicht berücksichtigt werden[78, 108]. Außerdem fehlt für die implizite oder explizite Gewichtung von Items eine empirische Grundlage. Entsprechend gibt es Hinweise aus Untersuchungen, dass durch numerische Gesamtbewertungen die Validität von Studien nicht korrekt gemessen wird[89]. Andere Autoren zeigen, dass unterschiedliche Skalen zu verschiedenen Einschätzungen der Qualität führen[144]: Bei der Anwendung von sechs verschiedenen Skalen zur Qualitätsbewertung einer Studie variieren die Gesamtscores von 23 % bis 74 % der möglichen Gesamtpunktzahl. Diese große Variation in der Gesamtbewertung kann dazu führen, dass in Abhängigkeit von der verwendeten Skala und ihrer Schwellenwerte für den Einschluss von Studien unterschiedliche Studien als Evidenzbasis z. B. für Metaanalysen herangezogen werden. Auf diese Weise kann es zu verschiedenen Einschätzungen von Zusammenhängen kommen. Whiting et al.[236] weisen für ihr für die Qualitätsbewertung diagnostischer Studien entwickeltes Instrument QUADAS ebenfalls nach, dass unterschiedliche Gewichtungen zu verschiedenen Einschätzungen der Studienqualität führen. Aus den genannten Gründen besteht unter Experten, HTA-Organisationen (u. a. LBI) und anderen Autoren systematischer Übersichtsarbeiten (u. a. Cochrane Collaboration), weitgehend Konsens, keine Skalen zur Qualitätsbewertung zu verwenden.

**Durchführung der Qualitätsbewertung**

Die Bewertung der methodischen Qualität von Studien kann aufgrund der Subjektivität der Bewertung der einzelnen Items zu einer Varianz der Bewertung unterschiedlicher Reviewer führen. Besonders durch Items, die nicht nach dem Vorhandensein eines bestimmten Studienaspekts, sondern nach der Angemessenheit eines Vorgehens fragen, kann es zu unterschiedlichen Einschätzungen kommen. Um eine gute Übereinstimmung (Interrater-Reliabilität) zu erzielen und damit Unsicherheiten bei der Qualitätsbewertung zu minimieren, sind ein möglichst standardisiertes Vorgehen sowie eine präzise, eindeutige und ausführliche Operationalisierung der einzelnen Qualitätsparameter erforderlich.

Es ist allgemeiner Konsens, dass die Qualitätsbewertung von mindestens zwei Reviewern unabhängig voneinander durchgeführt und bei Unstimmigkeiten eine Diskussion und Konsensbildung ggf. unter Einbeziehung weiterer Personen durchgeführt werden sollte. Die Reviewer sollten zudem ein profundes Methodenwissen mitbringen.

Diskutiert wird, inwieweit eine Verblindung der Reviewer für die Autoren der zu bewertenden Studien und ggf. auch für die Publikationsquelle geeignet und erforderlich ist, einen Reviewerbias zu minimieren. Während eine Untersuchung ergibt, dass die Verblindung mit einer insgesamt niedrigeren und konsistenteren Qualitätsbewertung assoziiert ist[106], kommt es in einer anderen Studie zu einer höheren Qualitätsbewertung unter Verblindung[145]. Zwei weitere Studien können keine wesentlichen Unterschiede zwischen verblindeter und unverblinderter Bewertung finden[45, 227]. Es muss berücksichtigt werden, dass eine Verblindung nicht möglich ist, wenn ein Reviewer mit der zu bewertenden Literatur gut vertraut ist.

### 6.2.7 Integration der Qualitätsbewertung in die Informationssynthese

Es gibt unterschiedliche Möglichkeiten, die ermittelte Studienqualität bei der Synthese der Daten zu berücksichtigen[55, 59, 110, 235]. Die Ergebnisse der Qualitätsbewertung können anhand a priori definierter Kriterien (z. B. Cutpoints in Skalen) zum Ein- bzw. Ausschluss von Studien in die Datengrundlage eingesetzt werden. Außerdem können quantitative Qualitätsbewertungen zur Gewichtung von Studien oder stratifizierter Informationssynthese genutzt oder als Einflussfaktor in einer Metaregression oder Sensitivitätsanalyse berücksichtigt werden. Wenn die Studienqualität nicht quantitativ integriert wird, sollte sie zumindest qualitativ in die Informationssynthese einfließen, indem methodische Schwächen eingeschlossener Studien sowie deren Impact auf die Schlussfolgerungen und Empfehlungen herausgestellt bzw. diskutiert werden[235].

Es gibt wenige Untersuchungen, die die Art der Berücksichtigung der Studienqualität empirisch analysieren. Wie bereits erwähnt, kann gezeigt werden, dass unterschiedliche Skalen zu verschiedenen Qualitätsbewertungen und bei Verwendung der Studienqualität als Ein- bzw. Ausschlusskriterium zu divergierenden Informationssynthesen führen können[89, 144]. Herbison et al.[89] teilen in Metaanalysen eingeschlossene Studien mit 43 unterschiedlichen Skalen in Studien hoher und niedriger Studienqualität ein. Mit keinem Instrument sind die Ergebnisse der Metaanalysen näher an dem Referenzstandard bzw. an einer großen Studie. Die Autoren folgern, dass kein Qualitätsscore valide die Qualität misst und raten von einer Integration der Ergebnisse der Qualitätsbewertung auf diese Weise ab.

## 6.3 Fragestellungen

Es ergeben sich folgende drei, auf einander aufbauende Forschungsfragen:

1. Welche Bewertungsinstrumente zur Studienqualität, insbesondere für nicht-randomisierte Studien und Beobachtungsstudien gibt es? (Bestandsaufnahme)
   Es werden Instrumente eingeschlossen, die Interventions-, Beobachtungs-, Diagnose- und gesundheitsökonomische Studien bewerten.
2. Wie unterscheiden sich vorhandene Bewertungsinstrumente voneinander? (Vergleich)
   Es werden formale und inhaltliche Charakteristika von Instrumenten verglichen.
3. Welche Schlussfolgerungen können für die systematische Bewertung der Studienqualität abgeleitet werden? (Schlussfolgerungen)

## 6.4 Methoden

Das Methodenkapitel beschreibt die Verfahren und Methoden, die für Bewertungsinstrumente für Interventions-, Diagnose-, gesundheitsökonomische Studien sowie für den Workshop verwendet werden.

### 6.4.1 Bewertung von Studien zur Wirksamkeit

Im Folgenden werden die Methoden der Literaturrecherche und -auswahl bzw. der Datenextraktion und -synthese erläutert.

#### 6.4.1.1 Literaturrecherche

Um einen möglichst umfassenden und vollständigen Überblick der existierenden Instrumente zur Bewertung der Studienqualität zu erhalten, werden vielfältige Strategien zur Identifikation relevanter Publikationen eingesetzt. Die Literaturrecherche wird stufenweise durchgeführt. Grundlage bildet eine systematische Datenbankrecherche. Darauf aufbauend werden HTA-Berichte der Deutschen Agentur für Health Technology Assessment (DAHTA) gesichtet. Anschließend wird eine Internetrecherche bei der Cochrane Collaboration sowie bei internationalen HTA-Organisationen durchgeführt. Sonstige Quellen umfassen Instrumente, die anhand von Referenzen gefunden werden. Bereits identifizierte Instrumente werden in den nachfolgenden Suchschritten nicht erneut eingeschlossen.

**Systematische Datenbankrecherche**

Am 06.01.2009 wird von Art & Data Communications im Auftrag des DIMDI eine systematische Literaturrecherche durchgeführt. Die detaillierte Suchstrategie kann im Anhang eingesehen werden.

Folgende elektronische Literaturdatenbanken werden berücksichtigt:

NHS-CRD-HTA (INAHTA), DAHTA, NHS Economic Evaluation Database (NHSEED), NHS-CRD-DARE (CDAR94), Cochrane Library (CDSR93), MEDLINE (ME83), EMBASE (EM83), AMED (CB 85), BIOSIS Previews (BA83), MEDIKAT (MK77), Cochrane Library – Central (CCTR93), German Medical Science (GA03), SOMED (SM78), CAB Abstracts (CV72), Index to Scientific and Technical Proceedings (II98), ETHMED (ED93), GLOBAL Health (AZ72), Deutsches Ärzteblatt (AR969, EMBASE Alert (EA08), SciSearch (IS00), CCMed (CC00), Social SciSearch (IN73), Karger-Verlagsdatenbank (KR03), Kluwer-Verlagsdatenbank (KL97), Springer-Verlagsdatenbank (SP97), Springer-Verlagsdatenbank PrePrint (SPPP), Thieme-Verlagsdatenbank (TV01), Derwent Drug File (DD83), IPA (IA70).

**Screening von HTA-Berichten**

In der Datenbank der DAHTA wird eine systematische Recherche von HTA-Berichten sowie darin enthaltener QBI für Primär- und Sekundärstudien durchgeführt. Dies dient zum einen der Identifikation von Bewertungsinstrumenten, zum anderen soll ein Überblick über den bisherigen Einsatz dieser Instrumente aus dem deutschsprachigen Raum dargestellt werden. Ausgeschlossen werden daher Berichte internationaler Organisationen (National Coordinating Centre for Health Technology Assessment (NCCHTA), Canadian Agency for Drugs and Technologies in Health (CADTH; früher: Canadian Coordinating Office for Health Technology Assessment (CCOHTA), Agence Nationale d'Accréditation et d'Evaluation en Santé (ANAES), Health Technology Assessment International (HTAI)) deren Volltext nicht aus dem deutschsprachigen Raum stammt. Aufgrund der Themenstellung werden Berichte zur Bewertung von Leitlinien ausgeschlossen. Methodenberichte werden für die Suche nach Instrumenten verwendet. Sie fließen in die Darstellung der Nutzung der Qualitätsbewertung jedoch nur ein, sofern aufgrund der Themenstellung eine Qualitätsbewertung von Primär- und/oder Sekundärstudien anwendbar ist. Tagungsbände werden ausgeschlossen.

**Internetrecherche**

Es wird eine Internetrecherche, primär auf den Internetseiten der Cochrane Collaboration sowie nationaler und internationaler HTA-Organisationen, durchgeführt, um weitere QBI zu identifizieren. Neben den Informationen auf den Webseiten werden Dokumente wie beispielsweise Handbücher heruntergeladen und nach möglichen QBI durchsucht. Eine Liste aller gesichteten Internetseiten ist im Anhang (Kapitel 8.2) dokumentiert.

#### 6.4.1.2 Literaturauswahl

Zur Auswahl relevanter Publikationen aus der stufenweisen Literaturrecherche erfolgt a priori eine Definition von Ein- und Ausschlusskriterien.

**Einschlusskriterien**

- Der Volltext der Publikation ist in deutscher oder englischer Sprache verfasst.
- Die Publikation enthält die vollständige Darstellung eines oder mehrerer Instrumente zur methodischen Qualitätsbewertung von Primär- oder Sekundärstudien.
- Die Publikation ist eine Übersichtsarbeit zu Instrumenten der Qualitätsbewertung von Primär- oder Sekundärstudien.

- Die Publikation untersucht die Qualität eines oder mehrerer Instrumente der Qualitätssicherung von Primär- oder Sekundärstudien.
- Publikationszeitraum: 1988 bis 2009.

**Ausschlusskriterien**

- Der Volltext der Publikation ist nicht in deutscher oder englischer Sprache verfasst.
- Die Publikation ist vor 1988 erschienen.
- Die Publikation steht nur in Form eines Abstracts oder Posters zur Verfügung.
- Die Publikation ist ein Editorial, Leserbrief oder Kommentar.
- Die Publikation thematisiert Fallberichte, Fallstudien, genetische, tierexperimentelle oder in-vitro-Studien.
- Die Publikation enthält nur eines oder mehrere Instrumente, die nicht zur Anwendung im medizinischen und/oder epidemiologischen Bereich entwickelt sind.
- Die Publikation enthält nur eines oder mehre Instrumente zu pharmakologischen Studien zum Zeitpunkt vor der Marktzulassung.
- Die Publikation enthält ausschließlich eines oder mehrere Instrumente zur Bewertung von Leit- oder Richtlinien.
- Die Publikation enthält ausschließlich eines oder mehrere Instrumente zur Entscheidung über die Aufnahme eines Artikels in eine Zeitschrift (für Reviewer/Editoren).
- Die Publikation enthält Hinweise, Vorgaben, Richtlinien, Empfehlungen hinsichtlich der Durchführung einer Studie, aber kein QBI.
- Die Publikation enthält nur eines oder mehrere Instrumente zur Bewertung qualitativer Studien.
- Die Publikation enthält nur eines oder mehrere Instrumente, die ausschließlich die Modellierung in gesundheitsökonomischen Studien untersuchen.
- Die Publikation enthält ausschließlich eines oder mehrere Instrumente zur Bewertung der Berichtsqualität von Studien.
- Die Publikation erwähnt die Anwendung eines oder mehrerer Instrumente, ohne diese(s) vollständig darzustellen
- Die Publikation ist die ältere Version eines Instruments, für das eine aktuellere Version verfügbar ist.
- Doppelt gefundene Publikationen.

Die Beschränkung auf Publikationen der letzten 20 Jahre ist prinzipiell willkürlich. Sie hat jedoch das Ziel, den Suchzeitraum sinnvoll und effizient zu reduzieren ohne wesentliche Instrumente zu übersehen. Laut Moher et al.[143] wird die erste Checkliste 1961 publiziert, bis 1993 seien insgesamt neun veröffentlicht worden, die erste Skala wird 1981 publiziert, bis 1993 24 weitere. Ältere Instrumente werden ggf. über die eingeschlossenen Übersichtsarbeiten abgebildet. Darüber hinaus sind sie mitunter auch in Form einer Modifikation in neueren Instrumenten enthalten.

Publikationen, die die Messung der Testgüte eines bestehenden Instruments thematisieren, werden verwendet, um das Originalinstrument zu suchen, sofern es nicht bereits vorliegt. Zudem werden Daten zur Validität und Reliabilität extrahiert und später mit den Informationen der Originalpublikation zusammengeführt. Publikationen, in denen ein bereits bestehendes Instrument angewandt wird, werden ebenfalls zur Identifikation der Originalpublikation genutzt.

Die Fundstellen aus der systematischen Literaturrecherche werden durch zwei voneinander unabhängige Gutachter gesichtet, die mit den Methoden der evidenzbasierten Medizin (EbM) vertraut sind. Anschließend werden anhand der Abstracts relevante Publikationen identifiziert. Die vom DIMDI erhaltenen Volltexte werden ebenfalls durch die beiden voneinander unabhängigen Gutachter gesichtet. Bei abweichenden Bewertungen erfolgen eine Absprache und Konsensbildung.

#### 6.4.1.3 Datenextraktion

**Systematische Übersichtsarbeiten**

Charakteristika und Inhalte von Übersichtsarbeiten, die mehrere Instrumente systematisch miteinander vergleichen, werden systematisch extrahiert (Tabelle 38: Übersicht über die eingeschlossenen systematischen Übersichtsarbeiten (n = 8)) sowie narrativ und tabellarisch dargestellt. Eine Qualitätsbewertung der eingeschlossenen Publikationen wird mit einer modifizierten Checkliste nach West et al.[235] von zwei voneinander unabhängigen Reviewern durchgeführt. Bei fehlender Übereinstimmung wird ein Konsens herbei geführt.

**Einzelne QBI: Formale Aspekte**

Die Datenextraktion von Primärstudien wird instrumentenbezogen durchgeführt. Extrahiert wird somit nicht das Dokument, sondern das bzw. die darin enthaltene(n) QBI. Das vollständige Formular der Datenextraktion ist im Anhang einsehbar (Kapitel 8.6). Zu den formalen Eigenschaften gehört zunächst eine Unterscheidung in originale bzw. modifizierte Instrumente. Instrumente, die exakt ein bestehendes System modifizieren, werden als modifizierte Instrumente bezeichnet. Systeme, die keine Modifikation erwähnen oder mehr als ein Bewertungsinstrument zugrunde legen, werden in Anlehnung an West et al.[235] als Originalinstrumente definiert.

Des Weiteren erfolgt eine Unterscheidung der Instrumente in Checklisten, Komponentensysteme sowie Skalen. Skalen werden als Systeme mit einer numerischen Bewertung der einzelnen Items sowie einer Bewertung der Gesamtqualität definiert. Checklisten werden bestimmt als Listen von Items mit Antwortvorgaben, die nicht numerisch bewertet werden. Die Komponentenbewertung enthält als Items Komponenten wie „Randomisierung“ und „Verblindung“, die nicht numerisch, sondern qualitativ bewertet werden. Die einzelnen Komponenten setzen sich in der Regel aus mehreren Studienaspekten zusammen. Sowohl bei Checklisten als auch bei der Komponentenbewertung gibt es Instrumente, die anhand definierter Kriterien zu einer qualitativen Gesamtbewertung kommen, z. B. hohes, mittleres oder niedriges Risiko für systematische Verzerrungen.

Sofern die Anwendung des Instruments von den Autoren für eine spezielle Exposition/Intervention, einen Test oder ein Outcome angegeben wird, das Instrument als „spezifisch“ eingestuft. Andernfalls wird es als generisches Instrument gesehen. Prinzipiell sind jedoch viele der spezifisch eingesetzten Instrumente auch in anderen Bereichen anwendbar, da sie keine spezifizierenden Items enthalten. Bei Skalen wird neben dem Wertebereich auch, falls vorhanden, ein Cutpoint zur Abgrenzung von guter vs. schlechter Qualität dokumentiert.

Weiterhin werden Definitionen des Konzepts Qualität, Informationen zum Entwicklungsprozess sowie Hinweise zur Anwendung des Instruments bzw. der Operationalisierung und der benötigte Zeitbedarf nach Autorenangaben extrahiert. Angaben zu Reliabilität und Validität werden dokumentiert, sofern sie sich auf das vollständige, finale Bewertungsinstrument und nicht auf einzelne Bestandteile des Instruments beziehen. Die Datenextraktion wird von zwei unabhängigen Reviewern vorgenommen, bei Differenzen erfolgen eine Diskussion und Konsensbildung.

**Einzelne QBI: Inhaltliche Aspekte**

Aus den Publikationen werden neben formalen Aspekten auch inhaltliche Elemente erhoben. Diese Extraktion erfolgt anhand eines für das jeweilige Studiendesign spezifischen Formulars, das (mit Ausnahme der Extraktion im Bereich Diagnosestudien) zum größten Teil auf dem Datenextraktionsformular von West et al.[235] basiert.

Der HTA-Bericht von West et al.[235] wird im Auftrag der Agency for Healthcare Research and Quality (AHRQ) durchgeführt. Die Domänen und Items sind spezifisch für vier Studiendesigns: Systematische Reviews, RCT, Beobachtungs- und diagnostische Studien. Bei der Datenextraktion bewerten West et al.[235] das Vorhandensein der einzelnen Items mit „ja“ oder „nein“ und fassen die Abdeckung der einzelnen Domäne, denen jeweils ein oder mehrere Elemente zugeordnet sind, mit „ja“, „teilweise“ oder „nein“ zusammen.

Abweichend davon wird im vorliegenden Bericht auf die zusammenfassende Bewertung auf der Ebene der Domänen verzichtet, da diese nicht in Gänze nachvollziehbar ist und die Formulare von West et al.[235] um zusätzliche Items ergänzt werden. Diese stammen in der Regel aus weiteren Übersichtsarbeiten zu Bewertungsinstrumenten[168, 190, 239]. Darüber hinaus werden aggregierte Items in

die einzelnen Aspekte aufgeteilt. Extrahiert wird auf der Basis des studiendesignspezifischen Datenextraktionsformulars das Vorhandensein (ja/nein) der einzelnen Items im zu extrahierenden Bewertungsinstrument. Die Datenextraktion wird von zwei unabhängigen Reviewern vorgenommen, bei Differenzen erfolgen eine Diskussion und eine Konsensbildung.

Die Items der Formulare zu inhaltlichen Kriterien dienen nicht ausschließlich der Erfassung möglicher Quellen von Bias und Confounding, sondern beinhalten in Anlehnung an die Methodik der aufgeführten Übersichtsarbeiten auch Items, die das wissenschaftliche Vorgehen dokumentieren und damit eher der Berichtsqualität entsprechen und weniger Bezug zur internen Validität haben (z. B. angemessene und präzise Studienfrage, Power-Berechnung).

Es wird explizit darauf hingewiesen, dass diese Formulare keine Checklisten zur Qualitätsbewertung von Primär- bzw. Sekundärstudien darstellen, sondern einzig der Datenextraktion dienen. So werden sowohl detaillierte als auch übergeordnete Items verwendet, um den sehr unterschiedlichen Bewertungsinstrumenten gerecht zu werden und die Items möglichst gut abbilden zu können. Die verwendeten Datenextraktionsformulare finden sich im Anhang (Kapitel 8.8 Datenextraktionsformular für systematische Reviews, 8.9 Datenextraktionsformular für Interventionsstudien, 8.10 Datenextraktionsformular für Beobachtungsstudien, 8.11 Datenextraktionsformular für Diagnosestudien). Nachfolgend wird ein Überblick über die Modifikation der Formulare gegeben.

**Systematische Reviews, HTA und Metaanalysen**

Für die inhaltliche Datenextraktion von QBI für systematische Reviews, HTA-Berichten und Metaanalysen wird das Formular von West et al.[235] als Grundlage herangezogen. Items aus den Domänen „Fragestellung“, „Datensynthese und -analyse“, „Ergebnisse“, „Diskussion“ sowie „Finanzielle Förderung/Auftraggeber“ werden übernommen. Die Domäne „Ein-/und Ausschlusskriterien“ wird in zwei Items aufgeteilt, sodass die Angemessenheit der Kriterien und ihre a priori Definition separat voneinander abgefragt werden können, da auf der Basis von West et al. unklar ist, wann das Item als vorhanden bzw. nicht vorhanden gewertet wird. Der Bereich der Suchstrategie wird umbenannt in „Literaturrecherche und -auswahl“ und inhaltlich ergänzt, u. a. um den Einbezug weiterer Datenquellen sowie die Extraktion durch zwei unabhängige Reviewer. In der Domäne „Datenextraktion“ werden die Extraktion von Informationen zu Intervention/Exposition und Outcome separiert. Die Anzahl der extrahierenden Reviewer wird auf mindestens zwei festgelegt. Die Domäne „Studienqualität und Validität“ wird ergänzt um eine Bewertung durch mindestens zwei Reviewer, die verblindete Durchführung der Bewertung sowie die Bewertung der Übereinstimmung der Reviewer. Das aggregierte Item von West et al.[235] zur Diskussion wird in zwei Items zu ergebnisbasierten Schlussfolgerungen und der Diskussion des Einflusses von Bias und Confounding aufgeteilt.

Das Datenextraktionsformular für die QBI für systematische Reviews, HTA und Metaanalysen beinhaltet die folgenden Domänen und Elemente:

Studienfrage

- Fragestellung präzise und angemessen

Ein- und Ausschlusskriterien

- A-priori-Definition von Kriterien
- Kriterien sind angemessen

Literaturrecherche und -auswahl

- Relevante Datenbanken einbezogen
- Einbezug weiterer Datenquellen (z. B. Handsuche, graue Literatur, Referenzen, persönlicher Kontakt)
- Kombination von Schlagworten/Thesaurus und Freitext
- Vielzahl von Synonymen
- Restriktionen bei der Suche sind akzeptabel (z. B. Sprache, Land, Zeitraum)
- Dokumentation der verwendeten Suchterme und Datenbanken
- Literaturauswahl unabhängig voneinander durch mindestens zwei Reviewer
- Ausschluss von Literatur begründet

- Ausreichend detailliert, um die Literaturrecherche/-auswahl zu reproduzieren

Datenextraktion

- Extraktion von Interventionen/Expositionen für alle relevanten (Sub-) Gruppen
- Extraktion von Outcomes für alle relevanten (Sub-) Gruppen
- Datenextraktion unabhängig voneinander durch mindestens zwei Reviewer
- Datenextraktion verblindet für Reviewer (z. B. Autoren, Zeitschrift, Jahr, Ergebnisse)
- Messung der Übereinstimmung der Reviewer
- Ausreichend detailliert, um die Datenextraktion zu reproduzieren

Studienqualität/interne Validität

- Bewertungsmethode wird beschrieben, ist angemessen
- Bewertung unabhängig voneinander durch mindestens zwei Reviewer
- Bewertung verblindet für Reviewer (z. B. Autoren, Zeitschrift, Jahr)
- Bewertung der Übereinstimmung der Reviewer
- Methode zur Integration der Ergebnisse der Qualitätsbewertung ist angemessen

Datensynthese und -analyse

- Angemessene qualitative und/oder quantitative Synthese
- Berücksichtigung der Robustheit der Ergebnisse und/oder mögliche Heterogenität
- Darstellung von Schlüsselelementen von Primärstudien, die ausreichend sind für eine kritische Bewertung und Wiederholung

Ergebnisse

- Narrative Zusammenfassung und/oder quantitative Zusammenfassung und Angabe der Präzision, wenn angemessen

Diskussion

- Schlussfolgerungen werden durch die Ergebnisse unterstützt
- Berücksichtigung möglicher Bias und anderen Limitationen

Finanzielle Förderung/Auftraggeber

- **Art und Quelle der Finanzierung**

Die fettgedruckten Elemente beinhalten modifiziert nach West et al.[235] Elemente, die als besonders relevant für die Bewertung der internen Validität eingeschätzt werden. Bei den Elementen handelt es sich bis auf „Extraktion von Interventionen/Expositionen für alle relevanten Gruppen“, „Extraktion von Outcomes für alle relevanten Gruppen“ und „Bewertung unabhängig voneinander durch mindestens zwei Reviewer“ um empirisch nachgewiesene Biasquellen[235] (s. Kapitel 4.2 Wissenschaflticher Hintergrund). Ergänzend zu West et al.[235] zeigt eine Studie, dass die Datenextraktion durch zwei unabhängige Reviewer weniger Fehler produziert als die Datenextraktion durch einen Reviewer, die durch einen zweiten lediglich kontrolliert wird[27]. Dieses Ergebnis wird auf die Durchführung der Qualitätsbewertung übertragen.

**Interventionsstudien**

Für den vorliegenden Bericht wird das Formular von West et al.[235] als Grundlage für die Datenextraktion von Instrumenten zur Bewertung von Interventionsstudien herangezogen und um Items von Olivo et al.[168] sowie Aspekte der externen Validität ergänzt. Items aus den Domänen „Studienfrage“, „Studienpopulation“, „Randomisierung“ und „Finanzielle Förderung/Auftraggeber“ werden von West et al.[235] übernommen. In der Domäne „Verblindung“ wird das Item „Doppelte Verblindung“ aufgeteilt in die Verblindung der Studienteilnehmer und die der Erheber sowie ergänzt um die Items „Verblindung des übrigen Studienpersonals“ von Olivo et al.[168] und die Überprüfung der Verblindung. Die Domäne „Interventionen“ wird ergänzt um drei Items aus dem Formular von Olivo et al.[168], in dem die Vermeidung bzw. Beschreibung von Kointerventionen sowie die Vermeidung bzw. Akzeptabilität von Kontamination erfragt wird. Hinzugefügt wird auch die Vergleichbarkeit von Placebo und Verum sowie der Aspekt der gleichzeitigen Kontrollgruppe. Im Bereich „Outcome“ werden die Validität und Reliabilität

von West et al.[235] aufgesplittet in zwei separate Items. Es werden zwei Items von Olivo et al.[168] hinzugefügt, die die Länge und Gleichzeitigkeit des Follow-up erheben. In den Domänen „Statistische Analyse“ sowie „Ergebnisse“ werden aus einem aggregierten Item bei West et al.[235] drei einzelne Items gebildet, die die Angemessenheit der statistischen Analyse, den Anteil der Studienabbrecher, den Loss-to-follow-up, fehlende Werte sowie die Intention-to-treat-Analyse betreffen. Hinzugefügt wird der Aspekt des multiplen Testens sowie das von Olivo et al.[168] stammende Item zu Ursachen von Drop-outs. Das aggregierte Item von West et al.[235] zur Diskussion wird auf zwei Items zu ergebnisbasierten Schlussfolgerungen und der Diskussion des Einflusses von Bias und Confounding aufgeteilt.

Das Datenextraktionsformular für die QBI für Interventionsstudien beinhaltet die folgenden Domänen und Elemente:

Studienfrage

- Fragestellung präzise und angemessen

Studienpopulation

- Beschreibung der Studienpopulation
- Spezifische Ein- und Ausschlusskriterien
- Angemessene Stichprobengröße für alle Gruppen (Power-Berechnung)

Randomisierung

- Methode der Randomisierung beschrieben und angemessen
- Gruppenzuweisung geheim
- Vergleichbarkeit der Gruppen zu Beginn

Verblindung

- Verblindung der Studienteilnehmer
- Verblindung der Untersucher/Erheber des Outcomes
- Verblindung des übrigen Studienpersonals (z. B. Betreuer, Behandler)
- Verblindung überprüft und ausreichend

Interventionen

- Interventionen eindeutig und detailliert für alle Gruppen beschrieben
- Gleichzeitige Kontrollgruppe
- Behandlungsgleichheit bis auf die Intervention
- Placebo vergleichbar mit Verum (Darreichungsform, Aussehen, Geschmack, Geruch)
- Kointerventionen vermieden
- Kointerventionen für alle Gruppen beschrieben
- Kontamination vermieden/akzeptabel
- Compliance akzeptabel in allen Gruppen

Outcomes

- Primäre und sekundäre Outcomes präzise definiert
- Verwendete Methoden sind valide
- Verwendete Methoden sind reliabel
- Follow-up mit angemessener Länge
- Follow-up gleichzeitig in allen Studiengruppen

Statistische Analyse

- Angemessene statistische Analyse
- Viele Vergleiche sind berücksichtigt worden (Multiples Testen)
- **Intention-to-treat-Analyse**
- **Angemessener Umgang mit fehlenden Werten**

- Berücksichtigung von Confounding
- Bewertung von Confounding angemessen
- Bewertung von Heterogenität, wenn anwendbar

Ergebnisse

- Effekte hinsichtlich des Outcomes mit Punktschätzer und Präzision angegeben
- Anteil Studienabbrecher/Loss-to-follow-up angegeben und akzeptabel
- Unterschiede Teilnehmer/Abbrecher geprüft und akzeptabel
- Ursachen für Drop-outs/Loss-to-follow-up dargestellt
- **Selektives Berichten von Outcomes (ungeplante Analysen oder geplante/erwartete Analysen werden nicht berichtet)**
- Vorzeitiger Abbruch der Studie (aufgrund von Zwischenergebnissen)

Diskussion

- Schlussfolgerung werden durch Ergebnisse unterstützt
- Möglicher Einfluss von Confounding und Bias wird diskutiert

Externe Validität

- Anteil Nichtteilnehmer angegeben und akzeptabel
- Unterschiede Teilnehmer/Nichtteilnehmer geprüft und akzeptabel
- Studienpopulation repräsentativ

Finanzielle Förderung/Auftraggeber

- **Art und Quelle der Förderung**

Die fettgedruckten Elemente beinhalten, modifiziert nach West et al.[235], Elemente, die als besonders relevant für die Bewertung der internen Validität eingeschätzt werden. Es handelt sich bei den ausgewählten Elementen um empirisch nachgewiesene Biasquellen (s. auch Kapitel 4.2 Wissenschaftlicher Hintergrund).

**Beobachtungsstudien**

Für den vorliegenden Bericht wird das Formular von West et al.[235] als Grundlage der Datenextraktion von Instrumenten zur Bewertung von Beobachtungsstudien herangezogen und um Items von Saunders et al.[190] sowie Aspekte der externen Validität ergänzt. Items für die Domänen „Studienfrage“, „Studienpopulation“, „Exposition“, „Diskussion“ und „Finanzielle Förderung/Auftraggeber“ werden aus dem Formular von West et al.[235] übernommen. Im Bereich „Outcome“ wird der Aspekt der identischen Outcomemessung in allen Studiengruppen aufgenommen. Die Domäne „Statistische Analyse“ wird um ein Item zu fehlenden Werten erweitert. Im Bereich „Ergebnisse“ werden zusätzlich die Ursachen von Drop-outs eingefügt sowie von Saunders et al.[190] Items zum Loss-to-follow-up und den Unterschieden Teilnehmern vs. Abbrechern übernommen. Die Domäne „Externe Validität“ beinhaltet den Anteil Nichtteilnehmer von West et al.[235], sowie die beidem Items zu den Unterschieden Teilnehmer vs. Nichtteilnehmer und der Repräsentativität von Saunders et al.[190].

Das Datenextraktionsformular für die QBI für Beobachtungsstudien beinhaltet die folgenden Domänen und Elemente:

Studienfrage

- Fragestellung präzise und angemessen

Studienpopulation

- Beschreibung der Studienpopulation
- Spezifische Ein- und Ausschlusskriterien für alle Gruppen
- Identische Ein- und Ausschlusskriterien für alle Gruppen
- Angemessene Stichprobengröße für alle Gruppen (Power-Berechnung)
- **Gleichzeitige Kontrollgruppe**

- **Vergleichbarkeit der Studiengruppen untereinander zu Beginn (Krankheitsstatus und prognostische Faktoren)**
- Fall-Kontrollstudie: Explizite Falldefinition
- Fall-Kontrollstudie: Kontrollen gleichen den Fällen bis auf das interessierende Outcome, Kontrollen haben die gleiche Expositionschance wie die Fälle

Exposition

- Exposition/Intervention präzise definiert
- Methoden zur Erhebung der Exposition sind valide
- Methoden zur Erhebung der Exposition sind reliabel
- **Expositionsmessung gleich in allen Studiengruppen**
- Fall-Kontrollstudie: Expositionserhebung verblindet für Outcomestatus

Outcome

- Primäre und sekundäre Outcomes präzise definiert
- Methoden zur Erhebung des Outcomes sind valide
- Methoden zur Erhebung des Outcomes sind reliabel
- Fall-Kontrollstudie: Diagnosesicherung unbeeinflusst von Expositionsstatus (verblindet)
- Erhebung des Outcomes verblindet für Expositions- oder Interventionsstatus
- **Outcomemessung gleich in allen Studiengruppen**
- Follow-up mit angemessener Länge
- Länge des Follow-up gleich für alle Studiengruppen

Statistische Analyse

- Angemessene statistische Analyse
- Viele Vergleiche sind berücksichtigt worden (Multiples Testen)
- **Modellierung und/oder multivariate Methoden angemessen (Confounderkontrolle)**
- Angemessener Umgang mit fehlenden Werten
- Bewertung von Confounding/Residual Confounding
- Bewertung von Heterogenität (Effektmodifikation/Interaktion), wenn anwendbar
- Bestimmung der Dosiswirkungsbeziehung wenn möglich

Ergebnisse

- Effekte hinsichtlich des Outcomes mit Punktschätzer und Präzision angegeben
- Anteil Studienabbrecher/Loss-to-follow-up angegeben und akzeptabel
- Unterschiede Teilnehmer/Abbrecher geprüft und akzeptabel
- Ursachen für Drop-outs/Loss-to-follow-up dargestellt
- **Selektives Berichten von Outcomes (ungeplante Analysen oder geplante/erwartete Analysen werden nicht berichtet)**

Diskussion

- Schlussfolgerung werden durch Ergebnisse unterstützt
- Möglicher Einfluss von Confounding und Bias wird diskutiert

Externe Validität

- Anteil Nichtteilnehmer angegeben und akzeptabel
- Unterschiede Teilnehmer/Nichtteilnehmer geprüft und akzeptabel
- Studienpopulation repräsentativ

Finanzielle Förderung/Auftraggeber

- **Art und Quelle der Förderung**

Die fettgedruckten Elemente beinhalten modifiziert nach West et al.[235] Elemente, die als besonders relevant für die Bewertung der internen Validität eingeschätzt werden. Bei den Elementen „Gleichzeitige Kontrollgruppe", „Selektives Berichten" und „Quelle und Art der Förderung" handelt es sich um empirisch nachgewiesene Biasquellen (s. Kapitel 4.2 Wissenschaftlicher Hintergrund). Die übrigen Elemente werden als allgemein akzeptierte Einflussfaktoren der internen Validität zur Auswahl hinzugefügt.

**Diagnosestudien**

Für die inhaltliche Extraktion von Elementen aus QBI für Diagnosestudien werden die Kriterien von Whiting et al.[237] aus einem umfangreichen HTA-Bericht zu QBI für Diagnosestudien genutzt. Diese Entscheidung basiert auf dem deutlich umfangreicheren Extraktionsinstrument von Whiting et al.[237] gegenüber dem von West et al.[235]. Eine geringfügige Modifikation am Datenextraktionsformular wird vorgenommen. So lautet das Item zum Reviewbias bei Whiting et al.[237] „Werden die Ergebnisse des Indextests ohne Kenntnis der Ergebnisse des Referenztests interpretiert und andersherum?". Aus dieser aggregierten Frage werden zwei Items gebildet: „Werden die Ergebnisse des Indextests ohne Kenntnis der Ergebnisse des Referenztests interpretiert?" sowie „Werden die Ergebnisse des Referenztests ohne Kenntnis der Ergebnisse des Referenztests interpretiert?".

Das Datenextraktionsformular für die QBI für Diagnosestudien beinhaltet die folgenden Domänen und Items:

Verzerrungspotenzial

- **Wird ein angemessener Referenztest verwendet um den Zielparameter zu erfassen?**
- Kann eine Änderung des Krankheitsstatus zwischen Durchführung des Index- und des Referenztests aufgetreten sein?
- **Wird bei allen Teilnehmern der Zielparameter mit dem gleichen Referenztest verifiziert? (partieller oder differenzieller Verifikationsbias)**
- Ist der Indextest Teil des Referenztests? (Sind die Tests nicht unabhängig voneinander?)
- Wird die Behandlung basierend auf dem Ergebnis des Indextests eingeleitet bevor der Referenztest erfolgte?
- **Wird das Ergebnis des Indextests verblindet gegenüber dem Resultat des Referenztests ausgewertet? (Reviewbias)**
- **Wird das Ergebnis des Referenztests verblindet gegenüber dem Resultat des Indextests ausgewertet? (Reviewbias)**
- **Sind klinische Informationen vorhanden bei der Auswertung der Testergebnisse vorhanden? (Klinischer Reviewbias)**
- **Ist es wahrscheinlich, dass eine Beobachter-/Instrumentenvariabilität Annahmen bei der Testausführung beeinflusst hat?**
- Werden nicht bewertbare Testergebnisse in die Analyse eingeschlossen?

Externe Validität

- **Ist die untersuchte Bevölkerung vergleichbar mit der interessierenden Population? (Klinische und demografische Subgruppen)**
- Ist die Rekrutierungsmethode angemessen um ein geeignetes Spektrum an Patienten einzuschließen?
- **Sind die Prävalenz der Erkrankung und das Spektrum der Krankheitsschwere in der Studienpopulation vergleichbar mit denen der interessierenden Population?**
- Ist es wahrscheinlich, dass die Methode des Tests im Laufe der Studie verändert wurde?

Studiendurchführung

- Sind Subgruppenanalysen angemessen und vorab spezifiziert worden?
- Wird eine angemessene Teilnehmerzahl in die Studie eingeschlossen? (Power)
- Sind die Studienziele relevant für die Studienfrage?
- Wird ein Studienprotokoll vor Studienbeginn entwickelt und wird dies befolgt?

Berichtsqualität

- Werden die Einschlusskriterien präzise dargestellt?
- Wird die Methodik des Indextests detailliert genug beschrieben, um die Replikation des Tests zu ermöglichen?
- Wird die Methodik des Referenztests detailliert genug beschrieben, um die Replikation des Tests zu ermöglichen?
- Haben die Autoren präzise dargestellt, was als „normales" Testergebnis bewertet wird?
- Werden adäquate Ergebnisse dargestellt? Z. B. Sensitivität, Spezifität, Likelihood ratios.
- Wird die Präzision der Ergebnisse dargestellt? Z. B. Konfidenzintervalle.
- Werden alle Studienteilnehmer bei der Analyse berücksichtigt?
- Wird eine Kreuztabelle zur Testdurchführung dargestellt?
- Werden Hinweise gegeben, wie nützlich der Test in der Praxis sein könnte?

Die fettgedruckten Fragen beinhalten Elemente, die als besonders relevant für die Bewertung der internen Validität eingeschätzt werden. Die Auswahl basiert auf Ergebnissen eines HTA-Berichts, der u. a. methodologische Literatur über Biasquellen in Diagnosestudien identifiziert hat[237]. Danach gibt es die meiste empirische Evidenz für folgende Biasquellen: Variation von klinischen und demografischen Subgruppen, Krankheitsprävalenz und/oder -schwere, einen partiellen Verifikationsbias, klinischer Reviewbias und die Variation bei der Testbewertung (Interobserver variation). Etwas weniger deutlich ist die empirische Evidenz für eine verzerrte Teilnehmerauswahl, nicht vorhandener oder unangemessener Referenzstandard, differenzieller Verifikationsbias und Reviewbias.

#### 6.4.1.4 Datensynthese

Ziel der Synthese ist es, aus der Vielzahl der QBI anhand der extrahierten Charakteristika Instrumente zu benennen, die besonders geeignet für die Qualitätsbewertung erscheinen.

Für diese Auswahl werden a priori Anforderungen an ein QBI formuliert:

- Es sollte ein **generisches Instrument** sein, das auf unterschiedliche Forschungsbereiche anwendbar ist.
- Die Auswahl erfolgt **designspezifisch** für systematische Reviews/HTA/Metaanalysen, Interventions-, Beobachtungs- und Diagnosestudien.
- Skalen werden berücksichtigt; da für die implizite Gewichtung der Items keine empirische Evidenz vorliegt, sollten sie jedoch wie Checklisten ohne numerische Bewertung angewandt werden.
- Es sollten **Ausfüllhinweise** vorhanden sein.
- Das Instrument sollte möglichst **viele Items zur internen Validität** abdecken.
- Die Bearbeitungszeit sollte angemessen sein.
- Das Instrument sollte gute Interrater-Reliabilitäten haben.

Aus diesen formulierten Kriterien an die Anforderung für ein geeignetes QBI wird folgendes Vorgehen abgeleitet: Es werden designspezifisch die generischen Instrumente und ihre Items zur internen Validität dargestellt und ausgezählt,

- wie viele Items zur internen Validität abgedeckt werden.
- wie viele Domänen mit mindestens einem Item abgedeckt werden.
- bei wie vielen Domänen mindestens die Hälfte der Items abgedeckt wird.
- wie viele als relevant definierte Elemente abgedeckt werden.

Für die Extraktion inhaltlicher Elemente liegen für Diagnosestudien keine Domänen wie bei den anderen Studiendesigns vor, die Oberbegriffe von Elementen bei der Planung, Durchführung und Auswertung einer Studie darstellen. Die einzelnen Elemente in Diagnosestudien werden vielmehr nach Whiting et al.[237] unter den Begriffen (1) Verzerrungspotenzial, (2) Studiendurchführung, (3) Berichts-

qualität und (4) externe Validität zusammengefasst. Für Diagnosestudien wird daher nur angegeben, wie viele Items insgesamt und wie viele der als relevant definierten Elemente abgedeckt werden.

Es werden keine Cutpoints a priori formuliert. Die Ergebnisdarstellung erfolgt qualitativ, es werden umfassendere von weniger umfassenden generischen Instrumenten unterschieden.

### 6.4.2 Bewertung gesundheitsökonomischer Studien

Im Folgenden werden die Methoden der Literaturrecherche und -auswahl sowie der Datenextraktion und -synthese erläutert.

#### 6.4.2.1 Literaturrecherche

Um einen möglichst umfassenden Überblick der existierenden Instrumente zur Bewertung der Studienqualität gesundheitsökonomischer Studien zu bekommen, wird eine schrittweise Literaturrecherche durchgeführt. Grundlage bildet eine systematische Datenbankrecherche. Darauf aufbauend werden HTA-Berichte der DAHTA gesichtet. Zusätzlich wird eine Internetrecherche, vorwiegend bei nationalen und internationalen HTA-Organisationen sowie der Cochrane Collaboration, durchgeführt. Instrumente aus einer initialen Literaturrecherche sowie aus der Suche in Referenzen werden unter sonstige Quellen subsumiert. Das Vorgehen bei der systematischen Literaturrecherche, dem Screening der HTA-Berichte und der Internetrecherche entspricht demjenigen bei epidemiologischen Studien. Erweitert wird die Internetrecherche um Webseiten, die spezifisch für den Bereich Gesundheitsökonomie sind. Die gesichteten Seiten sind im Anhang (Kapitel 8.4 Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Ökonomie)) dokumentiert.

#### 6.4.2.2 Literaturauswahl

Die Auswahl der Literatur erfolgt ebenfalls analog zu den Kriterien für epidemiologische Studien. Ausgeschlossen werden zudem Publikationen, die lediglich ein bestehendes Instrument nutzen. Die gefundenen Publikationen aus der systematischen Literaturrecherche werden durch zwei voneinander unabhängige Gutachter gesichtet, die mit den Methoden der Durchführung von gesundheitsökonomischen Studien vertraut sind. Relevante Publikationen werden anhand der Titel und Abstracts identifiziert. Die vom DIMDI erhaltenen Volltexte werden durch die beiden voneinander unabhängigen Gutachter gesichtet. Bei abweichenden Bewertungen erfolgen Diskussion und Konsensbildung.

#### 6.4.2.3 Datenextraktion und -synthese

Die Datenextraktion gesundheitsökonomischer Studien wird analog zum Vorgehen bei epidemiologischen Studien instrumentenbezogen durchgeführt. Dabei werden sowohl formale, als auch inhaltliche Elemente extrahiert.

Das Formular der Datenextraktion für formale Eigenschaften sowie die dazugehörige Operationalisierung entsprechen den im epidemiologischen Teil des Berichts verwendeten. Sie sind im Anhang (Kapitel 8.4 Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Ökonomie)) dokumentiert. Lediglich die Unterteilung in verschiedene Studiendesigns wird nicht übernommen.

Für die Datenextraktion der inhaltlichen Elemente wird ein Formular für gesundheitsökonomische Studien entwickelt, da keine Übersichtsarbeiten vorhanden sind, die als Referenz dienen können. Im ersten Schritt des Entwicklungsprozesses werden Standardlehrbücher[61, 64, 191] sowie aktuelle nationale und internationale Leitlinien zur Erstellung gesundheits- und pharmakoökonomischer Studien gesichtet und auf verschiedene Themenschwerpunkte (Elemente der gesundheitsökonomischen Evaluation) hin untersucht[1, 29, 74, 86, 87, 101, 172, 173, 212, 231]. Inhaltlich behandeln die Lehrbücher und Leitlinien weitgehend identische Themenschwerpunkte, wobei einige Leitlinien auf nationale oder institutionsspezifische Besonderheiten abgestimmt sind. In einem zweiten Schritt werden in einer Diskussionsrunde der Autoren die herausgearbeiteten Themenschwerpunkte auf den Bezug zur Studienqualität (interne Validität) gesundheitsökonomischer Studien hin untersucht. Es werden Domänen und Items entwickelt, die auf den Themenschwerpunkten der Lehrbücher und Leitlinien basieren und in ein Formular zur Extraktion von gesundheitsökonomischen QBI überführt, mit dessen Hilfe die verschiedenen

Bewertungsinstrumente extrahiert werden (Kapitel 8.12 Datenextraktionsformular für gesundheitsökonomische Studien). Bei der Entwicklung der Domänen und Items wird darauf geachtet, dass sich diese primär auf die interne Validität beziehen. Da eine eindeutige Abgrenzung zwischen Studien- und Berichtsqualität teilweise nicht möglich ist, beziehen sich einzelne Elemente auch auf die Berichtsqualität. Die Domänen und zugehörigen Items werden nachfolgend dargestellt und erläutert.

Studienfrage

- Wurde die Studienfrage präzise definiert?
- War die Art der ökonomischen Studie angemessen?

Interventionsalternativen

- Wurden die aus ökonomischer Sicht relevanten Interventionsalternativen einbezogen?
- Wurde ein direkter Vergleich der Interventionsalternativen vorgenommen?
- Falls nein, waren die indirekten Vergleiche angemessen?

Perspektive

- Passt die gewählte Perspektive zur Studienfrage?
- Ist die Perspektive konsistent?

Ressourcenverbrauch und Kosten

- Wurden Mengen und Preise getrennt voneinander erhoben?
- Wurden alle relevanten Ressourcen, die mit der Intervention in Zusammenhang stehen identifiziert?
- Sind angemessene Datenquellen genutzt worden?
- Wurden alle relevanten Ressourcen mengenmäßig erfasst?
- Wurden die Ressourcenverbräuche angemessen bewertet?
- Sind die Methoden der Inflationierung und Währungskonversion angemessen?

Outcome/Nutzen

- Sind die Outcomeparameter richtig gewählt?
- Sind geeignete Erhebungsinstrumente gewählt, falls die Lebensqualität erhoben wird?
- Sind die Datenquellen auf ihre Qualität überprüft?

Qualität der Daten

- Ist die Qualität der Primärdaten ausreichend um die Studienfrage beantworten zu können?
- Ist die Qualität der Methoden zur Identifikation, Extraktion und Synthese der Effektparameter (Metaanalyse) ausreichend zur Beantwortung der Studienfrage?

Zeitraum

- Ist der Beobachtungszeitraum so gewählt, dass alle relevanten Effekte und Kosten berücksichtigt werden?

Modellierung

- Wird das Modell nachvollziehbar dargestellt?
- Sind die Modellstruktur und die gewählten Parameter angemessen?

Diskontierung

- Wurden in der Studie alle zukünftigen Kosten und Nutzen diskontiert?
- Wenn ja, sind die Diskontraten angemessen?

Statistische Verfahren

- Waren die statistischen Verfahren angemessen?

Sensitivitätsanalyse

- Wurde eine Sensitivitätsanalyse durchgeführt?
- Wurden alle relevanten Parameter in die Sensitivitätsanalyse einbezogen?

- Ist die Methodik der Sensitivitätsanalyse angemessen?

Ergebnisse

- Wurden die Ergebnisse mit Punktschätzern und Präzision angegeben?

Diskussion der Ergebnisse

- Basieren die Schlussfolgerungen auf den Ergebnissen?
- Wurde der mögliche Einfluss von Confounding und Bias diskutiert?

Interessenkonflikte

- Werden Art und Quelle der Finanzierung genannt?

Die Studienfrage stellt den Ausgangspunkt der Erstellung einer gesundheitsökonomischen Studie dar. Sie muss präzise definiert sein und die Zielsetzung der Untersuchung berücksichtigen. Die Art der ökonomischen Analyse (CEA, CUA oder CBA) muss primär auf die Zielsetzung der Studie ausgerichtet sein. Bei der Identifizierung von Interventionsalternativen sind zunächst alle relevanten Interventionen zu berücksichtigen. Vergleichsalternativen sollten grundsätzlich die übliche Versorgung (eventuell auch einem nach Marktanteilen gewichtetem Interventionsmix), die bisher effektivste sowie die kosteneffektivste Intervention umfassen. Nationale und regionale Besonderheiten sollten berücksichtigt werden.

Die Perspektive sollte auf das Ziel der Untersuchung ausgerichtet sein. Grundsätzlich sollte (zumindest auch) die gesellschaftliche Perspektive ausgewiesen werden. Abweichungen sollten begründet werden. Bei der Durchführung der gesundheitsökonomischen Studien sollte streng darauf geachtet werden, dass die Messungen und Bewertungen von Kosten und Nutzen konsistent zur Perspektive erfolgen.

Alle Ressourcen und Kosten, die durch eine Intervention induziert werden und aufgrund der Perspektive relevant sind, sollten systematisch identifiziert, mengenmäßig erfasst und bewertet werden. Bei der Datenerhebung sollten Mengen und Preise stets getrennt voneinander bestimmt werden. Die Bewertung des Ressourcenkonsums sollte nach dem ökonomischen Konzept der Opportunitätskosten erfolgen. Sofern nur staatlich administrierte oder kollektiv-vertraglich vereinbarte Preise verfügbar sind, sollte geprüft werden, ob Preiskorrekturen (Zu- und Abschläge) erforderlich sind. Die Qualität der Kostendaten ist zu bewerten. Wenn Preisdaten aus unterschiedlichen Zeitperioden stammen, sollte eine Inflationsbereinigung erfolgen. Grundsätzlich sollten gesundheitsversorgungsspezifische Inflationsraten angewendet werden. Da spezifische Inflationsraten für die meisten Gesundheitsleistungen in Deutschland nicht verfügbar sind, wird empfohlen, den gesamtwirtschaftlichen Preisindex zu benutzen, der vom Statistischen Bundesamt bereitgestellt wird. Für Währungsumrechnungen sollte auf Kaufkraftparitäten abgestellt werden.

Bei der Bestimmung der Outcomes sollte überprüft werden, ob die – bezogen auf die Zielsetzung der Untersuchung – relevanten Effekte und Nutzen berücksichtigt werden. Wenn in der gesundheitsökonomischen Studie Lebensqualität erhoben wird, ist zu überprüfen, ob geeignete standardisierte Lebensqualitätsinstrumente gewählt (respektive Verfahren der direkten Nutzenbewertung angemessen durchgeführt) werden. Die Qualität der Outcomedaten ist zu bewerten. Insbesondere ist zu prüfen, ob die (eingeschränkte) Datenqualität die Validität der Analyse beeinträchtigt. Bei der Nutzung von Daten aus Primärstudien sollte ihre Qualität bewertet werden, wie es für die epidemiologischen Studien beschrieben ist. Metaanalysen sollten analog beurteilt werden.

Der Zeithorizont einer gesundheitsökonomischen Evaluation sollte hinreichend lang sein, um alle relevanten Kosten- und Outcomeunterschiede zwischen den Vergleichsalternativen abbilden zu können. Sofern Modellierungen erforderlich sind, ist auf Transparenz und Nachvollziehbarkeit zu achten. Modellstruktur, berücksichtigte Parameter und Qualität der Kosten und Outcomes (s. o.) sollten angemessen sein. Sofern der Analysezeitraum über ein Jahr hinausgeht, sollten Kosten und Nutzen diskontiert werden. Es sollte überprüft werden, ob die Diskontierungsraten in einer empirisch (Leitlinien) und theoretisch fundierten Bandbreite liegen.

Gesundheitsökonomische Evaluationen sollten eine inkrementelle Kosten-Effektivitäts-Relation (oder bei Kosten-Nutzen-Analysen die Nettonutzendifferenz) ausweisen. Um Unsicherheiten bei der Datenanalyse zu berücksichtigen, sollten angemessene statistische Verfahren bei Stichprobenunsicherheit sowie Sensitivitätsanalysen bei Parameterunsicherheit angewendet werden.

Wie im wissenschaftlichen Hintergrund dargestellt, beschreiben die gesundheitsökonomischen Standards einen theoretisch fundierten Konsens über zulässige Ansätze der Kostenanalyse und Outcomeevaluation. Diese Bandbreite der Standards müssen die QBI berücksichtigen. Im gesundheitsökonomischen Extraktionsformular wird für die Bewertung der Items der berücksichtigten QBI die folgende Abstufung vorgenommen:

- angemessen (●),
- begründet (◐),
- berichtet (○),
- fehlend.

Eine Bewertung „berichtet“ wird vergeben, wenn ein QBI lediglich abfragt, ob ein Item in einer gesundheitsökonomischen Studie berichtet wird (z. B. Perspektive der Analyse, einbezogene Outcomeparameter oder Diskontierungsrate). Ein Item heißt „begründet“, wenn das Qualitätsinstrument explizit nach Begründungen für die Ausprägung des Items fragt. Die Bewertung „angemessen“ besagt, dass ein QBI eine Überprüfung der Angemessenheit des Items fordert (unabhängig davon, ob eine Begründung in der gesundheitsökonomischen Studie erfolgt oder nicht). Die Überprüfung der Angemessenheit sollte an den Standards der gesundheitsökonomischen Evaluation orientiert sein. Nicht alle QBI erläutern, was sie unter „angemessen“ verstehen. Wenn Items in einem Instrument nicht berücksichtigt werden, gelten sie als „fehlend“.

Einige Items des gesundheitsökonomischen Extraktionsformulars erheben lediglich, ob ein Item berichtet wird (z. B. Wird eine inkrementelle Kosten-Effektivitäts-Analyse durchgeführt? Wird eine Sensitivitätsanalyse durchgeführt? Werden Art und Quelle der Finanzierung genannt?). Sofern ein QBI ein solches Item berücksichtigt, wird automatisch die Bewertung „angemessen“ vergeben, da die Verwendung der verschiedenen Abstufungen bei einer solchen Art von Frage nicht relevant scheint.

### 6.4.3 Workshop

Zwecks Austausch von praktischen Erfahrungen im Umgang mit QBI wird ein Workshop durchgeführt.

#### 6.4.3.1 Ziele

Ziele des Workshops sind der Austausch und die Diskussion von

1. Erfahrungen und Umgang mit Bewertungsinstrumenten zur Qualität von Interventionsstudien,
2. Anforderungen und Inhalte an/von Bewertungsinstrumente/n zur Qualität von Interventionsstudien,
3. der Art und Weise, wie die Ergebnisse der Qualitätsbewertung in die Datenauswertung integriert werden und
4. der Evidenzbewertung von nicht-randomisierten Interventionsstudien.

Der Workshop soll Erfahrungen und praktische Probleme bei der Anwendung von QBI beleuchten. Ziel ist eine Ergänzung von wissenschaftlichen Untersuchungen um praktische Aspekte, deren Stellenwert in Publikationen oft nicht thematisiert wird. Eine Konsensbildung zu einzelnen Aspekten wird nicht angestrebt.

#### 6.4.3.2 Zielgruppe

Zur primären Zielgruppe gehören folgende Expertengruppen aus dem deutschsprachigen Raum:

- Autoren von deutschsprachigen HTA-Berichten oder systematischen Reviews des DIMDI und IQWiG,
- Experten auf dem Gebiet der Methodik,
- Wissenschaftler (aus den Disziplinen Medizin, Public Health, Epidemiologie, Prävention, Gesundheitsökonomie), die mit gesundheitspolitisch relevanten Evaluationen befasst sind,
- Institute/Verbände, die systematische Reviews mit Qualitätsbewertung durchführen.

Als Referenten und Teilnehmer werden Mitarbeiter der folgenden nationalen und internationalen Einrichtungen eingeladen:

- Cochrane Metabolic and Endocrine Disorders Group, Düsseldorf
- Deutsches Cochrane Zentrum, Freiburg
- DIMDI, Köln
- Fachgebiet Management im Gesundheitswesen, Institut für Technologie und Management, Technische Universität, Berlin
- Fakultät für Gesundheitswissenschaften, Universität Bielefeld, Bielefeld
- G-BA, Siegburg
- GP Forschungsgruppe Institut für Grundlagen- und Programmforschung, München
- HTA Zentrum, Universität Bremen, Bremen
- Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung, Medizinische Hochschule Hannover, Hannover
- Institut für Medizinische Statistik, Informatik und Epidemiologie der Universität zu Köln (IMSIE), Köln
- Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsmedizin der Johannes Gutenberg- Universität Mainz, Mainz
- Institut für Pflegewissenschaft, Universität Witten/Herdecke, Witten
- IQWiG, Köln
- Institut für Sozialmedizin des Universitätsklinikums Schleswig-Holstein, Campus Lübeck, Lübeck
- Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie der Charité, Berlin
- Institut für Sozialmedizin, Epidemiologie und Gesundheitssystemforschung (ISEG), Hannover
- Klinikum und Fachbereich Medizin Johann Wolfgang Goethe-Universität Frankfurt am Main, Frankfurt am Main
- Stiftungslehrstuhl für Medizinmanagement, Universität Duisburg-Essen, Essen
- Universitätsklinikum Hamburg-Eppendorf, Hamburg
- Wirtschaftswissenschaftliche Fakultät, Universität Augsburg, Augsburg
- Zentrum für Innere Medizin, Universität Marburg, Marburg
- UMIT – Private Universität für Gesundheitswissenschaften, Medizinische Informatik und Technik, Hall in Tirol, Hall
- Institut für klinische Epidemiologie und Biostatistik, Universitätsspital Basel, Basel
- LBI für Health Technology Assessment, Wien
- Gesundheit Österreich GmbH, Wien

#### 6.4.3.3 Planung und Durchführung

Der Workshop mit dem Titel „Die Qualitätsbewertung von Interventionsstudien – randomisierte und nicht-randomisierte Studien im Vergleich“ findet am 05.06.2009 in der Medizinischen Hochschule Hannover statt. Die Teilnehmerzahl ist auf 25 Personen begrenzt, um einen intensiven Austausch und die Einbindung aller Teilnehmer zu ermöglichen.

In der weiteren inhaltlichen Vorbereitung werden die Inhalte des Workshops spezifiziert:

- Austausch zu praktischen Problemen im Umgang mit QBI für Interventionsstudien wie Zeitaufwand, Subjektivität der Bewertung, Kriterien für die Auswahl eines Instrumentes, Berichtsqualität vs. methodische Qualität, Synthese der Ergebnisse, externe Validität als Bestandteil von Instrumenten
- Austausch zu relevanten Inhalten (Komponenten/Domänen) von QBI und Vergleich dieser relevanten Inhalte bei randomisierten und nicht-randomisierten Studien

Der Workshop wird in drei inhaltliche Blöcke gegliedert (s. Tabelle 2: Detaillierter Ablaufplan des Workshops). Referenten werden mit ihren Vorträgen mit einer Zeitvorgabe von 15 bis 25 Minuten in die entsprechenden Themen einleiten. Im Anschluss an die Vorträge sind jeweils 20 bis 30 Minuten für eine moderierte Diskussion vorgesehen.

Der erste Block behandelt die Qualitätsbewertung von randomisierten Studien. Zu diesen Studien gibt es die meisten Checklisten und gesicherte empirische Kenntnisse zur Bedeutung einzelner Studienelemente wie Randomisierung und Verblindung für die Studienqualität. Leitfragen in diesem Block, die auch Grundlage für die Diskussionen bilden sollen, sind: Welches Instrument eignet sich zur Qualitätsbewertung? Welchen Anforderungen soll es genügen? Was sollte bei der Entwicklung eines Instruments beachtet werden? Wie werden die Ergebnisse der Qualitätsbewertung in die Datensynthese integriert?

Im Mittelpunkt des zweiten Blocks stehen nicht-randomisierte Studien. Ziel ist es, die methodischen Limitationen und Stärken von nicht-randomisierten Studien detailliert darzustellen sowie deren Bedeutung für die interne und externe Validität von Studienergebnissen zu diskutieren. Außerdem sollen inhaltliche Anforderungen an ein Instrument erarbeitet werden. Welche methodischen Verfahren bei der Datenanalyse sind geeignet, Confounding zu kontrollieren, wie gelingt eine Einschätzung des Residual confounding? Reicht ein Instrument für randomisierte und nicht-randomisierte Studien oder sind studienspezifische Instrumente erforderlich, die die unterschiedlichen Studiendesigns berücksichtigen (Kohortenstudien, Fall-Kontrollstudien etc.)?

Im abschließenden Block wird Grading of Recommendations Assessment, Development and Evaluation (GRADE)[116] vorgestellt, ein Instrument, das ursprünglich für die Erstellung von Leitlinien entwickelt wird[80]. Basierend auf der Einschätzung der methodischen Qualität von Studien wird die Stärke der vorliegenden Evidenz der Studienergebnisse beurteilt, die wiederum als Grundlage für Empfehlungen herangezogen wird. Das Instrument hat zwei Komponenten, die erste dient der Einschätzung der Qualität der Evidenz (vierstufig: hoch, moderat, niedrig, sehr niedrig), die zweite ermittelt anhand des Evidenzprofils die Stärke der Empfehlungen. Die erste Komponente geht über typische Evidenzhierarchien hinaus, die nur einen Teil der Qualitätsaspekte relevanter Studien berücksichtigen. Nichtrandomisierte Studien können unter bestimmten Voraussetzungen (Stärke des Effekts, Dosis-Wirkungs-Gradient, gute Confounderkontrolle) in ihrer Qualität aufgewertet werden. Es soll diskutiert werden, ob GRADE insbesondere für die Einschätzung der Validität von nicht-randomisierten Studien geeignet ist.

**Tabelle 2: Detaillierter Ablaufplan des Workshops**

| Zeit | Thema | Struktur und Anleitung |
|---|---|---|
| 10:00-10:30 | Ankunft, Kaffeetrinken | |
| **Block 1: Randomisierte Interventionsstudien<br>Moderation: E. M. Bitzer, M. Dreier** | | |
| 10:40 | **Qualitätsbewertungsinstrumente – ein Blick in die Praxis**<br>Dagmar Lühmann, Lübeck | Vortrag 15 Min. |
| 10:55 | Nachfragen | |
| | Wichtige Diskussionspunkte bei der Qualitätsbewertung | Sammeln im Auditorium, ergänzen von Problemfeldern (Flipchart) |
| | Soll externe Validität in die QB eingeschlossen werden? | Sollen QBI nur das „Verzerrungspotential“ bewerten, wenn ja, wie berücksichtigt man die externe Validität? |
| 11:20 | **Der Lübecker Journalclub – Entwicklung einer Checkliste**<br>Susanne Schramm, Lübeck | Vortrag 20-25 Min. |
| | Nachfragen | |
| | Welche methodische Stringenz ist notwendig? | |
| | Umgang mit Berichtsqualität | Sollte die Berichtsqualität in die QB integriert werden? |
| | Subjektivität, Interrater-Reliabilität | Wie sind Ihre Erfahrungen/Umgang mit der Subjektivität der Bewertungen? |
| 11:55 | 15 Min. Kaffeepause | |

**Tabelle 2: Detaillierter Ablaufplan des Workshops – Fortsetzung**

| | | |
|---|---|---|
| 12:10 | **Das *risk of bias* Instrument der Cochrane Collaboration**<br>Bernd Richter, Düsseldorf | Vortrag 20-25 Min |
| | Nachfragen | Wo sehen Sie Stärken und Schwächen beim Risk of Bias Tool?<br>Wie ist die Praktikabilität der Anwendung? |
| | Subjektivität | Wie sind Ihre Erfahrungen/Umgang mit der Subjektivität der Bewertungen? |
| 13:00 | Mittagspause | |
| **Block 2: Nicht-randomisierte Interventionsstudien<br>Moderation: A. Gerhardus** | | |
| 14:00 | **Wird die Validität von nicht-randomisierten Studien unterschätzt?**<br>Hajo Zeeb, Mainz | Vortrag 20-25 Min |
| | Welche Domänen/methodischen Aspekte sind auch für nicht-randomisierte Studien wichtig? | Weitere Überlegungen, Fortsetzung Vormittag |
| | Brauchen wir studienspezifische Instrumente? | |
| 14:45 | **Die Bewertung von nicht-randomisierten Studien aus Sicht des IQWiG**<br>Robert Großelfinger, Köln | Vortrag 20-25 Min |
| 15:30 | Kaffeepause | |
| **Block 3: Stärke der Evidenz<br>Moderation: D. Lühmann** | | |
| 16:00 | **GRADE<br>– ein Instrument für HTA?**<br>Monika Lelgemann, Bremen | Vortrag 20-25 Min |
| | Ist GRADE auch für nicht-randomisierte Studien geeignet? | |
| 16:45 | Abschlussworte | Dank<br>Zusammenfassung<br>Weitere Anregungen/Austausch |

GRADE = Grading of Recommendations Assessment, Development and Evaluation. HTA = Health Technology Assessment. IQWiG = Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. QB = Qualitätsbewertung. QBI = Qualitätsbewertungsinstrument.

Zur Dokumentation der Ergebnisse des Workshops werden folgende Materialien genutzt:

- Die Audio-Aufzeichnung des Workshops (vorbehaltlich der Einwilligung aller Teilnehmer) mit nachfolgender Transkription
- Die Folien der Vorträge
- Die Aufzeichnungen auf dem Flipchart

## 6.5 Ergebnisse

### 6.5.1 Bewertung von Studien zur Wirksamkeit

Nachfolgend werden die Ergebnisse der Literaturrecherche und -auswahl bzw. der Datenextraktion und -synthese erläutert.

#### 6.5.1.1 Literaturrecherche und -auswahl

Die systematische Datenbankrecherche gestaltet sich schwierig und erfordert mehrere Probeläufe, um eine akzeptable und praktikable Sensitivität und Spezifität zu erzielen. Problematisch ist vor allem eine ungenügende themenspezifische Indexierung in den einbezogenen Datenbanken. So existiert weder

ein deutsch- noch ein englischsprachiges Schlagwort zu „Studienqualität“. Dies untermauert die Relevanz der Recherchen, die ergänzend zur systematischen Datenbanksuche durchgeführt werden, wodurch unter anderem das Instrumentarium der Cochrane Collaboration identifiziert wird.

**Systematische Datenbankrecherche**

Die systematische Literaturrecherche ergibt 565 Treffer. Nach Durchsicht der Titel anhand der zuvor definierten Ein- und Ausschlusskriterien werden 387 Publikationen ausgeschlossen, darunter 17 Duplikate. Durch Screening der verbliebenen 178 Abstracts werden weitere 41 Dokumente ausgeschlossen. Die anschließende Literaturbestellung beim DIMDI umfasst 137 Volltexte. Von diesen Texten sind zwei nicht lieferbar[71, 153]. Auf der Basis der Volltexte werden weitere 102 Publikationen ausgeschlossen. Eine tabellarische Übersicht der Ausschlussgründe befindet sich im Anhang (Kapitel 8.3). Aus der systematischen Literaturrecherche werden insgesamt 33 Publikationen eingeschlossen.

**Screening von HTA-Berichten**

In der Recherche am 26.11.2008 werden in der Datenbank der DAHTA des DIMDI 213 Dokumente identifiziert. Unter den HTA-Berichten befinden sich 38 Publikationen, deren Volltext nicht aus dem deutschsprachigen Raum stammt und aufgrund der Fragestellung ausgeschlossen wird. Fünf der Berichte sind Tagungsbände und werden ausgeschlossen. Drei HTA-Berichte thematisieren die Bewertung von Leitlinien. Da dies nicht der Fragestellung des vorliegenden Berichts entspricht, werden sie ebenfalls ausgeschlossen. Die verbleibenden 167 Berichte werden nach QBI durchsehen. Für die Ermittlung des Anteils der HTA-Berichte, in denen eine Qualitätsbewertung der einbezogenen Studien vorgenommen wird, werden weitere zehn Berichte ausgeschlossen, da bei ihnen eine Qualitätsbewertung von Studien nicht anwendbar ist.

**Tabelle 3: Qualitätsbewertung in deutschsprachigen HTA-Berichten**

| **Deutschsprachige HTA-Berichte** | **n** | **%** |
|---|---|---|
| Gesamtzahl der Publikationen | 213 | 100 |
| Ausgeschlossene Publikationen | 56 | 26 |
| Eingeschlossene Publikationen | 157 | 74 |
| | | |
| Durchführung einer Qualitätsbewertung | 136 | 87 |
| Keine Qualitätsbewertung im Bericht erkennbar | 21 | 13 |
| | | |
| Einsatz einer Checkliste zur Qualitätsbewertung | 81 | 52 |
| Kein Einsatz einer Checkliste zur Qualitätsbewertung | 76 | 48 |

HTA = Health Technology Assessment.

In 136 der eingeschlossenen 157 HTA-Berichte wird eine Qualitätsbewertung der Primär- und/oder Sekundärstudien erwähnt. In fünf dieser Berichte findet die Bewertung der Studienqualität lediglich im ökonomischen Teil des Berichts statt. Instrumente in Form einer Checkliste werden in 81 HTA-Berichten genannt. Dies entspricht 56 % der HTA-Berichte, in denen eine Qualitätsbewertung durchgeführt wird und 52 % aller eingeschlossenen HTA-Berichte. Wird die Verwendung einer Checkliste zur Qualitätsbewertung erwähnt, so fehlt in einigen Berichten die Darstellung dieses Instruments. Das am häufigsten verwendete Instrument zur Qualitätsbewertung stellt das Instrumentarium der German Scientific Working Group Technology Assessment in Health Care (GSWG-TAHC)[69, 197] dar. Diese Checklisten werden in 56 HTA-Berichten verwendet, teilweise um weitere Instrumente ergänzt.

Aus der DAHTA-Datenbank werden fünf Publikationen eingeschlossen, die die Einschlusskriterien erfüllen. Diese enthalten insgesamt zwölf Instrumente. Zwei weitere Publikationen, die über die HTA-Berichte identifiziert werden[106, 238], werden nicht erneut eingeschlossen, da sie bereits über die Datenbankrecherche ermittelt worden sind.

**Tabelle 4: Eingeschlossene Instrumente aus HTA-Berichten**

| Publikation | Instrument |
|---|---|
| Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen[103] | Qualitätsbeurteilung: Diagnosestudien |
| Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen[102] | Keine Bezeichnung des Systems zur „Bewertung der Studien- und Publikationsqualität" |
| Thomas et al.[213] | Quality Assessment Tool for Quantitative Studies |
| Ekkernkamp et al.[69] | GSWG-Checkliste 1a: Kontextdokumente |
| | GSWG-Checkliste 1b: Systematische Reviews und Metaanalysen |
| | GSWG-Checkliste 2a: Primärstudien (RCT/Fall-Kontrollstudien/Kohortenstudien/ Längsschnittstudien/Fallserien) |
| | GSWG Checkliste 2b: Diagnosestudie |
| Scottish Intercollegiate Guidelines Network (SIGN 50)[194] | Methodology Checklist 1: Systematic Reviews and Meta-analyses |
| | Methodology Checklist 2: Randomised Controlled Trials |
| | Methodology Checklist 3: Cohort Studies |
| | Methodology Checklist 4: Case-control Studies |
| | Methodology Checklist 5: Diagnostic Studies |

GSWG = German Scientific Working Group. HTA = Health Technology Assessment. RCT = Randomisierte kontrollierte Studie.

### Internetrecherche

Die Internetrecherche, die insbesondere die Webseiten von HTA-Organisationen und der Cochrane Collaboration einbezieht, liefert nach Anwendung der Ein- und Ausschlusskriterien acht Publikationen, die insgesamt 21 Instrumente zur Qualitätsbewertung enthalten (Tabelle 7: Übersicht über systematische Reviews zu Bewertungsinstrumenten). Zwei Publikationen[69, 194], die bereits über die Suche in HTA-Berichten ermittelt werden, werden nicht erneut eingeschlossen. Ausgeschlossene Publikationen aus der Internetrecherche sind im Anhang (Kapitel 8.5) dokumentiert.

**Tabelle 5: Eingeschlossene Publikationen aus der Internetrecherche**

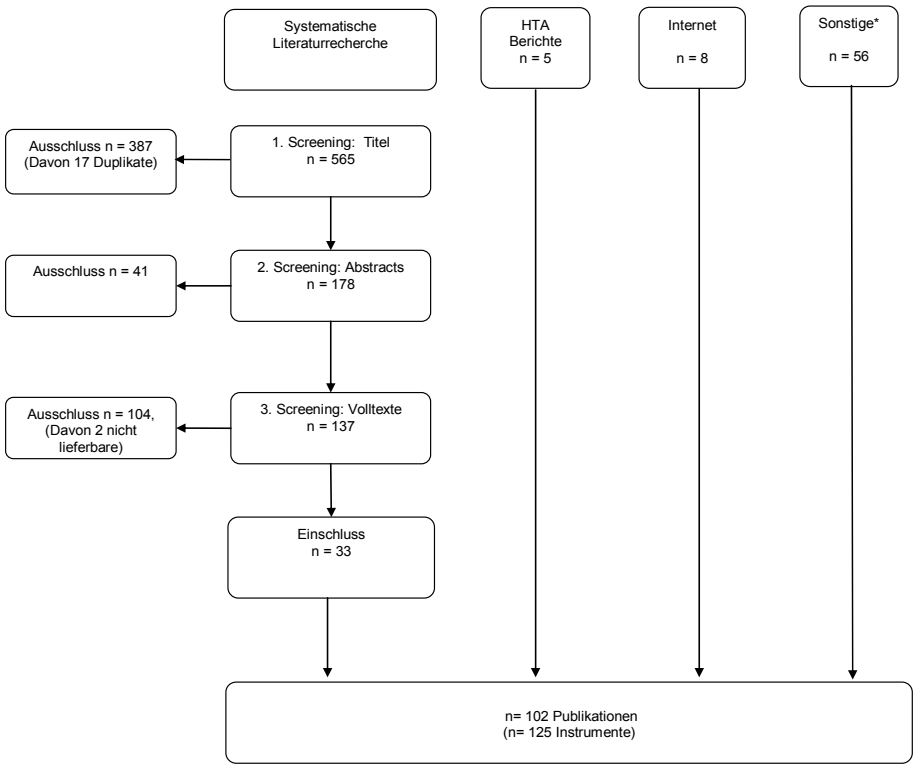
| Publikation | Instrument |
|---|---|
| Centre for Evidence-based Medicine (CEBM)[219] | Systematic Review Critical Appraisal Sheet |
| | RCT Critical Appraisal Sheet |
| | Diagnostic Critical Appraisal Sheet |
| Centre for Evidence-based Mental Health (CEBMH)[36] | Critical appraisal form for an overview |
| | Critical appraisal form for single therapy studies |
| | Critical appraisal form for a study of prognosis |
| | Critical appraisal form for a study of diagnosis |
| Delfini Group[57] | Short Critical Appraisal Checklist |
| Higgins & Green[92] | Risk of bias tool |
| Ludwig Boltzmann Institut (LBI)[133] | Formular zu Beurteilung der internen Validität von RCT |
| | Formular zu Beurteilung der internen Validität von Kohortenstudien |
| | Formular zu Beurteilung der internen Validität von systematischen Reviews und Metaanalysen |
| | Formular zu Beurteilung der internen Validität von diagnostischen Studien |
| NHS Public Health Research Unit (PHRU): Critical Appraisal Skills Programme (CASP)[161] | 12 questions to help you make sense of a diagnostic test study |
| | 11 questions to help you make sense of a case control study |
| | 12 questions to help you make sense of a cohort study |
| | 10 questions to help you make sense of randomised controlled trials |
| | 10 questions to help you make sense of reviews |
| Aggressive Research Intelligence Facility (ARIF)[220] | Critical Appraisal Checklist |
| Ottawa Health Research Institute; Wells et al.[234] | Newcastle-Ottawa Quality Assessment Scale Case Control Studies |
| | Newcastle-Ottawa Quality Assessment Scale Cohort Studies |

RCT = Randomisierte kontrollierte Studie.

Auf einigen Internetseiten, wie z. B. der des International Network of Agencies for Health Technology Assessment (INAHTA) finden sich Verweise zu nationalen HTA-Richtlinien. Eigene QBI können nicht identifiziert werden.

**Sonstige Quellen**

Auf der Basis der dargestellten Rechercheschritte können anhand von Referenzen sowie Dokumenten, die ein bereits bestehendes Instrument anwenden bzw. dessen Testgüte ermitteln, weitere 56 Instrumente zur Qualitätsbewertung identifiziert werden.




* Referenzen, Publikationen mit Anwendung oder Ermittlung der Testgüte eines bestehenden Instruments.

HTA = Health Technology Assessment.

**Abbildung 1: Stufenweise Literaturrecherche und -auswahl (Effektivität)**

Anhand der mehrstufigen Literaturrecherche (Abbildung 1: Stufenweise Literaturrecherche und -auswahl (Effektivität)) werden acht systematische Übersichtsarbeiten und 102 Primärstudien bzw. Dokumente mit Instrumenten zur Qualitätsbewertung von Primär- und/oder Sekundärstudien eingeschlossen. Insgesamt werden 125 Instrumente eingeschlossen. Die Mehrheit der Instrumente eignet sich für die Qualitätsbewertung von Interventionsstudien (Tabelle 6: Anzahl eingeschlossener Dokumente/Instrumente).

**Tabelle 6: Anzahl eingeschlossener Dokumente/Instrumente**

| Dokumente/Instrumente | Anzahl |
|---|---|
| Systematische Übersichtsarbeiten mit Instrumenten zur Qualitätsbewertung | 8 |
| Primärstudien sowie andere Dokumente mit einem oder mehreren Bewertungsinstrumenten | 102 |
| Gesamtzahl eingeschlossener Instrumente* | 125 |
| | |
| Bewertungsinstrumente für systematische Reviews, HTA, Metaanalysen | 15 |
| Bewertungsinstrumente für Interventionsstudien | 80 |
| Bewertungsinstrumente für Beobachtungsstudien | 30 |
| Bewertungsinstrumente für Diagnosestudien | 17 |

* Einige Instrumente eignen sich für mehrere Studiendesigns.

HTA = Health Technology Assessment.

Übersichtsarbeiten, Publikationen, die ein bestehendes Instrument nutzen sowie Publikationen, die die Testgüte bestehender Instrumente ermitteln, fließen indirekt in die Anzahl der gefundenen Publikationen ein, da sie zur Ermittlung der Originalversion des Instrumentes verwendet werden.

### 6.5.1.2 Systematische Übersichtsarbeiten zu Bewertungsinstrumenten

Es werden acht relevante systematische Übersichtsarbeiten identifiziert[56, 109, 131, 168, 189, 190, 235, 236]. Drei fokussieren auf Instrumenten zur Qualitätsbewertung von nicht-randomisierten Studien bzw. Beobachtungsstudien[56, 189, 190], ein Dokument untersucht Instrumente für diagnostische Studien[239], zwei Übersichtsarbeiten vergleichen Instrumente für randomisierte Studien[143, 168], letztere im Bereich physikalischer Therapie und Rehabilitationsforschung. Zwei Berichte schließen alle Studientypen ein[109, 235] (s. Tabelle 7: Übersicht über systematische Reviews zu Bewertungsinstrumenten).

**Tabelle 7: Übersicht über systematische Reviews zu Bewertungsinstrumenten**

| Erstautor | Jahr | Land | Titel | Studiendesign | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | SR, Metaanalysen | RCT | Kohortenstudien | Fall-Kontrollstudien | Querschnittstudien | Diagnosestudien |
| Deeks[56] | 2003 | UK | Evaluating non-randomised intervention studies | | | X | | | |
| Katrak[109] | 2004 | Australien | A systematic review of the content of critical appraisal tools | X | X | X | X | X | |
| Moher[143] | 1995 | Kanada | Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists | | X | | | | |
| Olivo[168] | 2008 | USA | Scales to assess the quality of randomized controlled trials: A systematic review | | X | | | | |
| Sanderson[189] | 2007 | UK | Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography | | | X | X | X | |
| Saunders[190] | 2003 | Kanada | Assessing the methodological quality of nonrandomized intervention studies | | | X | X | X | |
| West[235] | 2002 | USA | Systems to rate the strength of scientific evidence | X | X | X | X | X | X |
| Whiting[239] | 2005 | | A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools | | | | | | X |

RCT = Randomisierte kontrollierte Studie. SR = Systematischer Review.

Die methodische Qualität der Übersichtsarbeiten wird mit einer modifizierten Checkliste bewertet, die sechs Komponenten umfasst (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])).

Die Ergebnisse sind in Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235]) dargestellt. Eine ausführliche Tabelle der Ergebnisse der methodischen Details befindet sich im Anhang (s. Tabelle 38: Übersicht über die eingeschlossenen systematischen Übersichtsarbeiten (n = 8), Kapitel 8.13).

**Tabelle 8: Checkliste zur Qualitätsbewertung der systematischen Übersichtsarbeiten (mod. nach West et al.[235])**

| Kriterien | Ausprägung |
|---|---|
| A priori definierte Ein- und Ausschlusskriterien | √: Trifft zu.<br>-: Trifft nicht zu. |
| Vollständigkeit der Suche | √: Es wird mehr als eine Datenbank und eine zusätzliche Datenquelle (z. B. Internet, Handsuche, Expertenbefragung) herangezogen.<br>-: Es wird nur eine Datenbank durchsucht und keine weiteren Datenquellen. |
| Methode der Literaturauswahl | √: Die Literaturauswahl wird von mindestens zwei unabhängigen Reviewern vorgenommen.<br>-: Die Literaturauswahl wird nur von einem Reviewer vorgenommen. |
| Methode der Datenextraktion | √: Die Datenextraktion wird von mindestens zwei unabhängigen Reviewern vorgenommen.<br>- : Die Datenextraktion wird nur von einem Reviewer vorgenommen. |
| Datensynthese | √: Die Datensynthese ist transparent und nachvollziehbar.<br>-: Die Datensynthese ist nicht nachvollziehbar. |
| Schlussfolgerungen | √: Die Schlussfolgerungen basieren auf den Ergebnisse und sind nachvollziehbar.<br>-: Die Schlussfolgerungen sind nicht nachvollziehbar. |

**Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])**

| Erstautor | A priori def. Ein- und Ausschluss-kriterien | Vollständig-keit der Suche | Methode der Literatur-auswahl | Methode der Daten-extraktion | Daten-synthese | Schluss-folgerungen |
|---|---|---|---|---|---|---|
| Deeks[56] | √ | √ | K. A. | (√)[3] | √ | √ |
| Katrak[109] | √ | (√)[1] | √ | √ | √ | √ |
| Moher[143] | √ | (√)[2] | K. A. | K. A. | √ | √ |
| Olivo[168] | √ | √ | √ | √ | √ | √ |
| Sanderson[189] | √ | √ | K. A. | √ | √ | √ |
| Saunders[190] | √ | - | √ | √ | √ | √ |
| West[235] | √ | (√)[2] | √ | √ | √ | √ |
| Whiting[239] | √ | √ | (√)[3] | (√)[3] | √ | √ |

[1] Zahlreiche Datenbanken, aber 29 % der ausgewählten Dokumente nicht verfügbar.

[2] Nur eine Datenbank durchsucht, aber zusätzlich Handsuche und Expertenbefragung bzw. Internetrecherche.

[3] Die Literaturauswahl bzw. Datenextraktion wird von einem Reviewer vorgenommen und von einem zweiten kontrolliert.

K. A. = Keine Angabe.

Die Methodik und Inhalte der Übersichtsarbeiten werden im Folgenden narrativ dargestellt.

**Deeks et al. Evaluating non-randomised intervention studies[56]**

Der HTA-Bericht von Deeks et al. ist ein Auftrag des NHS R&D Health Technology Assessment Programme in Zusammenarbeit mit dem International Stroke Trial und der European Carotid Surgery Trial Collaborative Groups. Es werden Methoden und der Umgang mit der Evidenz in nicht-randomisierten Interventionsstudien untersucht. Hierzu werden drei systematische Reviews zur (1) vorhandenen Evidenz im Hinblick auf Bias in nicht-randomisierten Studien, (2) zur Evaluation von QBI und (3) zum Einsatz von QBI durchgeführt. Für die Fragestellung dieses Berichts ist nur der zweite Teil relevant.

Es wird eine systematische Literaturrecherche in mehr als zwölf Datenbanken durchgeführt, 8.326 Titel werden gescreent, von denen 212 Dokumente eingeschlossen werden. Um eingeschlossen zu werden, muss die Publikation ein Instrument zur Bewertung der methodischen Qualität oder der Validität von Primärstudien enthalten. Das QBI soll angewendet worden sein bzw. anwendbar sein auf nicht-randomisierte Interventionsstudien. Daher werden auch QBI eingeschlossen, die zwar nur zur Bewertung von RCT entworfen worden sind, wenn sie auch für die Bewertung von nicht-randomisierten Studien eingesetzt werden. Ausgeschlossen werden Instrumente zur Bewertung von Fall-

Kontrollstudien (da eher seltener Einsatz zur Untersuchung von Interventionen) und unkontrollierten Studien. „Modifizierte" Instrumente sind solche, die auf einem bereits bekannten QBI basieren, während Instrumente, die aus mehreren bestehenden QBI entwickelt worden sind und solche, deren Herkunft unbekannt ist, als „neu" eingestuft werden.

Aus den gefundenen QBI werden inhaltliche Charakteristika für zwölf vorab definierte Qualitätsdomänen, die Aspekte der internen Validität, der externen Validität und Berichtsqualität abdeckten, extrahiert. Die zwölf Domänen, die in einem modifizierten Delphiprozess innerhalb des Teams entstehen, umfassen (1) Hintergrund/Kontext, (2) Definition und Auswahl der Studienpopulation, (3) Interventionen, (4) Outcomes, (5) Zustandekommen der Studiengruppen, (6) Verblindung, (7) Zuverlässigkeit der Informationen, (8) Follow-up, (9) Analyse: Vergleichbarkeit der Gruppen, (10) Interpretation, (11) Interpretation sowie (12) Darstellung und Bericht. Den Domänen werden, sowohl a priori als auch post hoc, jeweils zwei bis sieben Items zugeordnet.

Sechs Domänen, also die Domänen (5) bis (10), beziehen sich überwiegend auf die Bewertung der internen Validität. Von diesen werden für die Bewertung von nicht-randomisierten Studien a priori zwei Domänen als zentral definiert: (5) Zustandekommen der Studiengruppen und (9) Analyse: Vergleichbarkeit der Gruppen. Innerhalb dieser beiden Domänen werden vier a priori definierte Items ebenfalls als zentral erachtet:

(5.3) Zuordnung der Intervention („How allocation occurred"),

(5.4) Balance der Gruppen durch Design („Any attempt to balance groups by design"),

(9.2) Identifizierung von prognostischen Faktoren („Identification of prognostic factors") und

(9.3) Case-mix-Adjustment.

Die QBI werden unterschieden in „Top tools" und in „Best tools". „Top tools" sind QBI die mindestens ein a priori festgelegtes Item in fünf der sechs Domänen zur internen Validität abdecken. Als „Best tools" gelten die QBI, die mindesten drei der vier zentralen Items enthalten. Die Brauchbarkeit der 14 „Best tools" für die Anwendung in systematischen Übersichtsarbeiten testen Mitglieder des Projektteams, indem jeder jedes Instrument mindestens zweimal an drei möglichen nicht-randomisierten Interventionsstudien ausprobiert und hinsichtlich Zeitaufwand, Einfachheit, widersprüchlicher oder schwierig zu interpretierender, fehlender und sonstiger Aspekte bewertet.

Es werden insgesamt 193 QBI identifiziert, die den Einschlusskriterien entsprechen. Von 182 ausreichend beschriebenen Instrumenten können 46 als „Top tools" und 14 als „Best tools" eingeordnet werden. Zu den „Best tools" gehören Instrumente, die publiziert werden von Bracken[22], Critical Appraisal Skills Programme (CASP)[52], Cowley[51], Downs & Black[60], DuRant[65], Fowkes & Fultan[73], Hadorn et al.[82], Wells et al.[234] (Autoren bezeichnen das Instrument als „Newcastle-Ottawa"), Reisch et al.[181], Spitzer et al.[204], Thomas[214], Vickers[229], Weintraub[233], Zaza et al.[245]. Jeweils etwa die Hälfte kann als Checkliste bzw. Skala charakterisiert werden. Unter den „Best tools" sind 79 % Checklisten. Abgesehen von den inhaltlichen Charakteristika sind die besten 14 Instrumente in ihrer methodischen Entwicklung, der Testung von Reliabilität und Validität den übrigen Instrumenten nicht überlegen. Von den 14 „Best tools" sind sechs für den Einsatz in systematischen Übersichtsarbeiten geeignet: ein QBI von Cowley et al.[51], Downs & Black[60], Reisch et al.[181], Thomas[214], Wells et al.[234] und Zaza et al.[245].

Die Autoren weisen darauf hin, dass Instrumente mit den vier zentralen Items nicht notwendigerweise auch nützlich sind, da zum Beispiel allein die Beschreibung der Zuordnung der Intervention keine Bewertung der Wahrscheinlichkeit von Bias bedeutet. Viele Instrumente schießen auch Items ein, die sich nicht auf die methodische, sondern auf die Berichtsqualität beziehen.

Die Autoren räumen ein, dass sie sehr viele Instrumente eingeschlossen haben, dadurch, dass nicht nur Instrumente, die für die Bewertung von nicht-randomisierten Studien vorgesehen sind, sondern alle Instrumente eingeschlossen werden, die zur Qualitätsbewertung von nicht-randomisierten Studien verwendet werden oder verwendet werden können. Dadurch wird eine Vielzahl von Instrumenten eingeschlossen, von denen einige ursprünglich für die Bewertung von RCT entwickelt werden.

Die Auswahl der zentralen Domänen und Items wird von den Autoren vorgenommen. Sie basiert auf dem Wissen von randomisierten Studien ergänzt um methodische Unterschiede von nicht-randomisierten Studien. Die Alternative, alle Items aus den Instrumenten zu extrahieren und anschließend zu

gruppieren, wird aufgrund der Vielzahl der Instrumente, unterschiedlicher Terminologie und Fragenformulierung als nicht machbar angesehen.

Die Autoren empfehlen weitere Untersuchungen der von ihnen ausgewählten Qualitätskriterien, die Entwicklung eines neuen oder die Revision eines existierenden Instruments und detailliertere Analysen zur Brauchbarkeit von Instrumenten.

Es handelt sich um eine Übersichtsarbeit mit umfassender Literaturrecherche, adäquater Synthese der Daten und nachvollziehbaren Schlussfolgerungen. Die Angemessenheit des Vorgehens bei der Literaturauswahl kann wegen fehlender Angaben nicht beurteilt werden. Ein Reviewerbias bei der Datenextraktion kann nicht ausgeschlossen werden (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])).

**Katrak et al. A systematic review of the content of critical appraisal tools[109]**

Die Publikation wird von Mitarbeitern des Centre for Allied Health Evidence und der Schule für Physiotherapie der Universität von Südaustralien, Melbourne erstellt. Ziel ist es, Instrumente zur kritischen Bewertung von Studien zu erfassen und zu vergleichen sowie deren Relevanz für die Anwendung von Interventionen aus dem Bereich von Gesundheitsfachberufen (Physiotherapie, Sprachtherapie, Familienplanung, Ernährungsberatung etc.) zu prüfen.

Es wird eine systematische Literaturrecherche ohne Sprachrestriktion in acht Datenbanken, auf den Internetseiten unterschiedlicher Institutionen sowie in den Literaturverzeichnissen gefundener Publikationen durchgeführt. Die Recherche ist nicht zeitlich eingegrenzt, ihr Datum wird nicht angegeben. Es werden nur Instrumente mit numerischem Summenscore berücksichtigt, die komplett und in englischer Sprache publiziert sind. Die Datenextraktion wird von zwei voneinander unabhängigen Reviewern durchgeführt, Diskrepanzen werden durch Diskussion mit einer dritten Person gelöst.

Die Items der gefundenen Instrumente werden jeweils einer von elf Kategorien (Domänen) zugeordnet, basierend auf Kriterien, die Clarke und Oxman im Cochrane Reviewer's Handbook (2003) beschreiben (s. Tabelle 10: Kategorien zur Klassifikation der Items bei Katrak et al.[109]).

**Tabelle 10: Kategorien zur Klassifikation der Items bei Katrak et al.[109]**

| | Kategorie | Erläuterung |
|---|---|---|
| 1 | Studienziel und Rechtfertigung | |
| 2 | Verwendete Methoden | Identifikation relevanter Studien und Befolgung des Studienprotokolls |
| 3 | Auswahl der Stichprobe | Ein- und Ausschlusskriterien, Homogenität der Gruppen |
| 4 | Randomisierung und verdeckte Zuordnung | Methode der Randomisierung |
| 5 | Verlust von Studienteilnehmern | Response und Loss-to-follow-up |
| 6 | Verblindung | Kliniker, Untersucher, Patient, Datenauswerter |
| 7 | Messung des Outcomes | |
| 8 | Details der Intervention oder Exposition | |
| 9 | Datenanalyse | |
| 10 | Potenzielle Biasquellen | |
| 11 | Externe Validität | Übertragbarkeit auf anderes Setting, Zusammenhang von Nutzen, Kosten und Schaden |
| | Zusätzliche Kategorie: Sonstiges | |

Aus der Suche resultierten 193 Publikationen, aus denen 121 Instrumente extrahiert werden. Sechzehn der Instrumente sind als generisch, die restlichen 105 als studiendesign-spezifisch kategorisiert. Für die Bewertung von Primärstudien ergeben sich 94, für die von systematischen Reviews 26 Instrumente. Die meisten designspezifischen Instrumente (n = 46) zielen auf die Bewertung von experimentellen Studien ab. Für Diagnose- und qualitative Studien werden jeweils sieben und für Beobachtungsstudien 19 Bewertungsinstrumente identifiziert.

**Systematische Reviews:** Die häufigsten Items können der Datenanalyse (Methoden der Datensynthese, Sensitivität der Ergebnisse, Beachtung von Heterogenität) und der externen Validität zugeordnet werden. Außerdem werden Items zur Identifikation relevanter Studien, zur verwendeten

Suchstrategie sowie zur Zahl der eingeschlossenen Studien erfasst. Items zu Randomisierung und Verblindung sind selten enthalten.

**Experimentelle Studien:** Die meisten Items gehören zur Datenanalyse, zur Verblindung, zur Randomisierung und zur Stichprobenauswahl.

**Beobachtungsstudien:** Die häufigsten Items zählen zur Datenanalyse (Berücksichtigung von Confoundern, Power-Berechnung, angemessene statistische Methoden) und zur Auswahl der Stichprobe.

**Diagnosestudien:** Die häufigsten Items sind spezifisch für Diagnosestudien (Definition diagnostischer Kriterien, Definition eines Goldstandards, Berechnung von Sensitivität und Spezifität) und gehören zur Datenanalyse, zur externen Validität oder zur Stichprobenauswahl.

**Qualitative Studien:** Die häufigsten Items können der externen Validität, den Studienzielen und der Datenanalyse zugeordnet werden. Items zu Verlustbias, Verblindung, Intervention oder Bias sind nicht vorhanden.

**Generische Instrumente (experimenteller und Beobachtungsstudien):** Die häufigsten Items gehören zur Auswahl der Stichprobe (Ein- und Ausschlusskriterien, Strukturgleichheit der Studiengruppen zu Beginn) und zur Datenanalyse (Angemessenheit, Berechnung der Power).

**Generische Instrumente (alle Studiendesigns, quantitativ und qualitativ):** Die häufigsten Items werden der Datenanalyse (Berücksichtigung von Confoundern, Power-Berechnung, angemessene statistische Methoden) und der externen Validität zugeschrieben.

Ein numerischer Summenscore wird in 58 Instrumenten gebildet, entweder durch die gleiche Gewichtung aller Items oder indem als wichtiger eingeschätzte Items höher gewichtet werden. Der Anteil von Instrumenten mit einem numerischen Summenscore ist bei den designspezifischen Instrumenten etwa gleich verteilt, nur Instrumente zur Bewertung von qualitativen Studien bilden überwiegend keinen Score. Angaben zu Reliabilität und Validität sind nur für sehr wenige Instrumente vorhanden.

Spezifische Instrumente für die Bewertung von Interventionen aus dem Bereich von Gesundheitsfachberufen werden nicht identifiziert.

Hauptergebnisse sind die Vielzahl und die große Variation von Bewertungsinstrumenten. Die Autoren weisen darauf hin, dass unterschiedliche Instrumente, die für die gleiche Literatur verwendet werden, verschiedene Ergebnisse produzieren können, sodass eine Qualitätsbewertung vorsichtig in Abhängigkeit von dem eingesetzten Instrument erfolgen sollte.

Die Autoren empfehlen kein bestimmtes Instrument, sondern die sorgfältige Auswahl eines Instruments. Das ausgewählte Instrument sollte mithilfe anerkannter Methoden entwickelt sowie seine Reliabilität und Validität getestet werden. Außerdem sollte es eine Anleitung zur Anwendung geben, um ein standardisiertes Vorgehen zu gewährleisten. Die Notwendigkeit eines Konsensus hinsichtlich wichtiger und zentraler Items für Instrumente zur kritischen Bewertung von Studien wird herausgestellt.

Es handelt sich um eine Übersichtsarbeit mit adäquater methodischer Qualität und nachvollziehbaren Schlussfolgerungen. Allerdings ist die Datenbasis möglicherweise nicht vollständig, da fast ein Drittel der ausgewählten Publikationen nicht beschafft werden kann (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])).

**Moher et al. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists[143]**

Die Publikation stammt von Wissenschaftlern der Universität Ottawa, Kanada in Zusammenarbeit mit britischen Kollegen. Ziel ist es, eine mit Erläuterungen versehene Übersicht über vorhandene Skalen und Checklisten zur Qualitätsbewertung von RCT zu geben.

Die Autoren unterscheiden zwischen Berichts- und Studienqualität. Die Studienqualität wird im Sinn von methodischer Qualität definiert als „the confidence that the trial design, conduct, and analysis has minimized or avoided biases in its treatment comparisons“. Davon abgegrenzt wird die Berichtsqualität als „providing information about the design, conduct, and analysis of a trial“. Die Autoren führen eine systematische Literaturrecherche ohne Sprachrestriktion in MEDLINE für den Zeitraum zwischen Januar 1966 und Dezember 1992 durch. Außerdem werden Experten nach Instrumenten befragt. Instrumente, die nur geringe Modifikationen eines bekannten Instruments aufweisen, werden nicht berücksichtigt. Neben formalen Charakteristika wird aus den gefundenen Skalen und Checklisten extrahiert, ob mindestens ein Item aus den folgenden vier inhaltlichen Bereichen der internen Validität

vorhanden sind: Zuordnung der Intervention, Verblindung, Patienten-Follow-up und statistische Analyse (s. Tabelle 11: Extrahierte Kriterien der QBI bei Moher et al.[143]). Bei Skalen werden zusätzlich noch die methodische Stringenz der Instrumentenentwicklung, die Interrater-Reliabilität, die Zahl der Items und der erreichbare Score dargestellt.

**Tabelle 11: Extrahierte Kriterien der QBI bei Moher et al.[143]**

| Extrahierte Kriterien | Extrahiert bei SK, CL oder bei beiden Instrumententypen | Erläuterung bzw. mögliche Kategorien |
|---|---|---|
| Art der Skala | SK | Generisch oder spezifisch |
| Qualität definiert | SK, CL | Ja, nein, teilweise |
| Qualitätstyp | SK, CL | Berichts-, methodische Qualität, beides |
| Wahl der Items | SK, CL | Akzeptierte Kriterien (Lehrbuch zu klinischen Studien), Itempool (großer Itempool wird bei Instrumentenentwicklung reduziert) |
| Zuordnung der Patienten | SK, CL | Ja, nein (Gab es ein Item zu …) |
| Verblindung | SK, CL | Ja, nein (Gab es ein Item zu …) |
| Patienten-Follow-up | SK, CL | Ja, nein (Gab es ein Item zu …) |
| Statistische Analyse | SK, CL | Ja, nein (Gab es ein Item zu …) |
| Anzahl der Items | SK, CL | |
| Skalenentwicklung | SK | Ja, nicht berichtet |
| Interrater-Reliabilität | SK | Wert (Kappa-, Intraclass- oder Korrelationskoeffizient nach Pearson) |
| Bearbeitungszeit | SK, CL | In Minuten |
| Wertebereich | SK | |
| Detaillierte Angaben zur Bewertung | SK | Ja, nein |
| Erzielter Score in Metaanalyse | SK | Erzielter Score in Metaanalyse, in der das Instrument verwendet wird |

CL = Checkliste. QBI = Qualitätsbewertungsinstrument. SK = Skala.

Die Autoren identifizieren 25 Skalen und neun Checklisten. Von den Skalen sind 60 % generische Instrumente. Drei (12 %) Skalen werden zur Bewertung der Berichtsqualität, 32 % zur Bewertung der methodischen Qualität und der Rest wird für beide Arten von Qualität entwickelt. Die vier inhaltlichen Kriterien werden jeweils von 80 % und mehr der Skalen abgedeckt, nur das Patienten-Follow-up wird lediglich von 44 % der Skalen abgedeckt. Bei einem Drittel wird eine Gewichtung zu Berechnung des Gesamtscores eingesetzt. Nur die Skala von Jadad (sowohl die 3- als auch die 6-Items-Skala) weist eine genügende methodische Entwicklung auf. Von den Checklisten untersuchen drei (33 %) die Berichtsqualität, vier (44 %) die methodische Qualität und zwei (22 %) beide Arten von Qualität. Die Abdeckung der vier inhaltlichen Kriterien ist vergleichbar mit der der Skalen.

Insgesamt sehen die Autoren große Unterschiede bei den Instrumenten. Es bestehen deutliche Schwächen bei der methodischen Entwicklung der Skalen, oftmals wird das Konstrukt Studienqualität nicht definiert. Dies erklärt möglicherweise, dass einige Skalen sowohl methodische als auch Berichtsqualität abfragen. Generische Skalen haben den Vorteil, für unterschiedliche Bereiche einsetzbar zu sein. Checklisten weisen viele der Probleme von Skalen auf. Die Autoren sehen den Einsatz von Checklisten als Unterstützung für Autoren bei der Erstellung von Manuskripten bzw. für Reviewer von Fachzeitschriften bei der Bewertung von Manuskripten. Es wird die Verwendung von generischen Skalen empfohlen. Sie sind nach strengen methodischen Kriterien entwickelt und einfach zu handhaben.

Es handelt sich um eine Übersichtsarbeit mit adäquater Datensynthese. Die Schlussfolgerungen mit der Empfehlung von Skalen sind nachvollziehbar, da zum Publikationszeitpunkt die Problematik der impliziten Gewichtung von Items in Skalen noch nicht wissenschaftlich untersucht worden ist. Die Methode der Literaturauswahl und Datenextraktion kann wegen fehlender Angaben nicht beurteilt werden. Die Literaturrecherche beschränkt sich auf eine Datenbank, wird aber ergänzt um eine

Handsuche, Expertenbefragung und Internetrecherche, sodass kein wesentlicher Publikationsbias erwartet wird (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])).

**Olivo et al. Scales to assess the quality of randomized controlled trials: a systematic review[168]**

Die Übersichtsarbeit wird von kanadischen und australischen Gesundheitswissenschaftlern verfasst. Ziel ist es, den Inhalt, die Entwicklung sowie die Reliabilität und die Validität von QBI für RCT zu dokumentieren und ein angemessenes Instrument für die Evaluation der methodischen Qualität von RCT in den Bereichen von physikalischer Therapie und Rehabilitationsforschung zu identifizieren.

Es wird eine umfassende systematische Literaturrecherche in einschlägigen Datenbanken ohne Sprachrestriktion für den Zeitraum von 1965 bis März 2007, ergänzt um eine gezielte Handsuche, durchgeführt. Es werden nur Publikationen eingeschlossen, die die systematische methodische Entwicklung eines Instruments beschreiben. Die Literaturauswahl erfolgt unabhängig von fünf Reviewern. Neben formalen Charakteristika der Instrumente werden weitere methodische Eigenschaften (Intra- und Interrater-Reliabilität, Kontentvalidität, Konstruktvalidität und interne Konsistenz) extrahiert.

Über eine aufwändige Suche werden über 10.000 Dokumente identifiziert, von denen 105 in die Studie eingeschlossen werden. Aus den Dokumenten werden insgesamt 21 Instrumente extrahiert. Die Mehrheit ist nicht anhand nachvollziehbarer Methoden entwickelt oder auf Reliabilität und Validität getestet worden. Sieben Instrumente sind für den Bereich physikalische Therapie entwickelt bzw. eingesetzt worden (Jadad scale, Maastricht scale, Delphi list, PEDro scale, Maastricht-Amsterdam list, van Tulder scale und Bizzini scale). Von diesen weisen die Delphi list sowie die Jadad scale eine bessere Reliabilität und Validität auf, allerdings fehlen beiden Instrumenten Angaben zur internen Konsistenz. Bei der Jadad scale ist zusätzlich die Konstruktvalidität untersucht und zeigt damit insgesamt den besten Nachweis für Reliabilität und Validität.

Die Jadad scale ist zwar das Instrument, dessen Reliabilität und Validität getestet wird, allerdings nicht im Bereich der physikalischen Therapie sondern für Schmerztherapie. Außerdem umfasst es nur wenige Qualitätskomponenten, wobei insbesondere die Doppelverblindung bei physiotherapeutischen Interventionen oft nicht anwendbar ist. Die Autoren halten keines der untersuchten Instrumente für geeignet. Sie empfehlen die Entwicklung eines validen und reliablen Instruments.

Die Übersichtsarbeit zeichnet sich durch eine hohe methodische Qualität aus (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])). Allerdings diskutieren die Autoren nicht die Problematik der impliziten Gewichtung von Items in Skalen, zu der es bereits empirische Untersuchungen zum Publikationszeitpunkt gibt.

**Sanderson et al. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography[189]**

Dieser Review wird von Mitarbeitern der Universität Cambridge, Großbritannien publiziert. Ziel ist es, eine Übersicht über Instrumente zur Bewertung der Qualität oder der Anfälligkeit für Bias in Beobachtungsstudien zu liefern und ggf. ein Instrument zur breiten Anwendung empfehlen zu können.

Es wird eine systematische Literaturrecherche in den drei Datenbanken MEDLINE, EMBASE und Dissertation Abstract bis März 2005 ohne Sprachrestriktion durchgeführt. Die Suchstrategie wird dargestellt, aber es gibt kein Fließdiagramm zur Literaturauswahl. Die Suche wird ergänzt von einer Internetrecherche mit der Google®-Suchmaschine im März 2005 und von einer Durchsicht der Referenzen. Die Datenextraktion führen zwei Autoren durch, Diskrepanzen werden durch Diskussion mit einem Dritten gelöst.

Es werden Instrumente eingeschlossen, die die Studienqualität oder das Verzerrungspotenzial in Kohorten-, Fall-Kontroll- oder Querschnittstudien beurteilten. Die Instrumente werden unterteilt in Checklisten, Skalen und Checklisten mit Gesamtbewertung. Checklisten mit Gesamtbewertung sind solche, die eine qualitative Gesamteinschätzung enthalten, z. B. „hoch“, „mittel“ oder „niedrig“. Die inhaltlichen Komponenten der Instrumente werden sechs vorab definierten Domänen zugeordnet, deren Auswahl sich am Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)-Statement (Instrument zur Berichtsqualität von Beobachtungsstudien) orientiert (s. Tabelle 12: Domänen zur Bewertung der Instrumenteninhalte bei Sanderson et al.[189]).

**Tabelle 12: Domänen zur Bewertung der Instrumenteninhalte bei Sanderson et al.[189]**

| Domänen | Erläuterung |
|---|---|
| Methoden zur Auswahl der Studienteilnehmer | Angemessene Quellpopulation (Fälle, Kontrollen, Kohorten) **und** Einschluss- **oder** Ausschlusskriterien |
| Methoden zur Messung von Expositions- und Outcomevariablen | Angemessene Messmethoden für Exposition(en) und Outcome(s) |
| Designspezifische Biasquellen (außer Confounding) | Angemessene Methoden zur Vermeidung designspezifischer Bias wie Recall-, Interviewerbias oder Verblindung |
| Methoden zur Kontrolle von Confounding | Angemessenes Design und/oder angemessene analytische Methoden |
| Statistische Methoden (außer Confounder-Kontrolle) | Angemessene Statistik für die Analyse des Effekts |
| Interessenkonflikt | Erklärung von Interessenkonflikten oder Angabe der Quelle der finanziellen Förderung |

Die Recherche ergibt insgesamt 86 Instrumente (72 % aus der Datenbankrecherche, 28 % aus der Internetrecherche), darunter 41 Checklisten (48 %), 33 Skalen (38 %) und zwölf Checklisten mit einer Gesamtbewertung (14 %). Ein Drittel der Instrumente ist für die einmalige Verwendung in einem spezifischen Review konzipiert worden, bei 15 % handelt es sich um ein generisches Instrument für systematische Reviews, 36 % sind für die kritische Bewertung (critical appraisal) bestimmt. Die Verteilung der Bestimmung ist zwischen den Instrumententypen sehr unterschiedlich: Einfache Checklisten sind überwiegend für die kritische Bewertung konzipiert, Checklisten mit einer Gesamtbewertung sind etwa zu gleichen Teilen für den einmaligen Gebrauch, als generisches Instrument oder für die kritische Bewertung vorgesehen, während die Mehrheit der Skalen für den einmaligen Gebrauch geplant ist. Bei 15 % der Instrumente ist die keine eindeutige Bestimmung möglich. Hinsichtlich der Domänen enthalten 92 % der Instrumente Items zu Methoden der Auswahl der Studienteilnehmer, 86 % jeweils Items zu Methoden der Messung von Exposition und Outcome sowie zu designspezifischen Bias und 78 % jeweils zu Methoden der Confounderkontrolle und statistischen Methoden. Ein Interessenkonflikt wird nur bei 3 % der Instrumente berücksichtigt. Es besteht eine große Variabilität der Instrumente hinsichtlich der Items pro Domäne und nach Art des Instruments. Die detaillierte methodische Entwicklung wird bei 54 % der Instrumente beschrieben.

Die Autoren geben zu bedenken, dass trotz umfassender Recherche wahrscheinlich nicht alle Instrumente für Beobachtungsstudien gefunden werden, da viele für spezifische Übersichtsarbeiten entwickelt werden, die sehr schwer in den elektronischen Datenbanken zu identifizieren sind. Die Autoren sind besorgt über Instrumente, deren Bestimmung nicht eindeutig ist und fordern die Unterscheidung von Berichtsqualität und der Qualität dessen, was tatsächlich in der Studie gemacht wird. Etwa die Hälfte aller Checklisten wird als geeignet für den zukünftigen Einsatz betrachtet, nämlich die Instrumente, die nach Einschätzung der Autoren die drei grundlegenden Domänen zur Auswahl der Studienteilnehmer, Messung der Variablen und Kontrolle von Confounding abdecken. (Eine Begründung für die Wahl der drei Schlüsseldomänen wird nicht gegeben.) Ein bestimmtes Instrument zur Qualitätsbewertung von Beobachtungsstudien wird jedoch nicht empfohlen, ohne zuvor seine Eigenschaften und Anwendbarkeit geprüft zu haben. Die allgemeine Empfehlung der Autoren lautet: Instrumente sollten (1) wenige Schlüsseldomänen enthalten, (2) so spezifisch wie möglich sein, (3) eine einfache Checkliste anstatt einer Skala darstellen und (4) sorgfältig entwickelt, reliabel und valide sein.

Es handelt sich um eine Übersichtsarbeit mit adäquater Methodik bei der Literaturrecherche, Datenextraktion und -synthese sowie nachvollziehbaren Schlussfolgerungen (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])). Zum Vorgehen bei der Literaturauswahl werden keine Angaben gemacht, sodass ein Reviewerbias nicht beurteilt werden kann.

**Saunders et al. Assessing the Methodological quality of Nonrandomized Intervention Studies[190]**

Die Autoren aus Kanada, Großbritannien und Norwegen wollen mit dieser Studie einen Beitrag zur zukünftigen Entwicklung eines Standardinstruments zur Bewertung von nicht-randomisierten Studien für systematische Reviews der Cochrane Collaboration leisten.

Mit einer systematischen Literaturrecherche in MEDLINE (Zeitraum: 1966 bis März 1999), begrenzt auf englischsprachige Dokumente, werden 18 Instrumente identifiziert[14, 19, 47, 60, 82, 88, 97, 130, 135, 139, 167, 176, 201, 208, 210, 211, 247], von denen zehn als Skalen und acht als Checklisten eingeordnet werden. Die Datenextraktion wird standardisiert mit einem Formular erhoben, das auf publizierten Vorgaben basiert (u. a. [143]). Die Items sind drei Bereichen zugeordnet: den (1) formalen Aspekten des Instruments, den (2) inhaltlichen Aspekten und der (3) Entwicklung des Instruments. Die formalen Aspekte umfassen die Bewertung von methodischer bzw. Berichtsqualität, Auswahl (aufgrund publizierter Kriterien oder standardisierter Techniken zur Entwicklung von Skalen) und Anzahl der Items. Die inhaltlichen Items decken die Bereiche Studienziel, Studienteilnehmer, Interventionen, Outcomes und statistische Analyse ab. Die zugehörigen Items sind in der folgenden Tabelle dargestellt. Zur Entwicklung der Instrumente wird extrahiert, ob detaillierte Instruktionen zur Anwendung des Instruments vorliegen, der Wertebereich des Instruments, Methodik der Skalenentwicklung sowie Reliabilität und Validität.

**Tabelle 13: Klassifikation von Items zum Inhalt der Instrumente bei Saunders et al.[190]**

| **Übergeordnete Items** | **Items** |
|---|---|
| Präzise Studienziele | |
| Studienteilnehmer | |
| | Beschreibung der Studienteilnehmer |
| | Repräsentativität der Studienteilnehmer |
| | Methode der Zuordnung der Intervention |
| | Teilnahmerate |
| | Vergleichbarkeit der Gruppen |
| | Sonstige |
| Interventionen | |
| | Eindeutig definierte/objektive Interventionen |
| | Messungen valide/reliabel und gleichermaßen in den Studiengruppen angewandt |
| | Kointerventionen |
| | Kontamination |
| | Compliance |
| | Verblindung der Teilnehmer |
| | Verblindung des medizinischen Personals |
| | Sonstiges |
| Outcomes | |
| | Klar definierte/objektive Outcomes |
| | Adäquate Länge des Follow-up |
| | Charakteristika von Studienabbrechern |
| | Verblindete Erhebung des Outcomes |
| | Sonstiges |
| Statistische Analyse | |
| | Intention-to-treat-Analyse |
| | Adjustierung für Confounder |
| | Vorab festgelegte Analysen |
| | Statistische Signifikanz |
| | Klinische Signifikanz |
| | Daten bereitgestellt zur Bestätigung der Ergebnisse |

Die Instrumente variieren stark im Hinblick auf den Anwendungsbereich, Zahl und Inhalt der Items sowie die methodische Stringenz ihrer Entwicklung. Die meisten Instrumente sind für die Anwendung für Kohortenstudien (n = 5) bzw. eine Vielzahl von Studiendesigns (n = 11) vorgesehen. Die Zahl der Items liegt zwischen 4 und 39. Alle Instrumente decken inhaltlich sowohl die methodische als auch die Berichtsqualität ab. Nur ein Instrument ist mithilfe nachvollziehbarer methodischer Prozesse entwickelt

und validiert[60], während die übrigen Instrumente ihre Items aufgrund von allgemein anerkannten/ publizierten Kriterien ausgewählt haben. Bei drei weiteren Instrumenten wird die Reliabilität bestimmt.

Die Autoren beschreiben ihre Untersuchung als einleitend (preliminary) und erkennen eine Limitation in der Beschränkung der Literatursuche auf eine Datenbank und die englische Sprache. Trotzdem erscheint das Ergebnis relevant, das eine große Varianz der Instrumente bei gleichzeitiger Schwäche ihrer methodischen Entwicklung zeigt. Die Autoren schätzen die von Downs und Black[60] entwickelte Skala als das beste Instrument ein, weil es basierend auf wissenschaftlichen Methoden entwickelt sowie seine Reliabilität und Validität getestet wird. Außerdem deckt es die meisten wichtigen Qualitätskriterien bis auf die Verblindung der Teilnehmer, Kointerventionen und Kontamination ab. Bis es ein besseres Instrument gibt, empfehlen die Autoren dieses Instrument, allerdings als Checkliste und nicht als Skala. Weitere Untersuchungen sind erforderlich, um Studiencharakteristika zu identifizieren, die die methodische Qualität von nicht-randomisierten Studien beeinflussen.

Es handelt sich um eine methodisch adäquate und transparente Übersichtsarbeit mit nachvollziehbaren Schlussfolgerungen. Da sich die Literatursuche auf eine Datenbank beschränkt, kann nicht ausgeschlossen werden, dass relevante Dokumente nicht berücksichtigt werden (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])).

**West et al. Systems to rate the strength of scientific evidence[235]**

Es handelt sich um einen HTA- bzw. Evidenzbericht im Auftrag der AHRQ, Rockville, Maryland, USA.

Es werden eine systematische Literaturrecherche in MEDLINE von 1995 bis Mitte 2000 nach englischsprachigen Dokumenten sowie eine ergänzende Handsuche durchgeführt. Von 1.602 gescreenten Publikationen werden 109 Dokumente ausgewählt. Zusammen mit zwölf Berichten von unterschiedlichen AHRQ-unterstützten evidenzbasierten Praxiszentren bilden 121 Quellen die Grundlage des Berichts.

Die Datenextraktion basiert zunächst auf einer Einteilung nach Studiendesign: systematischer Review und Metaanalysen, RCT, Beobachtungsstudien und Studien zu diagnostischen Tests. Zu jedem Studiendesign werden Domänen (s. Tabelle 14: Domänen der einzelnen Studiendesigns bei West et al.[235]) und innerhalb der Domänen Elemente zugeordnet, die relevant zur Bewertung der Studienqualität sind. Die gewählten Domänen und Elemente gründen überwiegend auf allgemein akzeptierten Kriterien, einige sind empirisch belegt. Es werden essenzielle Elemente definiert, die enthalten sein müssen, damit eine Domäne als abgedeckt gilt. Fragen nach der externen Validität der Instrumente sind nicht Bestandteil der Datenextraktion, da die Autoren darauf hinweisen, dass diese davon abhängt, auf welche Person/Subpopulation die Ergebnisse der Studie übertragen werden sollen.

**Tabelle 14: Domänen der einzelnen Studiendesigns bei West et al.[235]**

| **Studiendesign** | Systematischer Review | RCT | Beobachtungsstudien | Studien zu diagnostischen Tests |
|---|---|---|---|---|
| **Domänen** | • Studienfrage<br>• Suchstrategie*<br>• Ein- und Ausschlusskriterien<br>• Interventionen<br>• Outcomes<br>• Datenextraktion<br>• Studienqualität und -validität*<br>• Datensynthese und -analyse*<br>• Ergebnisse<br>• Diskussion<br>• Finanzielle Förderung* | • Studienfrage<br>• Studienpopulation<br>• Randomisierung*<br>• Verblindung*<br>• Interventionen<br>• Outcomes<br>• Statistische Analyse*<br>• Ergebnisse<br>• Diskussion<br>• Finanzielle Förderung* | • Studienfrage<br>• Studienpopulation<br>• Vergleichbarkeit der Teilnehmer*<br>• Exposition oder Intervention<br>• Messung der Outcomes<br>• Statistische Analyse<br>• Ergebnisse<br>• Diskussion<br>• Finanzielle Förderung* | • Studienfrage*<br>• Adäquate Beschreibung des Tests*<br>• Angemessener Referenzstandard*<br>• Verblindeter Vergleich von Test- und Referenzstandard*<br>• Vermeidung von Verifikationsbias* |

RCT = Randomisierte kontrollierte Studie.

* Basieren auf empirischen Erkenntnissen.

Es wird die folgende Anzahl von Instrumenten gefunden: elf für systematische Reviews, 32 für RCT, zwölf für Beobachtungsstudien und sechs für diagnostische Tests. Von diesen Instrumenten werden im Sinn einer „Best practice"-Orientierung diejenigen ausgewählt, die mindestens teilweise sogenannte Schlüsseldomänen abdecken. Auf diese Weise ergeben sich fünf Instrumente für systematische Reviews[10, 13, 104, 111, 188] acht für RCT[39, 54, 60, 84, 125, 181, 198, 222], sechs für Beobachtungsstudien[60, 76, 84, 181, 204, 245] und drei für diagnostische Studien[48, 126, 156]. Insgesamt handelt es sich um 19 Instrumente, da drei sowohl für die Bewertung von RCT als auch für Beobachtungsstudien verwendet werden können.

Die Autoren haben 19 Instrumente zur Qualitätsbewertung identifiziert, die die vorab festgelegten Qualitätsaspekte abdecken. Sie empfehlen weitere Untersuchungen zum Nachweis der Bedeutung von inhaltlichen Studienelementen und der Studienqualität bzw. Studienergebnissen.

Es handelt sich um eine methodisch adäquate und transparente Übersichtsarbeit mit nachvollziehbaren Schlussfolgerungen. Aufgrund der zeitlichen Beschränkung der Suche und Einschränkung auf eine Datenbank besteht die Möglichkeit, dass relevante Instrumente übersehen werden (s. Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])).

**Whiting et al. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools[239]**

Diese Übersichtsarbeit wird von britischen und niederländischen Wissenschaftlern verfasst. Sie basiert auf einem HTA-Bericht[237] des NHS R&D Health Technology Assessment Programms derselben Autoren. Ziel ist es, einen Überblick über existierende Instrumente zur Qualitätsbewertung von Studien zu diagnostischen Tests zu geben und darzustellen, auf welche Weise die Ergebnisse der Qualitätsbewertung in die Datensynthese integriert werden.

Es werden eine systematische Literaturrecherche in einschlägigen Datenbanken ohne Sprachrestriktion bis April 2001 durchgeführt und Experten nach zusätzlichen Studien befragt. Jedes Instrument wird nur mit einer Publikation eingeschlossen. Die Literaturauswahl nehmen zwei Reviewer vor. Folgende Daten werden extrahiert: Definition von Qualität, Ziel des Instruments (Qualitätsbewertung, Leitfaden für die Interpretation, Publikation oder Durchführung einer Studie zu diagnostischen Tests), Art des Instruments (Checkliste, Skala, Evidenzlevel), Inhalte der Items, Anwendungsbereich, Angaben zur Instrumentenentwicklung. Es werden 27 Items dokumentiert, die die vier Bereiche (1) Potenzial für Bias (n = 10), Übertragbarkeit auf die interessierende Population (n = 4), die Durchführung der Studie (n = 4) und die Berichtsqualität (n = 9) abdeckten (s. Tabelle 15: Verwendete Items zur Klassifikation von QBI für Studien zu diagnostischen Tests bei Whiting et al.[236]).

**Tabelle 15: Verwendete Items zur Klassifikation von QBI für Studien zu diagnostischen Tests bei Whiting et al.[236]**

| Items | Beschreibung |
|---|---|
| **Verzerrungspotenzial (10 Items)** | |
| Referenzstandard | Wird ein angemessener Referenztest verwendet um den Zielparameter zu erfassen? |
| Bias durch Krankheitsprogression | Kann eine Änderung des Krankheitsstatus zwischen Durchführung des Index- und des Referenztests aufgetreten sein? |
| Verifikationsbias | Wird bei allen Teilnehmern der Zielparameter mit dem gleichen Referenztest verifiziert? |
| Bias durch nicht-unabhängige Tests (Incorporation bias) | Ist der Indextest Teil des Referenztests? (Waren die Tests nicht unabhängig voneinander? |
| Behandlungsparadox | Wird die Behandlung basierend auf dem Ergebnis des Indextests eingeleitet bevor der Referenztest erfolgte? |
| Reviewbias | Werden die Ergebnisse des Indextests verblindet gegenüber dem Ergebnis des Referenztests ausgewertet und umgekehrt? |
| Klinischer Reviewbias | Sind klinische Informationen vorhanden, als die Testergebnisse ausgewertet werden? |
| Beobachter-/Instrumenten-variabilität | Ist es wahrscheinlich, dass eine Beobachter-/Instrumentenvariabilität Annahmen bei der Testausführung beeinflusst hat? |
| Umgang mit nicht bewertbaren Testergebnissen | Werden nicht bewertbare Testergebnisse in die Analyse eingeschlossen? |
| Zufällige Wahl des Grenzwertes | Wird der Grenzwert des Tests unabhängig von den Studienergebnissen gewählt? (D. h., er sollte nicht zur Verbesserung der Testperformance gewählt worden sein) |

**Tabelle 15: Verwendete Items zur Klassifikation von QBI für Studien zu diagnostischen Tests bei Whiting et al. – Fortsetzung**

| Items | Beschreibung |
|---|---|
| **Externe Validität (4 Items)** | |
| Spektrum der Teilnehmer | Ist die untersuchte Bevölkerung vergleichbar mit der interessierenden Population? |
| Rekrutierung | Ist die Rekrutierungsmethode angemessen um ein geeignetes Spektrum an Patienten einzuschließen? |
| Krankheitsprävalenz/-schwere | Sind Prävalenz der Erkrankung und Spektrum der Krankheitsschwere in der Studienpopulation vergleichbar mit der der interessierenden Population? |
| Methodenwechsel beim Indextest | Ist es wahrscheinlich, dass die Methode des Tests im Laufe der Studie verändert wird? |
| **Studiendurchführung (4 Items)** | |
| Subgruppenanalysen | Sind Subgruppenanalysen angemessen und vorab spezifiziert worden? |
| Stichprobengröße | Wird eine angemessene Teilnehmerzahl in die Studie eingeschlossen? |
| Studienziele | Sind die Studienziele relevant für die Studienfrage? |
| Protokoll | Wird ein Studienprotokoll vor Studienbeginn entwickelt und wird es befolgt? |
| **Berichtsqualität (9 Items)** | |
| Einschlusskriterien | Werden die Einschlusskriterien präzise dargestellt? |
| Indextestdurchführung | Wird die Methodik des Indextests genügend detailliert beschrieben, um die seine Replikation zu ermöglichen? |
| Referenztestdurchführung | Wird die Methodik des Referenztests genügend detailliert beschrieben, um seine Replikation zu ermöglichen? |
| Definition von „normalem“ Testergebnis | Haben die Autoren präzise dargestellt, was als „normales“ Testergebnis bewertet wird? |
| Angemessene Ergebnisse | Werden adäquate Ergebnisse dargestellt? Z. B. Sensitivität, Spezifität, Likelihood ratios |
| Genauigkeit der Ergebnisse | Wird die Präzision der Ergebnisse dargestellt? Z. B. Konfidenzintervalle |
| Studienabbrecher | Werden alle Studienteilnehmer bei der Analyse berücksichtigt? |
| Datentabelle | Wird eine Kreuztabelle zur Testdurchführung dargestellt? |
| Nützlichkeit des Tests | Werden Hinweise gegeben, wie nützlich der Test in der Praxis sein kann? |

QBI = Qualitätsbewertungsinstrument.

Insgesamt werden 67 Instrumente identifiziert, von denen nur zwei Auskunft über die Instrumentenentwicklung geben. Kein Instrument wird systematisch evaluiert. Inhaltlich decken fast alle Instrumente mindestens ein Item zum Biaspotenzial ab, darunter am häufigsten zum Reviewbias, zur Angemessenheit des Referenzstandards und zum Verifikationsbias. Items zur Übertragbarkeit der Ergebnisse werden von acht Instrumenten gar nicht eingesetzt. Am häufigsten werden die Zusammensetzung der Studienpopulation und deren Rekrutierung berücksichtigt. Weniger als die Hälfte der Instrumente deckt Items zur Studiendurchführung ab, am häufigsten wird die Größe der Studienpopulation bewertet. In etwa zwei Drittel der Instrumente wird die Berichtsqualität erfragt, aber keines der neun Items wird besonders oft verwendet. Häufigere Items sind die detaillierte Beschreibung des Indextests, die Definition eines normalen Testergebnisses und die angemessene Ergebnisdarstellung mit Angabe von Sensitivität, Spezifität etc.

Die Untersuchung zur Integration der Qualitätsbewertung zeigt, dass in insgesamt 114 Dokumenten nur bei 51 % eine Qualitätsbewertung durchgeführt wird. Eine fehlende Berücksichtigung einer durchgeführten Qualitätsbewertung in den Ergebnissen liegt bei 3 % der Studien vor. Bei 19 % der Studien wird die Qualitätsbewertung nur in einer Tabelle oder narrativ beschrieben. Alternativ werden die Ergebnisse der Qualitätsbewertung als Einschlusskriterium genutzt (13 %), für eine Sensitivitätsanalyse (11 %) oder Regressionsanalyse (6 %) eingesetzt oder in den Empfehlungen für zukünftige Forschung festgehalten (10 %).

Bei der großen Variation unter den verfügbaren Instrumenten und gleichzeitig mangelnden Angaben zur Instrumentenentwicklung und -evaluation ist die Wahl eines Bewertungsinstruments für Studien von diagnostischen Tests schwierig. Die Autoren verweisen auf das von ihnen entwickelte, evidenzbasierte Bewertungsinstrument Quality assessment of diagnostic accuracy studies (QUADAS).

Es handelt sich um eine methodisch adäquate und transparente Übersichtsarbeit mit nachvollziehbaren Schlussfolgerungen. Da die Auswahl der Literatur und die Datenextraktion nur von einem Reviewer vorgenommen und von einem zweiten kontrolliert werden, kann ein Reviewerbias nicht ausgeschlossen werden (Tabelle 9: Ergebnisse der Qualitätsbewertung von systematischen Übersichtsarbeiten (mod. nach West et al.[235])).

**Zusammenfassung**

Die Übersichtsarbeiten variieren hinsichtlich der Studiendesigns und der Bereiche, auf die die Instrumente angewendet werden. Jedoch stellen die Autoren übereinstimmend fest, dass es bei einer Vielzahl von Instrumenten keinen Goldstandard bei der Qualitätsbewertung gibt und dass die Mehrheit der Instrumente nicht auf wissenschaftlichen Methoden basierend entwickelt bzw. deren Reliabilität und Validität getestet wird.

### 6.5.1.3 Identifizierte QBI – Übersicht über formale Charakteristika

Insgesamt werden 125 Instrumente zur Qualitätsbewertung identifiziert. Es werden 15 Instrumente für systematische Reviews eingeschlossen, 80 Instrumente für Interventions-, 30 für Beobachtungs- und 17 für Diagnosestudien. Darunter befinden sich 16 Instrumente, die sich sowohl für die Bewertung von Interventions- als auch für die von Beobachtungsstudien eignen.

Ein Überblick über die formalen Charakteristika der QBI stratifiziert nach Studiendesign gibt Tabelle 16: Übersicht über formale Charakteristika nach Studiendesign. Der Vergleich der Bewertungsinstrumente über alle Studiendesigns hinweg zeigt, dass der größte Teil der Instrumente (80 bis 95 %) in englischer Sprache verfasst ist. Die meisten Instrumente (67 bis 77 %) sind Originale, die kein oder mehr als ein anderes Instrument zur Grundlage hatten.

Hinsichtlich des Anteils generischer vs. spezifischer Instrumente gibt es deutliche Unterschiede. So sind alle Instrumente für systematische Reviews generisch, während bei den Instrumenten für Interventionsstudien nur 52 % der Instrumente generisch sind. Bei Beobachtungs- und Diagnosestudien sind es 77 % bzw. 71 % der Instrumente. Die Anzahl der Items weist eine hohe Spannbreite auf und ist bei Interventionsstudien (drei bis 81 Items) am größten und bei Diagnosestudien (fünf bis 35 Items) am geringsten.

**Tabelle 16: Übersicht über formale Charakteristika nach Studiendesign**

| | Systematische Reviews | | Interventions-studien | | Beobachtungs-studien | | Diagnose-studien | |
|---|---|---|---|---|---|---|---|---|
| | 15 Instrumente | | 80 Instrumente | | 30 Instrumente | | 17 Instrumente | |
| | Anzahl (n) | Anteil (%) | Anzahl (n) | Anteil (%) | Anzahl (n) | Anteil (%) | Anzahl (n) | Anteil (%) |
| **Englische Sprache** | 12 | 80 % | 76 | 95 % | 28 | 93 % | 14 | 82 % |
| **Deutsche Sprache** | 3 | 20 % | 4 | 5 % | 2 | 7 % | 3 | 18 % |
| **Originalinstrument** | 15 | 67 % | 80 | 67 % | 23 | 77 % | 13 | 76 % |
| **Modifiziertes Instrument** | 5 | 33 % | 26 | 33 % | 7 | 23 % | 4 | 24 % |
| **Generisches Instrument** | 15 | 100 % | 42 | 52 % | 23 | 77 % | 12 | 71 % |
| **Spezifisches Instrument** | 0 | 0 % | 38 | 48 % | 7 | 23 % | 5 | 29 % |
| **Anzahl der Items** | 5-80 | | 3-81 | | 5-60 | | 5-35 | |
| **Checkliste** | 12 | 80 % | 18 | 23 % | 15 | 50 % | 12 | 71 % |
| **Komponentensystem** | 0 | 0 % | 10 | 13 % | 2 | 7 % | 1 | 6 % |
| **Skala** | 3 | 20 % | 52 | 65 % | 13 | 43 % | 4 | 24 % |
| **Qualitative Komponentenbewertung** | 0 | 0 % | 8 | 10 % | 5 | 17 % | 2 | 12 % |

**Tabelle 16: Übersicht über formale Charakteristika nach Studiendesign – Fortsetzung**

| | Systematische Reviews | | Interventions-studien | | Beobachtungs-studien | | Diagnose-studien | |
|---|---|---|---|---|---|---|---|---|
| | 15 Instrumente | | 80 Instrumente | | 30 Instrumente | | 17 Instrumente | |
| | Anzahl (n) | Anteil (%) | Anzahl (n) | Anteil (%) | Anzahl (n) | Anteil (%) | Anzahl (n) | Anteil (%) |
| **Definition von Qualität** | 2 | 13 % | 11 | 14 % | 4 | 13 % | 3 | 18 % |
| **Entwicklungsprozess beschrieben** | 3 | 20 % | 18 | 23 % | 8 | 27 % | 5 | 29 % |
| **Ausführliche Ausfüllhinweise** | 6 | 40 % | 23 | 29 % | 10 | 33 % | 9 | 53 % |
| **Kurze Ausfüllhinweise** | 6 | 40 % | 26 | 33 % | 12 | 40 % | 4 | 24 % |
| **Keine Ausfüllhinweise** | 3 | 20 % | 31 | 38 % | 8 | 27 % | 4 | 23 % |
| **Zeitbedarf angegeben** | 0 | 0 % | 3 | 4 % | 3 | 10 % | 0 | 0 % |
| **Testgüte berichtet** | 0 | 0 % | 21 | 26 % | 4 | 13 % | 4 | 24 % |

Für die Bewertung von systematischen Reviews werden vorrangig Checklisten entwickelt (80 % der Instrumente), ebenso bei Diagnosestudien (71 %). Für Interventionsstudien werden primär Skalen (65 %) entwickelt. Bei Beobachtungsstudien gibt es keine so deutliche Trennung, 50 % der Instrumente stellen Checklisten dar, 43 % Skalen. Komponentensysteme wie z. B. das „Risk of bias tool" der Cochrane Collaboration werden in sehr viel geringerer Zahl entwickelt. So werden keine Instrumente auf Komponentenbasis für systematische Reviews identifiziert (0 %). Am häufigsten finden sie sich im Bereich der Qualitätsbewertung von Interventionsstudien (13 %). Im Gegensatz zu Skalen mit einem rein quantitativen Ansatz ist bei Checklisten und Komponentensystemen eine qualitative Komponenten- und/oder qualitative Gesamtbewertung machbar. Diese ist jedoch nur selten Bestandteil von QBI. Während bei Beobachtungsstudien eher eine qualitative Komponentenbewertung (17 %) möglich ist, kann bei der Bewertung systematischer Reviews nur auf Systeme mit qualitativer Gesamtbewertung zurückgegriffen werden.

Das Verständnis von methodischer Qualität, Studienqualität bzw. interner Validität wird nur in einem geringen Teil der Publikationen dargestellt. Erläuterungen zum jeweiligen Qualitätskonzept liegen für 13 % der Instrumente für Übersichtsarbeiten, für 14 % der Instrumente für Interventions-, für 13 % der Instrumente für Beobachtungs- und für 18 % der Instrumente für Diagnosestudien vor. Die Mehrheit der Qualitätskonzepte definiert die methodische Qualität bzw. Studienqualität anhand der Fähigkeit, eine hohe interne Validität zu erzielen, also sozusagen die „therapeutische Wahrheit" (Jadad et al.[106]) abzubilden und das Risiko von Bias zu minimieren. Cho und Bero[44] weisen darüber hinaus darauf hin, dass die methodische Qualität nur in dem Maß bewertet werden kann, in dem eine ausreichend hohe Berichtsqualität dies ermöglicht. Das LBI[133] äußert, dass eine hohe interne Validität beinhaltet, dass die untersuchte Exposition/Intervention und nicht Bias oder nicht-systematische Fehler das Outcome hervorrufen.

Demgegenüber wird der Entwicklungsprozess des Instruments (20 bis 29 %) etwas häufiger erläutert. Ausführliche oder kurze Ausfüllhinweise bzw. Vorgaben zur Operationalisierung finden sich bei 71 bis 80 % der Bewertungsinstrumente. Der veranschlagte Zeitbedarf für die Durchführung der Qualitätsbewertung ist nur bei 0 bis 10 % der Instrumente angegeben. Hinsichtlich der Testgüte finden sich bei den Instrumenten für systematische Reviews keine dokumentierten Angaben zur Testgüte. Bei Beobachtungsstudien liegen diese Daten für 13 % der Instrumente vor, bei Interventions- sowie Diagnosestudien sind für immerhin 26 % bzw. 24 % der Instrumente Angaben zu Validität und/oder Reliabilität aufgeführt.

#### 6.5.1.4 QBI für systematische Reviews, HTA und Metaanalysen

Durch die Literaturrecherche werden 15 Instrumente zur Qualitätsbewertung von systematischen Reviews, HTA bzw. Metaanalysen identifiziert (Tabelle 39: Formale Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen, Tabelle 43: Inhaltliche Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen, Teil 1, Tabelle 44: Inhaltliche Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen, Teil 2).

**Formale Charakteristika**

Unter diesen 15 Instrumenten befinden sich drei deutschsprachige: ein Instrument des LBI[133] sowie zwei Instrumente der GSWG[69]. Fünf der Instrumente sind modifizierte Versionen eines anderen Instruments. Alle Instrumente sind generisch. Die Anzahl der Items reicht von acht bis 80. Zwölf der Instrumente sind Checklisten, drei Skalen. Unter den Checklisten befinden sich zwei Instrumente mit einer qualitativen Gesamtbewertung, alle anderen Instrumente machen keine Angaben zur Ausgestaltung einer qualitativen Bewertung der jeweiligen Studie.

Für zwei der 15 Instrumente zur Qualitätsbewertung von systematischen Übersichtsarbeiten (13 %) existieren Erläuterungen des Qualitätskonzepts, das dem Bewertungsinstrument bzw. dessen Anwendung zugrunde liegt (Tabelle 17: Qualitätskonzepte bei QBI für systematische Reviews, HTA und Metaanalysen).

**Tabelle 17: Qualitätskonzepte bei QBI für systematische Reviews, HTA und Metaanalysen**

| Publikation | Qualität/Validität |
|---|---|
| Ludwig Boltzmann Institut[133] | „Hohe interne Validität impliziert, dass die untersuchte Exposition bzw. Intervention und nicht ein Bias (systematischer Fehler) oder ein zufälliger (nicht-systematischer) Fehler für die Resultate verantwortlich ist.“ |
| Oxman et al.[170] | "We are attempting to measure the extent to which a review is likely to be valid; i. e. the extent to which an overview guards against bias (by which we can mean systematic deviation from the truth)." |

HTA = Health Technology Assessment. QBI = Qualitätsbewertungsinstrument.

Hinweise zur Anwendung des Instruments bzw. zur Operationalisierung finden sich in jeweils sechs Instrumenten ausführlich bzw. kurz. Der veranschlagte Zeitbedarf für die Anwendung des Instruments sowie Angaben zur Testgüte liegen für keines der Instrumente vor. Alle Instrumente sind ausschließlich für systematische Reviews, Metaanalysen bzw. HTA angedacht. Eine Übersicht der formalen Aspekte der QBI für systematische Reviews, Metaanalysen und HTA findet sich im Anhang in Tabelle 39: Formale Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen.

**Inhaltliche Charakteristika**

Eine angemessene und präzise Fragestellung wird bei 87 % der Instrumente abgefragt. Die a priori Definition von Ein- und Ausschlusskriterien ist bei 53 % der Instrumente enthalten, die Angemessenheit dieser Kriterien bei 27 % der Instrumente (s. Tabelle 43: Inhaltliche Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen, Teil 1).

Von den Instrumenten berücksichtigen 73 % relevante Datenbanken, 80 % weitere Datenquellen. Die Dokumentation der Suchstrategie ist in 60 % der Instrumente enthalten. Rund die Hälfte der Instrumente (47 %) beinhaltet ein Item zu Restriktionen bei der Suche, 40 % erfragen den Ausschluss von Literatur. Weniger häufig wird die ausreichende Detailliertheit der Literaturrecherche (20 %), die Literaturauswahl durch mindestens zwei voneinander unabhängige Reviewer (13 %) sowie die Kombination von Schlagworten und Suchworten (7 %, n = 1) dargestellt. Keines der Instrumente berücksichtigt, ob eine angemessene Anzahl von Synonymen verwendet wird.

Der Bereich Datenextraktion wird in den Instrumenten vergleichsweise selten berücksichtigt. Das häufigste Item ist die voneinander unabhängige Datenextraktion durch mindestens zwei Reviewer (27 %). Rund ein Viertel der Instrumente (20 %) erfragt die ausreichende Detailliertheit der Datenextraktion. Während ebenfalls 20 % der Instrumente die Extraktion von Outcomes einbeziehen, fragen nur 13 % nach der Extraktion der Interventionen bzw. Expositionen. Lediglich 7 % der Instrumente (n = 1) erheben die Messung der Übereinstimmung der Reviewer. Kein Instrument bezieht die Verblindung der Datenextraktion in die Qualitätsbewertung mit ein.

Ein großer Teil der Instrumente berücksichtigt die Angemessenheit der Methode der Qualitätsbewertung (80 %). Von den Instrumenten erheben 53 %, ob die Bewertung unabhängig durch mindestens zwei Reviewer erfolgt, 20 %, ob die Übereinstimmung der Reviewer bewertet wird und ebenfalls 20 % die Angemessenheit der Integration der Ergebnisse der Qualitätsbewertung. Lediglich 7 % der Instrumente (n = 1) erfragen die Verblindung der Reviewer (Tabelle 44: Inhaltliche Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen, Teil 2).

Die Berücksichtigung der Robustheit der Ergebnisse bzw. ihre mögliche Heterogenität wird von 80 % der Instrumente erfasst. Eine angemessene Datensynthese erfragen 47 %, während nur 13 % die Darstellung von Schlüsselelementen abfragen.

Nur rund ein Viertel der Instrumente (27 %) berücksichtigt eine Zusammenfassung und Angabe der Präzision der Ergebnisse.

Jeweils rund die Hälfte der Instrumente (47 %) erfasst, ob Schlussfolgerungen durch Ergebnisse gestützt werden sowie die Berücksichtigung möglicher Bias und Limitationen.

Nur ein einziges Instrument (7 %) beinhaltet ein Item zur finanziellen Förderung der Studie.

**Übersicht über generische QBI für systematische Reviews, HTA und Metaanalysen**

Als Basis für die Auswahl eines geeigneten Bewertungsinstruments werden nachfolgend die QBI für Übersichtsarbeiten abgebildet, die generischer Art sind, d. h. nicht für eine spezifische Fragestellung entwickelt wurden, sondern allgemein anwendbar sind. Dies trifft auf 13 der insgesamt 15 identifizierten Instrumente zu. Es werden nur Items dargestellt, die der internen Validität zugeordnet werden können.

Drei der Instrumente sind deutschsprachig (eins vom LBI[133] sowie zwei Instrumente von Ekkernkamp et al.[69]), alle übrigen sind in englischer Sprache verfasst. Für keines der Instrumente liegen Angaben zur Testgüte vor. Die QBI enthalten zwischen zwei und zehn der extrahierten 18 Items.

Die Instrumente von Assendelft et al.[9] sowie Shea et al.[196] enthalten jeweils zehn Items. Sie decken von allen Instrumenten den höchsten Anteil (56 %) der extrahierten Aspekte interner Validität ab. Unter den deutschsprachigen Instrumenten enthält das des LBI[133] die höchste Anzahl der extrahierten Items zur internen Validität (9/18). Für die relevanten Werkzeuge gilt: Sieben Instrumente decken sieben bis neun der insgesamt elf als relevant definierten Elemente ab. Davon haben vier Instrumente umfassende Ausfüllhinweise[9, 133, 159, 187], sodass diese am ehesten für eine Qualitätsbewertung in Betracht gezogen werden können.

Das von Assendelft et al.[9] publizierte Instrument enthält 14 Items, die quantitativ mit Punkten bewertet werden im Sinn einer Skala. Die Kriterien basieren auf methodischen Forschungsarbeiten und werden weiter operationalisiert aufgrund der Erfahrungen der Autoren. Das Instrument sollte als Checkliste und nicht als Skala eingesetzt werden, da, wie bereits zuvor dargestellt, es keine empirische Evidenz für die Gewichtung der Kriterien gibt.

Tabelle 18: Charakteristika der generischen Instrumente für systematische Reviews, HTA und Metaanalysen gibt einen Überblick über die generischen Instrumente zur Qualitätsbewertung für Übersichtsarbeiten, sortiert nach der Anzahl erfüllter Items der internen Validität.

**Tabelle 18: Charakteristika der generischen Instrumente für systematische Reviews, HTA und Metaanalysen**

| | | | | | | Ein- und Ausschluss-kriterien | Literaturrecherche und -auswahl | | | | | | Datenextraktion | | | | Studienqualität/ Interne Validität | | | | Datensynthese und -analyse | | Auftraggeber |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Anzahl Elemente, die erfüllt sind** | **Anzahl Domänen, die mit mind. 1 Item abgedeckt werden** | **Anzahl Domänen, die zu mind. 50 % abgedeckt sind** | **Anzahl relevanter Elemente** | **Ausfüllhinweise** | **Kriterien a priori** | **Relevante Datenbanken** | **Weitere Datenquellen** | Schlagworte/ Freitext | Synonyme | **Restriktionen** | ≥ 2 Reviewer | **Interventionen/ Expositionen** | **Outcomes** | **≥ 2 Reviewer** | Verblindung | **Bewertungsmethode** | **≥ 2 Reviewer** | Verblindung | Integration der Ergebnisse | **Angemessene Synthese** | **Robustheit/ Heterogenität** | **Finanzierung** |
| Assendelft et al.[9] | 10 | 5 | 4 | 9 | ● | ● | ● | ● | | | | | ● | ● | | | ● | ● | ● | | ● | ● | |
| Shea et al.[196] | 10 | 6 | 5 | 8 | ◐ | ● | ● | ● | | | | ● | ● | ● | ● | | | | | ● | | ● | ● |
| Sackett[187] | 9 | 5 | 4 | 8 | ● | ● | ● | ● | | | | ● | | | ● | | ● | ● | | | ● | ● | |
| LBI[133] | 8 | 4 | 4 | 7 | ● | ● | ● | ● | | | ● | | | | | | ● | ● | | ● | | ● | |
| PHRU[159] | 8 | 4 | 3 | 8 | ● | | ● | ● | | | ● | | | ● | | | ● | ● | | | ● | ● | |
| ARIF[220] | 7 | 3 | 3 | 7 | ◐ | | ● | ● | | | ● | | | | | | ● | ● | | | ● | ● | |
| CEBM[219] | 7 | 4 | 3 | 6 | ● | ● | ● | ● | ● | | ● | | | | | | ● | | | | | ● | |
| Rychetnik & Frommer[185] | 7 | 4 | 3 | 7 | ◐ | ● | ● | ● | | | ● | | | | | | ● | | | | ● | ● | |
| CEBMH[36] | 5 | 3 | 2 | 5 | | | ● | ● | | | ● | | | | | | ● | | | | | ● | |
| SIGN 50[194] | 5 | 3 | 2 | 4 | ● | | ● | ● | | | | | | | | | ● | | | ● | | ● | |
| Vigna-Taglianti et al.[230] | 5 | 3 | 2 | 5 | ◐ | | | ● | | | ● | | | | | | ● | ● | | | | ● | |
| Ekkernkamp et al.[69] (CL 1b) | 4 | 4 | 2 | 4 | ◐ | ● | | | | | | | | | ● | | | ● | | | | ● | |
| Oxman et al.[170] | 4 | 3 | 1 | 4 | | | ● | ● | | | | | | | | | ● | | | | ● | | |
| Ekkernkamp et al.[69] (CL 1a) | 3 | 3 | 1 | 3 | ◐ | ● | | | | | | | | | ● | | | ● | | | | | |
| Barnes & Bero[13] | 2 | 2 | 1 | 2 | | | | | | | | | | | | | ● | | | | ● | | |

● Erfüllt. ◐ Teilweise erfüllt. **Fett**: relevante Elemente.

HTA = Health Technology Assessment.

### 6.5.1.5 QBI für Interventionsstudien

Durch die Literaturrecherche werden 80 Instrumente zur Qualitätsbewertung von Interventionsstudien identifiziert (Tabelle 40: Formale Charakteristika von Instrumenten für Interventionsstudien, Tabelle 45: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 1, Tabelle 46: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 2).

**Formale Charakteristika**

Unter den 80 Instrumenten befinden sich insgesamt vier deutschsprachige Instrumente des IQWiG[102], des LBI[133] der GSWG[69] sowie von Petrak et al.[171]. Von allen Instrumenten sind 25 modifizierte Versionen eines anderen Instruments, 42 sind laut Publikation generisch und 38 spezifisch. Die Anzahl der Items reicht von drei bis 81. Achtzehn Instrumente sind Checklisten, zehn sind Komponentensysteme, 52 sind Skalen. Von den Checklisten und Komponentensystemen ermöglichen sieben eine qualitative Komponentenbewertung, fünf eine qualitative Gesamtbewertung, ein Instrument sieht beides vor. Für elf der 80 Instrumente zur Qualitätsbewertung von Interventionsstudien (14 %) existiert eine Beschreibung des dem Instrument bzw. dessen Anwendung zugrunde liegenden Qualitätskonzepts (Tabelle 19: Qualitätskonzepte von QBI für Interventionsstudien).

**Tabelle 19: Qualitätskonzepte von QBI für Interventionsstudien**

| Publikation | Qualität/Validität |
|---|---|
| Bath et al.[15] | "(…) quality (...) as judged by whether they give a minimum set of information describing the 'design, conduct, analysis, and generalizability of the trial.'" (angelehnt an CONSORT) |
| Borsody & Yamada[20] | "(…) RCT is internally valid when 'within the confines of the study, results appear to be accurate and interpretation of the investigators is supported'." |
| Cho & Bero[44] | "We defined 'methodologic quality' of a study as minimization of systematic bias and consistency of conclusions with results. We are able to determine methodologic quality of a study only to the extent that study design and analytic methods are reported." |
| Higgins & Green[92] | "The validity of a study may be considered to have two dimensions. The first dimension is whether the study is asking an appropriate research question. This is often described as 'external validity', and its assessment depends on the purpose for which the study is to be used. External validity is closely connected with the generalizability or applicability of a study's findings (...).The second dimension of a study's validity relates to whether it answers its research question 'correctly', that is, in a manner free from bias. This is often described as 'internal validity' (...)." |
| Jadad et al.[106] | "(…) was to assess quality, defined as the likelihood of the trial design to generate unbiased results and approach the 'therapeutic truth'." |
| Kmet et al.[114] | "'Quality' was defined in terms of the internal validity of the studies, or the extent to which the design, conduct and analyses minimized errors and biases." |
| Lamont[121] | "Validity was defined as the extent to which study design, execution and analyses minimized bias." |
| Ludwig Boltzmann Institut[133] | „Hohe interne Validität impliziert, dass die untersuchte Exposition bzw. Intervention und nicht ein Bias (systematischer Fehler) oder ein zufälliger (nicht-systematischer) Fehler für die Resultate verantwortlich ist." |
| Moncrieff & Drummond[149] | "Internal validity refers to the management of potential sources of bias which might produce misleading results. External validity is concerned with the application or generalizability and clinical relevance of study findings, which also depends on trial design and conduct." |
| Smith et al.[203] | "What we mean by validity (…) is the overall validity of the findings of each trail, taking into account all aspects of design and statistical interpretation which have a bearing on the accuracy of the efficacy estimate." |
| Verhagen et al.[227] | "Quality is a set of parameters in the design and conduct of a study related to effect sizes (…) Quality is a set of parameters in the design and conduct of a study that reflects the validity of the outcome, related to the external and internal validity and the statistical model used." |

QBI = Qualitätsbewertungsinstrument.

Für 23 der 80 Instrumente (29 %), die sich zur Bewertung von Interventionsstudien eignen, liegen Angaben zur Testgüte vor (Tabelle 20: Testgüte von QBI für Interventionsstudien). Dabei handelt es sich zum überwiegenden Teil (n = 16) um die Interrater-Reliabilität. Zu jeweils drei Instrumenten liegen Werte zur Intrarater-Reliabilität anhand von kappa und Intraclass-Correlation sowie zur internen Konsistenz auf der Basis von Cronbachs alpha vor, zu zwei Instrumenten Angaben der Test-Retest-Reliabilität und zu einem das Ausmaß der Kriteriumsvalidität.

**Tabelle 20: Testgüte von QBI für Interventionsstudien**

| Instrument | Inhaltliche Validität | Konstruktvalidität | Kriteriumsvalidität | Interrater-Reliabilität | Intrarater-Reliabilität | Test-Retest-Reliabilität | Interne Konsistenz |
|---|---|---|---|---|---|---|---|
| Balas et al.[11] | | | | | κ = 0,94 | | |
| Bizzini et al.[18] | | | | ICC = 0,97 | | | |
| Borsody & Yamada[20] | | | | κ = 0,74 | κ = 0,94 | | |
| Chalmers et al.[39] | | | | ICC = 0,66[17] | | r = 0,81[17] | |
| de Vet et al.[54] | | | | ICC = 0,77[228] | | | |
| Downs & Black[60] | | | r = 0,90 | | | r = 0.88 | 0,89 |
| Heneghan et al.[88] | | | | κ = 0,83 | | | |
| Heyland et al.[91] | | | | κ = 0,73 | | | |
| Hill et al.[93] | | | | κ = 0,80 | | | |
| Imperiale & McCullough[100] | | | | κ = 0,79 | | | |
| Jadad et al.[106] (3 Items) | | | | ICC = 0,66 | | | |
| Jadad et al.[106] (6 Items) | | | | ICC = 0,65 | | | |
| Kwakkel et al.[120] | | | | κ = 0,86 | | | |
| MacLehose et al.[134] | | | | | | | Cronbachs alpha 0,61 und 0,72 |
| Moncrieff & Drummond[149] | | | | | | | Cronbachs alpha 0,873 |
| Moncrieff et al.[148] | | | | κ = 0,51; 0,53; 0,54 (3 Rater) | | | Cronbachs alpha 0,65; 0,76; 0,778 |
| Moseley et al.[151] | | | | ICC = 0,53[152]; ICC = 0,56[136] | | | |
| Slim et al.[199] | | | | κ = 0,92 | | | |
| Staiger et al.[206] | | | | κ = 0,63 | | | |
| Thomas et al.[213] | | | | | κ = 0,74 und 0,61 | | |
| Yates et al.[243] | | | | ICC für 3 Rater 0,91; κ = 0,405 | | | |

ICC = Intraclass-Correlation. κ = kappa. QBI = Qualitätsbewertungsinstrument. r = Korrelationskoeffizient.

Der Entwicklungsprozess wird für 18 Instrumente dargestellt. Hinsichtlich der Anwendung des Instruments bzw. der Operationalisierung werden in 23 Instrumenten ausführliche Angaben gemacht, in 26 Instrumenten kurze Hinweise. Der angenommene Zeitbedarf wird bei drei Instrumenten mit jeweils zehn, 18 und 20 Minuten dargelegt. Sechzehn der dreißig Instrumente sind laut Publikation für die Bewertung von Beobachtungs- als auch von Interventionsstudien geeignet. Eine Übersicht der formalen Aspekte ist in Tabelle 40: Formale Charakteristika von Instrumenten für Interventionsstudien dargestellt.

**Inhaltliche Charakteristika**

Nur 21 % der Instrumente berücksichtigen die Studienfrage (s. Tabelle 45: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 1).

Ein- und Ausschlusskriterien sind Bestandteil von 54 % der Instrumente, die Stichprobengröße wird bei 45 % der Instrumente erfragt und 31 % gehen auf die Beschreibung der Studienpopulation ein.

Den Aspekt einer angemessenen Randomisierungsmethode beinhalten 74 % der Instrumente. Die Vergleichbarkeit der Gruppen zu Beginn wird bei 65 % der Instrumente abgefragt, die geheime Gruppenzuweisung wird bei 45 % der Bewertungsinstrumente gefordert.

Im Bereich Verblindung gibt es deutliche Unterschiede. Während die Verblindung der Erheber des Outcomes (86 %) häufig erwähnt wird, ist die Verblindung der Studienteilnehmer (69 %) bereits seltener Bestandteil der Instrumente, wohingegen nur ein Drittel die Verblindung des übrigen Studienpersonals (33 %) abfragt. Die Überprüfung der Verblindung wird bei 11 % aller Instrumente berücksichtigt.

Eine Beschreibung der Intervention ist bei 50 % der Instrumente enthalten. Rund ein Viertel der Instrumente erfragt die Compliance (23 %). Die Beschreibung möglicher Kointerventionen (23 %) sowie deren Vermeidung (20 %) werden nur bei einigen Instrumenten berücksichtigt, weitere Aspekte werden noch seltener genannt.

Eine Definition der Outcomes erheben 45 % der Instrumente. Deutlich weniger beinhalten Fragen zum Follow-up (je 14 %) sowie zur Reliabilität (13 %) und Validität (9 %) der Methoden.

Die Intention-to-treat-Analyse ist in 64 % der Instrumente enthalten, die statistische Analyse als solche bei 53 % der Instrumente. Confounding wird bei 21 % der Instrumente berücksichtigt, die Bewertung von Confounding oder Heterogenität in keinem der Instrumente (s. Tabelle 46: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 2).

Studienabbrecher berücksichtigen 61 % der Instrumente; Ursachen für Drop-outs 41 %. Rund ein Drittel erhebt die angemessene Darstellung der Effekte (35 %). Weitere Aspekte werden deutlich seltener untersucht.

Ob Schlussfolgerungen auf Ergebnissen basieren, wird bei 18 % der Instrumente abgefragt, eine Diskussion von Bias und Confounding lediglich bei 3 %.

Die Repräsentativität der Studienpopulation wird nur bei 15 % der Instrumente erfragt, der Anteil Nichtteilnehmer nur bei 5 % sowie die Unterschiede zwischen Teilnehmern und Nichtteilnehmern gar nicht.

Insgesamt beinhalten 5 % der Instrumente ein Item zur finanziellen Förderung bzw. dem Auftraggeber.

**Übersicht über generische QBI für Interventionsstudien**

Es werden nachfolgend die für Interventionsstudien geeigneten QBI dargestellt, die generischer Art sind, d. h., dass diese nicht für eine spezifische Fragestellung entwickelt wurden, sondern allgemein anwendbar sind. Dies gilt für 42 der insgesamt 80 gefundenen Instrumente für Interventionsstudien. Es werden ausschließlich Items der internen Validität abgebildet.

Drei der Instrumente sind deutschsprachig (Ekkernkamp et al.[69], LBI[133] IQWiG[103]), alle übrigen sind in englischer Sprache verfasst. Angaben zur Interrater-Reliabilität (Intraclass-Correlation (ICC) bzw. kappa) liegen für fünf der 42 generischen QBI vor.

Die Bewertungsinstrumente decken zwischen drei und 18 der extrahierten 30 Items zur internen Validität ab. Die höchste Anzahl Items enthält das Instrument der Delfini Working Group[57] mit 18/30. Ebenfalls viele Items decken die Instrumente von Reisch et al.[181] (16/30) und Downs & Black[60] (15/30) ab. Die QBI von Chalmers et al.[39], MacLehose et al.[134] und Sindhu et al.[198] enthalten jeweils 13/18 Items der internen Validität. Das Instrument der Cochrane Collaboration[92] deckt zwölf der extrahierten Items zur internen Validität ab. Unter den deutschsprachigen Instrumenten enthält das der GSWG[69] die höchste Anzahl extrahierter Items zur internen Validität (12/30).

Von den insgesamt acht Domänen decken zehn Instrumente sechs und drei Instrumente sieben Domänen ab. Die drei Instrumente, die sieben Domänen abdecken, sind gleichzeitig die mit den meisten erfüllten Elementen. Bei den Instrumenten, die sechs Domänen mit mindestens einem Item abdecken, reicht die Spannweite der abgedeckten Elemente von acht bis 13. Die größte Anzahl an Domänen, in denen mindestens die Hälfte der Elemente enthalten ist, wird von den drei Instrumenten erzielt, die auch die meisten Elemente abdecken.

Die Anzahl enthaltener relevanter Elemente hängt nicht so deutlich von der Gesamtzahl der Elemente ab wie die Anzahl mit einem Element bzw. zu mindestens 50 % abgedeckter abgedeckter Domänen. Die größte Anzahl relevanter Elemente erzielt das von der Cochrane Collaboration empfohlene „Risk of bias tool“[92] (9/11). Es folgen drei Instrumente mit jeweils acht relevanten Elementen sowie sechs weitere Instrumente mit sieben relevanten Elementen.

Unter Berücksichtigung der vorhandenen Ausfüllhinweise gibt es acht QBI[39, 60, 92, 133, 134, 158, 181, 213], die anhand der Anzahl der erfüllten Elemente insgesamt sowie der als relevant definierten erfüllten Elemente als umfassender im Vergleich zu den übrigen Instrumenten gelten können.

Tabelle 21: Charakteristika der generischen Instrumente für Interventionsstudien gibt einen Überblick über die generischen Instrumente zur Qualitätsbewertung von Interventionsstudien, sortiert nach der Anzahl erfüllter Items der internen Validität.

**Tabelle 21: Charakteristika der generischen Instrumente für Interventionsstudien**

| | | | | | | Studienpopulation | **Randomisierung** | | | **Verblindung** | | | | Interventionen | | | | | | Outcomes | | | | Statistische Analyse | | | | | | | Ergebnisse | | | | **Finanzierung** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Anzahl Elemente, die erfüllt sind** | **Anzahl Domänen, die erfüllt sind (mind. 1 Item)** | **Anzahl Domänen, die zu mind. 50 % abgedeckt sind** | **Anzahl relevanter Elemente** | **Ausfüllhinweise** | Stichprobengröße | **Methode der Randomisierung** | **Gruppenzuweisung geheim** | **Vergleichbarkeit der Gruppen** | **Studienteilnehmer** | **Erheber des Outcomes** | **Übriges Studienpersonal** | Ausreichend | Kontrollgruppe | Behandlungsgleichheit | Placebo vergleichbar mit Verum | Kointerventionen vermieden | Kontamination | Compliance | Valide Methoden | Reliable Methoden | Follow-up-Länge | Follow-up zeitgleich | Angemessene Analyse | Multiples Testen | **Intention-to-treat-Analyse** | **Fehlende Werte** | **Berücksichtigung von Confounding** | Bewertung von Confounding | Bewertung von Heterogenität | Studienabbrecher | Unterschiede Teilnehmer/ Abbrecher | **Selektives Berichten** | Vorzeitiger Abbruch der Studie | **Art und Quelle der Förderung** |
| Delfini Group[57] | 18 | 7 | 5 | 7 | | ● | ● | ● | ● | ● | ● | ● | | ● | ● | | ● | ● | ● | | | | ● | ● | ● | ● | | | | | ● | ● | | | |
| Reisch et al.[181] | 16 | 7 | 5 | 8 | ● | ● | ● | | ● | ● | ● | ● | | | ● | | | | ● | ● | | | ● | | ● | | ● | | | | ● | | ● | | |
| Downs & Black[60] | 15 | 7 | 5 | 7 | ● | ● | ● | ● | | ● | ● | | | | | | | ● | ● | ● | | | ● | | ● | | ● | | | | ● | | ● | | |
| Chalmers et al.[39] | 13 | 6 | 4 | 7 | ● | | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | | ● | | | | | | ● | | | | |
| MacLehose et al.[134] | 13 | 6 | 4 | 6 | ● | ● | ● | ● | | ● | ● | ● | | ● | | | | | ● | ● | ● | | ● | | ● | | | | | | | | | | |
| Sindhu et al.[198] | 13 | 6 | 3 | 8 | | ● | ● | ● | ● | ● | ● | ● | | | ● | | | ● | | | | | ● | | ● | | ● | | | | ● | | | | |
| Higgins & Green[92] | 12 | 4 | 3 | 9 | ● | | ● | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | ● | ● | | | | | ● | | ● | ● | |
| Thomas et al.[213] | 12 | 5 | 4 | 7 | ● | | ● | | ● | ● | ● | | | | | | ● | ● | ● | | | | | | ● | ● | | | | | ● | | ● | ● | |
| Ekkernkamp et al.[69] | 11 | 6 | 3 | 5 | ◐ | | ● | ● | ● | | ● | | | | ● | | | | | ● | ● | | | ● | | ● | | | | | ● | ● | | | |
| Levine[123] | 11 | 4 | 3 | 5 | ◐ | ● | | | | ● | ● | ● | | | | ● | ● | ● | ● | | | | | ● | | ● | | ● | | | | | | | |
| PHRU[158] | 11 | 6 | 3 | 7 | ● | ● | ● | | ● | ● | ● | ● | | | ● | ● | | | | | | | | ● | | | | ● | | | | | ● | | |
| Prendiville et al.[177] | 11 | 6 | 3 | 5 | ◐ | | ● | ● | | ● | ● | | | | | ● | | | | | | | ● | ● | ● | ● | | | | | ● | ● | | | |
| Balk et al.[12] | 10 | 4 | 3 | 8 | ◐ | ● | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | ● | | ● | | ● | | | | | | | |
| Hadorn et al.[82] | 10 | 5 | 3 | 6 | ◐ | | ● | | ● | ● | ● | | | | | | ● | ● | | | | | | ● | | | | ● | | | ● | | ● | | |
| LBI[133] | 10 | 5 | 4 | 7 | ● | ● | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | | | ● | | | | | ● | ● | | | |
| Rochon[184] | 10 | 6 | 3 | 6 | ● | ● | ● | ● | ● | ● | ● | | | | | ● | | | ● | | | | | | | ● | | | | | ● | | | | |
| SIGN 50[194] | 10 | 6 | 2 | 6 | ● | | ● | ● | ● | ● | ● | | | | ● | | | | | | ● | | | ● | | ● | | | | | ● | | | | |

**Tabelle 21: Charakteristika der generischen Instrumente für Interventionsstudien – Fortsetzung**

| | | | | | | Studienpopulation | **Randomisierung** | | | **Verblindung** | | | | Interventionen | | | | | | Outcomes | | | | Statistische Analyse | | | | | | | Ergebnisse | | | | **Finanzierung** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Anzahl Elemente, die erfüllt sind** | **Anzahl Domänen, die erfüllt sind (mind. 1 Item)** | **Anzahl Domänen, die zu mind. 50 % abgedeckt sind** | **Anzahl relevanter Elemente** | **Ausfüllhinweise** | Stichprobengröße | **Methode der Randomisierung** | **Gruppenzuweisung geheim** | **Vergleichbarkeit der Gruppen** | **Studienteilnehmer** | **Erheber des Outcomes** | **Übriges Studienpersonal** | Ausreichend | Kontrollgruppe | Behandlungsgleichheit | Placebo vergleichbar mit Verum | Kointerventionen vermieden | Kontamination | Compliance | Valide Methoden | Reliable Methoden | Follow-up-Länge | Follow-up zeitgleich | Angemessene Analyse | Multiples Testen | **Intention-to-treat-Analyse** | **Fehlende Werte** | **Berücksichtigung von Confounding** | Bewertung von Confounding | Bewertung von Heterogenität | Studienabbrecher | Unterschiede Teilnehmer/ Abbrecher | **Selektives Berichten** | Vorzeitiger Abbruch der Studie | **Art und Quelle der Förderung** |
| Balas et al.[11] | 9 | 5 | 3 | 6 | ◐ | ● | ● | | ● | ● | ● | ● | | | | | | | | | | | | ● | | ● | | | | | ● | | | | |
| CEBM[217] | 9 | 6 | 2 | 5 | ● | | ● | | ● | ● | ● | | | | ● | | | | | | | ● | | ● | | | | ● | | | ● | | | | |
| Pua et al.[178] | 9 | 6 | 3 | 3 | ◐ | ● | ● | | | | ● | | | | | | | | | ● | ● | | | ● | | ● | | | | | ● | ● | | | |
| Chalmers et al.[38] | 8 | 4 | 2 | 6 | ● | ● | | ● | ● | ● | ● | ● | ● | | | ● | | | ● | | | | | ● | | ● | | | | | ● | | | ● | |
| Geng[75] | 8 | 5 | 2 | 5 | | ● | ● | | ● | | ● | | | | | | | | | | | | | ● | | ● | | ● | | | ● | | | | |
| Gupta et al.[79] | 8 | 5 | 3 | 5 | ◐ | ● | ● | | ● | ● | ● | | | | | | | | ● | | | | | ● | | ● | | | | | | | | | |
| IQWiQ[103] | 8 | 5 | 4 | 5 | | ● | ● | ● | | ● | ● | | | | | | | | | | | | | | | ● | | | | | ● | ● | | | |
| Zaza et al.[245] | 8 | 4 | 2 | 3 | ● | | | | ● | | | | | | | | | | | | ● | | ● | ● | ● | ● | | | | | ● | | ● | | |
| Cho & Bero[44] | 7 | 5 | 2 | 5 | 3 | ● | ● | | | ● | ● | | | | | | | | | | | | | ● | | | | ● | | | | | ● | | |
| Hill et al.[93] | 7 | 5 | 3 | 6 | ● | | ● | ● | | ● | ● | | | | | ● | | | | | | | | | | ● | | | | | | | | | ● |
| Verhagen et al.[227] | 7 | 4 | 2 | 5 | | | | ● | ● | ● | ● | ● | | | | | | | | | | | | ● | | | | | | | | | | ● | |
| CEBMH[37] | 6 | 4 | 2 | 5 | | | | ● | ● | ● | ● | | | | ● | | | | | | | | | | | ● | | | | | | | | | |
| Kmet et al.[114] | 6 | 4 | 3 | 4 | ● | ● | ● | | ● | ● | ● | | | | | | | | | | | | | ● | | | | | | | | | | | |
| Spitzer et al.[204] | 6 | 5 | 3 | 4 | | ● | ● | | ● | | ● | | | | | | | | | | | | | | | | | | | | ● | | | | ● |
| Huwiler-Müntener et al.[99] | 5 | 4 | 2 | 2 | ◐ | ● | ● | ● | | | | | | | | | | | | | | | | ● | | | | | | | | | | ● | |

**Tabelle 21: Charakteristika der generischen Instrumente für Interventionsstudien – Fortsetzung**

| | | | | | | Studienpopulation | **Rando-misierung** | | | **Verblindung** | | | | Interventionen | | | | | | Outcomes | | | | Statistische Analyse | | | | | | | Ergebnisse | | | | **Finanzierung** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Anzahl Elemente, die erfüllt sind** | **Anzahl Domänen, die erfüllt sind (mind. 1 Item)** | **Anzahl Domänen, die zu mind. 50 % abgedeckt sind** | **Anzahl relevanter Elemente** | **Ausfüllhinweise** | Stichprobengröße | **Methode der Randomisierung** | **Gruppenzuweisung geheim** | **Vergleichbarkeit der Gruppen** | **Studienteilnehmer** | **Erheber des Outcomes** | **Übriges Studienpersonal** | Ausreichend | Kontrollgruppe | Behandlungsgleichheit | Placebo vergleichbar mit Verum | Kointerventionen vermieden | Kontamination | Compliance | Valide Methoden | Reliable Methoden | Follow-up-Länge | Follow-up zeitgleich | Angemessene Analyse | Multiples Testen | **Intention-to-treat-Analyse** | **Fehlende Werte** | **Berücksichtigung von Confounding** | Bewertung von Confounding | Bewertung von Heterogenität | Studienabbrecher | Unterschiede Teilnehmer/ Abbrecher | **Selektives Berichten** | Vorzeitiger Abbruch der Studie | **Art und Quelle der Förderung** |
| Jadad et al.[106] (6 Items) | 5 | 4 | 1 | 3 | ● | | ● | | | ● | ● | | | | | ● | | | | | | | | | | | | | | ● | | | | |
| Kleijnen et al.[113] | 5 | 4 | 1 | 3 | ◐ | | ● | | | ● | ● | | | | | ● | | | | | | | | ● | | | | | | | | | | |
| MacMillan et al.[135] | 5 | 4 | 1 | 4 | ◐ | | ● | | ● | | ● | | | | | | | | | | | | | | | ● | | | | | ● | | | | |
| Yates et al.[243] | 5 | 3 | 2 | 4 | ● | ● | ● | ● | ● | | ● | | | | | | | | | | | | | | | | | | | | | | | |
| Jadad et al.[106] (3 Items) | 4 | 3 | 1 | 3 | ● | | ● | | | ● | ● | | | | | ● | | | | | | | | | | | | | | | | | | |
| Rychetnik & Frommer[185] | 4 | 3 | 0 | 1 | ◐ | | | | | | | | | | ● | | ● | | | | | | ● | | | ● | | | | | | | | |
| Yuen & Pope[244] | 4 | 3 | 2 | 2 | | ● | | | | ● | ● | | | | | | | | | | | | | | | | | | | | ● | | | |
| Brown[26] | 3 | 3 | 0 | 1 | ◐ | | | | | ● | | | | | | | ● | | | | | | | | | | | | | | ● | | | |
| Colditz et al.[49] | 3 | 2 | 1 | 2 | ◐ | | | | | ● | ● | | | | | | | | | | | | | ● | | | | | | | | | | |
| Sackett[187] | 3 | 3 | 0 | 1 | | | ● | | | | | | | | | | | | | ● | | | | | | | | | | | ● | | | |

● Erfüllt. ◐ Teilweise erfüllt. **Fett**: relevante Elemente.

### 6.5.1.6 QBI für Beobachtungsstudien

Durch die Literaturrecherche werden 30 Instrumente zur Qualitätsbewertung von Beobachtungsstudien identifiziert (Tabelle 41: Formale Charakteristika von Instrumenten für Beobachtungsstudien, Tabelle 47: Inhaltliche Charakteristika von Instrumenten für Beobachtungsstudien, Teil 1, Tabelle 48: Inhaltliche Charakteristika von Instrumenten für Beobachtungsstudien, Teil 2).

**Formale Charakteristika**

Unter den 30 Instrumenten befinden sich die beiden deutschsprachigen Instrumente des LBI[133] sowie der GSWG[69]. Sieben der Instrumente sind modifizierte Versionen eines anderen Instruments. Von den Instrumenten sind 23 generisch und sieben spezifisch. Die Anzahl der Items reicht von fünf bis 60. Fünfzehn der Instrumente sind Checklisten, zwei sind Komponentensysteme, 13 sind Skalen. Von den Checklisten und Komponentensystemen ermöglicht ein Instrument eine qualitative Komponentenbewertung, vier eine qualitative Gesamtbewertung, ein Instrument sieht beides vor. Für vier der 30 Instrumente zur Qualitätsbewertung von Beobachtungsstudien (13 %) gibt es eine Darstellung des Qualitätsbegriffs, das dem Instrument bzw. dessen Anwendung zugrunde liegt (Tabelle 22: Qualitätskonzepte von QBI für Beobachtungsstudien).

**Tabelle 22: Qualitätskonzepte von QBI für Beobachtungsstudien**

| Publikation | Qualität/Validität |
|---|---|
| Cho & Bero[44] | "We defined 'methodologic quality' of a study as minimization of systematic bias and consistency of conclusions with results. We are able to determine methodologic quality of a study only to the extent that study design and analytic methods are reported." |
| Kmet et al.[114] | "'Quality' was defined in terms of the internal validity of the studies, or the extent to which the design, conduct and analyses minimized errors and biases." |
| Ludwig Boltzmann Institut[133] | „Hohe interne Validität impliziert, dass die untersuchte Exposition bzw. Intervention und nicht ein Bias (systematischer Fehler) oder ein zufälliger (nicht-systematischer) Fehler für die Resultate verantwortlich ist.“ |
| Spooner et al.[205] | "An assessment of quality provides an estimate of the likelihoods that the results of a clinical trial are a valid estimate of the truth and that the reported methodology, conduct, and analysis are free from systematic bias." |

QBI = Qualitätsbewertungsinstrument.

Der Entwicklungsprozess wird für acht Instrumente dargestellt. Hinsichtlich der Anwendung des Instruments bzw. der Operationalisierung finden sich bei zehn Instrumenten ausführliche Angaben, bei zwölf Instrumenten kurze. Der angenommene Zeitbedarf für die Anwendung des Instruments wird bei einem Instrument mit 20 Minuten und zwei Instrumenten mit je zehn Minuten angegeben. Sechzehn der dreißig Instrumente sind laut Publikation sowohl für die Bewertung von Beobachtungs- als auch von Interventionsstudien geeignet. Eine Übersicht der formalen Aspekte ist in Tabelle 41: Formale Charakteristika von Instrumenten für Beobachtungsstudien dargestellt.

Für vier der 30 Instrumente (13 %), die zur Bewertung von Beobachtungsstudien geeignet sind, können Angaben zur Testgüte identifiziert werden. Diese Werte liegen je nach Instrument für die Kriteriumsvalidität, die Interrater-Reliabilität, die Intrarater-Reliabilität, die Test-Retest-Reliabilität sowie die interne Konsistenz vor (Tabelle 23: Testgüte von QBI für Beobachtungsstudien).

**Tabelle 23: Testgüte von QBI für Beobachtungsstudien**

| Instrument | Inhaltliche Validität | Konstruktvalidität | Kriteriumsvalidität | Interrater-Reliabilität | Intrarater-Reliabilität | Test-Retest-Reliabilität | Interne Konsistenz |
|---|---|---|---|---|---|---|---|
| Downs & Black[60] | | | r = 0,90 | | | r = 0,88 | 0,89 |
| MacLehose et al.[134] | | | | | | | Cronbachs alpha = 0,61 und 0,72 |
| Spooner et al.[205] | | | | ICC 0,42 | | | |
| Thomas et al.[213] | | | | | $\kappa$ = 0,74 und 0,61 | | |

ICC = Intraclass-Correlation. $\kappa$ = kappa. QBI = Qualitätsbewertungsinstrument.

**Inhaltliche Charakteristika**

Etwas weniger als die Hälfte der Instrumente (41 %) berücksichtigt die Fragestellung.

Hinsichtlich der Studienpopulation fragen jeweils 48 % der Instrumente nach spezifischen Ein- und Ausschlusskriterien und 52 % nach der Vergleichbarkeit der Studiengruppen zu Beginn. 41 % der Instrumente beinhalten ein Item zur Power-Berechnung und 31 % zur Beschreibung der Studienpopulation. Eine gleichzeitige Kontrollgruppe wird bei 24 % der Instrumente abgefragt. Lediglich ein Instrument (3 %) erhebt, ob für alle Gruppen identische Ein- und Ausschlusskriterien zugrunde liegen. Für das Design der Fall-Kontrollstudien fragen 24 % der Instrumente nach einer expliziten Falldefinition und nur zwei Instrumente (7 %) erheben die Vergleichbarkeit von Fällen und Kontrollen mit Ausnahme des Outcomes.

Eine präzise Definition der Exposition bzw. Intervention wird bei 38 % der Instrumente erfragt. Die Validität der Methoden (28 %) wird etwas häufiger als die Reliabilität (21 %) einbezogen. Eine identische Messung des Outcomes in allen Gruppen wird bei 28 % der Instrumente erwähnt. Für Fall-Kontrollstudien fragen 21 % der Instrumente nach einer verblindeten Erhebung des Outcomes.

Im Bereich des Outcomes berücksichtigt ein Großteil der Instrumente (72 %) die verblindete Erhebung des Outcomes. Die Validität der Methoden (41 %) wird auch hier häufiger erfragt als die Reliabilität (34 %). 34 % der Instrumente beinhalten Aspekte der präzisen Definition von Outcomes. Auf die wieteren Indikatoren im Bereich des Outcomes gehen weniger Instrumente ein. Für Fall-Kontrollstudien beinhaltet rund die Hälfte der Instrumente (48 %) eine verblindete Diagnosesicherung.

Eine Modellierung bzw. multivariate Verfahren zur Kontrolle von Confounding werden bei 62 % der Instrumente abgefragt, 59 % der Instrumente gehen auf die statistische Analyse ein. Andere Items werden nur selten berücksichtigt. Eine Bewertung von Confounding oder Heterogenität sowie der angemessene Umgang mit fehlenden Werten werden bei keinem Instrument berücksichtigt.

Die angemessene Darstellung von Effekten wird bei 52 % der Instrumente erfasst. 48 % gehen auf den Anteil Studienabbrecher ein. Ursachen für Drop-outs werden bei 31 % der Instrumente abgefragt. Das selektive Berichten beziehen 21 % der Instrumente in die Qualitätsbewertung ein, die Unterschiede zwischen Teilnehmern und Abbrechern 17 %.

Nur 24 % der Instrumente erheben, ob die Schlussfolgerungen auf den Ergebnissen basieren. Bei 7 % der Instrumente wird berücksichtigt, ob der Einfluss von Confounding und Bias diskutiert wird.

Die Repräsentativität der Studienpopulation ist Bestandteil von 48 % der Instrumente, wohingegen nur 21 % auf den Anteil Nichtteilnehmer und 17 % auf die Unterschiede von Teilnehmern und Nichtteilnehmern eingehen.

Keines der Instrumente berücksichtigt die finanzielle Förderung bzw. den Auftraggeber der Studie bzw. Publikation.

**Übersicht über generische QBI für Beobachtungsstudien**

Nachfolgend werden die für Beobachtungsstudien geeigneten QBI dargestellt, die generischer Art sind, d. h. nicht für eine spezifische Fragestellung entwickelt wurden, sondern allgemein anwendbar sind (Tabelle 24: Charakteristika der generischen Instrumente für Beobachtungsstudien). Diese Anforderung erfüllen 21 der insgesamt 30 vorliegenden Instrumente. Es werden ausschließlich Items der internen Validität dargestellt.

Zwei der Instrumente sind deutschsprachig (LBI[133], Ekkernkamp et al.[69]), alle übrigen sind in englischer Sprache verfasst. Angaben zur Interrater-Reliabilität liegen für kein Instrument vor.

Die Instrumente enthalten ein bis 15 der 26 extrahierten Items zur internen Validität. Das deutschsprachige Instrument des LBI[133] enthält mit 15/26 die meisten der extrahierten Items zur internen Validität. Jeweils zwölf der 26 Items sind im deutschsprachigen Instrument der GSWG[69] sowie im Instrument von Kmet et al.[114] enthalten. Drei weitere Instrumente decken elf Elemente ab[60, 134, 185], darunter auch die Skala von Downs & Black[60]. Bei fast allen Instrumenten gibt es vier oder fünf der sechs Domänen, die mit mindestens einem Element abgedeckt sind.

Die meisten als relevant definierten Elemente enthält das deutschsprachige Instrument des LBI[133] (5/7) und ein weiteres von Spitzer et al.[204] (4/7). Für letzteres liegen allerdings keine Ausfüllhinweise vor.

Tabelle 24: Charakteristika der generischen Instrumente für Beobachtungsstudien gibt einen Überblick über die generischen Instrumente zur Qualitätsbewertung für Beobachtungsstudien, sortiert nach der Anzahl erfüllter Items der internen Validität.

**Tabelle 24: Charakteristika der generischen Instrumente für Beobachtungsstudien**

| | | | | | | Studienpopulation | | | | | Exposition | | | | Outcome | | | | | | | Statistische Analyse | | | | | | **Ergebnisse** | | | **Finanzierung** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Anzahl Elemente, die erfüllt sind | Anzahl Domänen, die erfüllt sind (mind. 1 Item) | Anzahl Domänen, die zu mind. 50 % abgedeckt sind | Anzahl relevanter Elemente | Ausfüllhinweise | Identische Ein- und Ausschlusskriterien | Stichprobengröße | **Kontrollgruppe** | **Vergleichbarkeit der Gruppen** | FKS: Gleichheit von Fällen und Kontrollen | Valide Methoden | Reliable Methoden | **Messung gleich in allen Studiengruppen** | FKS: verblindete Erhebung | Valide Methoden | Reliable Methoden | FKS: verblindete Erhebung | Verblindete Erhebung | **Messung gleich in allen Studiengruppen** | Follow-up-Länge | Follow-up zeitgleich | Angemessene Analyse | Multiples Testen | **Confounderkontrolle** | Fehlende Werte | Bewertung von Confounding | Heterogenität | Studienabbrecher | Unterschiede Teilnehmer/Abbrecher | **Selektives Berichten** | **Art und Quelle der Förderung** |
| LBI[133] | 15 | 5 | 4 | 5 | ● | | ● | ● | ● | | ● | | ● | | ● | ● | | ● | ● | ● | ● | ● | | ● | | | | ● | ● | | |
| Ekkernkamp et al.[69] | 12 | 5 | 3 | 3 | ◐ | | | ● | ● | ● | ● | ● | ● | | ● | ● | | ● | | | | ● | | | | | | ● | ● | | |
| Kmet et al.[114] | 12 | 5 | 3 | 3 | ● | | ● | | ● | | ● | ● | | | ● | ● | ● | ● | | | | ● | ● | ● | | | | | | ● | |
| Downs & Black[60] | 11 | 4 | 2 | 3 | ● | | ● | ● | | | | | | | ● | ● | ● | ● | | | ● | ● | | ● | | | | ● | | ● | |
| MacLehose et al.[134] | 11 | 4 | 2 | 3 | ● | | ● | ● | | | | | | | ● | ● | ● | ● | | | ● | ● | | ● | | | | ● | | ● | |
| Rychetnik & Frommer[185] | 11 | 5 | 2 | 3 | ◐ | | | | ● | ● | | ● | ● | | | ● | ● | ● | | ● | | ● | | ● | | | | | ● | | |
| Spitzer et al.[204] | 10 | 5 | 1 | 4 | | | ● | | ● | | | | ● | | | | ● | ● | ● | | | ● | ● | ● | | | | ● | | | |
| Thomas et al.[213] | 10 | 5 | 2 | 2 | ● | | | | ● | | ● | ● | | | ● | ● | ● | ● | | | | ● | | ● | | | | ● | | | |
| PHRU (KS)[157] | 9 | 4 | 2 | 3 | ● | | | | | | ● | | ● | | ● | ● | | ● | ● | ● | | | | ● | | | | | ● | | |
| SIGN 50 (FKS)[194] | 9 | 5 | 2 | 3 | ● | | | | ● | | | ● | | | ● | ● | | ● | ● | | | | | ● | | | | ● | ● | | |
| Slim et al.[200] | 8 | 4 | 1 | 2 | ◐ | | ● | ● | ● | | | | | | | | ● | ● | | ● | | ● | | | | | | ● | | | |
| Geng[75] | 7 | 4 | 1 | 2 | | | ● | | ● | | | | ● | ● | | | ● | ● | | | | ● | | | | | | | | | |
| Hadorn et al.[82] | 7 | 4 | 1 | 3 | ◐ | | | ● | ● | | | | | | | | ● | ● | | | | ● | | | | | | ● | | ● | |

**Tabelle 24: Charakteristika der generischen Instrumente für Beobachtungsstudien – Fortsetzung**

| | | | | | | Studienpopulation | | | | | Exposition | | | | Outcome | | | | | | | Statistische Analyse | | | | | | **Ergebnisse** | | | **Finanzierung** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Anzahl Elemente, die erfüllt sind | Anzahl Domänen, die erfüllt sind (mind. 1 Item) | Anzahl Domänen, die zu mind. 50 % abgedeckt sind | Anzahl relevanter Elemente | Ausfüllhinweise | Identische Ein- und Ausschlusskriterien | Stichprobengröße | **Kontrollgruppe** | **Vergleichbarkeit der Gruppen** | FKS: Gleichheit von Fällen und Kontrollen | Valide Methoden | Reliable Methoden | **Messung gleich in allen Studiengruppen** | FKS: verblindete Erhebung | Valide Methoden | Reliable Methoden | FKS: verblindete Erhebung | Verblindete Erhebung | **Messung gleich in allen Studiengruppen** | Follow-up-Länge | Follow-up zeitgleich | Angemessene Analyse | Multiples Testen | **Confounderkontrolle** | Fehlende Werte | Bewertung von Confounding | Heterogenität | Studienabbrecher | Unterschiede Teilnehmer/Abbrecher | **Selektives Berichten** | **Art und Quelle der Förderung** |
| SIGN 50 (KS)[194] | 7 | 4 | 1 | 1 | ● | ● | | | | | ● | ● | | ● | | | ● | ● | | | | | | ● | | | | | | | |
| Wells et al.[234] | 7 | 5 | 0 | 2 | ● | | | | ● | | ● | | | | ● | | | ● | | ● | | | | ● | | | | ● | | | |
| CEBMH[34] | 6 | 4 | 0 | 2 | | | | | ● | | | | | | ● | | | ● | | ● | | | | ● | | | | ● | | | |
| Cho & Bero[44] | 6 | 4 | 0 | 2 | 3 | | ● | | | | | | | | | | ● | ● | | | | ● | | ● | | | | | | ● | |
| Nguyen et al.[162] | 6 | 4 | 0 | 1 | ◐ | | ● | | | | | | | ● | | | ● | ● | | | | ● | | ● | | | | | | | |
| PHRU (FKS)[160] | 6 | 4 | 1 | 2 | ● | | ● | | | | | | ● | ● | ● | | | | | | | ● | | ● | | | | | | | |
| Colditz et al.[49] | 3 | 2 | 0 | 0 | ◐ | | | | | | | | | | | | ● | ● | | | | ● | | | | | | | | | |
| Brown[26] | 2 | 2 | 0 | 0 | ◐ | | | | | | | | | | ● | | | | | | | | | | | | | ● | | | |
| Carson et al.[30] | 1 | 1 | 0 | 1 | ◐ | | | | | | | | | | | | | | | | | | | ● | | | | | | | |

● Erfüllt. ◐ Teilweise erfüllt. **Fett**: relevante Elemente.

FKS = Fall-Kontrollstudie.

### 6.5.1.7 QBI für Diagnosestudien

Durch die Literaturrecherche werden 17 Instrumente zur Qualitätsbewertung von Diagnosestudien identifiziert, nachdem Instrumente, die bereits von Whiting et al.[237] in ihrem umfangreichen HTA-Bericht dargestellt werden, nicht berücksichtigt werden (Tabelle 42: Formale Charakteristika von Instrumenten für Diagnosestudien, Tabelle 49: Inhaltliche Charakteristika von Instrumenten für Diagnosestudien).

**Formale Charakteristika**

Unter den 17 Instrumenten befinden sich die deutschsprachigen Instrumente des IQWiG[103], des LBI[133] sowie der GSWG[69]. Vier Instrumente sind modifizierte Versionen eines anderen Instruments. Von den Instrumenten sind zwölf generisch und fünf spezifisch. Die Anzahl der Items reicht von fünf bis 35. Elf der Instrumente sind Checklisten, eines ist ein Komponentensystem, fünf sind Skalen. Von den Checklisten und Komponentensystemen ermöglichen jeweils zwei eine qualitative Komponenten- bzw. Gesamtbewertung.

Für drei der 17 Instrumente zur Qualitätsbewertung von Diagnosestudien (18 %) existieren Erläuterungen des dem Instrument bzw. dessen Anwendung zugrunde liegenden Qualitätskonzepts (Tabelle 25: Qualitätskonzepte von QBI für Diagnosestudien).

**Tabelle 25: Qualitätskonzepte von QBI für Diagnosestudien**

| Publikation | Qualität/Validität |
|---|---|
| de Vet et al.[55] | "Internal validity addresses the question the results presented in a study are unbiased (…) External validity addresses the question for which situation the conclusions hold." |
| Ludwig Boltzmann Institut[133] | „Hohe interne Validität impliziert, dass die untersuchte Exposition bzw. Intervention und nicht ein Bias (systematischer Fehler) oder ein zufälliger (nichtsystematischer) Fehler für die Resultate verantwortlich ist.“ |
| Whiting et al.[238] | "It is therefore essential that the quality of individual studies included in a systematic review is assessed in terms of potential for bias, lack of applicability, and, inevitably to a certain extent, the quality of reporting." |

QBI = Qualitätsbewertungsinstrument.

Der Entwicklungsprozess wird für fünf Instrumente dargestellt. Hinsichtlich der Anwendung des Instruments bzw. der Operationalisierung finden sich bei neun Instrumenten ausführliche Angaben, in vier Instrumenten kurze Hinweise. Der angenommene Zeitbedarf wird bei keinem Instrument dargestellt. Alle Instrumente speziell zur Bewertung von Diagnosestudien entwickelt worden. Eine Übersicht der formalen Aspekte der QBI für Diagnosestudien ist in Tabelle 42: Formale Charakteristika von Instrumenten für Diagnosestudien dargestellt.

Für vier[95, 223, 226, 238] der 17 Bewertungsinstrumente für Diagnosestudien (24 %) können Angaben zur Testgüte identifiziert werden. Dabei handelt es sich um die Interrater-Reliabilität mittels kappa sowie um die Intraclass-Correlation (Tabelle 26: Testgüte der QBI für Diagnosestudien).

**Tabelle 26: Testgüte der QBI für Diagnosestudien**

| | Inhaltliche Validität | Konstruktvalidität | Kriteriumsvalidität | Interrater-Reliabilität | Intrarater-Reliabilität | Test-Retest-Reliabilität | Interne Konsistenz |
|---|---|---|---|---|---|---|---|
| Hoffmann et al.[95] | | | | $\kappa$ = 0,61 | | | |
| Varela-Lema & Ruano-Ravina[226] | | | | ICC 0,98 | | | |
| Whiting et al.[238] | | | | $\kappa$ = 0,22[96] | | | |
| Wurff et al.[223] | | | | $\kappa$ = 0,63 | | | |

ICC = Intraclass-Correlation. $\kappa$ = kappa. QBI = Qualitätsbewertungsinstrument.

**Inhaltliche Charakteristika**

Am häufigsten berücksichtigen die Instrumente, ob die Ergebnisse des Index- (89 %) bzw. des Referenztests (78 %) verblindet ausgewertet werden. Die Anwendung desselben Referenztests bei allen Teilnehmern erheben 83 % der Instrumente. Die Verwendung eines angemessenen Referenzstandards

wird bei 72 % der Instrumente berücksichtigt. Weniger häufig (11 bis 38 %) umfassen die Instrumente Items aus den Bereichen nicht interpretierbare Werte, Incorporation-Bias, Krankheitsprogression, klinischer Reviewbias, Beobachter-/Instrumentenvariabilität sowie Behandlungsparadox (s. Tabelle 49: Inhaltliche Charakteristika von Instrumenten für Diagnosestudien).

Die Vergleichbarkeit der Studien- mit der Zielpopulation wird hinsichtlich des Teilnehmerspektrums bei 56 % und hinsichtlich von Prävalenz, Schweregrad der Erkrankung bei 50 % der Instrumente berücksichtigt. Die Angemessenheit der Rekrutierung wird bei 39 % der Instrumente erfragt. Zwei Instrumente (11 %) beziehen einen möglichen Wechsel der Methodik beim Indextest in die Bewertung ein.

Der Bereich Studiendurchführung wird relativ selten berücksichtigt. So wird eine ausreichende Power bei lediglich 22 % der Instrumente erhoben. Die Aspekte Studienprotokoll und Relevanz der Studienziele für die Fragestellung werden bei nur jeweils einem Instrument (6 %) abgefragt, die Subgruppenanalyse bei keinem Instrument.

Das am häufigsten enthaltene Item im Bereich Berichtsqualität ist die Beschreibung des Indextests (56 %), wohingegen die Beschreibung des Referenztests (28 %) nur halb so oft abgefragt wird. Die Darstellung der Einschlusskriterien berücksichtigen 44 % der Instrumente. Ebenso viele (44 %) beziehen die Ergebnisdarstellung in die Bewertung ein. Der Aspekt der Definition eines „normalen“ Testergebnisses ist in 39 % der Instrumente enthalten. Seltener beinhalten die Instrumente die Definition eines normalen Testergebnisses, Ergebnisdarstellung, Studienabbrecher, Genauigkeit der Ergebnisse sowie die Nützlichkeit des Tests.

**Übersicht über generische QBI für Diagnosestudien**

Es werden nachfolgend die QBI für Diagnosestudien dargestellt, die generisch sind, d. h. dass sie nicht für eine spezifische Fragestellung entwickelt wurden, sondern allgemein anwendbar sind. Dies betrifft zwölf der insgesamt 17 identifizierten Instrumente. Es werden ausschließlich Items der internen Validität abgebildet.

Drei der Instrumente sind deutschsprachig (LBI[133], IQWiG[102], Ekkernkamp et al.[69]), alle übrigen sind in englischer Sprache verfasst. Angaben zur Interrater-Reliabilität liegen ausschließlich für das Instrument von Whiting et al.[238] vor.

Die QBI enthalten zwischen vier und zwölf der extrahierten 18 Items. Die meisten Items decken das Instrument von Whiting et al.[238] (12/18) und die Instrumente des LBI[133] sowie von Scottish Intercollegiate Guidelines Network (SIGN)[194] ab (jeweils 10/18).

Die meisten als relevant definierten Elemente weisen neben QUADAS[238] das von van den Hoogen et al. publizierte Instrument[221] (jeweils 7/8) sowie fünf weitere Instrumente mit jeweils sechs von acht erfüllten Elementen auf, darunter auch das vom IQWiG eingesetzte[102] und die Checkliste der GSWG[69].

Tabelle 27: Charakteristika der generischen Instrumente für Diagnosestudien gibt einen Überblick über die generischen Instrumente zur Qualitätsbewertung von Diagnosestudien, sortiert nach der Anzahl erfüllter Items der internen Validität.

**Tabelle 27: Charakteristika der generischen Instrumente für Diagnosestudien**

| | | | | Interne Validität | | | | | | | | | | Studiendurch-führung | | | Berichts-qualität | Externe Validität | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Anzahl Elemente, die erfüllt sind | Anzahl relevanter Elemente | Ausfüllhinweise | **Referenzstandard** | Bias durch Krankheits-progression | **Verifikationsbias** | Bias durch nicht-unabhängi-ge Tests (Incorporation bias) | Behandlungsparadox | **Review bias Indextest** | **Review bias Referenztest** | **Klinischer Review bias** | **Beobachter/ Instrumentenvariabilität** | Umgang mit nicht bewert-baren Testergebnissen | Subgruppenanalysen | Stichprobengröße | Protokoll | Studienabbrecher | **Spektrum der Teilnehmer** | Rekrutierung | **Krankheitsprävalenz-schwere** | Methodenwechsel beim Indextest |
| Whiting et al.[238] | 12 | 7 | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | | | | ● | ● | ● | ● | |
| LBI[133] | 10 | 4 | ● | ● | ● | ● | ● | | ● | ● | | | ● | | | ● | | | ● | | ● |
| SIGN 50[194] | 10 | 6 | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | | | | | ● | ● | | |
| IQWiQ[102] | 9 | 6 | | ● | ● | ● | ● | | ● | ● | | | ● | | | | | ● | | ● | |
| van den Hoogen et al.[221] | 9 | 7 | ● | ● | | ● | | | ● | ● | | ● | | | ● | | | ● | ● | ● | |
| de Vet et al.[55] | 8 | 5 | ● | ● | | ● | | | ● | ● | | | ● | | ● | | | ● | ● | | |
| Mulrow et al.[155] | 8 | 6 | ◐ | ● | | ● | | | ● | ● | | | ● | | ● | | | ● | | ● | |
| CEBM [218] | 7 | 6 | ● | ● | | ● | ● | | ● | ● | | | | | | | | ● | | ● | |
| CEBMH[32] | 7 | 4 | | ● | | ● | ● | | ● | ● | | | | | | | | | ● | | ● |
| PHRU[161] | 7 | 5 | ● | ● | | ● | ● | ● | ● | | | | | | | | | ● | | ● | |
| Ekkernkamp et al.[69] | 6 | 6 | ◐ | ● | | ● | | | ● | ● | | | | | | | | ● | | ● | |
| Sackett[187] | 4 | 3 | | ● | | | | | ● | | | ● | | | | | | | | ● | |

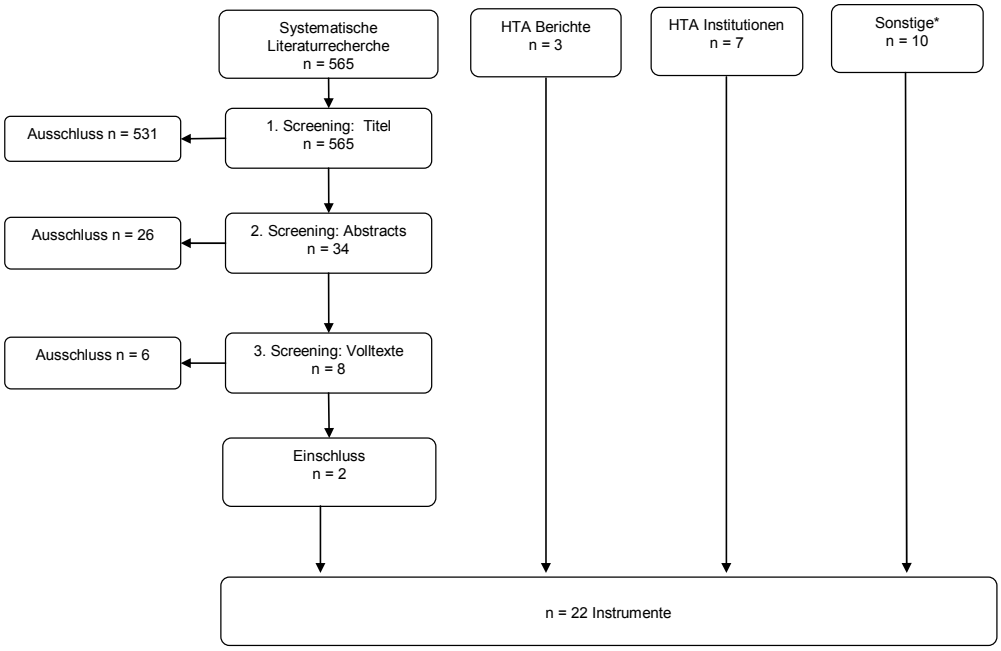
● Erfüllt. ◐ Teilweise erfüllt. **Fett**: relevante Elemente.

## 6.5.2 Bewertung gesundheitsökonomischer Studien

### 6.5.2.1 Literaturrecherche und -auswahl

**Systematische Datenbankrecherche**

Die Treffermenge der systematischen Datenbankrecherche umfasst insgesamt 565 Literaturquellen. Nach Durchsicht der Publikationstitel nach Studien, die sich mit einer gesundheitsökonomischen Fragestellung beschäftigen, werden 531 Studien ausgeschlossen. Die Durchsicht der 34 verbliebenen Abstracts ergibt einen Ausschluss von 26 weiteren Dokumenten. Insgesamt werden acht Volltexte nach Instrumenten zur Bewertung der Qualität gesundheitsökonomischer Studien gesichtet. Weitere sechs Dokumente werden ausgeschlossen. Die Ausschlussgründe sind im Anhang dokumentiert (s. 8.4). Insgesamt werden zwei Publikationen aus der systematischen Datenbankrecherche eingeschlossen und in der Analyse berücksichtigt.




* Initiale Suche, Referenzen etc.

HTA = Health Technology Assessment.

**Abbildung 2: Stufenweise Literaturrecherche und -auswahl (Ökonomie)**

**Screening deutschsprachiger HTA-Berichte**

Das Screening deutsprachiger HTA-Berichte in der Datenbank der DAHTA erfolgt auf der Basis der im epidemiologischen Teil durchgeführten Recherche. Aus den 154 eingeschlossenen HTA-Berichten werden drei Instrumente identifiziert[4, 64, 197].

**Internetrecherche**

Auf den Internetseiten von verschiedenen HTA-Organisationen sowie der Cochrane Collaboration werden sieben Instrumente zur Beurteilung der Qualität gesundheitsökonomischer Studien gefunden, die den Einschlusskriterien entsprechen (s. Tabelle 28: Eingeschlossene Instrumente aus der Internetrecherche).

**Tabelle 28: Eingeschlossene Instrumente aus der Internetrecherche**

| Publikation | Instrumentenname |
|---|---|
| Center for Disease Control and Prevention[31] | Reviewer Checklist for Health Economic Papers |
| European Collaboration in Health Technology Assessment[67] | Checklist for an economic analysis article |
| Larsen et al.[122] | Checklist for the assessment of economic evaluations carried out as part of health technology assessment |

**Tabelle 28: Eingeschlossene Instrumente aus der Internetrecherche – Fortsetzung**

| Publikation | Instrumentenname |
|---|---|
| Ludwig Boltzmann Institut[133] | Formular für Beurteilung der Qualität ökonomischer Studien |
| Public Health Research Unit[163] | Critical Appraisal Skills Programme (CASP) – 10 questions to help you make sense of economic evaluations |
| Pharmaceutical Management Agency[173] | PHARMAC Guidelines for Reviewing CUA |
| Scottish Intercollegiate Guidelines Network[194] | Methodology Checklist: Economic Evaluation |

Auf den Internetseiten des European Network for Health Technology Assessment (EUnetHTA) und der International Network of Agencies for Health Technology Assessment (INAHTA) werden keine eigenen Vorgaben oder Empfehlungen zur Anwendungen von Bewertungsinstrumenten für gesundheitsökonomische Studien gefunden. Auf den Internetseiten der INAHTA finden sich lediglich Verweise zu nationalen HTA-Richtlinien.

**Handsuche in Referenzen**

Anhand der Referenzen der Volltexte werden neun Dokumente mit Instrumenten zur Bewertung der methodischen Qualität gesundheitsökonomischer Studien identifiziert, aus denen zehn Instrumente extrahiert werden.

### 6.5.2.2 Datenextraktion und -synthese

Aus der schrittweisen Literaturrecherche werden insgesamt 21 Dokumente eingeschlossen, aus denen 22 Instrumente extrahiert und analysiert werden. Die überwiegende Zahl der Instrumente (n = 19) ist in englischer Sprache verfasst. Lediglich drei Instrumente sind deutschsprachig[4, 133, 197]. Fünf Instrumente stellen Modifikationen von bestehenden Instrumenten dar[4, 67, 133, 163, 185]. Die einzelnen Instrumente unterscheiden sich deutlich in der Anzahl der Items. Während das Instrument von Ramos et al.[179] mit fünf Items die geringste Anzahl umfasst, sind in dem Instrument von Larsen et al.[122] mit 63 Items die größte Anzahl an Items vorhanden. Bei den extrahierten Instrumenten handelt es sich überwiegend um Checklisten (n = 15). Bei sechs Instrumenten bilden die Studienqualität mittels Skalen ab. Lediglich ein Instrument bewertet die einzelnen Komponenten auf qualitativer Ebene, ohne jedoch eine Gesamtbewertung vorzunehmen[194]. Nur in dem Dokument des LBI findet sich eine Definition von Studienqualität[133]. Methodische Studienqualität wird als interne Validität definiert.

„Interne Validität wird dabei als Ausmaß methodologischer Qualität in Studiendesign und Durchführung definiert. Hohe interne Validität impliziert, dass die untersuchte Exposition bzw. Intervention und nicht ein Bias (systematischer Fehler) oder ein zufälliger (nicht-systematischer) Fehler für die Resultate verantwortlich ist. In anderen Worten, interne Validität ist die Wahrscheinlichkeit, dass Resultate möglichst nahe an die ‚Wahrheit' herankommen."[133].

Der Entwicklungsprozess zur Erstellung der QBI wird in fünf Publikationen dargestellt. In den übrigen Publikationen verweisen die Autoren größtenteils auf nationale oder internationale Richtlinien und gesundheitsökonomische Standardwerke für die Erstellung von gesundheitsökonomischen Studien. Zu lediglich vier Instrumenten lassen sich Operationalisierungen der Domänen und Items bzw. Ausfüllhinweise zur Datenextraktion finden. Keine der Publikationen gibt den Zeitbedarf für die Anwendung des Instruments an. Informationen zur Testgüte sind lediglich zu dem Pediatric Quality Appraisal Questionnaire (PQAQ) in der Publikation von Ungar et al.[216] vorhanden. Die Interrater-Reliabilität liegt bei ICC = 0,75 und die Test-Retest-Reliabilität bei ICC = 0,92.

**Tabelle 29: Formale Charakteristika von Instrumenten für gesundheitsökonomische Studien**

| Instrument | Name | Sprache | Modifikation von | Spezifisches Instrument | Items | Art des Instruments | Qualitative Bewertung | Range* | Cut-point* | Definition von Qualität | Entwicklungsprozess | Ausfüllhinweise | Zeitbedarf | Testgüte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Aidelsburger et al.[4] | Checkliste zur Beurteilung der methodischen Qualität im Rahmen gesundheitsökonomischer Kurz-HTA-Berichte | DE | Siebert 1999 | | 26 | SK | | 0-26 | | | | | | |
| Center for Disease Control and Prevention[31] | Reviewer Checklist for Health Economic Papers | EN | | | 35 | CL | | | | | | | | |
| Chiou et al.[43] | Quality of Health Economic Studies | EN | | | 16 | SK | | 0-100 | 4 Kategorien | | ● | | | |
| Drummond et al.[63] | BMJ Referees' Checklist | EN | | | 35 | CL | | | | | ● | ● | | |
| Drummond et al.[62] | Detecting Flaws in Economic Evaluation | EN | | | 18 | CL | | | | | | | | |
| Drummond et al.[64] | Checklist for assessing economic evaluation | EN | | | 29 | CL | | | | | | ● | | |
| European Collaboration in Health technology Assessment[67] | Checklist for an economic analysis article | EN | Drummond 1997/PHRU 2006 | | 10 | CL | | | | | | | | |
| Evers et al.[70] | Consensus on Health Economic Criteria (CHEC) | EN | | | 19 | CL | | | | | ● | | | |
| Hobbs et al.[94] | Health economic evaluation against the proforma | EN | | | 62 | CL | | | | | ● | | | |

**Tabelle 29: Formale Charakteristika von Instrumenten für gesundheitsökonomische Studien – Fortsetzung**

| Instrument | Name | Sprache | Modifikation von | Spezifisches Instrument | Items | Art des Instruments | Qualitative Bewertung | Range* | Cut-point* | Definition von Qualität | Entwicklungsprozess | Ausfüllhinweise | Zeitbedarf | Testgüte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Iskedjian et al.[105] | Checklist for Evaluation of Economic Studies | EN | | | 13 | SK | | 0-4 | | | | ◐ | | |
| Larsen et al.[122] | Checklist for the assessment of economic evaluations carried out as part of health technology assessment | EN | | | 63 | CL | | | | | | | | |
| Ludwig Boltzmann Institut[133] | Formular für Beurteilung der Qualität ökonomischer Studien | DE | Siebert 1999 | | 58 | CL | | | | ● | | | | |
| Public Health Research Unit[163] | Critical Appraisal Skills Programme (CASP) – 10 questions to help you make sense of economic evaluations | EN | Drummond 1997 | | 10 | CL | | | | | | | | |
| Pharmaceutical management Agency[173] | PHARMAC Guidelines for Reviewing CUAs | EN | | | 28 | CL | | | | | | | | |
| Ramos et al.[179] | Critical Appraisal Questions | EN | | | 5 | CL | | | | | | ◐ | | |
| Ramsberg et al.[180] | Pharmaceutical Benefits Board Checklist | EN | | | 23 | SK | | 0-1 | | | | | | |

**Tabelle 29: Formale Charakteristika von Instrumenten für gesundheitsökonomische Studien – Fortsetzung**

| Instrument | Name | Sprache | Modifikation von | Spezifisches Instrument | Items | Art des Instruments | Qualitative Bewertung | Range* | Cut-point* | Definition von Qualität | Entwicklungsprozess | Ausfüllhinweise | Zeitbedarf | Testgüte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rychetnik & Frommer[186] | Checklist for appraising economic evaluations | EN | Drummond 1997/PHRU 2006 | | 7 | CL | | | | | | | | |
| Sackett (Critical Appraisal)[187] | Critical appraisal of Journal articles | EN | | | 6 | CL | | | | | | | | |
| Sackett (Readers' Guides)[187] | Readers' Guides for Assessing an Economic Analysis of Clinical and other Health Care | EN | | | 8 | CL | | | | | | | | |
| Siebert et al.[197] | Qualitätskatalog | DE | | | 56 | SK | | 0-56 | | | ● | | | |
| Scottish Intercollegiate Guidelines Network[195] | Methodology Checklist: Economic Evaluation | EN | | | 20 | KO | QKB | | | | | ● | | |
| Ungar & Santos[216] | Pediatric Quality Appraisal Questionnaire (PQAQ) | EN | | ● | 57 | SK | | 0-92 | | | | | | ● |

* Bei Skalen. Ausfüllhinweise: ◐ = Kurz. ● = Ausführlich.

CL = Checkliste. DE = Deutsch. EN = Englisch. QKB = Qualitative Komponentenbewertung. KO = Komponentensystem. SK = Skala.

Auf Basis der Literaturrecherche werden 22 Instrumente zur Qualitätsbewertung von gesundheitsökonomischen Studien identifiziert. Die Ergebnisse der inhaltlichen Datenextraktion sind in Tabelle 30: Inhaltliche Charakteristika von Instrumenten für gesundheitsökonomische Studien, Teil 1, ausführlich dargestellt und werden nachfolgend kurz beschrieben.

**Studienfrage**

Im Bereich der Studienfrage berücksichtigen 73 % der Instrumente (n = 16) das Item zur präzisen Definition der Studienfrage. Ob die gewählte Evaluationsform zur Beantwortung der Studienfrage geeignet ist, berücksichtigt nur rund die Hälfte (55 %) der Instrumente.

**Interventionsalternativen**

Das Item zum Einbezug der aus ökonomischer Sicht relevanten Interventionsalternativen, wird bei fast allen Instrumenten (86 %) berücksichtigt. Keines deckt die Fragestellung ab, ob ein direkter oder indirekter Vergleich der Interventionsalternativen vorgenommen wird und ob ein indirekter Vergleich angemessen ist.

**Perspektive**

Die Wahl der Evaluationsperspektive und ob diese passend zur Studienfrage gewählt wird, berücksichtigen 59 % der Instrumente. Ob die gewählte Perspektive im Verlauf der Studie durchgängig eingehalten wird, beachten zwei Instrumente (9 %).

**Ressourcenverbrauch und Kosten**

Im Bereich des Ressourcenverbrauchs und der Kosten decken 82 % der Instrumente das Thema der angemessenen Bewertung der relevanten Ressourcenverbräuche ab. Die Identifikation der relevanten Ressourcenverbräuche sowie die mengenmäßige Erfassung werden bei 17 Instrumenten (77 %) berücksichtigt. Das Kriterium der getrennten Erhebung von Mengen und Preisen ist bei 55 % der Instrumente vorhanden. Die Frage zur Nutzung angemessener Datenquellen für die Kostenberechnung behandeln lediglich fünf Instrumente (23 %). Ob die Methoden der Inflationierung und Währungskonversion angemessen sind, untersuchen vier Instrumente (18 %).

**Outcome/Nutzen**

Die Wahl der richtigen Outcomeparameter berücksichtigen 77 % der Instrumente. 55 % beachten die Angemessenheit der Erhebungsinstrumente bei der Betrachtung von Lebensqualität. Die Überprüfung von Datenquellen auf ihre Qualität wird ebenfalls bei 55 % der Fälle betrachtet.

**Qualität der Daten**

Das Item hinsichtlich der Qualität von genutzten Primärdaten ist bei 36 % und hinsichtlich der methodischen Qualität von Metaanalysen bei 41 % der Instrumente enthalten.

**Modellierung**

Den Bereich Modellierung berücksichtigt weniger als die Hälfte der Instrumente. So wird die Darstellung der Modelle bei Verwendung von Modellen in gesundheitsökonomischen Studien bei lediglich 36 % der Instrumente erhoben. Fragen zur Angemessenheit der Modellstruktur und zu den gewählten Parametern sind nur bei 32 % der Instrumente enthalten.

**Diskontierung**

Die Diskontierung von zukünftigen Kosten und Nutzen wird in fast allen Instrumenten (82 %) berücksichtigt, die Verwendung adäquater Diskontraten bei 68 %.

**Analyse**

Die Fragestellung ob eine inkrementelle Kosten-Effektivitäts-Analyse durchgeführt wird, wird bei 77 % der Instrumente berücksichtigt. Ob die in der Studie angewendeten statistischen Verfahren angemessen sind, ermitteln hingegen nur 45 %.

**Sensitivitätsanalyse**

Am häufigsten untersuchen die Instrumente die Durchführung einer Sensitivitätsanalyse (86 %). Die Angemessenheit der Methodik der durchgeführten Sensitivitätsanalyse erheben 45 %. Ob alle relevanten Parameter in die Sensitivitätsanalyse berücksichtigt werden, betrachten 41 %.

**Ergebnisse**

Der Bereich Ergebnisdarstellung wird relativ selten berücksichtigt. Lediglich 32 % der Instrumente gehen der Frage nach, ob die Ergebnisse mittels angemessener statistischer Kenngrößen dargestellt werden.

**Diskussion**

Ob die Schlussfolgerungen auf den Ergebnissen basieren wird bei 59 % der Instrumente erfragt. Den Einfluss von Confounding und Bias berücksichtigen 45 %.

**Interessenkonflikte**

Interessenkonflikte berücksichtigen lediglich 23 %. Sie untersuchen, ob die Art und Quelle der Finanzierung in der Studie genannt werden.

**Tabelle 30: Inhaltliche Charakteristika von Instrumenten für gesundheitsökonomische Studien, Teil 1**

| | | | | | | Studienfrage | | Interventionsalternativen | | | Perspektive | | Zeitraum | Ressourcenverbrauch und Kosten | | | | | | Outcome/Nutzen | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Anzahl „Angemessen“** | **Anzahl „Begründet“** | **Anzahl „Berichtet“** | **Berücksichtigte Items, gesamt** | **Anteil „Berücksichtigt“ in %** | Wird die Studienfrage präzise definiert? | Ist die Art der ökonomischen Studie angemessen? | Werden die aus ökonomischer Sicht relevanten Interventions-alternativen einbezogen? | Wird ein direkter Vergleich der Interventionsalternativen vorgenommen? | Falls nein, sind die indirekten Vergleiche angemessen? | Passt die gewählte Perspektive zur Studienfrage? | Ist die Perspektive konsistent? | Ist der Beobachtungszeitraum so gewählt, dass alle relevanten Effekte und Kosten berücksichtigt wurden? | Werden Mengen und Preise getrennt voneinander erhoben? | Werden alle relevanten Ressourcen, die mit der Intervention in Zusammenhang stehen identifiziert? | Sind angemessene Datenquellen genutzt worden? | Werden alle relevanten Ressourcen mengenmäßig erfasst? | Werden die Ressourcenverbräuche angemessen bewertet? | Sind die Methoden der Inflationierung und Währungskonversion angemessen? | Sind die Outcomeparameter richtig gewählt worden? | Sind geeignete Erhebungsinstrumente gewählt worden, falls LQ erhoben wurde? | Sind die Datenquellen auf ihre Qualität überprüft worden? |
| Aidelsburger et al. 2003[4] | 18 | 0 | 2 | **20** | **63 %** | ● | ○ | ● | | | ○ | | ● | ● | | | ● | ● | ● | ● | ● | ● |
| Center for Disease Control and Prevention 2008[31] | 10 | 3 | 6 | **19** | **59 %** | ● | ◐ | ● | | | | | ○ | ● | | | | ○ | | ○ | | ○ |
| Chiou et al. 2003[43] | 14 | 3 | 4 | **21** | **66 %** | ● | | | | | ◐ | | ● | ● | | ○ | ● | ● | | ○ | ● | ● |
| Drummomd et al. 2005b[62] | 13 | 3 | 1 | **17** | **53 %** | ● | | ● | | | ● | | | | ● | | ● | ● | | ● | | |
| Drummond et al. 1996[63] | 9 | 6 | 11 | **26** | **81 %** | ● | ◐ | ◐ | | | ◐ | | ○ | ● | | | ● | ○ | ○ | ○ | ○ | ○ |
| Drummond et al. 2005a[64] | 6 | 3 | 1 | **10** | **31 %** | | | ● | | | | | ◐ | | ● | | | | | | | |
| European Collaboration in Health technology Assessment 2001[67] | 2 | 0 | 3 | **5** | **16 %** | | | ○ | | | | | | | | | ○ | | | | | |
| Evers et al. 2005[70] | 18 | 0 | 1 | **19** | **59 %** | ● | ● | ○ | | | ● | ● | ● | ● | ● | | ● | ● | | ● | ● | |
| Hobbs et al. 1997[94] | 5 | 1 | 4 | **10** | **31 %** | ● | | ● | | | | | | | ○ | | ○ | ○ | | | ● | |
| Iskedjian et al. 1997[105] | 11 | 0 | 0 | **11** | **34 %** | ● | ● | ● | | | ● | | | | ● | | | | | ● | | |
| Larsen et al. 2003[122] | 9 | 2 | 14 | **25** | **78 %** | ● | ○ | ○ | | | ○ | ● | ○ | ● | ● | ○ | ○ | ○ | | ○ | ○ | ○ |
| Ludwig Boltzmann Institut 2007[133] | 23 | 2 | 2 | **27** | **84 %** | ● | ○ | ◐ | | | ○ | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Public Health Research Unit 2006[163] | 13 | 0 | 1 | **14** | **44 %** | ● | | ○ | | | | | | ● | ● | | ● | ● | | ● | ● | |
| Pharmaceutical management Agency 2007[173] | 10 | 3 | 0 | **13** | **41 %** | | | ● | | | | | ● | | ● | | | ● | | ● | | ◐ |
| Ramos et al. 2004[179] | 2 | 0 | 6 | **8** | **25 %** | | | | | | ○ | | | | ○ | | | | | ○ | | ○ |
| Ramsberg et al. 2004[180] | 12 | 1 | 2 | **15** | **47 %** | | ◐ | ● | | | ● | | ● | ● | ● | | ● | ○ | | | ● | ○ |
| Rychetnik & Frommer 2002[186] | 9 | 0 | 1 | **10** | **31 %** | ● | | ○ | | | | | | ● | ● | | ● | ● | | | | |
| Sackett 1991 (Crit. Appraisal)[187] | 5 | 0 | 2 | **7** | **22 %** | | | ○ | | | ○ | | | | ● | | ● | ● | | ● | | |
| Sackett 1991 (Readers' Guides)[187] | 10 | 1 | 0 | **11** | **34 %** | ● | ● | ◐ | | | | | | | ● | | ● | ● | | ● | | ● |
| Siebert et al. 1999[197] | 23 | 2 | 2 | **27** | **84 %** | ● | ○ | ◐ | | | ○ | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Scottish Intercollegiate Guidelines Network 2004[195] | 12 | 1 | 0 | **13** | **41 %** | ● | ◐ | | | | | | | | ● | | ● | ● | | ● | ● | |
| Ungar et al. 2003[216] | 11 | 4 | 9 | **24** | **75 %** | ● | ○ | ◐ | | | ○ | | ◐ | ● | ● | ○ | ○ | ● | | ● | ○ | ○ |

**Tabelle 30: Inhaltliche Charakteristika von Instrumenten für gesundheitsökonomische Studien, Teil 1 – Fortsetzung**

| | | | | | | Studienfrage | | Interventionsalternativen | | | Perspektive | | Zeitraum | Ressourcenverbrauch und Kosten | | | | | | Outcome/Nutzen | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Anzahl „Angemessen“ | Anzahl „Begründet“ | Anzahl „Berichtet“ | Berücksichtigte Items, gesamt | Anteil „Berücksichtigt“ in % | Wird die Studienfrage präzise definiert? | Ist die Art der ökonomischen Studie angemessen? | Werden die aus ökonomischer Sicht relevanten Interventions-alternativen einbezogen? | Wird ein direkter Vergleich der Interventionsalternativen vorgenommen? | Falls nein, sind die indirekten Vergleiche angemessen? | Passt die gewählte Perspektive zur Studienfrage? | Ist die Perspektive konsistent? | Ist der Beobachtungszeitraum so gewählt, dass alle relevanten Effekte und Kosten berücksichtigt wurden? | Werden Mengen und Preise getrennt voneinander erhoben? | Werden alle relevanten Ressourcen, die mit der Intervention in Zusammenhang stehen identifiziert? | Sind angemessene Datenquellen genutzt worden? | Werden alle relevanten Ressourcen mengenmäßig erfasst? | Werden die Ressourcenverbräuche angemessen bewertet? | Sind die Methoden der Inflationierung und Währungskonversion angemessen? | Sind die Outcomeparameter richtig gewählt worden? | Sind geeignete Erhebungsinstrumente gewählt worden, falls LQ erhoben wurde? | Sind die Datenquellen auf ihre Qualität überprüft worden? |
| **Anzahl „Angemessen“** | | | | | | 16 | 3 | 8 | 0 | 0 | 4 | 2 | 7 | 12 | 15 | 2 | 13 | 13 | 3 | 12 | 9 | 5 |
| **Anzahl „Begründet“** | | | | | | 0 | 4 | 5 | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **Anzahl „Berichtet“** | | | | | | 0 | 5 | 6 | 0 | 0 | 7 | 0 | 3 | 0 | 2 | 3 | 4 | 5 | 1 | 5 | 3 | 6 |
| **Items berücksichtigt Gesamt** | | | | | | **16** | **12** | **19** | **0** | **0** | **13** | **2** | **12** | **12** | **17** | **5** | **17** | **18** | **4** | **17** | **12** | **12** |
| **Anteil in %** | | | | | | **73 %** | **55 %** | **86 %** | **0 %** | **0 %** | **59 %** | **9 %** | **55 %** | **55 %** | **77 %** | **23 %** | **77 %** | **82 %** | **18 %** | **77 %** | **55 %** | **55 %** |

● Angemessen. ◐ Begründet. ○ Berichtet.

LQ = Lebensqualität.

**Tabelle 31: Inhaltliche Charakteristika von Instrumenten für gesundheitsökonomische Studien, Teil 2**

| | Qualität der Daten | | Modellierung | | Diskontierung | | Analyse | | Sensitivitätsanalyse | | | Ergebnisse | Diskussion der Ergebnisse | | Interessenkonflikte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ist die Qualität der Primärdaten ausreichend um die Studienfrage beantworten zu können? | Ist die Qualität der Methoden zur Identifikation, Extraktion und Synthese der Effektparameter (Meta-Analyse) ausreichend zur Beantwortung der Studienfrage? | Wird das Modell nachvollziehbar dargestellt? | Sind die Modellstruktur und die gewählten Parameter angemessen? | Werden in der Studie alle zukünftigen Kosten und Nutzen diskontiert? | Wenn ja, sind die Diskontraten angemessen? | Sind die stat. Verfahren angemessen? | Wird eine inkrementelle Kosten-Effektivitäts-Analyse durchgeführt? | Wird eine Sensitivitätsanalyse durchgeführt? | Werden alle relevanten Parameter in die Sensitivitätsanalyse einbezogen? | Ist die Methodik der Sensitivitätsanalyse angemessen? | Werden die Ergebnisse mit Punktschätzern und Präzision angegeben? | Basieren die Schlussfolgerungen auf den Ergebnissen? | Wird der mögliche Einfluss von Confounding und Bias diskutiert? | Werden Art und Quelle der Finanzierung genannt? |
| Aidelsburger et al. 2003[4] | | ● | ● | ● | ● | ● | | ● | | | | | ● | ● | |
| Center for Disease Control and Prevention 2008[31] | | ○ | ● | | ● | ◐ | ○ | ● | ● | ◐ | | ● | ● | ● | |
| Chiou et al. 2003[43] | ○ | | ● | ◐ | ● | ◐ | ○ | ● | ● | | | | ● | ● | ● |
| Drummomd et al. 2005b[62] | ● | ○ | | | ● | ◐ | ● | ● | ● | ◐ | ◐ | | ● | | |
| Drummond et al. 1996[63] | ○ | ○ | ● | ◐ | ○ | ◐ | ○ | ● | ● | ◐ | ○ | ● | ● | ● | |
| Drummond et al. 2005a[64] | ◐ | ○ | | | | | | ● | ● | | ◐ | ● | ● | | |
| European Collaboration in Health technology Assessment 2001[67] | | | | | | | ○ | ● | ● | | | | | | |
| Evers et al. 2005[70] | | | | | ● | ● | | ● | ● | ● | | | ● | | ● |
| Hobbs et al. 1997[94] | | | | | ● | ◐ | | | ● | | ○ | | | | |
| Iskedjian et al. 1997[105] | | | | | ● | | ● | | ● | | | | ● | ● | |
| Larsen et al. 2003[122] | ○ | ○ | ○ | | ● | ● | ● | ● | ● | ◐ | ○ | | | ● | |
| Ludwig Boltzmann Institut 2007[133] | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | |
| Public Health Research Unit 2006[163] | | | | | ● | | | ● | ● | ● | ● | | ● | | |
| Pharmaceutical management Agency 2007[173] | | | ● | ● | ● | ● | | | ● | | ◐ | | ◐ | | |

**Tabelle 31: Inhaltliche Charakteristika von Instrumenten für gesundheitsökonomische Studien, Teil 2 – Fortsetzung**

| | Qualität der Daten | | Modellierung | | Diskontierung | | Analyse | | Sensitivitätsanalyse | | | Ergebnisse | Diskussion der Ergebnisse | | Interessenkonflikte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ist die Qualität der Primärdaten ausreichend um die Studienfrage beantworten zu können? | Ist die Qualität der Methoden zur Identifikation, Extraktion und Synthese der Effektparameter (Meta-Analyse) ausreichend zur Beantwortung der Studienfrage? | Wird das Modell nachvollziehbar dargestellt? | Sind die Modellstruktur und die gewählten Parameter angemessen? | Werden in der Studie alle zukünftigen Kosten und Nutzen diskontiert? | Wenn ja, sind die Diskontraten angemessen? | Sind die stat. Verfahren angemessen? | Wird eine inkrementelle Kosten-Effektivitäts-Analyse durchgeführt? | Wird eine Sensitivitätsanalyse durchgeführt? | Werden alle relevanten Parameter in die Sensitivitätsanalyse einbezogen? | Ist die Methodik der Sensitivitätsanalyse angemessen? | Werden die Ergebnisse mit Punktschätzern und Präzision angegeben? | Basieren die Schlussfolgerungen auf den Ergebnissen? | Wird der mögliche Einfluss von Confounding und Bias diskutiert? | Werden Art und Quelle der Finanzierung genannt? |
| Ramos et al. 2004[179] | ○ | ○ | | | | | | ● | ● | | | | | | |
| Ramsberg et al. 2004[180] | | | | ● | ● | ● | | ● | | | | | | | ● |
| Rychetnik & Frommer 2002[186] | | | | | ● | | | ● | ● | ● | | | | | |
| Sackett 1991 (Crit. Appraisal)[187] | | | | | | | | | ● | | | | | | |
| Sackett 1991 (Readers' Guides)[187] | | | | | ● | ● | | | | | | | | ● | |
| Siebert et al. 1999[197] | | ● | ● | ● | ● | ◐ | ● | ● | ● | ● | ● | ● | ● | ● | |
| Scottish Intercollegiate Guidelines Network 2004[195] | | | | | ● | ● | | ● | ● | | | ● | | | ● |
| Ungar et al. 2003[216] | ○ | | | | ○ | ◐ | ○ | ● | ● | | ◐ | ● | ● | ● | ● |
| **Anzahl „Angemessen“** | 1 | 3 | 7 | 5 | 16 | 6 | 5 | 17 | 19 | 5 | 3 | 7 | 12 | 10 | 5 |
| **Anzahl „Begründet“** | 1 | 0 | 0 | 2 | 0 | 9 | 0 | 0 | 0 | 4 | 4 | 0 | 1 | 0 | 0 |
| **Anzahl „Berichtet“** | 5 | 6 | 1 | 0 | 2 | 0 | 5 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| **Items berücksichtigt Gesamt** | **7** | **9** | **8** | **7** | **18** | **15** | **10** | **17** | **19** | **9** | **10** | **7** | **13** | **10** | **5** |
| **Anteil in %** | **32 %** | **41 %** | **36 %** | **32 %** | **82 %** | **68 %** | **45 %** | **77 %** | **86 %** | **41 %** | **45 %** | **32 %** | **59 %** | **45 %** | **23 %** |

● Angemessen. ◐ Begründet. ○ Berichtet.

### 6.5.3 Workshop

Es haben insgesamt 27 Personen am Workshop teilgenommen. Neben drei Projektmitgliedern und zwei Mitarbeiterinnen des DIMDI waren 22 Personen der Einladung gefolgt (s. Tabelle 32: Übersicht über Institutionen und Einrichtungen, denen die Teilnehmenden angehören).

**Tabelle 32: Übersicht über Institutionen und Einrichtungen, denen die Teilnehmenden angehören**

| |
|---|
| Abteilung Allgemeinmedizin des Universitätsklinikums der Heinrich-Heine-Universität Düsseldorf, Düsseldorf |
| Basel Institute of Clinical Epidemiology, Basel |
| Cochrane Metabolic and Endocrine Disorders Group, Düsseldorf |
| Deutsche Gesellschaft für Public Health, Bielefeld |
| Deutsches Institut für Medizinische Dokumentation und Information, Köln |
| Deutsches Netzwerk Evidenzbasierte Medizin, Berlin |
| Fachgebiet Management im Gesundheitswesen, Institut für Technologie und Management, Technische Universität, Berlin |
| Fakultät für Gesundheitswissenschaften, Universität Bielefeld, Bielefeld |
| Fachhochschule Fulda, Fachbereich Pflege und Gesundheit, Fulda |
| GRADE-Arbeitsgruppe, ohne Ort |
| HTA Zentrum, Universität Bremen, Bremen |
| Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung, Hannover |
| Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsmedizin der Johannes Gutenberg-Universität Mainz, Mainz |
| Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln |
| Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie der Charité, Berlin |
| Institut für Sozialmedizin, Epidemiologie und Gesundheitssystemforschung, Hannover |
| Institut für Sozialmedizin, Universitätsklinikum Schleswig-Holstein, Lübeck |
| Lehrstuhl für Medizinmanagement, Universität Duisburg-Essen, Essen |
| Ludwig Boltzmann Institut für Health Technology Assessment, Wien |
| MDK Sachsen, Dresden |

GRADE = Grading of Recommendations Assessment, Development and Evaluation. HTA = Health Technology Assessment. MDK = Medizinischer Dienst der Krankenversicherung.

Im ersten Block des Workshops werden folgende Themen als Diskussionspunkte von den Teilnehmern vorgeschlagen bzw. ergänzt:

- Externe Validität: Bestandteil von QBI?
- Subjektivität der Bewertung
- QBI
  - o Geringe Berichtsqualität: Umgang mit fehlenden Informationen?
  - o Begriff Studienqualität: Verzerrungspotenzial?
  - o QBI: Endpunktbezogen? Beurteilungseinheit Studie vs. Endpunkt
- Integration der Ergebnisse der Bewertung

Die Diskussionsbeiträge werden nachfolgenden themenspezifisch zusammengefasst dargestellt. Es handelt sich jeweils um Einzelmeinungen, die nicht notwendigerweise die Meinungen aller Workshop-Teilnehmer widerspiegeln. Die Wortbeiträge werden nicht wortwörtlich, sondern inhaltlich wiedergegeben.

**Externe Validität als Bestandteil der Qualitätsbewertung**

Die externe Validität hängt ab von dem Kontext, für den eine Entscheidung gefällt werden soll. Es wird kein festes Kriterienset zu Bewertung der externen Validität verwendet. Es wird auf die Schwierigkeit hingewiesen, externe Validität mit einer Checkliste abzugreifen. Es ist schwierig, die externe Validität mit einem generischen Instrument zur bewerten, sondern wahrscheinlich notwendig, die Kriterien zur Bewertung vorab jeweils themenspezifisch zu definieren. Mehrere Teilnehmende äußern sich dahingehend, die externe getrennt von der internen Validität zu bewerten.

**Subjektivität der Bewertung**

Allgemein wird dargelegt, dass Kriterien, die einen größeren Spielraum für eine subjektive Interpretation lassen, zu einem großen Diskussionsbedarf infolge fehlender Übereinstimmungen bei einer Doppelbewertung führen. Bei Mangel an Übereinstimmung der Bewertung erhöht sich der Zeitaufwand durch den erhöhten Diskussionsbedarf. Die Übereinstimmung hängt u. a. ab von der methodischen Kompetenz der Reviewer. Qualitätsbewertungen werden auch von Reviewern durchgeführt, die keine Methodiker sind oder eine fundierte epidemiologisch-statistische Ausbildung haben. Allerdings ist eine Symbiose aus klinischem und methodischem Sachverstand für die Erstellung von HTA-Berichten erstrebenswert. Es ist zu berücksichtigen, dass klinische Experten häufig wenig methodologische Kenntnisse haben. Daher sollten die Instrumente auch für Nicht-Methodiker anwendbar sein.

**Instrumente zur Bewertung der Studienqualität**

Als Probleme bei der Anwendung der Checklisten der GSWG wird das Fehlen eines Handbuchs genannt, sodass bei eher subjektiven Kriterien ein großer Diskussionsbedarf aufgrund mangelnder Übereinstimmung der Bewertung resultiert. Als weiterer problematischer Punkt in der Praxis hat sich herausgestellt, dass die designspezifischen Kriterien nicht immer anwendbar sind. Auch die Vermischung von interner und externer Validität bei den Checklisten der GSWG wird kritisch angemerkt.

Der Begriff Qualitätsbewertung/Bewertung der Studienqualität sollte zugunsten der Bewertung der internen Validität verlassen werden. Dies wird so innerhalb der Cochrane Collaboration gehandhabt. Der Begriff Qualität wird bei GRADE nur noch für die Bewertung der Qualität der Evidenz verwendet.

Im Zusammenhang mit der Berichtsqualität wird als gute Lösung die Erfassung in den Antwortkategorien des Instruments wie beispielsweise in Instrumenten von SIGN über eine Differenzierung in „nicht berichtet“ und „nicht zutreffend für den Kontext“ angesehen. Dadurch kann die Problematik Berichts- vs. Studienqualität besser eingeschätzt werden.

Skalen können empirisch nachgewiesenermaßen nicht valide zwischen niedriger und höher Studienqualität unterscheiden und werden daher nicht angewandt.

Einer der Hauptvorteile des „Risk of bias tool“ der Cochrane Collaboration ist die Transparenz der Bewertung. Diese wird dadurch erreicht, dass die Autoren ihre Bewertung der einzelnen Domänen mit Originalzitaten aus den jeweiligen Studien belegen.

Es wird auf das Biasrisiko durch ‚selective outcome reporting’ hingewiesen.

Auf ein neueres Prinzip wird aufmerksam gemacht: die Bewertung nicht auf Studienebene, sondern auf Ebene der Endpunkte (Outcomes), da sich das Biasrisiko je nach Endpunkt unterscheiden kann.

**Nicht-randomisierte Studien**

Die Frage nach der Unterschätzung der Validität von nicht-randomisierten Studien kann nicht generell beantwortet werden. Es scheint, dass in ersten Studien zu einer bestimmten Fragestellung die Validität häufiger überschätzt wird oder verfrüht Schlussfolgerungen auf der Basis einer geringen Anzahl von Studien gezogen werden. Aber es gibt sehr gute Fälle, wo der Verzicht auf gute nicht-randomisierte kontrollierte Studien mit einem Erkenntnisverlust einhergeht. Ein pragmatischer diskutabler Ansatz ist, nicht ausschließlich qualitativ hochwertige (randomisierte) Studien, sondern auch eine kritische Anzahl qualitativ hochwertiger Studien zu berücksichtigen im Sinn von „best evaluated evidence“. Es wird kritisch hinterfragt, ob RCT immer die „best available evidence“ widerspiegeln. Es wird auch darauf hingewiesen, dass nicht-randomisierte Studien nicht für einen klaren Kausalitätshinweis herangezogen werden können.

**Integration der Ergebnisse der Qualitätsbewertung**

Bei der Qualitätsbewertung werden viele Informationen erhoben und es ist schwierig, diese in die Ergebnissynthese zu integrieren. Als praktikabel wird eine qualitative (beispielsweise dreistufige) Gesamtbewertung angesehen. Dieses Vorgehen wird auch als semi-quantitativ bezeichnet. Alternativ können problematische Aspekte qualitativ in narrativer Form berichtet werden.

In einer Metaanalyse werden nur Studien mit niedrigem Verzerrungspotenzial berücksichtigt. In eine anschließende Sensitivitätsanalyse werden alle Studien einbezogen. Auch das umgekehrte Vorgehen, eine Metaanalyse mit allen Studien und eine Sensitivitätsanalyse der Studien mit niedrigem Verzerrungspotenzial, wird berichtet.

## 6.6 Diskussion

Die Qualitätsbewertung von Studien, aus deren Ergebnissen Empfehlungen für evidenzbasiertes Handeln abgeleitet werden, ist obligatorischer Bestandteil bei der Erstellung systematischer Übersichtsarbeiten. Der vorliegende Forschungsbericht hat das Ziel, einen Überblick über QBI für Primär- und Sekundärstudien zu geben, diese Instrumente zu vergleichen und ggf. aus den Ergebnissen Schlussfolgerungen für die Durchführung von Qualitätsbewertungen abzuleiten. Es werden Instrumente zur Bewertung von Studien, die die Wirksamkeit bzw. Güte von gesundheitsbezogenen Verfahren und Interventionen untersuchen sowie Instrumente zur gesundheitsökonomischen Bewertung eingeschlossen, die im Weiteren getrennt voneinander diskutiert werden.

Durch eine umfassende und ausgedehnte Literaturrecherche, die u. a. die Internetrecherche bei HTA- und EbM-Organisationen einbezieht, werden zahlreiche Instrumente zur Qualitätsbewertung gefunden. Die systematische Datenbankrecherche gestaltet sich aufgrund der ungenügenden themenspezifischen Indexierung in den einbezogenen Datenbanken schwierig. So existiert weder ein deutsch- noch ein englischsprachiges Schlagwort zu „Studienqualität". Auch andere Autoren beschreiben vergleichbare Schwierigkeiten bei der Suche nach methodischer Literatur[23, 56, 235]. Aufgrund der Vielzahl der weltweit eingesetzten Instrumente im Rahmen der in den letzten Jahren stark steigenden Zahl an systematischen Übersichtsarbeiten[182] können sicher nicht alle jemals verwendeten Instrumente aufgedeckt werden. Jedoch wird die Möglichkeit, bedeutsame und häufig eingesetzte Instrumente übersehen zu haben, als sehr gering angesehen.

Ein intiales Screening von HTA-Berichten zum Einsatz von QBI in der DIMDI-Datenbank zeigt, dass in der großen Mehrheit der Berichte (87 %) die Durchführung einer Qualitätsbewertung der Studien angegeben wird. Inwieweit es sich bei Berichten ohne Angabe einer Qualitätsbewertung lediglich um mangelnde Dokumentation handelt, ist unklar. Ältere, internationale Untersuchungen fanden, dass in systematischen Reviews eine Qualitätsbewertung bei bis zu 50 % der Studien nicht durchgeführt wird[56, 141, 236]. Insofern erscheint das Ergebnis von 87 % deutlich positiver, darf jedoch nicht darüber hinwegtäuschen, dass die Durchführung einer Qualitätsbewertung ein obligatorischer Bestandteil von systematischen Reviews und HTA-Berichten ist und die Nichtdurchführung einen deutlichen methodischen Mangel darstellt. Die weitere Durchsicht der Berichte mit Angabe einer Qualitätsbewertung zeigt, dass eine Dokumentation des verwendeten Instrumentes nur bei der Hälfte der Berichte vorliegt. Dies ist Ausdruck einer ungenügenden Berichtsqualität und entsprechend verbesserungswürdig.

### 6.6.1 Bewertung von Studien zur Wirksamkeit

Durch die umfangreiche Recherche werden 125 Instrumente zur Qualitätsbewertung gefunden, die eine große Variation hinsichtlich ihrer Charakteristika aufweisen. Die QBI werden nach Studiendesign den vier Gruppen Intervention-, Beobachtungs-, Diagnosestudien und systematischen Reviews/HTA-Berichte/Metaanalysen zugeordnet. Ihre formalen und inhaltlichen Charakteristika werden extrahiert und tabellarisch dargestellt. Anhand der Abdeckung von Elementen der internen Validität können umfassendere von weniger umfassenden Instrumenten unterschieden werden.

Studienqualität kann unterschiedlich operationalisiert werden. Meist wird Studienqualität als interne Validität definiert, also als die Glaubwürdigkeit der Studienergebnisse unter Berücksichtigung von möglichen systematischen Fehlern oder Verzerrungen durch das Design, die Durchführung oder die Auswertung einer Studie[117, 132, 143]. Mitglieder der Cochrane Collaboration grenzen den Begriff der Studienqualität (methodological quality) von dem des Verzerrungspotenzials (risk of bias) ab[92]: Sofern eine Studie hohe Qualitätsstandards erfülle, eine Verblindung jedoch beispielsweise aufgrund der Studienfrage nicht durchführbar sei (z. B. operative vs. konservative Therapie), bestehe trotzdem ein erhöhtes Verzerrungspotenzial. Sie raten daher von dem Begriff der Qualitätsbewertung zugunsten der Beschreibung als Bewertung des Verzerrungspotenzials ab. Entsprechend lautet das von der Cochrane Collaboration empfohlene Instrument „risk of bias tool". Dieser Definition folgend wird im vorliegenden Bericht der Begriff der Qualitätsbewertung mit dem der Bewertung der internen Validität gleichgesetzt. Dieser Aspekt wird ebenfalls auf dem Workshop thematisiert.

Die Bestandsaufnahme der zahlreichen Instrumente zeigt, dass viele Instrumente eine Mischung aus Items zur Berichtsqualität, sowie denen zur internen und externen Validität enthalten. Die Vermischung

von Berichts- und Studienqualität wird u. a. bereits von Deeks et al.[56] in ihrem HTA-Bericht beschrieben, in dem sie fast 200 Instrumente zur Bewertung der Studienqualität von nicht-randomisierten Studien untersuchen. Diese Vermischung insbesondere von Berichtsqualität und interner Validität kann zu einer Fehleinschätzung der Studienqualität führen, beispielsweise, wenn nach der Angabe der Höhe der Drop-out-Rate gefragt wird anstatt nach der Angemessenheit (appropriateness) der Höhe der Rate. Wird nur nach der Angabe gefragt und diese berichtet, so wird bei Erfüllung dieses Items eventuell fälschlicherweise auf eine adäquate Studienqualität geschlossen, wenn die Drop-out-Rate zwar berichtet, aber vielleicht sehr hoch ist und auf diese Weise die Studienqualität überschätzt. Wird dagegen die Angemessenheit der Drop-out-Rate erfasst, die idealerweise präzise operationalisiert sein sollte (z. B. angemessen, wenn < 20 % und nicht differenziell), so kann bei Erfüllung des Items korrekterweise eine adäquate Studienqualität mit diesem Punkt assoziiert werden.

Eine adäquate Berichtsqualität ist Voraussetzung für die Bewertung der internen Validität. Werden relevante methodische Aspekte nicht berichtet, können sie nicht zur Bewertung der Studienqualität herangezogen werden. Bei schlechter Berichtsqualität besteht die Gefahr einer Unterschätzung der tatsächlichen Studienqualität[143, 228]. Eine Nachfrage einzelner nicht berichteter Studienelemente bei den Autoren ist jedoch nicht empfehlenswert, da die Validität der Angaben nicht kontrolliert werden kann und es Hinweise gibt, dass die Studiencharakteristika zu positiv dargestellt werden[81]. Um zwischen einer schlechten Studienqualität aufgrund einer schlechten Berichtsqualität oder einer schlechter methodischen Qualität differenzieren zu können, kann die Berichtsqualität durch eine zusätzliche Antwortkategorie „nicht berichtet“ erfasst werden, wie dies beispielsweise bei Instrumenten von SIGN der Fall ist. Dieses Vorgehen wird auch von Workshop-Teilnehmern als vorteilhaft angesehen.

Auch die Bewertung der externen und internen Validität soll nach Ansicht von Workshop-Teilnehmern nicht mit einem gemeinsamen Instrument, sondern getrennt voneinander erhoben werden. In diesem Zusammenhang wird darauf hingewiesen, dass die externe Validität immer nur für den gewählten Kontext bewertet werden kann. Daher ist ein generisches Instrument weniger geeignet, stattdessen sind die Kriterien zur Bewertung der externen Validität spezifisch auf das jeweilige Thema abzustimmen.

Der Ansatz, die Qualitätsbewertung rein auf die Bewertung der internen Validität zu beschränken, kann in diesem Bericht nicht durchgehend umgesetzt werden. Items von Instrumenten werden zur besseren Transparenz der Berichtsqualität, der internen oder der externen Validität zugeordnet. Für einzelne Aspekte kann die Zuordnung nicht eindeutig getroffen werden, beispielsweise zielt die Frage nach einer Power-Berechnung zunächst weniger auf die interne Validität, sondern auf die Präzision der Ergebnisse bzw. die Berichtsqualität. Allerdings hat die Präzision der Effektschätzer Einfluss auf die Signifikanz der Ergebnisse. Ein weiteres Beispiel für den Einfluss der Präzision auf Studienergebnisse ist die Nichtbeachtung einer Cluster-Randomisierung bei der Datenauswertung. Diese führt fälschlich zu einer hohen Präzision und damit eher zu signifikanten Studienergebnissen.

Ein weiteres Beispiel für die schwierige Zuordnung einzelner Elemente von Qualitätsbewertungsinstrumenten zu den Bereichen interne Validität, externe Validität und Berichtsqualität ist die Beschreibung von Ein- und Ausschlusskriterien. Diese sind am ehesten der Berichtsqualität zuzuordnen bzw. für die Einschätzung der externen Validität wichtig. Dagegen wird die a priori Definition von Ein- und Ausschlusskriterien in systematischen Reviews als Element der internen Validität verortet, da Änderungen nach Studienbeginn die Ergebnisse systematisch verzerren können.

Wie in vergleichbaren Übersichtsarbeiten[56, 109, 143, 168, 189, 190, 235, 236] werden für den inhaltlichen Vergleich der Instrumente Elemente definiert, deren Zutreffen oder Vorhandensein in den jeweiligen Instrumenten geprüft und extrahiert wird. Die verwendeten inhaltlichen Elemente werden aus dem HTA-Bericht der AHRQ von West et al.[235] für QBI für systematische Reviews/HTA-Berichte/Metaanalysen sowie Interventions- und Beobachtungsstudien übernommen, teilweise modifiziert (wenn ein Parameter mehrere Aspekte umfasst, werden diese in mehreren Parametern abgebildet) und nach Sichtung weiterer methodischer Literatur ergänzt. Die von West et al.[235] aufgelisteten Elemente sind aus Konstrukten, die die Studienqualität beeinträchtigen können, sowie mithilfe einer Expertenberatergruppe, der u. a. Mitglieder der Cochrane Collaboration, SIGN und dem NHS Centre for Reviews and Dissemination angehören, abgeleitet worden und basieren zum Teil auf nachgewiesenen empirischen Erkenntnissen. Die Zusammenfassung mehrerer Elemente zu einem übergeordneten Themenbereich (Domäne) und dessen Bewertung (ja, nein, teilweise erfüllt) wird nicht übernommen, da der Bewertungsvorgang nicht vollständig nachvollziehbar ist und Elemente hinzugefügt werden.

Davon abweichend werden QBI für Diagnosestudien anhand extrahierter inhaltlicher Elemente aus einem von Whiting et al.[237] publizierten HTA-Bericht verglichen.

Die Auswahl von inhaltlichen Kriterien zur Bewertung bzw. zum Vergleich von QBI ist methodisch anspruchsvoll. Es existiert kein Konsens über geeignete Kriterien und die untersuchten Elemente in den methodischen Übersichtsarbeiten variieren. Das Vorgehen zur Auswahl von als geeignet angesehenen Kriterien reicht über Dephi-Verfahren mit Experten, anderweitige Expertenbefragungen, die Verwendung allgemein anerkannter Kriterien, die Verwendung von Kriterien mit empirisch nachgewiesenem Verzerrungspotenzial bis zur Kombination der aufgezählten Verfahren.

Anhand der tabellarischen Darstellung der abgedeckten inhaltlichen Elemente können die identifizierten QBI verglichen werden. Als Basis für die Auswahl eines geeigneten Bewertungsinstruments werden im Weiteren nur generische Instrumente betrachtet, die themenübergreifend anwendbar sind. Dieses Vorgehen wählen auch andere Autoren[143, 238]. Es werden für jedes Instrument die Gesamtzahl erfüllter Elemente sowie die Anzahl an Domänen, in denen mindestens ein bzw. mindestens 50 % der Elemente abgedeckt ist, beziffert (Ausnahme: Diagnosestudien). Es zeigen sich designspezifisch große Unterschiede in der Anzahl abgedeckter Elemente und Domänen. Allerdings ist eine möglichst hohe Zahl an abgedeckten Elementen nicht notwendigerweise ein Hinweis auf ein besonders gutes Instrument. Ziel kann es ebenso sein, im Sinn der Effizienz ein Instrument mit möglichst wenigen, aber ausreichenden Elementen anzuwenden[143, 189]. Inwieweit die verwendeten inhaltlichen Elemente alle ein notwendiger Bestandteil einer Qualitätsbewertung sein sollten, ist unklar.

Zur weiteren Differenzierung der inhaltlichen Elemente werden als besonders relevant eingeschätzte Elemente definiert und deren Erfüllung ebenfalls instrumentenbezogen dargestellt. Die relevanten Elemente für Interventions- und Diagnosestudien sind evidenzbasierte Biasquellen, während dies nur für einen Teil der als relevant erachteten Elemente für Beobachtungsstudien und systematischen Reviews/HTA-Berichten/Metaanalysen gilt. Daher sind insbesondere für die beiden letztgenannten Studiendesigns weitere methodische Studien notwendig, um Studiencharakteristika zu identifizieren, die die Höhe der Studienergebnisse systematisch beeinflussen und daher Auswirkung auf die interne Validität einer Studie haben. Hierzu gehören insbesondere die Elemente der Qualitätsbewertung durch zwei unabhängige Reviewer bei der Bewertung von systematischen Reviews/HTA-Berichten/Metaanalysen und bei Beobachtungsstudien die Bewertung von Verfahren zur Kontrolle von Confounding.

Die gewählten inhaltlichen Elemente und deren Abdeckung sind eingeschränkt geeignet, die Güte von QBI zu vergleichen. Nicht alle sind evidenzbasierte Biasquellen und diejenigen, für die keine Evidenz vorliegt, können dennoch Einfluss auf die interne Validität haben, der nur bislang nicht systematisch untersucht wird. Die Abdeckung der inhaltlichen Elemente und insbesondere der als relevant definierten Elemente kann daher nur als erste Einschätzung dienen, um Instrumente zu identifizieren, die mehr oder weniger umfassend sind. Je nach Themenbereich, für den ein Qualitätsbewertungsinstrument eingesetzt werden soll, sollte geprüft werden, ob alle Items des Instruments relevant sind bzw. themenspezifisch zusätzliche Items einbezogen werden sollten[240].

Die Methode, Qualitätsinstrumente anhand des Vorhandenseins bestimmter inhaltlicher Parameter zu vergleichen, hat weitere Schwächen. Der Nachweis der Abfrage bestimmter Studiencharakteristika (Randomisierung, Vergleichbarkeit der Studiengruppen etc.) zur Abschätzung der Studienqualität vernachlässigt die notwendige Operationalisierung der einzelnen Parameter, die jedoch für eine korrekte Einschätzung erforderlich ist. Für die Frage nach der Angemessenheit der Randomisierung beispielsweise müssen verschiedenste Verfahren der Sequenzgenerierung als angemessen oder nicht angemessen beurteilt werden. Je präziser und ausführlicher dieser Parameter in einer begleitenden Anleitung operationalisiert wird, desto korrekter kann die Frage nach der Angemessenheit durch das jeweilige Instrument eingeschätzt werden. (Auf die Bedeutung der Operationalisierung der Items wird im Zusammenhang mit Subjektivität und Reviewerbias nochmals eingegangen.)

Ein wichtiger methodischer Aspekt bei der Durchführung von systematischen Übersichtsarbeiten ist die Subjektivität der Bewertung, die zu einer Verzerrung der Studienergebnisse führen kann (Reviewerbias). Der Grad der Übereinstimmung der Bewertung unterschiedlicher Reviewer (Interrater-Reliabilität) gibt einen Hinweis, wie gut die Subjektivität der Bewertung durch Standardisierung mit präziser Operationalisierung von Inhalten und/oder ggf. Methodenschulung kontrolliert werden kann. Diese Einschätzung wird unterstützt durch eine aktuelle Publikation, die die Raterübereinstimmung für das von der Cochrane Collaboration seit 2008 empfohlene „Risk of bias tool“[92] untersucht[85]. Das

Instrument besteht aus sechs Komponenten, deren angemessene Erfüllung mit ja, nein oder unklar bewertet wird. Für die Bewertung liegt eine ausführliche Anleitung vor. Die Kappa-Werte der einzelnen Komponenten variierten von ϰ = 0,13 bis ϰ = 0,74 mit geringerer Übereinstimmung bei Items, die mehr Urteilsvermögen erfordern. Für die Anwendung des Instruments sind daher ein sorgfältiges Training sowie Entscheidungsregeln notwendig. Auch Oxman et al.[176] finden eine höhere Übereinstimmung bei Items, die weniger Interpretationsmöglichkeiten bieten. Dies untermauert die Forderung, Items bzw. Komponenten möglichst präzise zu operationalisieren, um die Subjektivität bzw. den Bewertungsspielraum zu minimieren.

Für die Güte der Operationalisierung der Items oder Komponenten von QBI gilt, je weniger Items oder Komponenten ein Instrument umfasst, desto präziser und ausführlicher sollte die Operationalisierung gestaltet sein. Diese Einschätzung wird auch von Workshop-Teilnehmern geteilt. Bei einem Instrument mit sehr vielen Items sind die einzelnen Items sehr detailliert und bedürfen eher keiner oder nur geringer Erläuterung. In diesem Forschungsbericht wird die Güte der Operationalisierung nicht bewertet, sondern erfasst, ob und wie ausführlich die Items erläutert sind. Bis zu 40 % der Instrumente halten ausführliche Erläuterungen zur Operationalisierung einzelner Items und Komponenten vor. Hier besteht Verbesserungsbedarf. Auch Whiting et al.[240] weisen bei der Evaluation ihres evidenzbasierten Instruments für diagnostische Studien QUADAS auf die Bedeutsamkeit der Ausfüllhinweise und Operationalisierungen bei der Anwendung hin. Es wird als essenziell angesehen, dass die Reviewer die Ausfüllhinweise ggf. entsprechend anpassen, um eine eindeutige Bewertungsgrundlage für alle Reviewer sicherzustellen.

Die Cochrane Collaboration zeichnet sich bekanntlich durch hohe Anforderungen an die methodische Qualität aus. Dies betrifft nicht nur die Einschlusskriterien von Studien in systematische Reviews, sondern auch die Durchführung und transparente Darstellung von systematischen Reviews. In diesem Zusammenhang soll eine Besonderheit des „Risk of bias tool“ der Cochrane Collaboration herausgestellt werden, die bei der Qualitätsbewertung nicht Standard ist. Um die Bewertung der Reviewer möglichst transparent und nachvollziehbar zu gestalten, sind für jede der sechs bewerteten Komponenten Originalzitate aus den entsprechenden Studien anzugeben, auf denen die jeweilige Bewertung basiert. Aber auch Mitglieder der Cochrane Collaboration geben zu bedenken: „The most realistic assessment of the validity of a study may involve subjectivity: for example an assessment of whether lack of blinding of patients might plausibly have affected recurrence of a serious condition such as cancer.“[92]

Zur Wirksamkeit einer Verblindung der Reviewer für Autor und Publikationsquelle als Maßnahme zur Minimierung eines Reviewerbias gibt es diskrepante Studienergebnisse[45, 106, 145, 227], aus denen keine Empfehlung abgeleitet werden kann. Außerdem ist der mit der Verblindung verbundene Zeitaufwand zu bedenken und dass eine Verblindung nicht möglich ist, wenn die Reviewer die Literatur bereits kennen.

Allgemein akzeptiertes Vorgehen bei der Bewertung der Studienqualität ist die Bewertung durch zwei unabhängige Reviewer und die Klärung der diskrepanten Punkte ggf. unter Hinzuziehung weiterer Personen. Eine Studie zur Datenextraktion zeigt, dass die Datenextraktion durch zwei unabhängige Reviewer weniger Fehler produziert als in dem Fall, in dem die Datenextraktion des ersten Reviews durch den zweiten Reviewer lediglich kontrolliert wird[27]. Eine diesbezügliche Untersuchung für die Qualitätsbewertung wird nicht gefunden, die Ergebnisse zur Datenextraktion geben jedoch Hinweise für den Prozess der Qualitätsbewertung. Für die Qualitätsbewertung ist mit eindeutigeren Unterschieden zu rechnen, da die Qualitätsbewertung eher ein höheres Maß subjektiver Einschätzungen erfordert als die Datenextraktion.

### 6.6.2 Bewertung gesundheitsökonomischer Studien

Die Bewertung der Qualität gesundheitsökonomischer Studien, aus deren Ergebnissen Empfehlungen für evidenzbasiertes Handeln abgeleitet werden, ist ein zwingend erforderlicher Bestandteil bei der Erstellung von HTA-Berichten und anderen systematischen Übersichtsarbeiten. Insgesamt werden 22 gesundheitsökonomische Instrumente identifiziert. Die Zielsetzung eines QBI ist es, die methodische Qualität gesundheitsökonomischer Studien in einem vorgegebenen Kontext standardisiert zu bewerten. Die Qualität gesundheitsökonomischer Studien wird bestimmt durch (a) die Validität der Studien-

ergebnisse, (b) die Einhaltung methodischer Standards der gesundheitsökonomischen Evaluation und (c) den Zugang zu belastbaren Kosten- und Outcomedaten.

Zwischen den untersuchten QBI gibt es deutliche Unterschiede bezüglich

- Anzahl der untersuchten Items aus dem Extraktionsformular (Themenschwerpunkte)
- Bewertungsqualität: angemessen – begründet – berichtet
- Differenziertheit der Qualitätsabfragen.

Keines der untersuchten Instrumente deckt die gesamte Bandbreite der Themenschwerpunkte (Elemente der gesundheitsökonomischen Evaluation) ab. Die Instrumente von Siebert et al.[197], Drummond et al.[63] und Larsen et al.[122] befassen sich mit den meisten Themenschwerpunkten. Die Instrumente von Sackett (critical appraisal), von Ramos et al.[179] und der European Collaboration for Assessment of Health Interventions (ECHTA) berücksichtigen nur wenige Themenschwerpunkte.

Deutliche Unterschiede zwischen den QBI bestehen auch in der Differenziertheit der Qualitätsabfragen. Wie differenziert ein Bewertungsinstrument die Themenschwerpunkte erfragt, wird über die Anzahl der Items abgebildet. Die Instrumente von Ramos et al.[179], Sackett[187] und Rychetnik & Frommer[185] bestimmen die Studienqualität mit wenigen Items (fünf bis zehn Items). Demgegenüber sind die Instrumente von Siebert et al.[197], Larsen et al.[122] und Ungar & Santos[216] deutlich umfangreicher (56 bis 63 Items). Wenn sich die Qualitätsbewertung auf wenige Items stützt, müssen die Fragen global gestellt werden. Reviewern bleiben dann große Spielräume bei der Interpretation von Items – mit der Folge, dass große Unterschiede in der Bewertung zwischen Reviewern auftreten können. Bei umfangreicheren Instrumenten mit großer Itemanzahl lassen sich Items stärker operationalisieren, sodass die Interpretationsspielräume deutlich eingeschränkt und objektivere Bewertungen unterstützt werden.

Die gesundheitsökonomischen Bewertungsinstrumente sind überwiegend generische Instrumente. Es wird nur ein spezifisches Bewertungsinstrument identifiziert: Das Instrument von Ungar & Santos[216] ist für die Qualitätsbewertung ökonomischer Studien in der pädiatrischen Versorgung entwickelt. Relevante Aspekte bei der Versorgung von Kindern werden hier besonders betont (z. B. Produktivitätsverluste der Eltern durch informelle Pflege, Lebensqualität des Kindes und der Eltern). Spezifische Instrumente werden tendenziell stärker operationalisiert sein. Grundsätzlich scheint es jedoch hinreichend, generische Instrumente zu verwenden. Sie sollten aber hinreichend differenziert sein, um die Bewertungen zu operationalisieren.

Insgesamt bestehen zwischen den verschiedenen Instrumenten deutliche Unterschiede bezüglich der Abdeckung der verschiedenen gesundheitsökonomischen Themenschwerpunkte, Bewertungsqualität und Differenziertheit der Qualitätsabfragen. Eine standardisierte Qualitätsbewertung von gesundheitsökonomischen Studien sollte ein integraler Bestandteil bei der Erstellung von HTA-Berichten und systematischen Übersichtsarbeiten sein. Hierzu sollten Instrumente verwendet werden, die ein möglichst breites Themenspektrum abbilden sowie die Qualität möglichst differenziert erheben. Umfangreichste QBI sind die Instrumente von Siebert et al.[197] und Drummond et al.[63]. Sie berücksichtigen die überwiegende Mehrheit der Themenschwerpunkte (84 % bzw. 81 %) und weisen aufgrund der hohen Anzahl an Items (56 bzw. 35 Items) eine hohe Differenziertheit und Operationalisierung auf. Das Instrument von Siebert et al. erfragt zudem die Angemessenheit des Vorgehens bei der Durchführung der Studien, sodass der Aspekt der Qualität stärker berücksichtigt wird. Trotz der hohen Anzahl an Items decken beide Instrumente nicht alle Themenschwerpunkte ab. Beispielsweise wird die wichtige Frage nach Art und Quelle der Finanzierung der Studie von den Instrumenten nicht erhoben.

Es lässt sich festhalten, dass QBI (weiter-) entwickelt werden sollten, die (1) die gesamten Themenschwerpunkte abbilden, (2) die angemessene Umsetzung von Items in gesundheitsökonomischen Studien überprüfen und (3) die Themenschwerpunkte hinreichend differenziert abfragen. Die Angemessenheit sollte sich an den Standards der gesundheitsökonomischen Evaluation orientieren. Es sollten Erläuterungen und Ausfüllhinweise zu den Bewertungsinstrumenten entwickelt werden, in denen beschrieben wird, wie Angemessenheit definiert ist. Erläuterungen fördern die Operationalisierbarkeit der Bewertungsinstrumente und lassen sich leicht anpassen, wenn Standards der gesundheitsökonomischen Evaluation verändert werden. Zudem schränken sie den Ermessensspielraum der Reviewer bei der Bewertung ein und tragen somit zu einer standardisierten Qualitätsbewertung bei. Die vorhandenen Bewertungsinstrumente weisen überwiegend keine Erläuterungen oder Ausfüllhinweise auf.

Detaillierte Informationen zum genauen Vorgehen oder konkrete Empfehlungen zur Bewertung von Studien in Form von Handbüchern sind bei keinem Instrument vorhanden.

## 6.7 Schlussfolgerungen

Gesundheitspolitische Entscheidungen basieren auf wissenschaftlichen Erkenntnissen, aus denen Empfehlungen für evidenzbasiertes Handeln abgeleitet werden. Ziel der Qualitätsbewertung ist die Einschätzung der Glaubwürdigkeit von Studienergebnissen. Die Qualitätsbewertung von Studien ist daher obligatorischer Bestandteil bei der Erstellung systematischer Übersichtsarbeiten und transparent darzustellen. Die Durchführung einer Qualitätsbewertung ist ein anspruchsvoller Prozess, der profunde methodische Kenntnisse erfordert.

Aus dem vorliegenden Forschungsbericht können die folgenden Schlussfolgerungen abgeleitet werden.

**Auswahl eines Instruments zur Qualitätsbewertung**

Es kann keine Empfehlung für die Verwendung eines bestimmten Instruments aus der Vielzahl der vorhandenen Instrumente gegeben werden. Es können jedoch anhand der tabellarisch dargestellten inhaltlichen Elemente Instrumente für die Qualitätsbewertung ausgewählt werden, die Elemente umfassender als andere abdecken bzw. die Elemente abdecken, die für spezifische Fragestellungen als wichtig eingeschätzt werden.

Instrumente mit einem quantitativen Bewertungssystem (Skalen) sollten nicht bzw. ohne quantitative Gesamtbewertung eingesetzt werden.

Bei der Wahl eines QBI sollten folgende Aspekte mit berücksichtigt werden:

- Ausführliche und präzise Operationalisierung der zu bewertenden Items bzw. Komponenten a priori.
- Möglichst keine Vermischung von Parametern zur Berichtsqualität und interner Validität.
- Getrennte Bewertung der externen Validität, da hier andere Aspekte als für die interne Validität berücksichtigt werden. Eine Ausnahme bildet die Qualitätsbewertung von diagnostischen Studien, die einen möglichen Spektrumbias, der der externen Validität zugeordnet wird, einbezieht, da soziodemografische und klinische Charakteristika sowie Krankheitsprävalenz und -schwere einer Studienpopulation sich auf die Höhe von Sensitivität und Spezifität diagnostischer Tests auswirken können.

**Durchführung der Qualitätsbewertung**

Wenn möglich, sollten die ausgewählten Instrumente zuvor an ausgewählten Studien getestet werden und bei Bedarf die Operationalisierung der Items ergänzt bzw. präzisiert werden, um die Subjektivität der Bewertung zu minimieren und eine gute Übereinstimmung der Bewertungen sicherzustellen. Auch eine Schulung der Reviewer hinsichtlich der spezifischen methodischen Anforderungen an eine Qualitätsbewertung ist je nach Vorkenntnissen und Erfahrung zu erwägen.

Die Qualitätsbewertung sollte möglichst von mindestens zwei Reviewern unabhängig voneinander durchgeführt und diskrepante Einschätzungen durch Diskussion und Konsensbildung, ggf. unter Einbezug weiterer Personen, gelöst werden.

Reviewer sollten profunde Methodenkenntnisse aus der Epidemiologie und/oder EbM sowie für den gesundheitsökonomischen Berichtsteil Kenntnisse der gesundheitsökonomischen Konzepte und Methoden haben. Für die Beurteilung der Angemessenheit von komplexeren Auswertungsverfahren, beispielsweise multivariaten Modellen, ist zu prüfen, ob zusätzlich eine statistische Expertise erforderlich ist.

Es sind ausreichend zeitliche und personelle Ressourcen für die Qualitätsbewertung einzuplanen bzw. sicherzustellen. Dies betrifft sowohl die Vorbereitung (Auswahl von Instrumenten, Operationalisierung der Parameter, a priori Festlegung der Art der Integration der Ergebnisse) als auch die Durchführung der Qualitätsbewertung sowie Diskussion der Ergebnisse und ggf. Konsensbildung.

**Forschungsbedarf**

Es sind methodische Studien erforderlich, um Studiencharakteristika identifizieren, die die Höhe der Studienergebnisse systematisch beeinflussen, um auf diese Weise evidenzbasierte Biasquellen zu definieren und als Grundlage für die Weiterentwicklung von QBI verwenden zu können. Dies ist besonders wichtig für Elemente von Beobachtungsstudien, vor allem um Verfahren zur Kontrolle von Confounding besser bewerten zu können.

Die Durchführung der Qualitätsbewertung wird üblicherweise von zwei unabhängigen Reviewern durchgeführt. Alternativ kann die Bewertung durch einen Reviewer erfolgen und von einem weiteren lediglich kontrolliert werden, was als zeitsparender und damit effizienter einzuschätzen ist. Während für die Datenextraktion aus einer Studie bekannt ist, dass das letztere Vorgehen mehr Fehler produziert, werden für den Prozess der Qualitätsbewertung keine Untersuchungen identifiziert. Daher sind Studien erforderlich, die beide Vorgehensweisen bei der Qualitätsbewertung vergleichen.

Für die Verwendung von Skalen als QBI kann nachgewiesen werden, dass durch die numerische Bewertung, die implizit oder explizit eine Gewichtung der Items beinhaltet, die Höhe der internen Validität nicht korrekt abgebildet wird. Aber auch Checklisten oder Komponentensysteme mit einer Komponenten- bzw. Gesamtbewertung kommen zu einer Abstufung der Studien hinsichtlich ihrer internen Validität. Es ist jedoch unklar, inwieweit eine qualitative Gesamtbewertung die Höhe der internen Validität korrekt abbildet.

Bei der Qualitätsbewertung gesundheitsökonomischer Studien sollte ein Instrument verwendet werden, das möglichst alle Themenschwerpunkte der gesundheitsökonomischen Evaluation abdeckt. Hierzu sollten neue Instrumente entwickelt, bzw. bestehende Bewertungsinstrumente weiterentwickelt werden. Zu den Instrumenten sollten Erläuterungen und Ausfüllhinweise bereitgestellt werden, um eine standardisierte und adäquate Bewertung der Studien vornehmen zu können. Es sollte darauf geachtet werden, dass die verwendeten Instrumente dem aktuellen Stand der gesundheitsökonomischen Forschung entsprechen.

# 7 Literaturverzeichnis

1. Academy of Managed Care Pharmacy. The AMCP format for formulary submissions – a format for submission of clinical and economic data in support of formulary consideration by health care systems in the United States. 2005.
2. AG Reha-Ökonomie im Förderschwerpunkt Rehabilitationswissenschaften, Hessel F, Kohlmann T, Krauth C, Nowy R, Seitz R, Siebert U, Wasem J. Gesundheitsökonomische Evaluation in der Rehabilitation. Teil I: Prinzipien und Empfehlungen für die Leistungserfassung. In: Verband Deutscher Rentenversicherungsträger (Ed). Förderschwerpunkt „Rehabilitationswissenschaften“: Empfehlungen der Arbeitsgruppen „Generische Methoden“, „Routinedaten“ und „Reha-Ökonomie“. Frankfurt am Main, 1999, 106-147.
3. Ah-See KW, Molony NC. A qualitative assessment of randomized controlled trials in otolaryngology. The Journal of Laryngology & Otology 1998; 112(5): 460-463.
4. Aidelsburger P, Felder S, Siebert U, Wasem J. Gesundheitsökonomische „Kurz-HTA-Berichte“. Eine systematische Übersichtsarbeit zur Methodik und Implementation. Deutsche Agentur für Health Technology Assessment des Deutschen Instituts für Medizinische Dokumentation und Information (DAHTA@DIMDI). 2003. Health Technology Assessment, Bd. 6.
5. Altman DG. Statistical reviewing for medical journals. Statistics in Medicine 1998; 17(23): 2661-2674.
6. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials. Explanation and elaboration. Annals of internal medicine 2001; 134(8): 663-694.
7. Andrew E, Eide H, Fuglerud P, Hagen EK, Kristoffersen DT, Lambrechts M, Waaler A, Weibye M. Publications on clinical trials with X-ray contrast media. Differences in quality between journals and decades. European Journal of Radiology 1990; 10(2): 92-97.
8. Assendelft WJ, Hay EM, Adshead R, Bouter LM. Corticosteroid injections for lateral epicondylitis. A systematic overview. British Journal of General Practice 1996; 46(405): 209-216.
9. Assendelft WJJ, Koes BW, Knipschild PG, Bouter LM. The relationship between methodological quality and conclusions in reviews of spinal manipulation. JAMA 1995; 274(24): 1942-1948.
10. Auperin A, Pignon JP, Poynard T. Review article. Critical review of meta-analyses of randomized clinical trials in hepatogastroenterology. Alimentary Pharmacology and Therapeutics 1997; 11(2): 215-225.
11. Balas EA, Austin SM, Ewigman BG, Brown GD, Mitchell JA. Methods of randomized controlled clinical trials in health services research. Medical care 1995; 33(7): 687-699.
12. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, Lau J. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002; 287(22): 2973-2982.
13. Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. JAMA 1998; 279(19): 1566-1570.
14. Bass JL, Christoffel KK, Widome M, Boyle W, Scheidt P, Stanwick R, Roberts K. Childhood injury prevention counseling in primary care settings. A critical review of the literature. Pediatrics 1993; 92(4): 544-550.
15. Bath FJ, Owen VE, Bath PM. Quality of full and final publications reporting acute stroke trial. A systematic review. Stroke: a journal of cerebral circulation 1998; 29(10): 2203-2210.
16. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996; 276(8): 637-639.
17. Bérard A, Andreu N, Tétrault J, Niyonsenga T, Myhal D. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. Annals of epidemiology 2000; 10(8): 498-503.
18. Bizzini M, Childs JD, Piva SR, Delitto A. Systematic review of the quality of randomized controlled trials for patellofemoral pain syndrome. The Journal of orthopaedic and sports physical therapy 2003; 33(1): 4-20.
19. Boers M, Ramsden M. Long acting drug combinations in rheumatoid arthritis: a formal overview. Journal of rheumatology 1991; 18(3): 316-324.

20. Borsody MK, Yamada C. Effects of the search technique on the measurement of the change in quality of randomized controlled trials over time in the field of brain injury. BMC medical research methodology 2005; 5(7).
21. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. British Medical Journal 2003; 326(7379): 41-44.
22. Bracken MB. Reporting observational studies. British Journal of Obstetrics and Gynaecology 1989; 96(4): 383-388.
23. Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. Health Technology Assessment 1999; 3(9): I-164.
24. Briggs AH. Handling uncertainty in economic evaluation and presenting the results. In: Drummond M and McGuire A (Eds). Economic evaluation in health care: merging theory with practice. New York, 2001, pp 172-214.
25. Brouwer W, Rutten F, Koopmanschap M. Costing in economic evaluations. In: Drummond M and McGuire A (Eds). Economic evaluation in health care: merging theory with practice. New York, 2001, pp 68-93.
26. Brown SA. Measurement of quality of primary studies for meta-analysis. Nursing Research 1991; 40(6): 352-355.
27. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J.Clin.Epidemiol. 2006; 59(7): 697-703.
28. Campbell MK, Elbourne DR, Altman DG. CONSORT statement. Extension to cluster randomised trials. British Medical Journal 2004; 328(7441): 702-708.
29. Canadian Agency for Drugs and Technologies in Health: Guidelines for the economic evaluation of health technologies: Canada, 3. Auflage ed. Ottawa, 2006.
30. Carson CA, Fine MJ, Smith MA, Weissfeld LA, Huber JT, Kapoor WN. Quality of published reports of the prognosis of community-acquired pneumonia. Journal of general internal medicine : official journal of the Society for Research and Education in Primary Care Internal Medicine 1994; 9(1): 13-19.
31. Center for Disease Control and Prevention. Preventing chronic disease. Reviewer checklist for health economic papers. www.cdc.gov/Pcd/for_reviewers/checklists/health_economics.pdf cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm“ (26.06.2009).
32. Centre for Evidence Based Mental Health. Critical appraisal form for a study of diagnosis. cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm (26.06.2009).
33. Centre for Evidence Based Mental Health. Critical appraisal form for a study of diagnosis. cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm (26.06.2009).
34. Centre for Evidence Based Mental Health. Critical appraisal form for a study of prognosis. cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm (26.06.2009).
35. Centre for Evidence Based Mental Health. Critical appraisal form for a study of prognosis. cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm (26.06.2009).
36. Centre for Evidence Based Mental Health. Critical appraisal form for an overview. cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm (26.06.2009).
37. Centre for Evidence Based Mental Health. Critical appraisal form for single therapy studies. cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm (26.06.2009).
38. Chalmers I, Adams M, Dickersin K, Hetherington J, Tarnow-Mordi W, Meinert C, Tonascia S, Chalmers TC. A cohort study of summary reports of controlled trials. JAMA 1990; 263(10): 1401-1405.
39. Chalmers TC, Smith H, Jr., Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. Controlled clinical trials 1981; 2(1): 31-49.
40. Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed. Review of publications and survey of authors. British Medical Journal 2005; 330(7494): 753.
41. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche 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.

42. Chan AW, Krleza-Jeric K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. Canadian Medical Association journal 2004; 171(7): 735-740.
43. Chiou CF, Hay JW, Wallace JF, Bloom BS, Neumann PJ, Sullivan SD, Yu HT, Keeler EB, Henning JM, Ofman JJ. Development and validation of a grading system for the quality of cost-effectiveness studies. Medical care 2003; 41(1): 32-44.
44. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. JAMA 1994; 272(2): 101-104.
45. Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, Fergusson D, Laupacis A. Assessing the quality of randomized trials. Reliability of the Jadad scale. Controlled clinical trials 1999; 20(5): 448-452.
46. Clark JP. How to peer review a qualitative manuscript. In: Godlee F and Jefferson T (Eds). Peer review in health sciences. 2nd ed. London, 2003, 219-235.
47. Clemens JD, Chuong JJ, Feinstein AR. The BCG controversy. A methodological and statistical reappraisal. JAMA 1983; 249(17): 2362-2369.
48. Cochrane Working Group on Systematic Review of Screening and Diagnostic Tests. Recommended Methods. 1996.
49. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Statistics in Medicine 1989; 8(4): 441-454.
50. Cook DJ, Laine LA, Guyatt GH, Raffin TA. Nosocomial pneumonia and the role of gastric pH. A meta-analysis. Chest 1991; 100(1): 7-13.
51. Cowley DE. Prostheses for primary total hip replacement. A critical appraisal of the literature. International journal of technology assessment in health care 1995; 11(4): 770-778.
52. Critical Appraisal Skills Programme: 12 questions to help you make sense of a cohort study. Oxford, 1999.
53. Dardennes R, Even C, Bange F, Heim A. Comparison of carbamazepine and lithium in the prophylaxis of bipolar disorders. A meta-analysis. British Journal of Psychiatry 1995; 166(3): 378-381.
54. De Vet HCW, de Bie RA, van der Heijden GJ, Verhagen AP, Sijpkes P. Systematic reviews on the basis of methodological criteria. Physiotherapy 1997; 83(6): 284-289.
55. De Vet HCW, Weijden T.van der, Muris JW, Heyrman J, Buntinx F, Knottnerus JA. Systematic reviews of diagnostic research. Considerations about assessment and incorporation of methodological quality. European journal of epidemiology 2001; 17(4): 301-306.
56. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG. Evaluating non-randomised intervention studies. Health Technology Assessment 2003; 7(27): 1-173.
57. Delfini Group. Short critical appraisal checklist: interventions for prevention, screening & therapy. www.delfini.org/Delfini_Tool_StudyValidity_Short.pdf (26.06.2009).
58. Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions. The TREND statement. American Journal of Public Health 2004; 94(3): 361-366.
59. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbe KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. Journal of clinical epidemiology 1992; 45(3): 255-265.
60. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. Journal of Epidemiology & Community Health 1998; 52(6): 377-384.
61. Drummond M, McGuire A. Economic evaluation in health care. Merging theory with practice. 2001. New York, Oxford Univ. Press.
62. Drummond M, Sculpher M. Common methodological flaws in economic evaluations. Medical care 2005; 43(7 Suppl.): 5-14.
63. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. British Medical Journal 1996; 313(7052): 275-283.

64. Drummond MF, Sculpher MJ, Torrance GW, O'Brien B, Stoddart GL: Methods for the economic evaluation of health care programmes, 3rd ed. New York, 2005.
65. DuRant RH. Checklist for the evaluation of research articles. Journal of adolescent health 1994; 15(1): 4-8.
66. Earle C, Hebert PC. A reader's guide to the evaluation of screening studies. Postgraduate Medical Journal 1996; 72(844): 77-83.
67. ECHTA. Best practice in undertaking and reporting HTA. 2001.
68. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technology Assessment 2003; 7(1): 1-76.
69. Ekkernkamp M, Lühmann D, Raspe H. Methodenmanual für „HTA-Schnellverfahren“ und exemplarisches „Kurz-HTA“. Die Rolle der quantitativen Ultraschallverfahren zur Ermittlung des Risikos für osteoporotische Frakturen. 2003. Sankt Augustin, Asgard-Verlag. Health Technology Assessment/Schriftenreihe des Deutsches Institus für Medizinische Dokumentation und Information, Bd. 34.
70. Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. International journal of technology assessment in health care 2005; 21(2): 240-245.
71. Fernández-de-las-Peñas C, Alonso-Blanco C, San-Roman J, Miangolarra-Page JC. Methodological quality of randomized controlled trials of spinal manipulation and mobilization in tension-type headache, migraine, and cervicogenic headache. The Journal of orthopaedic and sports physical therapy 2006; 36(3): 160-169.
72. Forsetlund L, Reinar L. Quality of reporting and of methodology of studies on interventions for trophic ulcers in leprosy. A systematic review. Indian Journal of Dermatology, Venereology and Leprology 2008; 331-337.
73. Fowkes FG, Fulton PM. Critical appraisal of published research. Introductory guidelines. British Medical Journal 1991; 302(6785): 1136-1140.
74. French Health Economists Association: French Guidelines for the Economic Evaluation of Health Care Technologies. 2004.
75. Geng V. Literatur: Qualitäts- und Beurteilungskriterien für Literatur. Krankenhaushygiene und Infektionsverhütung 2007; 29(1): 19-21.
76. Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at Annals of Internal Medicine. Annals of internal medicine 1994; 121(1): 11-21.
77. Graf J, Doig GS, Cook DJ, Vincent JL, Sibbald WJ. Randomized, controlled clinical trials in sepsis. Has methodological quality improved over time? Critical care medicine 2002; 30(2): 461-472.
78. Greenland S. Invited commentary. A critical look at some popular meta-analytic methods. American Journal of Epidemiology 1994; 140(3): 290-296.
79. Gupta AK, Chaudhry MM. Evaluating the quality of rosacea studies. Implications for the patient and physician. Dermatology 2003; 207(2): 173-177.
80. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ. GRADE. An emerging consensus on rating quality of evidence and strength of recommendations. British Medical Journal 2008; 336(7650): 924-926.
81. Haahr MT, Hrobjartsson A. Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors. Clinical trials 2006; 3(4): 360-365.
82. Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. Journal of clinical epidemiology 1996; 49(7): 749-754.
83. Hammerschlag R, Morris MM. Clinical trials comparing acupuncture with biomedical standard care. A criteria-based evaluation of research design and reporting. Complementary Therapies in Medicine 1997; 5(3): 133-140.
84. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. British Medical Journal 2001; 323(7308): 334-336.
85. Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Krebs SJ, Klassen TP. Risk of bias versus quality assessment of randomised controlled trials. Cross sectional study. British Medical Journal 2009; 339(191): b4012.

86. Health Care Insurance Board. Guidelines for pharmacoeconomic research, updated version. 2006.
87. Health Technology Board for Scotland. Guidance to Manufacturers for Completion of New Product Assessment Form (NPAF). 2007.
88. Heneghan AM, Horwitz SM, Leventhal JM. Evaluating intensive family preservation programs. A methodological review. Pediatrics 1996; 97(4): 535-542.
89. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. Journal of clinical epidemiology 2006; 59(12): 1249-1256.
90. Hettinga DM, Hurley DA, Jackson A, May S, Mercer C, Roberts L. Assessing the effect of sample size, methodological quality and statistical rigour on outcomes of randomised controlled trials on mobilisation, manipulation and massage for low back pain of at least 6 weeks duration. Physiotherapy 2008; 94(2): 97-104.
91. Heyland DK, Cook DJ, King D, Kernerman P, Brun-Buisson C. Maximizing oxygen delivery in critically ill patients. A methodologic appraisal of the evidence. Critical care medicine 1996; 24(3): 517-524.
92. Higgins JPT, Green S, The Cochrane Collaboration. Cochrane handbook for systematic reviews of interventions. 2008; Version 5.0.2, updated September 2008. Chichester, Wiley. Cochrane book series.
93. Hill CL, La Valley MP, Felson DT. Secular changes in the quality of published randomized clinical trials in rheumatology. Arthritis and Rheumatism 2002; 46(3): 779-784.
94. Hobbs FD, Delaney BC, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH, Earl-Slater AS, Jowett S, Tobias RS. A review of near patient testing in primary care. Health Technology Assessment 1997; 1(5): i-229.
95. Hoffman RM, Kent DL, Deyo RA. Diagnostic accuracy and clinical utility of thermography for lumbar radiculopathy. A meta-analysis. Spine 1991; 16(6): 623-628.
96. Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, Comstock B, Jarvik JG. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. Academic radiology 2006; 13(7): 803-810.
97. Horwitz RI, Viscoli CM, Clemens JD, Sadock RT. Developing improved observational methods for evaluating therapeutic effectiveness. American Journal of Medicine 1990; 89(5): 630-638.
98. Huebner RH, Park KC, Shepherd JE, Schwimmer J, Czernin J, Phelps ME, Gambhir SS. A meta-analysis of the literature for whole-body FDG PET detection of recurrent colorectal cancer. Journal of Nuclear Medicine 2000; 41(7): 1177-1189.
99. Huwiler-Müntener K, Jüni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. JAMA 2002; 287(21): 2801-2804.
100. Imperiale TF, McCullough AJ. Do corticosteroids reduce mortality from alcoholic hepatitis? A meta-analysis of the randomized trials. Annals of internal medicine 1990; 113(4): 299-307.
101. Institut für Pharmaökonomische Forschung. Guidelines zur gesundheitsökonomischen Evaluation. 2006.
102. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen: Fixe Kombinationen aus Kortikosteroiden und lang wirksamen Beta-2-Rezeptoragonisten zur inhalativen Anwendung bei Patienten mit Asthma bronchiale – Ergänzungsauftrag. Abschlussbericht A07-01. Köln, 2008.
103. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Früherkennungsuntersuchung von Sehstörungen bei Kindern bis zur Vollendung des 6. Lebensjahres: Abschlussbericht. 2008. Köln, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. IQWiG-Berichte, Nr. 32.
104. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. Annals of internal medicine 1994; 120(8): 667-676.
105. Iskedjian M, Trakas K, Bradley CA, Addis A, Lanctôt K, Kruk D, Ilersich AL, Einarson TR, Quality assessment of economic evaluations published in ParmacoEconomics the first four years (1992 to 1995). PharmacoEconomics 1997; 685-694.
106. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ. Assessing the quality of reports of randomized clinical trials. Is blinding necessary? Controlled clinical trials 1996; 17(1): 1-12.

107. Jonas WB, Linde K. Conducting and evaluating clinical research on complementary and alternative medicine. In: Gallin JI and Ognibene FP (Eds). Principles and practice of clinical research. San Diego, 2002, 401-426.
108. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association 1999; 282(11): 1054-1060.
109. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC medical research methodology 2004; 4(22).
110. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests. European journal of obstetrics & gynecology and reproductive biology 2001; 95(1): 6-11.
111. Khan KS, Ter Riet G, Glanville J, Sowden AJ, Kleijnen J. Undertaking systematic reviews of research on effectiveness. CRD's guidance for carrying out or commissioning reviews. University of York, NHS Centre for Reviews and Dissemination. 2000. York, England.
112. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Annals of internal medicine 2001; 135(11): 982-989.
113. Kleijnen J, Knipschild P, Ter Riet G. Clinical trials of homoeopathy. British Medical Journal 1991; 302(6772): 316-323.
114. Kmet LM, Lee RC, Cook LS. Standard quality assessment criteria for evaluating primary research papers from a variety of fields (Brief record). 2004; 22. Edmonton, Alberta Heritage Foundation for Medical Research (AHFMR).
115. Koes BW, Assendelft WJ, van der Heijden GJ, Bouter LM, Knipschild PG. Spinal manipulation and mobilisation for back and neck pain. A blinded review. British Medical Journal 1991; 303(6813): 1298-1303.
116. Kunz R, Burnand B, Schunemann HJ. The GRADE System. An international approach to standardize the graduation of evidence and recommendations in guidelines. Internist 2008; 49(6): 673-680.
117. Kunz R, Khan KS, Kleijnen J, Antes G: Systematische Übersichtsarbeiten und Meta-Analysen. Einführung in Instrumente der evidenzbasierten Medizin für Ärzte, klinische Forscher und Experten im Gesundheitswesen., 2. ed. Bern, 2009.
118. Kunz R, Oxman AD. The unpredictability paradox. Review of empirical comparisons of randomised and non-randomised clinical trials. British Medical Journal 1998; 317(7167): 1185-1190.
119. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. Cochrane database of systematic reviews 2007; (2): MR000012.
120. Kwakkel G, Wagenaar RC, Koelman TW, Lankhorst GJ, Koetsier JC. Effects of intensity of rehabilitation after stroke. A research synthesis. Stroke: a journal of cerebral circulation 1997; 28(8): 1550-1556.
121. Lamont RF, Khan KS, Beattie B, Cabero RL, Di Renzo GC, Dudenhausen JW, Helmer H, Svare J, van Geijn HP. The quality of nifedipine studies used to assess tocolytic efficacy: a systematic review. Journal of perinatal medicine 2005; 33(4): 287-295.
122. Larsen RJ, Asmussen M, Christensen T, Olsen J, Poulsen PB, Sorensen J. Economic evaluations in international health technology assessments. A study of methodologies. 2003; 5. Danish Centre for Evaluation and Health Technology Assessment. Danish Health Technology Assessment, Vol. 1.
123. Levine J. Trial assessment procedure scale (TAPS). In: Spilker B (Ed). Guide to clinical trials. New York, 1991, 780-786.
124. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions. Explanation and elaboration. PLoS Medicine 2009; 6(7): 1-28.
125. Liberati A, Himmel HN, Chalmers TC. A quality assessment of randomized control trials of primary treatment of breast cancer. Journal of clinical oncology 1986; 4(6): 942-951.
126. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11): 1061-1066.

127. Linde K, Scholz M, Melchart D, Willich SN. Should systematic reviews include non-randomized and uncontrolled studies? The case of acupuncture for chronic headache. Journal of clinical epidemiology 2002; 55(1): 77-85.
128. Lipscomb J, Weinstein MC, Torrance GW. Time preference. In: Gold MR, Siegel JE, Russell LB et al. (Eds). Cost effectiveness in health and medicine. New York, 1996, pp 214-235.
129. Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, Khoury MJ, Cohen B, Davey-Smith G. STrengthening the REporting of Genetic Association Studies (STREGA). An extension of the STROBE statement. PLoS Medicine 2009; 6(2): 151-163.
130. Loevinsohn BP. Health education interventions in developing countries. A methodological review of published articles. International journal of epidemiology 1990; 19(4): 788-794.
131. Lohr KN. Rating the strength of scientific evidence. Relevance for quality improvement programs. International Journal for Quality in Health Care 2004; 16(1): 9-18.
132. Lohr KN, Carey TS. Assessing „best evidence“. Issues in grading the quality of studies for systematic reviews. The Joint Commission Journal on Quality Improvement 1999; 25(9): 470-479.
133. Ludwig Boltzmann Institut HTA. (Internes) Manual – Abläufe und Methoden, Tl. 2. 2007. Wien, Ludwig Bolzmann Institut, Health Technology Assessment. HTA-Projektbericht, Nr. 006.
134. MacLehose RR, Reeves BC, Harvey IM, Sheldon TA, Russell IT, Black AM. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technology Assessment 2000; 4(34): 1-154.
135. MacMillan HL, MacMillan JH, Offord DR, Griffith L, MacMillan A. Primary prevention of child physical abuse and neglect: a critical review. Part I. Journal of Child Psychology and Psychiatry 1994; 35(5): 835-856.
136. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. Physical therapy 2003; 83(8): 713-721.
137. Mallen C, Peat G, Croft P. Quality assessment of observational studies is not commonplace in systematic reviews. Journal of clinical epidemiology 2006; 59(8): 765-769.
138. Mandelblatt JS, Fryback DG, Weinstein MC, Russel LB, Gold MR, Hadorn DC. Assessing the effectiveness of health interventions. In: Gold MR, Siegel JE, Russell LB et al. (Eds). Cost effectiveness in health and medicine. New York, 1996, 135-175.
139. Maziak DE, Meade MO, Todd TR. The timing of tracheotomy. A systematic review. Chest 1998; 114(2): 605-609.
140. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet 1999; 354(9193): 1896-1900.
141. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP. Assessing the quality of reports of randomised trials. Implications for the conduct of meta-analyses. Health Technology Assessment 1999; 3(12): 1-98.
142. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869.
143. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials. An annotated bibliography of scales and checklists. Controlled clinical trials 1995; 16(1): 62-73.
144. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. International journal of technology assessment in health care 1996; 12(2): 195-208.
145. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998; 352(9128): 609-613.
146. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR, Liberati A. What contributions do languages other than English make on the results of meta-analyses? Journal of clinical epidemiology 2000; 53(9): 964-972.
147. Moher D, Schulz KF, Altman DG. The CONSORT statement. Revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 2001; 357(9263): 1191-1194.

148. Moncrieff J, Churchill R, Colin DD, McGuire H. Development of a quality assessment instrument for trials of treatments for depression and neurosis. International Journal of Methods in Psychiatric Research 2001; 10(3): 126-133.
149. Moncrieff J, Drummond DC. The quality of alcohol treatment research: an examination of influential controlled trials and development of a quality rating system. Addiction 1998; 93(6): 811-823.
150. Morley JA, Finney JW, Monahan SC, Floyd AS. Alcoholism treatment outcome studies, 1980-1992: methodological characteristics and quality. Addictive Behaviors 1996; 21(4): 429-443.
151. Moseley AM, Herbert RD, Sherrington C, Maher CG. Evidence for physiotherapy practice: A survey of the Physiotherapy Evidence Database (PEDro). Australian Journal of Physiotherapy 2002; 48(1): 43-49.
152. Moseley AM, Maher C, Herbert RD, Sherrington C. Reliability of a Scale for Measuring the Methodological Quality of Clinical Trials. Rome 1999 PA30 1999.
153. Moyer A, Finney JW. Rating methodological quality. Toward improved assessment and investigation. Accountability in research 2005; 12(4): 299-313.
154. Mullins MD, Becker DM, Hagspiel KD, Philbrick JT. The role of spiral volumetric computed tomography in the diagnosis of pulmonary embolism. Archives of Internal Medicine 2000; 160(3): 293-298.
155. Mulrow CD, Linn WD, Gaul MK, Pugh JA. Assessing quality of a diagnostic test evaluation. Journal of General Internal Medicine 1989; 4(4): 288-295.
156. National Health and Medical Research Council (NHMRC). How to review the evidence. Systematic identification and review of the scientific literature. NHMRC. 2000. Canberra, Australia.
157. National Health Service Public Health Resource Unit. Critical Appraisal Skills Programme: making sense of evidence. 12 questions to help you make sense of a cohort study www.phru.nhs.uk/Pages/PHD/resources.htm (26.06.2009).
158. National Health Service Public Health Resource Unit. Critical Appraisal Skills Programme: making sense of evidence. 10 questions to help you make sense of randomised controlled studies www.phru.nhs.uk/Pages/PHD/resources.htm (26.06.2009).
159. National Health Service Public Health Resource Unit. Critical Appraisal Skills Programme: making sense of evidence. 10 questions to help you make sense of reviews www.phru.nhs.uk/Pages/PHD/resources.htm (26.06.2009).
160. National Health Service Public Health Resource Unit. Critical Appraisal Skills Programme: making sense of evidence. 11 questions to help you make sense of a case control study (26.06.2009).
161. National Health Service Public Health Resource Unit. Critical Appraisal Skills Programme: making sense of evidence. 12 questions to help you make sense of a diagnostic test study www.phru.nhs.uk/Pages/PHD/resources.htm (26.06.2009).
162. Nguyen QV, Bezemer PD, Habets L, Prahl-Andersen B. A systematic review of the relationship between overjet size and traumatic dental injuries. European Journal of Orthodontics 1999; 21(5): 503-515.
163. National Health Service Public Health Resource Unit. Critical Appraisal Skills Programme: making sense of evidence. 10 questions to help you make sense of economic evaluations (26.06.2009).
164. Nüesch E, Trelle S, Reichenbach S, Rutjes AWS, Bürgi E, Scherer M, Altman DG, Jüni P. Ausschluss von Patienten aus der Analyse. Auswirkungen auf die Ergebnisse randomisierter kontrollierter Studien. Eine meta-epidemiologische Studie. Deutsches Ärzteblatt 2009; 106(39): A1893-A1898.
165. Ogilvie D, Egan M, Hamilton V, Petticrew M. Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go? Journal of Epidemiology & Community Health 2005; 59(10): 886-892.
166. Ogilvie D, Hamilton V, Egan M, Petticrew M. Systematic reviews of health effects of social interventions: 1. Finding the evidence: how far should you go? Journal of Epidemiology & Community Health 2005; 59(9): 804-808.
167. Ogilvie-Harris DJ, Gilbart M. Treatment modalities for soft tissue injuries of the ankle. A critical review. Clinical Journal of Sport Medicine 1995; 5(3): 175-186.
168. Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials. A systematic review. Physical therapy 2008; 88(2): 156-175.

169. Onghena P, Van Houdenhove B. Antidepressant-induced analgesia in chronic non-malignant pain. A meta-analysis of 39 placebo-controlled studies. Pain 1992; 49(2): 205-219.
170. Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, Streiner DL. Agreement among reviewers of review articles. Journal of clinical epidemiology 1991; 44(1): 91-98.
171. Petrak F, Hardt J, Nickel R, Egle UT. Checkliste zur Bewertung der wissenschaftlichen Qualität kontrollierter psychotherapeutischer Interventionsstudien (CPI). Psychotherapeut 1999; 44(6): 390-393.
172. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.0). 2006.
173. Pharmaceutical Management Agency. Prescription for Pharmacoeconomic Analysis – Methods for Cost-Utility-Analysis. 2007.
174. Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. International journal of epidemiology 2007; 36(4): 847-857.
175. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. New England Journal of Medicine 1987; 317(7): 426-432.
176. Powe NR, Tielsch JM, Schein OD, Luthra R, Steinberg EP. Rigor of research methods in studies of the effectiveness and safety of cataract extraction with intraocular lens implantation. Cataract patient outcome research team. Archives of ophthalmology 1994; 112(2): 228-238.
177. Prendiville W, Elbourne D, Chalmers I. The effects of routine oxytocic administration in the management of the third stage of labour. An overview of the evidence from controlled trials. British Journal of Obstetrics and Gynaecology 1988; 95(1): 3-16.
178. Pua HL, Lerman J, Crawford MW, Wright JG. An evaluation of the quality of clinical trials in anesthesia. Anesthesiology 2001; 95(5): 1068-1073.
179. Ramos MLT, Ferraz MB, Sesso R. Critical appraisal of published economic evaluations of home care for the elderly. Archives of Gerontology and Geriatrics 2004; 39(3): 255-267.
180. Ramsberg J, Odeberg S, Engström A, Lundin D. Examining the quality of health economic analyses sub-mitted to the Pharmaceutical Benefits Board in Sweden. European Journal of Health Economics 2004; 49(4): 351-356.
181. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics 1989; 84(5): 815-827.
182. Ressing M, Blettner M, Klug SJ. Systematic literature reviews and meta-analyses. Part 6 of a series on evaluation of scientific publications. Deutsches Ärzteblatt International 2009; 106(27): 456-463.
183. Robeer GG, Brandsma JW, Vann Heuvel SP, Smit B, Oostendorp RAB, Wittens CHA. Exercise therapy for intermittent claudication. A review of the quality of randomised clinical trials and evaluation of predictive factors. European Journal of Vascular and Endovascular Surgery 1998; 15(1): 36-43.
184. Rochon PA, Gurwitz JH, Cheung CM, Hayes JA, Chalmers TC. Evaluating the quality of articles published in journal supplements compared with the quality of those published in the parent journal. JAMA 1994; 272(2): 108-113.
185. Rychetnik L, Frommer M. A schema for evaluating evidence on public health interventions (Version 4). Melbourne, 2002.
186. Rychetnik L, Frommer M. A schema for evaluating evidence on public health interventions (Version 4). Melbourne, 2002.
187. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. A basic science for clinical medicine, 2nd ed. Boston, 1991.
188. Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: an update. The Mount Sinai journal of medicine 1996; 63(3-4): 216-224.
189. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology. A systematic review and annotated bibliography. International journal of epidemiology 2007; 36(3): 666-676.
190. Saunders LD, Soomro GM, Buckingham J, Jamtvedt G, Raina P. Assessing the methodological quality of nonrandomized intervention studies. Western journal of nursing research 2003; 25(2): 223-237.

191. Schöffski O, von der Schulenburg JM. Gesundheitsökonomische Evaluationen. 2008. Heidelberg, Springer-Verlag.
192. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994; 272(2): 125-128.
193. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273(5): 408-412.
194. Scottish Intercollegiate Guidelines Network (SIGN). SIGN 50 – A guideline developer's handbook. 2008.
195. Scottish Intercollegiate Guidelines Network (SIGN). SIGN 50 – A guideline developer's handbook. 2008.
196. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC medical research methodology 2007; 7(10).
197. Siebert U, Behrend C, Mühlberger N, Wasem J, Greiner W, von der Schulenburg JM, Welte R, Leidl R. Entwicklung eines Kriterienkataloges zur Beschreibung und Bewertung ökonomischer Evaluationsstudien in Deutschland. In: Leidl R, von der Schulenburg JM, and Wasem J (Eds). Ansätze und Methoden der ökonomischen Evaluation – eine internationale Perspektive. Baden-Baden, 1999, 156-170.
198. Sindhu F, Carpenter L, Seers K. Development of a tool to rate the quality assessment of randomized controlled trials using a Delphi technique. Journal of advanced nursing 1997; 25(6): 1262-1268.
199. Slim K, Bousquet J, Kwiatkowski F, Pezet D, Chipponi J. Analysis of randomized controlled trials in laparoscopic surgery. British journal of surgery 1997; 84(5): 610-614.
200. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument. ANZ journal of surgery 2003; 73(9): 712-716.
201. Smeenk FW, van Haastregt JC, de Witte LP, Crebolder HF. Effectiveness of home care programmes for patients with incurable cancer on their quality of life and time spent in hospital. Systematic review. British Medical Journal 1998; 316(7149): 1939-1944.
202. Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation. A meta-analysis. American review of respiratory disease 1992; 145(3): 533-539.
203. Smith LA, Oldman AD, McQuay HJ, Moore RA. Teasing apart quality and validity in systematic reviews. An example from acupuncture trials in chronic neck and back pain. Pain 2000; 86(1-2): 119-132.
204. Spitzer WO, Lawrence V, Dales R, Hill G, Archer MC, Clark P, Abenhaim L, Hardy J, Sampalis J, Pinfold SP. Links between passive smoking and disease: a best-evidence synthesis. A report of the Working Group on Passive Smoking. Clinical and investigative medicine 1990; 13(1): 17-42.
205. Spooner CH, Pickard AS, Menon D. Edmonton Quality Assessment Tool for Drug Utilization Reviews. EQUATDUR-2 – The development of a scale to assess the methodological quality of a drugutilization review. Medical care 2000; 38(9): 948-958.
206. Staiger TO, Gaster B, Sullivan MD, Deyo RA. Systematic review of antidepressants in the treatment of chronic low back pain. Spine 2003; 28(22): 2540-2545.
207. Stelfox HT, Chua G, O'Rourke K, Detsky AS. Conflict of interest in the debate over calcium-channel antagonists. The New England Journal of Medicine 1998; 338(2): 101-106.
208. Stieb DM, Frayha HH, Oxman AD, Shannon HS, Hutchinson BG, Crombie FS. Effectiveness of Haemophilus influenzae type b vaccines. Canadian Medical Association journal 1990; 142(7): 719-733.
209. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. Meta-analysis of observational studies in epidemiology. A proposal for reporting. JAMA 2000; 283(15): 2008-2012.
210. Talley NJ, Owen BK, Boyce P, Paterson K. Psychological treatments for irritable bowel syndrome. A critique of controlled treatment trials. American Journal of Gastroenterology 1996; 91(2): 277-283.

211. ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria-based meta-analysis. Journal of clinical epidemiology 1990; 43(11): 1191-1199.
212. The Center of Expertise (KCE). The Draft Pharmacoeconomic Belgian Guidelines. 2008; KCE Reports (28).
213. Thomas BH, Ciliska D, Dobbins M, Micucci S. A process for systematically reviewing the literature. Providing the research evidence for public health nursing interventions. Worldviews on evidence-based nursing 2004; 1(3): 176-184.
214. Thomas H. Quality assessment tool for quantitative studies. Effective public health practice project. 2000. McMaster University, Toronto.
215. Torrance GW, Siegel JE, Luce BR. Framing and designing the cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB et al. (Eds). Cost effectiveness in health and medicine. New York, 1996, 54-81.
216. Ungar WJ, Santos MT. The pediatric quality appraisal questionnaire. An instrument for evaluation of the pediatric health economics literature. Value in Health 2003; 6(5): 584-594.
217. University of Oxford CfE-BM. Critical appraisal for therapy articles. www.cebm.net/index.aspx?o=1097 (26.06.2009).
218. University of Oxford CfE-BM. Diagnostic critical appraisal sheet. www.cebm.net/index.aspx?o=1096 (26.06.2009).
219. University of Oxford CfE-BM. Systematic review appraisal sheet. www.cebm.net/index.aspx?o=1567 (26.06.2009).
220. University of Birmingham. ARIF Critical Appraisal Checklist. www.arif.bham.ac.uk/critical-appraisal-checklist.shtml (26.06.2009).
221. van den Hoogen HM, Koes BW, van Eijk JT, Bouter LM. On the accuracy of history, physical examination, and erythrocyte sedimentation rate in diagnosing low back pain in general practice. A criteria-based review of the literature. Spine 1995; 20(3): 318-327.
222. van der Heijden GJ, van der Windt DA, Kleijnen J, Koes BW, Bouter LM. Steroid injections for shoulder disorders: a systematic review of randomized clinical trials. The British journal of general practice 1996; 46(406): 309-316.
223. van der Wurff P, Hagmeijer RH, Meyne W. Clinical tests of the sacroiliac joint. A systematic methodological review. Part 1: Reliability. Manual therapy 2000; 5(1): 30-36.
224. van Nieuwenhoven CV, Buskens E, Tiel FV, Bonten MJM. Relationship between methodological trial quality and the effects of selective digestive decontamination on pneumonia and mortality in critically ill patients. JAMA 2001; 286(3): 335-340.
225. Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Annals of internal medicine 2007; 147(8): W163-W194.
226. Varela-Lema L, Ruano-Ravina A. Development and use of a quality scale for assessing studies that analyze the diagnostic capacity of capsule endoscopy. Endoscopy 2006; 38(12): 1261-1267.
227. Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Bouter LM, Knipschild PG. The Delphi list. A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. Journal of clinical epidemiology 1998; 51(12): 1235-1241.
228. Verhagen AP, de Vet H, de Bie RA, Kessels AGH, Boers M, Knipschild PG. Balneotherapy and quality assessment. Interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. Journal of clinical epidemiology 1998; 51(4): 335-341.
229. Vickers A. Critical appraisal. How to read a clinical research paper. Complementary Therapies in Medicine 1995; 3(3): 158-166.
230. Vigna-Taglianti F, Vineis P, Liberati A, Faggiano F. Quality of systematic reviews used in guidelines for oncology practice. Annals of Oncology 2006; 17(4): 691-701.
231. von der Schulenburg JM, Greiner W, Jost F, Klusen Nea. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation – dritte und aktualisierte Fassung des Hannoveraner Konsens. Gesundheitsökonomie und Qualitätsmanagement 2007; 12: 285-290.

232. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. Guidelines for reporting observational studies. PLoS Medicine 2007; 4(10): 1628-1654.
233. Weintraub M. How to critically assess clinical drug trials. Drug therapy 1982; 12: 131-148.
234. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses (26.06.2009).
235. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, Lux L. Systems to rate the strength of scientific evidence. Evidence report – technology assessment (summary) 2002; (47): 1-11.
236. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC medical research methodology 2005; 5:(19): 1-9.
237. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technology Assessment 2004; 8(25): 1-234.
238. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS. A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC medical research methodology 2003; 3(25): 1-13.
239. Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. Journal of clinical epidemiology 2005; 58(1): 1-12.
240. Whiting PF, Weswood ME, Rutjes AWS, Reitsma JB, Bossuyt PNM, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC medical research methodology 2006; 6:9.
241. Wong WC, Cheung CS, Hart GJ. Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. Emerging themes in epidemiology 2008; 5(23).
242. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes. Meta-epidemiological study. British Medical Journal 2008; 336(7644): 601-605.
243. Yates SL, Morley S, Eccleston C, de C Williams AC. A scale for rating the quality of psychological trials for pain. Pain 2005; 117(3): 314-325.
244. Yuen SY, Pope JE. Learning from past mistakes. Assessing trial quality, power and eligibility in non-renal systemic lupus erythematosus randomized controlled trials. Rheumatology 2008; 47(9): 1367-1372.
245. Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, Sosin DM, Anderson L, Carande-Kulis VG, Teutsch SM, Pappaioanou M. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. American Journal of Preventive Medicine 2000; 18(1 Suppl): 44-74.
246. Zentner A, Velasco-Garrido M, Busse R. Methoden zur vergleichenden Bewertung pharmazeutischer Produkte. DIMDI. 2005. Köln, Deutsches Institut für Medizinische Dokumentation und Information (DIMDI). Schriftenreihe Health Technology Assessment, Bd. 13.
247. Zola P, Volpe T, Castelli G, Sismondi P, Nicolucci A, Parazzini F, Liberati A. Is the published literature a reliable guide for deciding between alternative treatments for patients with early cervical cancer? International Journal of Radiation Oncology 1989; 16(3): 785-797.

# 8 Anhang

## 8.1 Suchstrategie

**Tabelle 33: Suchstrategie**

| | No | Hits | Search Expression |
|---|---|---|---|
| C = | 1 | 68412788 | INAHTA; DAHTA; NHSEED; CDAR94; CDSR93; ME83; EM83; CB85; BA83; MK77; CCTR93; GA03; SM78; CV72; II98; ED93; AZ72; AR96; EA08; IS00; CC00; IN73; KR03; KL97; SP97; SPPP; TV01; DD83; IA70 |
| S = | 2 | 513507 | QUALITY/TI |
| | 3 | 70875 | QUALIT##T?/TI |
| | 4 | 9 | STUDIENQUALIT##T/TI |
| | 5 | 255 | QUALIT##TSKRITERIEN/TI |
| | 6 | 55 | QUALIT##TSBEWERTUNG##/TI |
| | 7 | 574871 | 2 OR 3 OR 4 OR 5 OR 6 |
| | 8 | 559222 | TRIAL#/TI |
| | 9 | 2517602 | STUDY/TI |
| | 10 | 947592 | STUDIE%/TI |
| | 11 | 56813 | META-ANALYS#S/TI |
| | 12 | 572497 | REVIEW#/TI |
| | 13 | 4558964 | 8 OR 9 OR 10 OR 11 OR 12 |
| | 14 | 4527790 | 13 NOT STUDENT% |
| | 15 | 12 | (EVALUATING # # # # # 14)/TI |
| | 16 | 579 | (EVALUATION # # # # # # 14)/TI |
| | 17 | 324 | BEURTEILUNG/TI AND 14 |
| | 18 | 347 | BEWERTUNG/TI AND 14 |
| | 19 | 1259 | 15 OR 16 OR 17 OR 18 |
| | 20 | 576083 | 7 OR 19 |
| | 21 | 87931 | (ASSESSMENT# AND QUALITY)/SAME SENT |
| | 22 | 107139 | (ASSESS## AND QUALITY)/SAME SENT |
| | 23 | 18967 | (ASSESSING AND QUALITY)/SAME SENT |
| | 24 | 67709 | (EVALUATION# AND QUALITY)/SAME SENT |
| | 25 | 93932 | (EVALUATE# AND QUALITY)/SAME SENT |
| | 26 | 14253 | (EVALUATING AND QUALITY)/SAME SENT |
| | 27 | 45935 | (MEASUREMENT# AND QUALITY)/SAME SENT |
| | 28 | 114578 | (MEASURE# AND QUALITY)/SAME SENT |
| | 29 | 14527 | (MEASURING AND QUALITY)/SAME SENT |
| | 30 | 282 | (BEWERTUNG AND QUALIT##T)/SAME SENT |
| | 31 | 45 | BEWERTUNGSINSTRUMENT? |
| | 32 | 80 | QUALIT##TSBEWERTUNG |
| | 33 | 13732 | (RATING# AND QUALITY)/SAME SENT |
| | 34 | 12246 | (GRADE# AND QUALITY)/SAME SENT |
| | 35 | 443 | (CRITICAL APPRAISAL# AND QUALITY)/SAME SENT |
| | 36 | 441486 | (21 OR 22 OR 23 OR 24 OR 25 OR 26 OR 27 OR 28 OR 29 OR 30 OR 31 OR 32 OR 33 OR 34 OR 35)/(TI;AB) |
| | 37 | 2992 | (CHECKLIST## AND QUALITY)/SAME SENT |
| | 38 | 48343 | (SCORE# AND QUALITY)/SAME SENT |
| | 39 | 45390 | (SCAL## AND QUALITY)/SAME SENT |
| | 40 | 27 | (SKAL## AND QUALIT##T)/SAME SENT |
| | 41 | 28933 | (TOOL# AND QUALITY)/SAME SENT |
| | 42 | 22183 | (INSTRUMENT# AND QUALITY)/SAME SENT |

**Tabelle 33: Suchstrategie – Fortsetzung**

| | | | |
|---|---|---|---|
| | 43 | 46326 | (CRITERI? AND QUALITY)/SAME SENT |
| | 44 | 43 | BEWERTUNGSINSTRUMENT## |
| | 45 | 168277 | (37 OR 38 OR 39 OR 40 OR 41 OR 42 OR 43 OR 44)/(TI;AB) |
| | 46 | 37939 | 20 AND 36 AND 45 |
| | 47 | 3442 | IT=QUALITY CONTROL |
| | 48 | 4796 | UT=QUALITY CONTROL |
| | 49 | 209158 | CT=REPRODUCIBILITY OF RESULTS |
| | 50 | 762 | IT=REPRODUCIBILITY OF RESULTS |
| | 51 | 117 | UT=REPRODUCIBILITY OF RESULTS |
| | 52 | 59918 | CT=QUALITY ASSURANCE |
| | 53 | 2846 | IT=QUALITY ASSURANCE |
| | 54 | 5479 | UT=QUALITY ASSURANCE |
| | 55 | 5715 | CT=QUALITY INDICATORS |
| | 56 | 507 | IT=QUALITY INDICATORS |
| | 57 | 236 | UT=QUALITY INDICATORS |
| | 58 | 39373 | CT=STANDARDS |
| | 59 | 0 | CTG=STANDARDS |
| | 60 | 0 | ITG=STANDARDS |
| | 61 | 11 | UTG=STANDARDS |
| | 62 | 2384 | IT=STANDARDS |
| | 63 | 4508 | UT=STANDARDS |
| | 64 | 37751 | CT=QUALITY ASSESSMENT, HEALTH CARE |
| | 65 | 1 | IT=QUALITY ASSESSMENT, HEALTH CARE |
| | 66 | 0 | UT=QUALITY ASSESSMENT, HEALTH CARE |
| | 67 | 51338 | CT=DATA QUALITY |
| | 68 | 683 | IT=DATA QUALITY |
| | 69 | 783 | UT=DATA QUALITY |
| | 70 | 12480 | CT=TECHNOLOGY ASSESSMENT, BIOMEDICAL |
| | 71 | 7 | IT=TECHNOLOGY ASSESSMENT, BIOMEDICAL |
| | 72 | 2 | UT=TECHNOLOGY ASSESSMENT, BIOMEDICAL |
| | 73 | 12481 | CT=BIOMEDICAL TECHNOLOGY ASSESSMENT |
| | 74 | 7 | IT=BIOMEDICAL TECHNOLOGY ASSESSMENT |
| | 75 | 7 | UT=BIOMEDICAL TECHNOLOGY ASSESSMENT |
| | 76 | 52012 | CT=EPIDEMIOLOGIC STUDY CHARACTERISTICS |
| | 77 | 4 | IT=EPIDEMIOLOGIC STUDY CHARACTERISTICS |
| | 78 | 0 | UT=EPIDEMIOLOGIC STUDY CHARACTERISTICS |
| | 79 | 70253 | CT=EPIDEMIOLOGIC METHODS |
| | 80 | 1209 | IT=EPIDEMIOLOGIC METHODS |
| | 81 | 577 | UT=EPIDEMIOLOGIC METHODS |
| | 82 | 79126 | CT=EVALUATION METHODOLOGY |
| | 83 | 387 | IT=EVALUATION METHODOLOGY |
| | 84 | 360 | UT=EVALUATION METHODOLOGY |
| | 85 | 38749 | CT=METHODOLOGICAL STUDIES |
| | 86 | 1252 | IT=METHODOLOGICAL STUDIES |
| | 87 | 1247 | UT=METHODOLOGICAL STUDIES |
| | 88 | 38749 | CT=METHODOLOGICAL STUDY |
| | 89 | 16 | IT=METHODOLOGICAL STUDY |
| | 90 | 12 | UT=METHODOLOGICAL STUDY |
| | 91 | 63069 | CT=METHODOLOGY, RESEARCH |

**Tabelle 33: Suchstrategie – Fortsetzung**

| | | | |
|---|---|---|---|
| | 92 | 0 | IT=METHODOLOGY, RESEARCH |
| | 93 | 2 | UT=METHODOLOGY, RESEARCH |
| | 94 | 141240 | CT=METHODS |
| | 95 | 2054 | IT=METHODS |
| | 96 | 57564 | UT=METHODS |
| | 97 | 427808 | CT=METHODOLOGY |
| | 98 | 3805 | IT=METHODOLOGY |
| | 99 | 68529 | UT=METHODOLOGY |
| | 100 | 7192 | CT=VALIDITY |
| | 101 | 4691 | IT=VALIDITY |
| | 102 | 27548 | UT=VALIDITY |
| | 103 | 0 | CTG=VALIDIT##T |
| | 104 | 0 | ITG=VALIDIT##T |
| | 105 | 68 | UTG=VALIDIT##T |
| | 106 | 36037 | CTG=METHODE# |
| | 107 | 222 | ITG=METHODE# |
| | 108 | 827 | UTG=METHODE# |
| | 109 | 36040 | CTG=METHODIK |
| | 110 | 0 | ITG=METHODIK |
| | 111 | 14 | UTG=METHODIK |
| | 112 | 493638 | 47 OR 48 OR 49 OR 50 OR 51 OR 52 OR 53 OR 54 OR 55 OR 56 OR 57 OR 58 OR 59 OR 60 OR 61 OR 62 OR 63 OR 64 OR 65 OR 66 OR 67 OR 68 OR 69 OR 70 OR 71 OR 72 OR 73 OR 74 OR 75 OR 76 OR 77 OR 78 OR 79 OR 80 |
| | 113 | 857220 | 81 OR 82 OR 83 OR 84 OR 85 OR 86 OR 87 OR 88 OR 89 OR 90 OR 91 OR 92 OR 93 OR 94 OR 95 OR 96 OR 97 OR 98 OR 99 OR 100 OR 101 OR 102 OR 103 OR 104 OR 105 OR 106 OR 107 OR 108 OR 109 OR 110 OR 111 |
| | 114 | 1321066 | 112 OR 113 |
| | 115 | 6172 | 46 AND 114 |
| | 116 | 3740 | 115 NOT (QUALITY # LIFE) |
| | 117 | 3740 | 116 NOT QUALITY-OF-LIFE |
| | 118 | 3740 | 117 NOT LEBENSQUALIT##T? |
| | 119 | 3611 | 118 NOT (IMAGE QUALITY) |
| | 120 | 3611 | 119 NOT (DISPLAY QUALITY) |
| | 121 | 3611 | 120 NOT (PRINTING QUALITY) |
| | 122 | 3536 | 121 NOT (SERVICE QUALITY) |
| | 123 | 3394 | 122 NOT (DATA QUALITY) |
| | 124 | 3383 | 123 NOT (WEBSITE QUALITY) |
| | 125 | 3366 | 124 NOT (AIR QUALITY) |
| | 126 | 3292 | 125 NOT (WATER QUALITY) |
| | 127 | 3290 | 126 NOT (DRINKING WATER) |
| | 128 | 3254 | 127 NOT POLLUTION |
| | 129 | 3246 | 128 NOT (PROTEIN QUALITY) |
| | 130 | 3236 | 129 NOT (QUALITY # # # # PROTEIN) |
| | 131 | 3235 | 130 NOT (BONE QUALITY) |
| | 132 | 3206 | 131 NOT (SLEEP### QUALITY) |
| | 133 | 3206 | 132 NOT (QUALITY # SLEEP###) |
| | 134 | 3150 | 133 NOT (QUALITY # # # # SERVICE) |
| | 135 | 3150 | 134 NOT (SERVICE QUALITY) |
| | 136 | 3148 | 135 NOT (VIDEO QUALITY) |

**Tabelle 33: Suchstrategie – Fortsetzung**

| | | | |
|---|---|---|---|
| | 137 | 3141 | 136 NOT TELEMEDICINE |
| | 138 | 3115 | 137 NOT SPECTROMETRY |
| | 139 | 3101 | 138 NOT ULTRASONOGRAPHY |
| | 140 | 2954 | 139 NOT INTERNET |
| | 141 | 2954 | 140 NOT (QUALITY # WEB?) |
| | 142 | 2954 | 141 NOT (QUALITY # WEB-BASED) |
| | 143 | 2888 | 141 NOT (QUALITY # # # # INFORMATION#) |
| | 144 | 2829 | 143 NOT IMAGING |
| | 145 | 4 | 144 NOT (QUALITY # # # # END # LIFE) |
| | 146 | 2806 | 144 NOT (NURSING HOME QUALITY) |
| | 147 | 1685 | 146 NOT (QUALITY # # CARE) |
| | 148 | 1540 | 147 NOT (CARE QUALITY) |
| | 149 | 1540 | 148 NOT (HEALTHCARE QUALITY) |
| | 150 | 1540 | 149 NOT (HOSPITAL QUALITY) |
| | 151 | 1538 | 150 NOT (QUALITY # HOSPITAL) |
| | 152 | 1537 | 151 NOT DOGS |
| | 153 | 1530 | 152 NOT VETERINARY |
| | 154 | 1147 | 153 NOT (QUALITY ASSURANCE) |
| | 155 | 1101 | 154 NOT (QUALITY IMPROVEMENT) |
| | 156 | 1033 | 155 NOT HOSPITAL |
| | 157 | 896 | 156 NOT (QUALITY CONTROL)/TI |
| | 158 | 896 | 157 NOT DT=REPORT |
| | 159 | 892 | 158 NOT DT=LETTER |
| | 160 | 892 | 159 NOT DT=COMMENT |
| | 161 | 892 | 160 NOT DT=CASE REPORT |
| | 162 | 890 | 161 NOT DT=EDITORIAL |
| | 163 | 890 | 162 NOT DT=POSTER |
| | 164 | 855 | 163 NOT DT=CONFERENCE PAPER |
| | 165 | 782 | 164 AND LA=(GERM OR ENGL) |
| | 166 | 762 | 165 AND PY>=1988 |
| | 167 | 565 | check duplicates: unique in s=166 |

## 8.2 Gesichtete Internetseiten

**Tabelle 34: Gesichtete Internetseiten**

| Kürzel | Name | URL |
|---|---|---|
| AATM | Agència d'Avaluació de Tecnologia Mèdica | www.gencat.cat/salut/depsan/units/aatrm |
| AEETS | Asociación Española de Evaluación de Tecnologías Sanitarias | www.aeets.org |
| AETMIS | Agence d'évaluation des technologies et des modes d'intervention en santé | www.aetmis.gouv.qc.ca |
| AETSA | Agencia de Evaluación de Tecnologías Sanitarias de Andalucía | www.juntadeandalucia.es/salud/aetsa |
| AHRQ | The Agency for Healthcare Research and Quality | www.ahrq.gov |
| AHTA | Adelaide Health Technology Assessment | www.adelaide.edu.au/ahta |
| ANAES | Agence Nationale d'Accréditation et d'Evaluation en Santé | www.anae.fr |
| ANZHSN | Australia and New Zealand Horizon Scanning Network | www.horizonscanning.gov.au |
| ARIF | University of Birmingham | www.arif.bham.ac.uk/critical-appraisal-checklist.shtml |

**Tabelle 34: Gesichtete Internetseiten – Fortsetzung**

| | | |
|---|---|---|
| ASERNIP-S | Australian Safety and Efficacy Register of New Interventional Procedures – Surgical | www.surgeons.org/asernip-s |
| Avalia-t | Galician Agency for Health Technology Assessment | avalia-t.sergas.es |
| CAHTA | Catalan Agency for Health Technology Assessment and Research | www.gencat.cat/salut/depsan/units/aatrm/html/en/Du8/index.html |
| CADTH | The Canadian Agency for Drugs and Technologies in Health | www.cadth.ca |
| CCOHTA | Canadian Coordinating Office for Health Technology Assessment | www.cadth.ca |
| CDC | Center for Disease Control and Prevention | www.cdc.gov |
| CEBM | Oxford Centre for Evidence-based Medicine | www.cebm.net/index.aspx?o=1913 |
| CEBM | Centre for Evidence-Based Medicine, Toronto | www.cebm.utoronto.ca/teach/materials/caworksheets.htm |
| CEBMH | Oxford Centre for Evidence-based Mental Health | cebmh.warne.ox.ac.uk/cebmh/education_critical_appraisal.htm |
| CEDIT | Comité d'Evaluation et de Diffusion des Innovations Technologiques Assistance Publique Hôpitaux de Paris | cedit.aphp.fr |
| CEESTAHC | Central and Eastern European Society of Technology Assessment in Health Care | www.ceestahc.org/en |
| CIHR | Canadian Institutes of Health Research | www.cihr-irsc.gc.ca |
| CMT | Center for Medical Technology Assessment | www.cmt.liu.se |
| Cochrane | The Cochrane Collaboration | www.cochrane-handbook.org |
| CRD | The Centre for Reviews and Dissemination | www.york.ac.uk/inst/crd/systematic_reviews_book.htm |
| CTFPHC | Canadian Task Force on Preventive Health Care | www.ctfphc.org |
| CVZ | Health Care Insurance Board | www.cvz.nl |
| DACEHTA | Danish Centre for Health Technology Assessment | www.sst.dk/~/media/Planlaegning%20og%20kvalitet/MTV%20metode/HTA_Handbook_net_final.ashx |
| DCC | Dutch Cochrane Centre | www.cochrane.nl/nl/newPage1.html |
| Delfini | Delfini Group | www.delfini.org/Delfini_Tool_StudyValidity_Short.pdf |
| DIMDI | Deutsches Institut für Medizinische Dokumentation und Information | www.dimdi.de/de/hta/index.htm |
| DSI | Danish Institute for Health Services Research and Development | www.dsi.dk/frz_about.htm |
| ECHTA | European Collaboration for Assessment of Health Interventions | www.oeaw.ac.at/ita/ebene4/e2-2b13.htm |
| EUnetHTA | European network for Health Technology Assessment | www.eunethta.net |
| EuroScan | International Information Network on New and Emerging Health Technologies | www.euroscan.bham.ac.uk |
| FINOHTA | Finnish Office for Health Care Technology Assessment | finohta.stakes.fi/EN/index.htm |
| G-BA | Gemeinsamer Bundesausschuss | www.g-ba.de |
| GOEG | Gesundheit Österreich GmbH | www.goeg.at |
| GR | Health Council of the Netherlands | www.gr.nl |
| HAS | La Haute Autorité de Santé | www.has-sante.fr |
| HTAi | Health Technology Assessment International | www.htai.org |
| HunHTA | Unit of Health Economics and Technology Assessment in Health Care | hecon.uni-corvinus.hu |
| ICTAHC | Israeli Center for Technology Assessment in Health Care, The Gertner Institute | www.gertnerinst.org.il/e/health_policy_e/technology |
| INAHTA | International Network of Agencies for Health Technology Assessment | www.inahta.org/HTA/Checklist |
| IQWiG | Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen | www.iqwig.de |
| ISCII | Healthcare Technology Evaluation Agency | www.isciii.es/htdocs/en/investigacion/Agencia_quees.jsp |

**Tabelle 34: Gesichtete Internetseiten – Fortsetzung**

| | | |
|---|---|---|
| ISTAHC | International Society of Technology Assessment in Health Care | www.istahc.org |
| KBV | Arbeitsgruppe HTA bei der Kassenärztlichen Bundesvereinigung | www.kbv.de/hta/199.html |
| KCE | Belgian Health Care Knowledge Centre | www.kce.fgov.be/Download.aspx?ID=873 |
| LBI | Ludwig Boltzmann Institut | hta.lbg.ac.at |
| MDS | Medizinischer Dienst der Spitzenverbände der Krankenkassen | www.mds-ev.org/index2.html |
| MSAC | Medical Services Advisory Committee | www.msac.gov.au |
| NETSCC | NIHR Evaluation, Trials and Studies Coordinating Centre | www.netscc.ac.uk |
| NHS QIS | National Health Service Quality Improvement Scotland | www.nhshealthquality.org |
| NHS PHRU | National Health Service Public Health Resource Unit | www.phru.nhs.uk/casp/critical_appraisal_tools.htm |
| NHSC | National Horizon Scanning Centre | www.pcpoh.bham.ac.uk/publichealth/horizon |
| NICE | National Institute for Health and Clinical Excellence | www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf |
| NOKC | Norwegian Knowledge Centre for the Health Services | www.nokc.no |
| NZHTA | New Zealand Health Technology Assessment | nzhta.chmeds.ac.nz |
| OHTAC | Ontario Health Technology Advisory Committee | www.health.gov.on.ca/english/providers/program/ohtac/ohtac_mn.html |
| OHRI | Ottawa Health Research Institute | www.ohri.ca/programs/clinical_epidemiology/oxford.htm |
| SIGN | Scottish Intercollegiate Guidelines Network | www.sign.ac.uk |
| SBU | The Swedish Council on Technology Assessment in Health Care | www.sbu.se |
| SMM | Senter for Medisinsk Metodevurdering | www.kunnskapssenteret.no |
| SNHTA | Swiss Network for Health Technology Assessment | www.snhta.ch |
| TAB | Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag | www.tab.fzk.de |
| TA-SWISS | Zentrum für Technologiefolgen-Abschätzung | www.ta-swiss.ch |
| TEC | Blue Cross/Blues Shield Association Technology Evaluation Center | www.bcbs.com/tec/index.html |
| TNO | TNO Prevention and Health | www.tno.nl |
| UETS | Unidad de Evaluacion de Tecnologias Sanitarias | www.madrid.org/cs |
| UVT | HTA-Unit der Universitätspoliklinik der Katholischen Universität Rom/Unità di Valutazione delle Tecnologie | www.policlinicogemelli.it/area/?s=206 |
| VATAP | VA Technology Assessment Programm | www.va.gov/vatap |
| WMHTAC | West Midlands Health Technology Assessment Collaboration | www.wmhtac.bham.ac.uk/evidence.shtml |

## 8.3 Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Effektivität)

**Tabelle 35: Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Effektivität)**

| Publikation | Ausschlussgrund |
|---|---|
| Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S, Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Controlled clinical trials 1995; 62-73. | Aktualisierte Version vorhanden |
| Abraham NS, Moayyedi P, Daniels B, Veldhuyzen Van Zanten SJO. Systematic review: The methodological quality of trials affects estimates of treatment efficacy in functional (non-ulcer) dyspepsia. Alimentary Pharmacology and Therapeutics 2004; 631-641. | Nutzung eines bestehenden Instruments |
| Al-Jader LN, Newcombe RG, Hayes S, Murray A, Layzell J, Harper PS. Developing a quality scoring system for epidemiological surveys of genetic disorders. Clinical Genetics 2002; 230-234. | Genetische Studien |
| Barbui C, Cipriani A, Malvini L, Tansella M. Validity of the impact factor of journals as a measure of randomized controlled trial quality. The Journal of clinical psychiatry 2006; 67(1): 37-40. | Kein Instrument dargestellt |
| Bérard A, Andreu N, Tétrault J, Niyonsenga T, Myhal D. Reliability of Chalmers' scale to assess quality in meta-analyses on pharmacological treatments for osteoporosis. Annals of epidemiology 2000; 10(8): 498-503. | Ermittlung der Testgüte eines bestehenden Instruments |
| Bereza BG, Machado M, Einarson TR. Assessing the reporting and scientific quality of meta-analyses of randomized controlled trials of treatments for anxiety disorders. The Annals of pharmacotherapy 2008; 42(10): 1402-1409. | Nutzung eines bestehenden Instruments |
| Berghmans T, Meert AP, Mascaux C, Paesmans M, Lafitte JJ, Sculier JP. Citation indexes do not reflect methodological quality in lung cancer randomised trials. Annals of Oncology 2003; 715-721. | Instrumente nicht ausreichend dargestellt |
| Bhogal SK, Teasell RW, Foley NC, Speechley MR. Quality of the stroke rehabilitation research. Topics in stroke rehabilitation 2003; 10(1): 8-28. | Nutzung eines bestehenden Instruments |
| Bhogal SK, Teasell RW, Foley NC, Speechley MR. The PEDro scale provides a more comprehensive measure of methodological quality than the Jadad scale in stroke rehabilitation literature. Journal of clinical epidemiology 2005; 58(7): 668-673 | Methodenbericht |
| Bothe AK, Davidow JH, Bramlett RE, Franic DM, Ingham RJ. Stuttering treatment research 1970-2005: II. Systematic review incorporating trial quality assessment of pharmacological approaches. American journal of speech-language pathology/ American Speech-Language-Hearing Association 2006; 15(4): 342-352. | Kein Instrument dargestellt |
| Bothe AK, Davidow JH, Bramlett RE, Ingham RJ. Stuttering treatment research 1970-2005: I. Systematic review incorporating trial quality assessment of behavioral, cognitive, and related approaches. American journal of speech-language pathology/American Speech-Language-Hearing Association 2006; 15(4): 321-341. | Kein Instrument dargestellt |
| Boulware LE, Daumit GL, Frick KD, Minkovitz CS, Lawrence RS, Powe NR, Frick KD. Quality of clinical reports on behavioral interventions for hypertension. Preventive Medicine 2002; 463-475. | Instrumente nicht ausreichend dargestellt |
| Brinkhaus B, Pach D, Lüdtke R, Willich SN. Who controls the placebo? Introducing a Placebo Quality Checklist for pharmacological trials. Contemporary clinical trials 2008; 29(2): 149-156. | Kein Instrument dargestellt |
| Brouwers MC, Johnston ME, Charette ML, Hanna SE, Jadad AR, Browman GP. Evaluating the role of quality assessment of primary studies in systematic reviews of cancer practice guidelines. BMC medical research methodology 2005. | Vergleich von drei Instrumenten |
| Chiou CF, Hay JW, Wallace JF, Bloom BS, Neumann PJ, Sullivan SD, Yu HT, Keeler EB, Henning JM, Ofman JJ. Development and validation of a grading system for the quality of cost-effectiveness studies. Medical care 2003; 41(1): 32-44. | Gesundheitsökonomie |
| Cho MK, Bero LA. The quality of drug studies published in symposium proceedings. Annals of internal medicine 1996; 124(5): 485-489. | Kein Instrument dargestellt |
| Clark HD, Wells GA, Huët C, McAlister FA, Salmi LR, Fergusson D, Laupacis A. Assessing the quality of randomized trials: reliability of the Jadad scale. Controlled clinical trials 1999; 20(5): 448-452 | Ermittlung der Testgüte eines bestehenden Instruments |
| Cook C, Cleland J, Huijbregts P. Creation and critique of studies of diagnostic accuracy: Use of the STARD and QUADAS methodological quality assessment tools. Journal of Manual and Manipulative Therapy 2007; 93-102. | Nutzung von bestehenden Instrumenten |

**Tabelle 35: Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Effektivität) – Fortsetzung**

| | |
|---|---|
| Crawford CC, Sparber AG, Jonas WB. A systematic review of the quality of research on hands-on and distance healing: clinical and laboratory studies. Alternative therapies in health and medicine 2003; 9(3 Suppl): A96-104. | Nutzung eines bestehenden Instruments |
| De Vito C, Manzoli L, Marzuillo C, Anastasi D, Boccia A, Villari P. A systematic review evaluating the potential for bias and the methodological quality of meta-analyses in vaccinology. Vaccine 2007; 25(52): 8794-8806. | Kein Instrument dargestellt |
| Delgado-Rodriguez M, Ruiz-Canela M, De Irala-Estevez J, Llorca J, Martinez-Gonzalez A. Participation of epidemiologists and/or biostatisticians and methodological quality of published controlled clinical trials. Journal of epidemiology and community health 2001; 569-572. | Berichtsqualität |
| Dhingra V, Chittock DR, Ronco JJ. Assessing methodological quality of clinical trials in sepsis: Searching for the right tool. Critical care medicine 2002; 30(N2): 487-488. | Kein Instrument dargestellt |
| Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B, Bonas S, Booth A, Jones D. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. Journal of health services research & policy 2007; 12(1): 42-47. | Kein Instrument dargestellt |
| Dulai SK, Slobogean BL, Beauchamp RD, Mulpuri K. A quality assessment of randomized clinical trials in pediatric orthopaedics. Journal of pediatric orthopedics 2007; 27(5): 573-581. | Kein Instrument dargestellt |
| Dunkelberg S. Wie gut ist eine qualitative Studie? 10 Hilfreiche Fragen für den Leser von Aufsätzen. Zeitschrift fur Allgemeinmedizin 2005; 248-251. | Qualitative Studien |
| Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled clinical trials 1990; 11(5): 339-352. | Nutzung eines bestehenden Instruments |
| Farrington DP. Methodological quality standards for evaluation research. Annals of the American academy of poiiitical and social science 2003; 587: 49-68. | Kriminalistik |
| Fenton JE, O'Connor A, Ullah I, Ahmed I, Shaikh M. Do citation classics in rhinology reflect utility rather than quality? Rhinology 2005; 43(3): 221-224. | Kein Instrument dargestellt |
| Fernández-de-las-Peñas C, onso-Blanco C, San-Roman J, Miangolarra-Page JC. Methodological quality of randomized controlled trials of spinal manipulation and mobilization in tension-type headache, migraine, and cervicogenic headache. The Journal of orthopaedic and sports physical therapy 2006; 36(3): 160-169. | Nicht lieferbar |
| Flores C, del MP-Y, Villar J. A quality assessment of genetic association studies supporting susceptibility and outcome in acute lung injury. Critical Care 2008. | Genetische Studien |
| Foley NC, Bhogal SK, Teasell RW, Bureau Y, Speechley MR. Estimates of quality and reliability with the physiotherapy evidence-based database scale to assess the methodology of randomized controlled trials of pharmacological and non-pharmacological interventions. Physical therapy 2006; 86(6): 817-824. | Nutzung eines bestehenden Instruments |
| Friedman DS, Bass EB, Lubomski LH, Fleisher LA, Kempen JH, Magaziner J, Sprintz M, Robinson K, Schein OD. The methodologic quality of clinical trials on regional anesthesia for cataract surgery. Ophthalmology 2001; 530-541. | Kein Instrument dargestellt |
| Gerkens S, Crott R, Cleemput I, Thissen JP, Closon MC, Horsmans Y, Beguin C. Comparison of three instruments assessing the quality of economic evaluations: a practical exercise on economic evaluations of the surgical treatment of obesity. International journal of technology assessment in health care 2008; 24(3): 318-325. | Gesundheitsökonomie |
| Greenfield ML, Rosenberg AL, O'Reilly M, Shanks AM, Sliwinski MJ, Nauss MD. The quality of randomized controlled trials in major anesthesiology journals. Anesthesia and analgesia 2005; 100(6): 1759-1764 | Kein Instrument dargestellt |
| Gupta AK, Chaudhry MM. Evaluating the quality of rosacea studies: implications for the patient and physician. Dermatology (Basel, Switzerland) 2003; 207(2): 173-177. | Nutzung eines bestehenden Instruments |
| Gupta AK, Nicol K, Johnson A. Pityriasis versicolor: quality of studies. The Journal of dermatological treatment 2004; 15(1): 40-45. | Nutzung eines bestehenden Instruments |
| Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. Journal of clinical epidemiology 2006; 59(12): 1249-1256. | Methodenbericht |
| Hollerwöger D. Methodological quality and outcomes of studies addressing manual cervical spine examinations: a review. Manual therapy 2006; 11(2): 93-98. | Nutzung eines bestehenden Instruments |

**Tabelle 35: Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Effektivität) – Fortsetzung**

| | |
|---|---|
| Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, Comstock B, Jarvik JG. Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A preliminary assessment. Academic radiology 2006; 13(7): 803-810. | Ermittlung der Testgüte eines bestehenden Instrumentes |
| Jannink MJ, van Dijk H, de Vries J, Groothoff JW, Lankhorst GJ. A systematic review of the methodological quality and extent to which evaluation studies measure the usability of orthopaedic shoes. Clinical rehabilitation 2004; 18(1): 15-26. | Nutzung eines bestehenden Instruments |
| Junhua Z, Hongcai S, Xiumei G, Boli Z, Yaozu X, Hongbo C, Ming R. Methodology and reporting quality of systematic review/meta-analysis of traditional Chinese medicine. Journal of Alternative and Complementary Medicine 2007; 797-805. | Nutzung eines bestehenden Instruments |
| Jüni P, Witschi A, Bloch R, Egger M, Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association 1999; 1054-1060. | Übersichtsarbeit |
| Justice LM, Nye C, Schwarz J, McGinty A, Rivera A. Methodological quality of intervention research in speech-language pathology: Analysis of 10 years of group-design studies. Evidence-Based Communication Assessment and Intervention 2008; 46-59. | Instrument entspricht bis auf zwei ausgelassenen Items dem von Downs & Black |
| Kelly KD, Travers A, Dorgan M, Slater L, Rowe BH. Evaluating the quality of systematic reviews in the emergency medicine literature. Annals of emergency medicine 2001; 518-526. | Nutzung eines bestehenden Instruments |
| Kjaergard LL, Nikolova D, Gluud C. Randomized clinical trials in hepatology: predictors of quality. Hepatology (Baltimore, Md.) 1999; 30(5): 1134-1138. | Nutzung eines bestehenden Instruments |
| Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Annals of internal medicine 2001; 135(11): 982-989. | Nutzung eines bestehenden Instruments |
| Lahtinen E, Koskinen-Ollonqvist P, Rouvinen-Wilenius P, Tuominen P, Mittelmark MB. The development of quality criteria for research: A Finnish approach. Health Promotion International 2005; 306-315. | Kein Instrument dargestellt |
| Lamont RF. A quality assessment tool to evaluate tocolytic studies. BJOG: An International Journal of Obstetrics and Gynaecology 2006; 96-99. | Instrumente nicht ausreichend dargestellt |
| Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. JAMA: the journal of the American Medical Association 2002; 287(21): 2805-2808. | Nutzung eines bestehenden Instruments |
| Lentine KL, Brennan DC. Statin use after renal transplantation: A systematic quality review of trial-based evidence. Nephrology Dialysis Transplantation 2004; 2378-2386. | Nutzung eines bestehenden Instruments |
| Linde K, Jonas WB, Melchart D, Willich S. The methodological quality of randomized controlled trials of homeopathy, herbal medicines and acupuncture. International journal of epidemiology 2001; 30(3): 526-531. | Nutzung eines bestehenden Instruments |
| Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, Jonas WB. Impact of study quality on outcome in placebo-controlled trials of homeopathy. Journal of clinical epidemiology 1999; 52(7): 631-636. | Nutzung eines bestehenden Instruments |
| Linde K, Ter Riet G, Hondras M, Melchart D, Willich SN. Characteristics and quality of systematic reviews of acupuncture, herbal medicines, and homeopathy. Forschende Komplementarmedizin und Klassische Naturheilkunde 2003; 88-94. | Kein Instrument dargestellt |
| Lozano-Calderón S, Anthony S, Ring D. The quality and strength of evidence for etiology: example of carpal tunnel syndrome. The Journal of hand surgery 2008; 33(4): 525-538. | Kein Instrument dargestellt |
| Lumbreras B, Porta M, Márquez S, Pollán M, Parker LA, Hernández-Aguado I. QUADOMICS: An adaptation of the Quality Assessment of Diagnostic Accuracy Assessment (QUADAS) for the evaluation of the methodological quality of studies on the diagnostic accuracy of "-omics"-based technologies. Clinical Biochemistry 2008; 1316-1325. | Biotechnologie |
| Machado M, Iskedjian M, Einarson TR. Quality assessment of published health economic analyses from South America. Annals of Pharmacotherapy 2006; 943-949. | Gesundheitsökonomie |
| Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. Physical therapy 2003; 83(8): 713-721. | Ermittlung der Testgüte eines bestehenden Instruments |

**Tabelle 35: Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Effektivität) – Fortsetzung**

| | |
|---|---|
| Mallen C, Peat G, Croft P, Mallen C, Peat G, Croft P. Quality assessment of observational studies is not commonplace in systematic reviews. Journal of clinical epidemiology 2006; 765-769. | Kein Instrument dargestellt |
| Mann R, Hewitt CE, Gilbody SM. Assessing the quality of diagnostic studies using psychometric instruments: applying QUADAS. Social Psychiatry and Psychiatric Epidemiology – The International Journal for Research in Social and Genetic Epidemiology and Mental Health Services 2008. | Nutzung eines bestehenden Instruments |
| Manzoli L, Schioppa F, Boccia A, Villari P. The efficacy of influenza vaccine for healthy children: a meta-analysis evaluating potential sources of variation in efficacy estimates including study quality. The Pediatric infectious disease journal 2007; 26(2): 97-106. | Kein Instrument dargestellt |
| McAlindon TE, La Valley MP, Gulin JP, Felson DT. Glucosamine and chondroitin for treatment of osteoarthritis: A systematic quality assessment and meta-analysis. Journal of the American Medical Association 2000; 1469-1475. | Nutzung eines bestehenden Instruments |
| Meert AP, Berghmans T, Branle F, Lemaître F, Mascaux C, Rubesova E, Vermylen P, Paesmans M, Sculier JP. Phase II and III studies with new drugs for non-small cell lung cancer: A systematic review of the literature with a methodology quality assessment. Anticancer Research 1999; 4379-4390. | Phase II Studien |
| Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP. Assessing the quality of reports of randomised trials: Implications for the conduct of meta-analyses. Health Technology Assessment 1999; iii-90. | Hintergrund |
| Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials. Current issues and future directions. International journal of technology assessment in health care 1996; 12(2): 195-208. | Aktualisierte Version vorhanden |
| Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998; 352(9128): 609-613. | Berichtsqualität |
| Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ (Clinical research ed.) 2005; 330(7499): 1053. | Kein Instrument dargestellt |
| Moyer A, Finney JW, Swearingen CE. Methodological characteristics and quality of alcohol treatment outcome studies, 1970-98: an expanded evaluation. Addiction (Abingdon, England) 2002; 97(3): 253-263. | Nutzung eines bestehenden Instruments |
| Moyer A, Finney JW. Rating methodological quality: toward improved assessment and investigation. Accountability in research 2005; 12(4): 299-313. | Nicht lieferbar |
| Nomura K, Nakao M, Morimoto T. Effect of smoking on hearing loss: Quality assessment and meta-analysis. Preventive Medicine 2005; 138-144. | Nutzung eines bestehenden Instruments |
| Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: A systematic review. Physical therapy 2008; 156-175. | Übersichtsarbeit |
| Oremus M, Wolfson C, Perrault A, Demers L, Momoli F, Moride Y. Interrater reliability of the modified Jadad quality scale for systematic reviews of Alzheimer's disease drug trials. Dementia and geriatric cognitive disorders 2001; 12(3): 232-236. | Ermittlung der Testgüte eines bestehenden Instrumentes |
| Parés D, Norton C, Chelvanayagam S. Fecal incontinence: the quality of reported randomized, controlled trials in the last ten years. Diseases of the colon and rectum 2008; 51(1): 88-95. | Nutzung eines bestehenden Instruments |
| Patel M. A meta-evaluation, or quality assessment, of the evaluations in this issue, based on the African Evaluation Guidelines: 2002. Evaluation and program planning 2002; 25(N4): 329-332. | Kein Instrument dargestellt |
| Peng X, Zhao Y, Liang X, Wu L, Cui S, Guo A, Wang W. Assessing the quality of RCT on the effect of beta-elemene, one ingredient of a Chinese herb, against malignant tumors. Contemporary clinical trials 2006; 27(1): 70-82. | Beichtsqualität |
| Powe NR, Kinnison ML, Steinberg EP. Quality assessment of randomized controlled trials of contrast media. Radiology 1989; 377-380. | Kein Instrument dargestellt |
| Quiñones D, Llorca J, Dierssen T, gado-Rodríguez M. Quality of published clinical trials on asthma. The Journal of asthma: official journal of the Association for the Care of Asthma 2003; 40(6): 709-719. | Nutzung eines bestehenden Instruments |

**Tabelle 35: Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Effektivität) – Fortsetzung**

| | |
|---|---|
| Reed DA, Beckman TJ, Wright SM, Levine RB, Kern DE, Cook DA. Predictive validity evidence for medical education research study quality instrument scores: quality of submissions to JGIM's Medical Education Special Issue. Journal of general internal medicine: official journal of the Society for Research and Education in Primary Care Internal Medicine 2008; 23(7): 903-907. | Kein Instrument dargestellt |
| Romeiser-Logan LR, Hickman RR, Harris SR, Heriza CB. Single-subject research design: Recommendations for levels of evidence and quality rating. Developmental medicine and child neurology 2008; 99-103. | Einzelfallstudien |
| Saunders LD, Soomro GM, Buckingham J, Jamtvedt G, Raina P. Assessing the methodological quality of nonrandomized intervention studies. Western journal of nursing research 2003; 25(2): 223-237. | Übersichtsarbeit |
| Schwenk W, Haase O, Günther N, Neudecker J. Methodological quality of randomised controlled trials comparing short-term results of laparoscopic and conventional colorectal resection. International Journal of Colorectal Disease 2007; 1369-1376. | Identischer Inhalt zu Evans & Pollock 1985 |
| Schwerla F, Hass-Degg K, Schwerla B. Evaluierung und kritische Bewertung von in der europäischen Literatur veröffentlichten, osteopathischen Studien im klinischen Bereich und im Bereich der Grundlagenforschung. Forschende Komplementärmedizin 1999; 6(6): 302-310. | Instrumente nicht ausreichend dargestellt |
| Shakespeare TP, Thiagarajan A, Gebski V. Evaluation of the quality of radiotherapy randomized trials for painful bone metastases. Cancer 2005; 103(9): 1976-1981. | Nutzung eines bestehenden Instruments |
| Shea B, Boers M, Grimshaw JM, Hamel C, Bouter LM. Does updating improve the methodological and reporting quality of systematic reviews? BMC medical research methodology 2006; 6: 27. | Nutzung eines bestehenden Instruments |
| Smith V, Devane D, Begley CM, Clarke M, Higgins S. A systematic review and quality assessment of systematic reviews of fetal fibronectin and transvaginal length for predicting preterm birth. European Journal of Obstetrics Gynecology and Reproductive Biology 2007; 134-142. | Instrument unzureichend beschrieben |
| Sonis J. The quality of clinical trials published in The Journal of Family Practice, 1974-1991. The Journal of family practice 1994; 39(3): 225-235. | Nutzung eines bestehenden Instruments |
| Spiegel BMR, Targownik LE, Kanwal F, Derosa V, Dulai GS, Gralnek IM, Chiou CF. The quality of published health economic analyses in digestive diseases: A systematic review and quantitative appraisal. Gastroenterology 2004; 403-411. | Gesundheitsökonomie |
| Steuten LMG, Vrijhoef HJM, van Merode GG, Severens JL, Spreeuwenberg C. The Health Technology Assessment-disease management instrument reliably measured methodologic quality of health technology assessments of disease management. Journal of clinical epidemiology 2004; 57(N9): 881-888. | Vollständiges Instrument und Erläuterungen nicht dargestellt |
| Taylor BJ, Taylor Brian, Dempster M, Donnelly M. Grading gems: Appraising the quality of research for social work and social care. British journal of social work 2007; 335-354. | Sozialarbeit |
| Tooth L, Bennett S, McCluskey A, Hoffmann T, McKenna K, Lovarini M. Appraising the quality of randomized controlled trials: Inter-rater reliability for the OTseeker evidence database. Journal of Evaluation in Clinical Practice 2005; 547-555. | Berichtsqualität |
| Treloar C, Champness S, Simpson PL, Higginbotham N. Critical appraisal checklist for qualitative research studies. Indian journal of pediatrics 2000; 67(5): 347-351. | Qualitative Studien |
| Tuech JJ, Pessaux P, Moutel G, Thoma V, Schraub S, Herve C. Methodological quality and reporting of ethical requirements in phase III cancer trials. Journal of medical ethics 2005; 31(5): 251-255. | Phase III Studien |
| Umbehr M. Qualitätsmessungskriterien im Gesundheitswesen müssen mit Bedacht und auf Evidenz basierend gewählt werden. Schweizerische Rundschau fur Medizin – Praxis 2008; 1307-1308. | Kein Instrument dargestellt |
| Ungar WJ, Santos MT. Quality appraisal of pediatric health economic evaluations. International journal of technology assessment in health care 2005; 21(2): 203-210. | Gesundheitsökonomie |
| Verhagen AP, de Bie RA, Lenssen AF, de Vet HC, Kessels AG, Boers M, van den Brandt PA. Impact of quality items on study outcome. Treatments in acute lateral ankle sprains. International journal of technology assessment in health care 2000; 16(4): 1136-1146. | Nutzung von bestehenden Instrumenten |
| Verhagen AP, de Bie RA, Lenssen AF, de Vet HC, Kessels AG, Boers M, Van DB. Quality assessment of trials: a comparison of three criteria lists. Phys Ther Rev 2000; 5(1): 49-58. | Kein Instrument dargestellt |

**Tabelle 35: Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Effektivität) – Fortsetzung**

| | |
|---|---|
| Verhagen AP, de Vet HCW, de Bie RA, Kessels AGH, Boers M, Knipschild PG. Balneotherapy and quality assessment: Interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. Journal of clinical epidemiology 1998; 335-341. | Ermittlung der Testgüte eines bestehenden Instruments |
| Wells K, Littell JH, Littell Julia. Study Quality Assessment in Systematic Reviews of Research on Intervention Effects. Research on social work practice 2009; 19(N1): 52-62. | Sozialarbeit |
| Westwood ME, Whiting PF, Kleijnen J, Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? BMC medical research methodology 2005. | Nutzung eines bestehenden Instruments |
| Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC medical research methodology 2005. | Vergleich unterschiedlicher Gewichtungen der QUADAS-Items |
| Whiting P, Rutjes AW, Dinnes J, Reitmsa JB, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies (Structured abstract). Health Technology Assessment 2004; 248. | Instrument identisch zu Whiting 2003 |
| Whiting P, Rutjes AWS, Dinnes J, Reitsma JB, Bossuyt PMM, Kleijnen J. A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. Journal of clinical epidemiology 2005; 1-12. | Übersichtsarbeit |
| Whiting PF, Weswood ME, Rutjes AWS, Reitsma JB, Bossuyt PNM, Kleijnen J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC medical research methodology 2006. | Instrument identisch zu Whiting 2003 |
| Zhang D, Yin P, Freemantle N, Jordan R, Zhong N, Cheng KK. An assessment of the quality of randomised controlled trials conducted in China. Trials 2008. | Berichtsqualität |

## 8.4 Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Ökonomie)

**Tabelle 36: Ausgeschlossene Publikationen der systematischen Datenbankrecherche (Ökonomie)**

| Publikation | Ausschlussgrund |
|---|---|
| Amin S, Tolley K, Harrison G, Amin S, Tolley K, Harrison G. Improving quality in economic evaluations of the management of schizophrenia. Journal of Medical Economics 1998; 163-176. | Nutzung eines bestehenden Instruments |
| Gerard K, Seymour J, Smoker I. A tool to improve quality of reporting published economic analyses. International journal of technology assessment in health care 2000; 16(1): 100-110. | Nutzung eines bestehenden Instruments |
| Gerkens S, Crott R, Cleemput I, Thissen JP, Closon MC, Horsmans Y, Beguin C. Comparison of three instruments assessing the quality of economic evaluations: a practical exercise on economic evaluations of the surgical treatment of obesity. International journal of technology assessment in health care 2008; 24(3): 318-325. | Vergleich von drei Instrumenten |
| Machado M, Iskedjian M, Einarson TR, Machado M, Iskedjian M, Einarson TR. Quality assessment of published health economic analyses from South America. Annals of Pharmacotherapy 2006; 943-949. | Nutzung eines bestehenden Instruments |
| Spiegel BMR, Targownik LE, Kanwal F, Derosa V, Dulai GS, Gralnek IM, Chiou C-F. The quality of published health economic analyses in digestive diseases: A systematic review and quantitative appraisal. Gastroenterology 2004; 403-411. | Nutzung eines bestehenden Instruments |
| Ungar WJ, Santos MT. Quality appraisal of pediatric health economic evaluations. International journal of technology assessment in health care 2005; 21(2): 203-210. | Nutzung eines bestehenden Instruments |

## 8.5 Ausgeschlossene Publikationen der Internetrecherche

**Tabelle 37: Ausgeschlossene Publikationen der Internetrecherche**

| Kürzel | Verfasser | Titel | Ausschlussgrund |
|---|---|---|---|
| AETSA | Andalusian Agency for Health Technology Assessment | Update of the Guide for Aquisition of New Technologies (GANT) | Kein Instrument dargestellt |
| AHRQ | Agency for Healthcare Research and Quality | Systems to Rate the Strength of Scientific Evidence Evidence Report/Technology Assessment Number 47 | Instrument dient der Datenextraktion, entspricht dem Vorgehen von West et al. 2000 |
| CADTH | The Canadian Agency for Drugs and Technologies in Health | Guidelines for Authors of CADTH Health Technology Assessment Reports | Kein Instrument dargestellt |
| CRD | The Centre for Reviews and Dissemination | CRD's guidance for undertaking reviews in health care | Es werden Instrumente nur aufgezählt |
| DACETHA | Danish Centre for Health Technology Assessment | Health Technology Assessment Handbook 2007 | Es wird nur ein Beispiel für eine Checkliste dargestellt |
| DIMDI | Deutsches Institut für Medizinische Dokumentation und Information | Methoden der Metaanalyse von diagnostischen Genauigkeitsstudien | Es werden Instrumente nur aufgezählt |
| DIMDI | Deutsches Institut für Medizinische Dokumentation und Information | Handbuch für Autoren zur Erstellung von HTA-Berichten | Kein Instrument dargestellt |
| ECHTA, ECAHI | European Collaboration for Health Technology Assessment – Assessment of Health Interventions | Best practice in undertaking and reporting HTA | Es werden Instrumente nur aufgezählt |
| EUnetHTA | European Network for Health Technology Assessment | HTA Core Model for medical and surgical interventions | Kein Instrument zur Bewertung der methodischen Qualität |
| EUnetHTA | European Network for Health Technology Assessment | HTA Core Model for diagnostic technologies | Kein Instrument zur Bewertung der methodischen Qualität |
| GOEG | Gesundheit Österreich GmbH | Prozesshandbuch für Health Technology Assessment | Kein Instrument dargestellt |
| INAHTA | International Network of Agencies for Health Technology Assessment | Checklist for health technology assessment reports | Instrument zur Berichtsqualität |
| IQWiG | Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen | Allgemeine Methoden Version 3.0 | Es werden Instrumente nur aufgezählt |
| LBI | Ludwig Boltzmann Institut | (Externes) Manual: Selbstverständnis und Arbeitsweise | Kein Instrument dargestellt |
| MSAC | Medical Services Advisory Committee | Guidelines for the assessment of diagnostic technologies | Es werden Instrumente nur aufgezählt |
| MSAC 2005 | Medical Services Advisory Committee | Funding for new medical technologies and procedures: application and assessment guidelines | Kein Instrument dargestellt |
| NICE | National Institute for Health and Clinical Excellence | Guide to the Methods of Technology Appraisal | Kein Instrument dargestellt |
| Toronto CEBM | Centre for Evidence-Based Medicine, Toronto | Critical Appraisal Worksheets | Fokus ist Relevanz für Patienten |

## 8.6 Datenextraktionsformular formale Kriterien

| | | | | |
|---|---|---|---|---|
| 1 | Publikation: | ……. | | |
| 2 | Ausfüllender: | ……. | | |
| 3 | Datum | ……. | | |
| 4 | Sprache | □ Englisch | □ Deutsch | |
| 5 | Name des Instrumentes | ……. | | |
| 6 | Abkürzung | ……. | | |
| 7 | | □ Original | □ Modifikation von 1 Instrument<br>→ Welches? …… | |
| 8 | Instrument für welche Studiendesigns? | □ Systematische Reviews/HTA/Meta-Analysen<br>□ Diagnosestudien<br>□ Interventionsstudien<br>□ Beobachtungsstudien | | |
| 9 | | □ Krankheits/Diagnosespezifisch | □ Generisch<br>→ für welche Krankheit/Diagnose? …… | |
| 10 | Anzahl Items<br>Art des Instrumentes<br>(Freitext) | ……<br>□ Checkliste | □ Komponenten-Bewertung<br>→ □ qualit. Komponentenbewertung<br>→ □ qualit. Gesamtbewertung<br>→ □ ohne Gesamt-Bewertung | □ Skala<br>↓<br>Range …<br>Cutpoint … |
| 11 | Definition Qualität | □ | Ja | □ Nein |
| 12 | Entwicklungsprozess | □ | Ja | □ Nein |
| 13 | Ausfüllhinweise | □ | Ausführlich □ Kurz | □ Nein |
| 14 | Zeitbedarf | …… | Min | □ Nein |
| 15 | Testgüte | □ | Ja | □ Nein |
| | Inhaltliche Validität | …… | | |
| | Konstruktvalidität | …… | | |
| | Kriteriumsvalidität | …… | | |
| | Interrater-Reliabilität | …… | | |
| | Intrarater-Reliabilität | …… | | |
| | Test-Retest-Reliabilität | …… | | |
| | Interne Konsistenz | …… | | |

## 8.7 Operationalisierung des Datenextraktionsformulars formale Kriterien

| | |
|---|---|
| 1<br>2<br>3<br>4<br>5<br>6 | Erstautor, Jahr<br>Name des Ausfüllenden<br>Datum der Datenextraktion<br>Sprache des **Volltextes** |
| 7 | Mehr als 1 Instrument modifiziert = Original<br>Welches Instrument wurde **laut Publikation** modifiziert? |
| 8 | Für welche Designs ist das Instrument **laut Publikation** geeignet?<br><br>Auch nicht randomisierte Interventionsstudien |
| 9 | Für welche Krankheit oder Diagnose ist das Instrument laut Publikation geeignet?<br>Für welche Krankheit oder Diagnose wird das Instrument in der Publikation angewandt? |
| 10 | Anzahl der **Items**, nicht der Domänen<br>Checkliste: Antwortvorgaben, nicht-numerische Bewertung der Items, mit/ohne Gesamtbewertung<br>Komponenten-Bewertung: Instrumente ohne konkrete fragende Items, sondern bestehend aus Stichwörtern wie „Randomisierung“; Antwortvorgaben in der Regel vorhanden<br>qualitative Gesamt- oder Komponentenbewertung: z. B. hoch-mittel-niedrig<br>Skala: Bewertung der Items, numerische Gesamtbewertung<br><br>Range: Wertebereich des Gesamtscores<br>Cutpoint: Wert, ab dem die Studie mit „gut“ o.ä. bewertet wird. |
| 11 | Studienqualität/methodische Qualität definiert? Definition von Validität genügt nicht |
| 12 | Entwicklung des Instrumentes beschrieben? z. B. Delphi-Prozess |
| 13 | Anwendungshinweise/Operationalisierung gegeben? z. B. Kriterien für Wertung als vorhanden/erfüllt/gut. Nicht: Hintergrundinfos |
| 14 | Zeitbedarf der Anwendung des Instrumentes angegeben? |
| 15 | Testgüte des **gesamten**, **endgültigen** Instrumentes getestet? Ja, wenn mind. einer der folgenden Parameter bestimmt wurde<br>Inhaltliche Validität (content-v., logical v.): inhaltliche Analyse dessen, was das Instrument misst: Herstellungsprozess, Definition der Items, Expertenbefragung; Wert ja oder nein<br>Konstruktvalidität: Grad der Übereinstimmung mit einem gleichen Konstrukt, z. B. Vergleich mit anderem Instrument; ggf. Wert eintragen: Korrelationskoeffizient r<br>Kriteriumsvalidität: Grad der Übereinstimmung mit einem unabhängigen Verfahren; ggf. Wert eintragen: Korrelationskoeffizient r<br>Inter-Rater-Reliabilität: Intraclass-correlation ICC oder kappa-Statistik (nach Cohen oder Fleiss), Wert eintragen (0-1)<br>Intra-Rater-Reliabilität: Wert eintragen und Bezeichnung angeben<br>Re-Test-Reliabilität: Wert eintragen: Korrelationskoeffizient r |

## 8.8 Datenextraktionsformular für systematische Reviews

| Publikation:<br>Ausfüllender:<br>Datum | |
|---|---|
| **Domäne** | **Elemente** |
| Studienfrage | Fragestellung präzise und angemessen |
| Ein- und Ausschlusskriterien | Kriterien wurden a priori definiert |
| | Kriterien sind angemessen |
| Literaturrecherche und -auswahl | Relevante Datenbanken einbezogen |
| | Einbezug weiterer Datenquellen (z.B. Handsuche, graue Literatur, Referenzen, pers. Kontakt) |
| | Kombination von Schlagworten/Thesaurus und Freitext |
| | Vielzahl von Synonymen |
| | Restriktionen bei der Suche sind akzeptabel (z.B. Sprache, Land, Zeitraum) |
| | Dokumentation der verwendeten Suchterme und Datenbanken |
| | Literaturauswahl unabhängig voneinander durch mind. 2 Reviewer |
| | Ausschluss von Literatur begründet |
| | Ausreichend detailliert, um die Literaturrecherche/-auswahl zu reproduzieren |
| Datenextraktion | Extraktion von Interventionen/Expositionen für alle relevanten (Sub)gruppen |
| | Extraktion von Outcomes für alle relevanten (Sub)gruppen |
| | Datenextraktion unabhängig voneinander durch mind. 2 Reviewer |
| | Datenextraktion verblindet für Reviewer (z.B. Autoren, Zeitschrift, Jahr, Ergebnisse) |
| | Messung der Übereinstimmung der Reviewer |
| | Ausreichend detailliert, um die Datenextraktion zu reproduzieren |
| Studienqualität/ Interne Validität | Bewertungsmethode wird beschrieben, ist angemessen |
| | Bewertung unabhängig voneinander durch mind. 2 Reviewer |
| | Bewertung verblindet für Reviewer (z.B. Autoren, Zeitschrift, Jahr) |
| | Bewertung der Übereinstimmung der Reviewer |
| | Methode zur Integration der Ergebnisse der Qualitätsbewertung ist angemessen |
| Datensynthese und -analyse | Angemessene qualitative und/oder quantitative Synthese |
| | Berücksichtigung der Robustheit der Ergebnisse und/oder mögl. Heterogenität |
| | Darstellung von Schlüsselelementen von Primärstudien, die ausreichend sind für eine kritische Bewertung und Wiederholung |
| Ergebnisse | Narrative Zusammenfassung und/oder quantitative Zusammenfassung und Angabe der Präzision, wenn angemessen |
| Diskussion | Schlussfolgerungen werden durch die Ergebnisse unterstützt |
| | Berücksichtigung möglicher Bias und anderen Limitationen |
| Finanzielle Förderung/ Auftraggeber | Art und Quelle der Finanzierung |

## 8.9 Datenextraktionsformular für Interventionsstudien

| Publikation:<br>Ausfüllender:<br>Datum | |
|---|---|
| **Domäne** | **Elemente** |
| Studienfrage | Fragestellung präzise und angemessen |
| Studienpopulation | Beschreibung der Studienpopulation |
| | Spezifische Ein- und Ausschlusskriterien |
| | Angemessene Stichprobengröße für alle Gruppen (Power-Berechnung) |
| Randomisierung | Methode der Randomisierung beschrieben und angemessen |
| | Gruppenzuweisung geheim |
| | Vergleichbarkeit der Gruppen zu Beginn |
| Verblindung | Verblindung der Studienteilnehmer |
| | Verblindung der Untersucher/Erheber des Outcomes |
| | Verblindung des übrigen Studienpersonals (z.B. Betreuer, Behandler) |
| | Verblindung überprüft und ausreichend |
| Interventionen | Interventionen eindeutig und detailliert für alle Gruppen beschrieben |
| | Gleichzeitige Kontrollgruppe |
| | Behandlungsgleichheit bis auf die Intervention |
| | Placebo vergleichbar mit Verum (Darreichungsform, Aussehen, Geschmack, Geruch) |
| | Co-Interventionen vermieden |
| | Co-Interventionen für alle Gruppen beschrieben |
| | Kontamination vermieden/akzeptabel |
| | Compliance akzeptabel in allen Gruppen |
| Outcomes | Primäre und sekundäre Outcomes präzise definiert |
| | Verwendete Methoden sind valide |
| | Verwendete Methoden sind reliabel |
| | Follow-Up mit angemessener Länge |
| | Follow-Up gleichzeitig in allen Studiengruppen |
| Statistische Analyse | Angemessene statistische Analyse |
| | Viele Vergleiche sind berücksichtigt worden (Multiples Testen) |
| | Intention-to-treat Analyse |
| | angemessener Umgang mit fehlenden Werten |
| | Berücksichtigung von Confounding |
| | Bewertung von Confounding |
| | Bewertung von Heterogenität, wenn anwendbar |
| Ergebnisse | Effekte hinsichtlich des Outcomes mit Punktschätzer und Präzision angegeben |
| | Anteil Studienabbrecher/loss-to-follow-up angegeben und akzeptabel |
| | Unterschiede Teilnehmer/Abbrecher geprüft und akzeptabel |
| | Ursachen für Drop-outs/loss-to-follow-up dargestellt |
| | Selektives Berichten von Outcomes (ungeplante Analysen oder geplante/erwartete A. werden nicht berichtet) |
| | Vorzeitiger Abbruch der Studie (aufgrund von Zwischenergebnissen) |
| Diskussion | Schlussfolgerung werden durch Ergebnisse unterstützt |
| | Möglicher Einfluss von Confounding und Bias wird diskutiert |
| Externe Validität | Anteil Nichtteilnehmer angegeben und akzeptabel |
| | Unterschiede Teilnehmer/Nichtteilnehmer geprüft und akzeptabel |
| | Studienpopulation repräsentativ |
| Finanzielle Förderung/ Auftraggeber | Art und Quelle der Förderung |

## 8.10 Datenextraktionsformular für Beobachtungsstudien

| | |
|---|---|
| Publikation: | |
| Ausfüllender: | |
| Datum | |
| **Domäne** | **Elemente** |
| Studienfrage | Fragestellung präzise und angemessen |
| Studienpopulation | Beschreibung der Studienpopulation |
| | Spezifische Ein- und Ausschlusskriterien für alle Gruppen |
| | Identische Ein- und Ausschlusskriterien für alle Gruppen |
| | Angemessene Stichprobengröße für alle Gruppen (Power-Berechnung) |
| | Gleichzeitige Kontrollgruppe |
| | Vergleichbarkeit der Studiengruppen untereinander zu Beginn (Krankheitsstatus und prognostische Faktoren) |
| | FKS: Explizite Falldefinition |
| | FKS: Kontrollen gleichen den Fällen bis auf das interessierende Outcome, Kontrollen haben die gleiche Expositionschance wie die Fälle |
| Exposition | Exposition/Intervention präzise definiert |
| | Methoden zur Erhebung der Exposition sind valide |
| | Methoden zur Erhebung der Exposition sind reliabel |
| | Expositionsmessung gleich in allen Studiengruppen |
| | FKS: Expositionserhebung verblindet für Outcomestatus |
| Outcome | Primäre und sekundäre Outcomes präzise definiert |
| | Methoden zur Erhebung des Outcomes sind valide |
| | Methoden zur Erhebung des Outcomes sind reliabel |
| | FKS: Diagnosesicherung unbeeinflusst von Expositionsstatus (verblindet) |
| | Erhebung des Outcomes verblindet für Expositions- oder Interventionsstatus |
| | Outcomemessung gleich in allen Studiengruppen |
| | Follow-Up mit angemessener Länge |
| | Länge des Follow-Up gleich für alle Studiengruppen |
| Statistische Analyse | Angemessene statistische Analyse |
| | Viele Vergleiche sind berücksichtigt worden (Multiples Testen) |
| | Modellierung und/oder multivariate Methoden angemessen (Confounderkontrolle) |
| | angemessener Umgang mit fehlenden Werten |
| | Bewertung von Confounding/Residual Confounding |
| | Bewertung von Heterogenität (Effektmodifikation/Interaktion), wenn anwendbar |
| | Bestimmung der Dosiswirkungsbeziehung wenn möglich |
| Ergebnisse | Effekte hinsichtlich des Outcomes mit Punktschätzer und Präzision angegeben |
| | Anteil Studienabbrecher/loss-to-follow-up angegeben und akzeptabel |
| | Unterschiede Teilnehmer/Abbrecher geprüft und akzeptabel |
| | Ursachen für Drop-outs/loss-to-follow-up dargestellt |
| | Selektives Berichten von Outcomes (ungeplante Analysen oder geplante/erwartete A. werden nicht berichtet) |
| Diskussion | Schlussfolgerung werden durch Ergebnisse unterstützt |
| | Möglicher Einfluss von Confounding und Bias wird diskutiert |
| Externe Validität | Anteil Nichtteilnehmer angegeben und akzeptabel |
| | Unterschiede Teilnehmer/Nichtteilnehmer geprüft und akzeptabel |
| | Studienpopulation repräsentativ |
| Finanzielle Förderung/ Auftraggeber | Art und Quelle der Förderung |

## 8.11 Datenextraktionsformular für Diagnosestudien

| | |
|---|---|
| Publikation: | |
| Ausfüllender: | |
| Datum | |
| **Domäne** | **Elemente** |
| **Verzerrungspotential (10 Items)** | |
| Referenzstandard | Wurde ein angemessener Referenztest verwendet um den Zielparameter zu erfassen? |
| Bias durch Krankheitsprogression | Könnte eine Änderung des Krankheitsstatus zwischen Durchführung des Indextests und des Referenztests aufgetreten sein? |
| Verifikationsbias | Wurde bei allen Teilnehmern der Zielparameter mit dem gleichen Referenztest verifiziert? |
| Bias durch nicht-unabhängige Tests (Incorporation bias) | War der Indextest Teil des Referenztests? (Waren die Tests nicht unabhängig voneinander?) |
| Behandlungsparadox | Wurde die Behandlung basierend auf dem Ergebnis des Indextests eingeleitet bevor der Referenztest erfolgte? |
| Review bias | Wurden die Ergebnisse des Indextests verblindet gegenüber dem Ergebnis des Referenztests ausgewertet? |
| Review bias | Wurden die Ergebnisse des Referenztests verblindet gegenüber dem Ergebnis des Indextests ausgewertet |
| Klinischer Review bias | Waren klinische Informationen vorhanden, als die Testergebnisse ausgewertet wurden? |
| Beobachter-/ Instrumentenvariabilität | Ist es wahrscheinlich, dass eine Beobachter-/Instrumentenvariabilität Annahmen bei der Testausführung beeinflusst haben? |
| Umgang mit nicht bewertbaren Testergebnissen | Wurden nicht bewertbare Testergebnisse in die Analyse eingeschlossen? |
| **Externe Validität (4 Items)** | |
| Spektrum der Teilnehmer (spectrum composition) | War die untersuchte Bevölkerung vergleichbar mit der interessierenden Population? |
| Rekrutierung | War die Rekrutierungsmethode angemessen um ein geeignetes Spektrum an Patienten einzuschließen? |
| Krankheitsprävalenz/ -schwere | Waren Prävalenz der Erkrankung und Spektrum der Krankheitsschwere in der Studienpopulation vergleichbar mit der der interessierenden Population? |
| Methodenwechsel beim Indextest | Ist es wahrscheinlich, dass die Methode des Tests im Laufe der Studie verändert wurde? |
| **Studiendurchführung (4 Items)** | |
| Subgruppenanalysen | Waren Subgruppenanalysen angemessen und vorab spezifiziert worden? |
| Stichprobengröße | Wurde eine angemessene Teilnehmerzahl in die Studie eingeschlossen? (Power?) |
| Studienziele | Waren die Studienziele relevant für die Studienfrage? |
| Protokoll | Wurde ein Studienprotokoll vor Studienbeginn entwickelt und wurde dies befolgt? |
| **Berichtsqualität (9 Items)** | |
| Einschlusskriterien | Wurden die Einschlusskriterien präzise dargestellt? |
| Indextestdurchführung | Wurde die Methodik des Indextests genügend detailliert beschrieben, um die Replikation des Tests zu ermöglichen? |
| Referenztestdurchführung | Wurde die Methodik des Referenztests genügend detailliert beschrieben, um die Replikation des Tests zu ermöglichen? |
| Definition von „normalem“ Testergebnis | Haben die Autoren präzise dargestellt, was als „normales“ Testergebnis bewertet wurde? |
| Angemessene Ergebnisse | Wurden adäquate Ergebnisse dargestellt? Z.B. Sensitivität, Spezifität, Likelihood Ratios |
| Genauigkeit der Ergebnisse | Wurde die Präzision der Ergebnisse dargestellt? Z.B. Konfidenzintervalle |
| Studienabbrecher | Wurden alle Studienteilnehmer bei der Analyse berücksichtigt? |
| Datentabelle | Wurde eine *n* x *n* Tabelle zur Testdurchführung dargestellt? |
| Nützlichkeit des Tests | Wurden Hinweise gegeben, wie nützlich der Test in der Praxis sein könnte? |

## 8.12 Datenextraktionsformular für gesundheitsökonomische Studien

| Publikation:<br>Ausfüllender:<br>Datum<br>**Domäne** | | | | **Elemente** |
|---|---|---|---|---|
| | **Angemessen** | **Begründet** | **Berichtet** | |
| Studienfrage | | | | Wurde die Studienfrage präzise definiert? |
| | | | | Ist die Art der ökonomischen Studie angemessen? |
| Interventions-alternativen | | | | Wurden die aus ökonomischer Sicht relevanten Interventionsalternativen einbezogen? |
| | | | | Wurde ein direkter Vergleich der Interventionsalternativen vorgenommen? |
| | | | | Falls nein, waren die indirekten Vergleiche angemessen? |
| Perspektive | | | | Passt die gewählte Perspektive zur Studienfrage? |
| | | | | Ist die Perspektive konsistent? |
| Ressourcen-verbrauch und Kosten | | | | Wurden Mengen und Preise getrennt voneinander erhoben? |
| | | | | Wurden alle relevanten Ressourcen, die mit der Intervention in Zusammenhang stehen identifiziert? |
| | | | | Sind angemessene Datenquellen genutzt worden? |
| | | | | Wurden alle relevanten Ressourcen mengenmäßig erfasst? |
| | | | | Wurden die Ressourcenverbräuche angemessen bewertet? |
| | | | | Sind die Methoden der Inflationierung und Währungskonversion angemessen? |
| Outcome/Nutzen | | | | Sind die Outcomeparameter richtig gewählt worden? |
| | | | | Sind geeignete Erhebungsinstrumente gewählt worden, falls LQ erhoben wurde? |
| | | | | Sind die Datenquellen auf ihre Qualität überprüft worden? |
| Qualität der Daten | | | | Ist die Qualität der Primärdaten ausreichend um die Studienfrage beantworten zu können? |
| | | | | Ist die Qualität der Methoden zur Identifikation, Extraktion und Synthese der Effektparameter (Meta-Analyse) ausreichend zur Beantwortung der Studienfrage? |
| Zeitraum | | | | Ist der Beobachtungszeitraum so gewählt, dass alle relevanten Effekte und Kosten berücksichtigt wurden? |
| Modellierung | | | | Wurde das Modell nachvollziehbar dargestellt? |
| | | | | Sind die Modellstruktur und die gewählten Parameter angemessen? |
| Diskontierung | | | | Wurden in der Studie alle zukünftigen Kosten und Nutzen diskontiert? |
| | | | | Wenn ja, waren die Diskontraten angemessen? |
| Statistische Verfahren | | | | Waren die statistischen Verfahren angemessen? |
| | | | | Wurde eine inkrementelle Kosten-Effektivitäts-Analyse durchgeführt? |
| Sensitivitäts-analyse | | | | Wurde eine Sensitivitätsanalyse durchgeführt? |
| | | | | Wurden alle relevanten Parameter in die Sensitivitätsanalyse einbezogen? |
| | | | | Ist die Methodik der Sensitivitätsanalyse angemessen? |
| Ergebnisse | | | | Wurden die Ergebnisse mit Punktschätzern und Präzision angegeben? |
| Diskussion der Ergebnisse | | | | Basieren die Schlussfolgerungen auf den Ergebnissen? |
| | | | | Wurde der mögliche Einfluss von Confounding und Bias diskutiert? |
| Interessens-konflikte | | | | Werden Art und Quelle der Finanzierung genannt? |

## 8.13 Übersicht über die eingeschlossenen systematischen Übersichtsarbeiten

**Tabelle 38: Übersicht über die eingeschlossenen systematischen Übersichtsarbeiten (n = 8)**

| Erstautor | Deeks[56] | Katrak[109] | Moher[143] | Olivo[168] | Sanderson[189] | Saunders[190] | West[235] | Whiting[239] |
|---|---|---|---|---|---|---|---|---|
| Jahr | 2003 | 2004 | 1995 | 2008 | 2007 | 2003 | 2002 | 2005 |
| Land | UK | Australien | Kanada | USA | UK | Kanada | USA | UK |
| Titel | Evaluating non-randomised intervention studies | A systematic review of the content of critical appraisal tools | Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists | Scales to assess the quality of randomized controlled trials: A systematic review | Tools for assessing quality an susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography | Assessing the methodological quality of non-randomized intervention studies | Systems to rate the strength of scientific evidence | A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools |
| Fokus | Nicht-randomisierte Interventionsstudien | Interventionen im Bereich von Gesundheitsfachberufen | RCT | RCT | Beobachtungsstudien | Nicht-randomisierte Interventionsstudien | Alle Studientypen | Studien zu diagnostischen Tests |
| Art der Publikation | HTA, INAHTA | Methodenpaper | Methodenpaper | Methodenpaper | Methodenpaper | Methodenpaper | Evidence Report/ Technology Assessment, AHRQ | Methodenpaper, basierend auf HTA-Bericht[144] des NHS R&D Health Technology Assessment Programms |
| **Adressiertes Studiendesign** | | | | | | | | |
| SR, Metaanalysen | Nein | Ja | Nein | Nein | Nein | Nein | Ja | Nein |
| RCT | Nein | Ja | Ja | Ja | Nein | Nein | Ja | Nein |
| Kohortenstudien | Ja | Ja | Nein | Nein | Ja | Ja | Ja | Nein |
| Fall-Kontrollstudien | Nein | Ja | Nein | Nein | Ja | Ja | Ja | Nein |
| Querschnittstudien | Nein | Ja | Nein | Nein | Ja | Ja | Ja | Nein |
| Qualitative Studien | Nein | Ja | Nein | Nein | Nein | Nein | Nein | Nein |
| Studien zu diagnostischen Tests | Nein | Nein | Nein | Nein | Nein | Nein | Ja | Ja |
| **Methode Literatursuche** | | | | | | | | |
| Zeitraum | Bis Dezember 1999 | Unklar | 1966 bis Ende 1992 | 1965 bis März 2007 | Bis März 2005 | 1966 bis März 1999 | 1995 bis 2000 | 1966 bis April 2001 |
| Restriktion Sprache | Unklar | Englisch | Nein | Nein | Nein | Englisch | Englisch | Nein |
| Suchstrategie angegeben | Ja | Tw | Ja | Ja | Ja | Ja | Ja | Ja |
| Reproduzierbarkeit der Suche | Ja | Nein | Ja | Ja | Nein | Ja | Ja | Ja |

**Tabelle 38: Übersicht über die eingeschlossenen systematischen Übersichtsarbeiten (n = 8) – Fortsetzung**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Zahl der DB (11 =>10) | 11 | 8 | 1 | 11 | 3 | 1 | 1 | 5 |
| DB 1 | MEDLINE | MEDLINE | MEDLINE | MEDLINE | MEDLINE | MEDLINE | MEDLINE | MEDLINE |
| DB 2 | EMBASE | EMBASE | | EMBASE | EMBASE | | | EMBASE |
| DB 3 | PsycLit | CINAHL | | CINAHL | Dissertation Abstracts | | | Biosis |
| DB 4 | DARE | DARE | | DARE | | | | CRD |
| DB 5 | | | | CDSR | | | | CDSR |
| Referenzlisten | Ja | Ja | Ja | Ja | Nein | Nein | Ja | Nein |
| Handsuche | Ja | Nein | Nein | Ja | Nein | Nein | Nein | Nein |
| Internet | Nein | Ja | Nein | Nein | Ja (Google) | Nein | Ja | Nein |
| Experten | Ja | Ja | Ja | Nein | Nein | Nein | Ja | Ja |
| Sonstiges | | | | | | | | |
| Zahl der Reviewer | K. A. | 2 | K. A. | 5 | K. A. | 2 | 2 | 1 (Kontrolle durch 2. Reviewer) |
| Teilnehmer Konsens | K. A. | 3 | K. A. | 5 | K. A. | 3 | 2 | 2 |
| Ein- (Ein) und Ausschlusskriterien (Aus) definiert? | Ja | Ja | Ja | Ja | Ja | Tw. | Tw. | Ja |
| Ein 1 | Eigene Publikation von QBI | QBI mit klaren und eindeutigen Kriterien | Skalen und CL | Publizierte Skalen zur QB von RCT | QBI für Kohorten-, Fall-Kontroll- oder Querschnittsstudien | Artikel, die die Entwicklung und Testung eines Instruments untersuchen | Systeme zur Bewertung der Studienqualität | Publizierte CL zur QB von Studien zu diagnostischen Tests |
| Ein 2 | Im Kontext eines SR | QBI mit numerischem Qualitätsscore | | Alle Gesundheits-bereiche | Skalen, CL, CL mit qualitativer Gesamt-bewertung | Artikel, die ein Instrument nutzen | | Anleitungen zum Berichten, zur Studiendurch-führung oder zur Interpretation von Diagnosestudien |
| Ein 3 | Explizit QBI für RCT, wenn für Nicht-RCT genutzt | | | | | | | Erste Publikation, wenn das gleiche Instrument in mehreren Publikationen verwendet wird |
| Ein 4 | Neu oder modifiziert | | | | | | | |
| Aus 1 | Fall-Kontrollstudien | QBI nicht vollständig publiziert | QBI, die Modifikationen darstellten | Methodische Entwicklung nicht angegeben | | | | |
| Aus 2 | Nicht kontrollierte Studien | QBI für diagnostische Instrumente | | Reliabilität und Validität nicht getestet | | | | |
| Aus 3 | | QBI für klinische Leitlinien | | CL | | | | |

**Tabelle 38: Übersicht über die eingeschlossenen systematischen Übersichtsarbeiten (n = 8) – Fortsetzung**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Aus 4 | | Nicht auf Englisch | | Nicht publizierte Instrumente | | | | |
| Fließdiagramm/ umfassende Ergebnisse der LR | Nein | Tw. | Nein | Ja | Nein | Nein | Ja | Nein |
| Sonstiges | | 29 % der Dokumente nicht verfügbar | | | | | | |
| **Datenextraktion** | | | | | | | | |
| Zahl der Reviewer | 1 (Kontrolle durch 2. Reviewer) | 2 | K. A. | 5 | 2 | 2 | 2 | 1 (Kontrolle durch 2. Reviewer) |
| Reviewerkonsens | 3 | 3 | K. A. | 5 | 3 | 3 | 3 | 3 |
| Reviewer verblindet? | K. A. | K. A. | K. A. | Nein | Nein | Nein | Nein | Nein |
| Interrater-Reliabilität | K. A. | K. A. | K. A. | κ = 0,90 | K. A. | K. A. | K. A. | K. A. |
| **Bewertungsmethode/Kriterien** | | | | | | | | |
| A priori festgelegte Kriterien | Ja | Ja | Ja | Unklar | Ja | Ja | Ja | Unklar |
| Entstehungsprozess der Kriterien | Kriterien durch Delphi-Prozess | Kriterien basieren auf Cochrane Reviewer Handbook | K. A. | Kriterien zur Güte der eingesetzten Instrumente basieren auf Definitionen von Streiner & Norman (2004) und Leitlinien zur QB der Messeigenschaften von Instrumenten (Terwee 2007) | Domänen basieren auf STROBE-Statement | Kriterien basieren auf Moher 1995 u. a. | Technische Expertenberatergruppe, peer-review, Studienkonstrukte, die Studienqualität beeinträchtigen können, Methoden der Cochrane Collaboration, SIGN, NHS Centre for Reviews and Dissemination, New Zealand Guidelines Group | K. A. |
| Externe Validität | Ja | Ja | Nein | Nein | Nein | Nein | Nein | Ja |
| Interne Validität | Ja | Ja | Ja | Ja | Ja | Ja | Ja | Ja |
| Berichtsqualität | Ja | Nein | Nein | Ja | Nein | Ja | Nein | Ja |
| **Ergebnisse** | | | | | | | | |
| Instrumente (n) | 193 | 121 | 34 | 21 | 86 | 18 | 121 | 67 |
| CL | | 63 | 9 | | 41 | 8 | | Unklar (67 %) |
| Skalen (n) | | 58 | 25 | 21 | 33 | 10 | | Unklar |
| Komponenten (n) | | | | | 12 | | | |
| Sonstiges | Best tools, top tools | | | | | | | |
| Empfehlung | 6 QBI | Nein | Nein | Nein | Nein | 1 QBI (Downs & Black als CL) | 19 QBI | 1 QBI (QUADAS) |
| Empfehlung begründet? | Ja | Ja | Ja | Ja | Ja | Ja | Ja | Ja |

**Tabelle 38: Übersicht über die eingeschlossenen systematischen Übersichtsarbeiten (n = 8) – Fortsetzung**

| **Diskussion (Schlussfolgerungen werden durch Ergebnisse unterstützt, Limitationen)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Limitationen | Ja | | Vollständigkeit der Suche | | Vollständigkeit der Suche | Vollständigkeit der Suche | Vollständigkeit der Suche | |
| Schlussfolgerung 1 | 6 QBI (best tools) sind potenziell nützlich für SR | Kein Goldstandard | Nur wenige Instrumente werden mit Standardtechniken entwickelt | Es gibt eine Vielzahl von Instrumenten, die meisten sind nicht adäquat entwickelt und ihre Validität und Reliabilität getestet worden | Es kann kein Instrument empfohlen werden ohne es für mehrere Studien angewendet, seine Eigenschaften analysiert und seine einfache Anwendung untersucht zu haben | Skala von Downs & Black wird als CL empfohlen | Empirische Lücke | Die große Variation unter den verfügbaren Instrumenten bei gleichzeitig mangelnden Angaben zur Instrumententenentwicklung und -evaluation machen die Wahl eines Bewertungsinstruments für Studien von diagnostischen Tests schwierig. |
| Schlussfolgerung 2 | Vermischung von Studien- und Berichtsqualität | Kriterien für Auswahl QBI: methodische Entwicklung, Reliabilität, Validität, Manual | Untersuchen, ob unterschiedliche Instrumente gleiche Ergebnisse bringen (dann das kürzere nutzen) | Eine valide und reliable Skala für die methodische QB von physikalischen Therapiestudien sollte entwickelt werden | Anforderungen an Instrument: – kleine Zahl an Schlüsseldomänen – so spezifisch wie möglich – CL statt Skala – methodische Entwicklung, Reliabilität, Validität | Forschung, um Studiencharakteristika zu identifizieren, die die methodische Qualität von nicht-randomisierten Studien beeinflussen können | Viele Instrumente werden prinzipiell empfohlen | QUADAS, ein evidenzbasiertes valides und reliables Instrument wird empfohlen. |
| Schlussfolgerung 3 | Weitere Forschung: neues QBI oder Revision eines bekannten | Konsens für wichtige und Kernitems ist notwendig | Zukünftige QBI sollen methodisch stringent entwickelt werden, generisch sein, einfach zu nutzen | | | | | |
| Schlussfolgerung 4 | Weitere Forschung: Anwendbarkeit (usability) | Unterschiedliche QBI erzielen unterschiedliche Ergebnisse | | | | | | |

AHRQ = Agency for Healthcare Research and Quality. BIOSIS = BIOSIS Datenbank. CDSR = Cochrane Database of Systematic Reviews. CINAHL = Cumulative Index to Nursing and Allied Health Literature. CL = Checkliste. CRD = Centre for Reviews and Dissemination. DARE = Database of Abstracts of Reviews of Effects. DB = Datenbank. EMBASE = Excerpta Medica Database. HTA = Health Technology Assessment. INAHTA = International Network of Agencies for Health Technology Assessment. K. A. = Keine Angabe. LR = Literaturrecherche. MEDLINE = Medical Literature Analysis and Retrieval System Online. NHS = National Health Service. Psyclit = Ehemalige Bezeichnung der Datenbank PsycINFO. QB = Qualitätsbewertung. QBI = Qualitätsberwertungsinstrument. QUADAS = Quality assessment of diagnostic accuracy studies. RCT = Randomisierte kontrollierte Studie. SIGN = Scottish Intercollegiate Guidelines Network. SR = Systematischer Review. STROBE = Strengthening the Reporting of Observational Studies in Epidemiology. Tw = Teilweise. UK = United Kingdom. USA = United States of America.

## 8.14 Formale Charakteristika der QBI (Effektivität)

**Tabelle 39: Formale Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen**

| | Name/ Abkürzung des Instruments | Sprache | Modifikation von | Spezifisches Instrument | Items | Art des Instruments | Qualitative Bewertung | Range* | Cutpoint* | Definition von Qualität | Entwicklungsprozess | Ausfüllhinweise | Zeitbedarf (Min) | Testgüte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARIF[220] | Critical Appraisal Checklist | EN | | | 14 | CL | | | | | | ◐ | | |
| Assendelft et al.[9] | | EN | | | 14 | SK | | 0-100 | | | | ● | | |
| Barnes & Bero[13] | | EN | Oxman 1988 | | 12 | SK | | 0-1 | | | | | | |
| CEBM[219] | | EN | | | 5 | CL | | | | | | ● | | |
| CEBMH[36] | | EN | | | 9 | CL | | | | | | | | |
| Ekkernkamp et al.[69] | GSWG Checkliste 1a | DE | | | 40 | CL | | | | | | ◐ | | |
| Ekkernkamp et al.[69] | GSWG Checkliste 1b | DE | | | 22 | CL | | | | | | ◐ | | |
| LBI[133] | | DE | | | 10 | CL | QGB | | | ● | | ● | | |
| Oxman et al.[170] | QQAQ | EN | | | 10 | CL | | | | ● | ● | | | |
| PHRU[159] | CASP | EN | Guyatt 1994 | | 10 | CL | | | | | | ● | | |
| Rychetnik & Frommer[185] | | | EN | | | 6 | CL | | | | | | ◐ | |
| Sackett[187] | | EN | Oxman o. J. | | 8 | CL | | | | | | ● | | |
| Shea et al.[196] | AMSTAR | EN | | | 11 | CL | | | | | ● | ◐ | | |
| SIGN 50[194] | | EN | Liddle 1996 | | 5 | CL | QGB | | | | | ● | | |
| Vigna-Taglianti et al.[230] | | EN | Moher 1999 | | 80 | SK | | 0-50 | | | ● | ◐ | | |
| Anzahl (n) | | EN: n = 12 (80 %) | n = 5 (33 %) | n = 0 (0 %) | 5-80 Items | CL: n = 12 (80 %) | QKB: n = 0 (0 %) | | n = 0 (0 %) | n = 2 (13 %) | n = 3 (20 %) | ● n = 6 (40 %) | n = 0 (0 %) | n = 0 (0 %) |
| Anteil (%) | | DE: n = 3 (20 %) | | | | KO: n = 0 (0 %) | QGB: n = 2 (13 %) | | | | | ◐ n = 6 (40 %) | | |
| | | | | | | SK: n = 3 (20 %) | | | | | | | | |

AMSTAR = Assessment of multiple systematic reviews. CASP = Critical Appraisal Skills Programme. CL = Checkliste. DE = Deutsch. EN = Englisch. GSWG = German Scientific Working Group. HTA = Health Technology Assessment. KO = Komponentensystem. QGB = Qualitative Gesamtbewertung. QKB = Qualitative Komponentenbewertung. QQAQ = Overview Quality Assessment Questionnaire. SK = Skala.

* Bei Skalen. Ausfüllhinweise: ◐ = Kurz. ● = Ausführlich.

**Tabelle 40: Formale Charakteristika von Instrumenten für Interventionsstudien**

| | Name/ Abkürzung des Instruments | Sprache | Modifikation von | Spezifisches Instrument | Items | Art des Instru-ments | Qualitative Bewertung | Range* | Cutpoint* | Definition von Qualität | Entwicklungs-prozess | Ausfüllhinweise | Zeitbedarf (Min) | Testgüte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ah-See & Molony[3] | | EN | Begg et al. 1996 | ● | 12 | SK | | 0-12 | | | | | | |
| Andrew et al.[7] | | EN | | ● | 11 | SK | | 0-22 | | | | | | |
| Assendelft et al.[8] | | EN | | ● | 16 | SK | | 0-100 | | | | ● | | |
| Balas et al.[11] | | EN | | | 20 | SK | | 0-100 | | | ● | ◐ | 18 | ● |
| Balk et al.[12] | | EN | | | 27 | CL | | | | | ● | ◐ | | |
| Bath et al.[15] | | EN | | ● | 7 | SK | | 0-16 | | ● | | ◐ | | |
| Bizzini et al.[18] | | EN | Clarke & Oxman 2001 | ● | 14 | SK | | 0-100 | 50 | | ● | ● | | ● |
| Borsody & Yamada[20] | | EN | Moher 1995 | ● | 7 | SK | | 0-7 | | ● | | ◐ | | ● |
| Brown[26] | | EN | Sackett & Haynes 1976 | | 6 | SK | | 0-21 | | | ● | ◐ | | |
| CEBM[217] | | EN | | | 7 | CL | | | | | | ● | | |
| CEBMH[37] | | EN | | | 6 | CL | | | | | | | | |
| Chalmers et al.[38] | | EN | | | | SK | | 0-9 | | | | ● | | |
| Chalmers et al.[39] | | EN | | | 30 | SK | | 0-1 | | | | ● | | ● |
| Cho & Bero[44] | | EN | | | 24 | SK | | 0-1 | | ● | ● | Nicht publiziert | | |
| Colditz et al.[49] | | EN | DerSimonian et al. 1982 | | 12 | SK | | | | | | ◐ | | |
| Cook et al.[50] | | EN | | ● | 5 | SK | | 0-10 | | | | | | |
| Dardennes et al.[53] | | EN | King 1990 | ● | 16 | SK | | 0-16 | | | | | | |
| de Vet et al.[54] | | EN | | ● | 51 | SK | | 0-80 | | | | | | ● |
| Delfini Group[57] | | EN | | | 22 | CL | | | | | | | | |
| Downs & Black[60] | | EN | | | 27 | SK | | 0-32 | | | ● | ● | 20 | ● |
| Earle & Hébert[66] | | EN | | ● | 9 | SK | | | | | | | | |

**Tabelle 40: Formale Charakteristika von Instrumenten für Interventionsstudien – Fortsetzung**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ekkernkamp et al.[69] | GSWG CL | DE | | | 30 | CL | | | | | | ◐ | | |
| Forsetlund & Rainer[72] | | EN | Cochrane | ● | 9 | KO | QGB | | | | | ◐ | | |
| Geng[75] | | EN | | | | CL | QGB | | | | | | | |
| Graf et al.[77] | MQAS | EN | Begg et al. 1996 | ● | 11 | SK | | 0-57 | | | ● | ● | | |
| Gupta et al.[79] | | EN | Begg et al. 1996 | | 12 | SK | | 0-20 | 11 | | | ◐ | | |
| Hadorn et al.[82] | | EN | | | 8 | KO | QKB | | | | ● | ◐ | | |
| Hammerschlag & Morris[83] | | EN | | ● | 25 | KO | QKB | | | | | | | |
| Heneghan et al.[88] | | EN | Chalmers 1981 | ● | 15 | SK | | 0-15 | | | | | | ● |
| Hettinga et al.[90] | | EN | van Tulder et al. 1997 | ● | 10 | SK | | 0-10 | | | | | | |
| Heyland et al.[91] | | EN | | ● | 9 | SK | | 0-13 | | | | | | ● |
| Higgins & Green[92] | | EN | Risk of bias tool | | 6 | KO | QKB | | | ● | ● | ● | | ● |
| Hill et al.[93] | | EN | Jadad et al. 1996 | | 6 | KO | QKB | | | | | ● | | |
| Huwiler-Müntener et al.[99] | | EN | | | 3 | KO | QKB | | | | | ◐ | | |
| Imperiale & McCullough[100] | | EN | | ● | 5 | SK | | 0-5 | | | | | | ● |
| IQWiQ[103] | | DE | | | 5 | KO | QGB | | | | | | | |
| Jadad et al. (3 Items)[106] | | EN | | | 3 | SK | | 0-5 | | ● | ● | ● | 10 | ● |
| Jadad et al. (6 Items)[106] | | EN | | | 6 | SK | | 0-8 | | ● | ● | ● | | ● |
| Jonas & Linde[107] | | EN | | ● | 60 | CL | | | | | | | | |
| Kleijnen et al.[113] | | EN | | | 7 | SK | | 0-100 | ≥55 | | | ◐ | | |
| Kmet et al.[114] | Quality Scoring of Quantitative Studies | EN | | | 14 | SK | | 0-28 | | ● | ● | ● | | |
| Koes et al.[115] | | EN | ter Riet 1990 | ● | 17 | SK | | 0-100 | 50 | | | ◐ | | |
| Kwakkel et al.[120] | | EN | | ● | 16 | SK | | 0-16 | | | | ◐ | | ● |
| Lamont[121] | | EN | | ● | 40 | KO | QKB | | | ● | ● | | | |
| LBI[133] | | DE | | | 9 | CL | QGB | | | ● | | ● | | |
| Levine[123] | TAPS | EN | | | 29 | SK | | 0-100 | | | | ◐ | | |

**Tabelle 40: Formale Charakteristika von Instrumenten für Interventionsstudien – Fortsetzung**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Linde et al.[127] | | EN | | ● | 10 | CL | | | | | | | | |
| MacLehose et al.[134] | | EN | Downs & Black 1998 | | 44 | SK | | 0-18 | 9 | | ● | ● | | ● |
| MacMillan et al.[135] | | EN | Chalmers 1981 | | 9 | SK | | 0-25 | | | | ◐ | | |
| Moncrieff & Drummond[149] | | EN | | ● | 30 | SK | | 0-60 | | ● | | ◐ | | ● |
| Moncrieff et al.[148] | | EN | | ● | 23 | SK | | 0-46 | | | ● | ● | | ● |
| Morley et al.[150] | | EN | | ● | 19 | SK | | 0-28,5 | | | | | | |
| Moseley et al.[151] | PEDro Scale | EN | | ● | 11 | SK | | 0-10 | | | | | | ● |
| Ogilvie-Harris & Gilbart[167] | | EN | Weiler 1992 | ● | 10 | SK | | 0-10 | | | | ◐ | | |
| Onghena & Van Houdenhove[169] | | EN | | ● | 10 | SK | | 0-10 | | | | | | |
| Petrak et al.[171] | | DE | | ● | 23 | SK | | 0-109 | | | | | | |
| PHRU[158] | CASP | EN | Guyatt 1994 | | 10 | CL | | | | | | ● | | |
| Prendiville et al.[177] | | EN | | | 3 | KO | QKB | | | | | ◐ | | |
| Pua et al.[178] | | EN | DerSimonian et al. 1982 | | 10 | SK | | 0-10 | | | | ◐ | | |
| Reisch et al.[181] | | EN | | | 81 | CL | | | | | | ● | | |
| Robeer et al.[183] | | EN | Bouter 1994 | ● | | SK | | 0-100 | 60 | | | ◐ | | |
| Rochon[184] | | EN | Antczak 1986 | | 14 | SK | | 0-1 | | | | ● | | |
| Rychetnik & Frommer[185] | | EN | | | 5 | CL | | | | | | ◐ | | |
| Sackett[187] | | EN | | | 6 | CL | | | | | | | | |
| SIGN 50[194] | | EN | Liddle 1996 | | 10 | CL | QGB | | | | | ● | | |
| Sindhu et al.[198] | | EN | | | 53 | SK | | 0-100 | | | ● | | | |
| Slim et al.[199] | | EN | Hall 1996 | ● | 11 | SK | | 0-22 | | | | ◐ | | ● |
| Smeenk et al.[201] | | EN | | ● | 34 | SK | | 0-100 | | | | | | |
| Smith et al.[202] | | EN | | ● | 8 | SK | | 0-40 | | | | | | |
| Smith et al.[203] | OPVS | EN | | ● | 8 | SK | | 0-16 | | ● | | ● | | |
| Spitzer et al.[204] | | EN | | | 32 | CL | | | | | | | | |
| Staiger et al.[206] | | EN | van Tulder 1997a | ● | 22 | SK | | 0-22 | | | | | | ● |

**Tabelle 40: Formale Charakteristika von Instrumenten für Interventionsstudien – Fortsetzung**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Stieb et al.[208] | | EN | | ● | 6 | CL | | | | | | ◐ | | |
| ter Riet et al.[211] | | EN | | ● | 18 | SK | | 0-100 | 50 | | | ◐ | | |
| Thomas et al.[213] | EPHPP | EN | | | 22 | KO | QKB, QGB | | | | | ● | | ● |
| van Nieuwenhoven et al.[224] | | EN | Cook 1991 | ● | 6 | SK | | 0-13 | ≥ 7 | | | ◐ | | |
| Verhagen et al.[227] | Delphi list | EN | | | 9 | CL | | | | ● | ● | | | |
| Yates et al.[243] | | EN | | | 20 | SK | | 0-26 | | | ● | ● | | ● |
| Yuen & Pope[244] | | EN | Campbell et al. 2004 | | 5 | SK | | 0-5 | 4 | | | ● | | |
| Zaza et al.[245] | | EN | | | 23 | CL | | | | | | | | |
| Anteil (%) | | DE: n = 4 (5 %) | | | | KO: n = 10 (13 %) | QGB: n = 6 (8 %) | | | | | ◐ n = 26 (33 %) | | |
| Anzahl (n) | | EN: n = 76 (95 %) | n = 26 (33 %) | n = 38 (48 %) | 3-81 Items | CL: n = 18 (23 %) | QKB: n = 8 (10 %) | | n = 9 (11 %) | n = 13 (16 %) | n = 18 (23 %) | ● n = 23 (29 %) | n = 3 (4 %) | n = 21 (26 %) |

CASP = Critical Appraisal Skills Programme. CL = Checkliste. DE = Deutsch. EN = Englisch. EPHPP = Effective Public Health Practice Project. GSWG = German Scientific Working Group. KO = Komponentensystem. MQAS = Methodological Quality Assessment Score. OPVS = Oxford Pain Validity Scale. PEDro = Physiotherapy Evidence Database Scale. QGB = Qualitative Gesamtbewertung. QKB = Qualitative Komponentenbewertung. SK = Skala. TAPS = Trial Assessment Procedure Scale.

Bei Skalen. Ausfüllhinweise: ◐ = Kurz. ● = Ausführlich.

**Tabelle 41: Formale Charakteristika von Instrumenten für Beobachtungsstudien**

| | Name/ Abkürzung des Instruments | Sprache | Modifikation von | Spezifisches Instrument | Items | Art des Instruments | Qualitative Bewertung | Range* | Cutpoint* | Definition von Qualität | Entwicklungs-prozess | Ausfüllhinweise | Zeitbedarf (Min) | Testgüte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Andrew et al.[7] | | EN | | ● | 11 | SK | | 0-22 | | | | | | |
| Brown[26] | | EN | Sackett & Haynes 1976 | | 6 | SK | | 0-21 | | | ● | ◐ | | |
| Carson et al.[30] | | EN | McMaster University 1981 | | 10 | SK | | 0-1 | | | | ◐ | | |
| CEBMH[35] | | EN | | | 5 | CL | | | | | | | | |
| Cho & Bero[44] | | EN | | | 24 | SK | | 0-1 | | ● | ● | Nicht publiziert | | |
| Colditz et al.[49] | | EN | DerSimonian et al. 1982 | | 12 | SK | | | | | | ◐ | | |
| Downs & Black[60] | | EN | | | 27 | SK | | 0-32 | | | ● | ● | 20 | ● |
| Earle & Hébert[66] | | EN | | ● | 9 | SK | | | | | | | | |
| Ekkernkamp et al.[69] | GSWG Checkliste 2a | DE | | | 30 | CL | | | | | | ◐ | | |
| Geng[75] | | EN | | | 39 | CL | QGB | | | | | | | |
| Hadorn et al.[82] | | EN | | | 8 | KO | QKB | | | | ● | ◐ | | |
| Jonas & Linde[107] | | EN | | | ● | 60 | CL | | | | | | | |
| Kmet et al.[114] | Quality Scoring of Quantitative Studies | EN | | | 14 | SK | | 0-28 | | ● | ● | ● | | |
| LBI[133] | | DE | Deeks et al. 2003 | | 11 | CL | QGB | | | ● | | ● | | |
| Linde et al.[127] | | EN | | ● | 10 | CL | | | | | | | | |
| MacLehose et al.[134] | | EN | Downs & Black 1998 | | 44 | SK | | 0-18 | 9 | | ● | ● | | ● |
| Nguyen et al.[162] | | EN | | | 18 | SK | | 0-100 | | | | ◐ | | |
| PHRU (FKS)[160] | CASP | EN | | | 11 | CL | | | | | | ● | | |
| PHRU (KS)[157] | CASP | EN | | | 14 | CL | | | | | | ● | | |
| Rychetnik & Frommer[185] | | EN | Liddle 1996 | | 19 | CL | QGB | | | | | ◐ | | |

**Tabelle 41: Formale Charakteristika von Instrumenten für Beobachtungsstudien – Fortsetzung**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SIGN 50 (FKS)[194] | | EN | | | 11 | CL | | | | | | ● | | |
| SIGN 50 (KS)[194] | | EN | Liddle 1996 | | 14 | CL | QGB | | | | | ● | | |
| Slim et al.[200] | MINORS | EN | | | 12 | SK | | 0-24 | | | ● | ◐ | | |
| Spitzer et al.[204] | | EN | | | 32 | CL | | | | | | | | |
| Spooner et al.[205] | EQUATDUR-2 | EN | | ● | 5 | SK | | 0-10 | | ● | ● | ◐ | 10 | ● |
| Stieb et al.[208] | | EN | | ● | 5 | CL | | | | | | ◐ | | |
| Thomas et al.[213] | EPHPP | EN | | | 22 | KO | QKB, QGB | | | | | ● | | ● |
| Wells et al., FKS[234] | NOS | EN | | | 7 | CL | | | | | | ● | | |
| Wells et al., KS[234] | NOS | EN | | | 8 | CL | | | | | | ◐ | | |
| Wong et al.[241] | QATSO | EN | | ● | 5 | SK | | 0-1 | | | | ◐ | 10 | |
| Anzahl (n) | | EN: n = 28 (93 %) | n = 7 (23 %) | n = 7 (23 %) | 5-60 Items | CL: n = 15 (50 %) | QGB: n = 5 (17 %) | | n = 1 (3 %) | n = 4 (13 %) | n = 8 (27 %) | ● n = 10 (33 %) | n = 3 (10 %) | n = 4 (13 %) |
| Anteil (%) | | DE: n = 2 (7 %) | | | | KO: n = 2 (7 %) | QKB: n = 2 (7 %) | | | | | ◐ n = 12 (40 %) | | |
| | | | | | | SK n = 13 (43 %) | | | | | | | | |

CASP = Critical Appraisal Skills Programme. CL = Checkliste. DE = Deutsch. EN = Englisch. EPHPP = Effective Public Health Practice Project. EQUATDUR-2 = Edmonton Quality Assessment Tool for Drug Utilization Reviews. FKS = Fall-Kontrollstudie. GSWG = German Scientific Working Group. KO = Komponentensystem. KS = Kohortenstudie. MINORS = Methodological index für non-randomized studies. NOS = Newcastle Ottawa Scale. QATSO = Quality assessment tool for systematic reviews of observational studies. QGB = Qualitative Gesamtbewertung. QKB = Qualitative Komponentenbewertung. SK = Skala.

Bei Skalen. Ausfüllhinweise: ◐ = Kurz. ● = Ausführlich.

**Tabelle 42: Formale Charakteristika von Instrumenten für Diagnosestudien**

| | Name/ Abkürzung des Instruments | Sprache | Modifikation von | Spezifisches Instrument | Items | Art des Instruments | Qualitative Bewertung | Range* | Cutpoint* | Definition von Qualität | Entwicklungs-prozess | Ausfüll-hinweise | Zeitbedarf (Min) | Testgüte |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CEBM [218] | | EN | | | 5 | CL | | | | | | ● | | |
| CEBMH[33] | | EN | | | 6 | CL | | | | | | | | |
| de Vet et al.[55] | | EN | Mulrow 1989 | | 17 | CL | | | | ● | | ● | | |
| Ekkernkamp et al.[69] | GSWG Checkliste 2b | DE | | | 14 | CL | | | | | | ◐ | | |
| Hoffmann et al.[95] | | EN | | ● | 6 | KO | QKB | | | | | ◐ | | ● |
| Huebner et al.[98] | | EN | | ● | 35 | CL | QKB | | | | | ◐ | | |
| IQWIQ[102] | | DE | | | 14 | CL | | | | | | | | |
| LBI[133] | | DE | Whiting et al. 2003 | | 9 | CL | QGB | | | ● | | ● | | |
| Mullins et al.[154] | | EN | | ● | 11 | CL | | | | | | ● | | |
| Mulrow et al.[155] | | EN | | | 19 | SK | | -100-100 | | | ● | ◐ | | |
| PHRU[161] | CASP | EN | Jaeschke 1994a | | 12 | CL | | | | | | ● | | |
| Sackett[187] | | EN | | | 6 | CL | | | | | | | | |
| SIGN 50[194] | | EN | Liddle 1996 | | 14 | CL | QGB | | | | | ● | | |
| van den Hoogen et al.[221] | | EN | | | 9 | SK | | 0-100 | | | ● | ● | | |
| Varela-Lema & Ruano-Ravina[226] | | EN | | ● | 10 | SK | | 0-100 | | | ● | ● | | ● |
| Whiting et al.[238] | QUADAS | EN | | | 14 | CL | | | | ● | ● | ● | | ● |
| Wurff et al.[223] | | EN | | ● | 20 | SK | | 0-100 | 50 | | ● | | | ● |
| Anzahl (n) | | EN: n = 14 (82 %) | n = 4 (24 %) | n = 5 (29 %) | 5-35 Items | CL: n = 12 (71 %) | QKB: n = 2 (12 %) | | n = 1 (6 %) | n = 3 (18 %) | n = 5 (29 %) | ● n = 9 (53 %) | n = 0 (0 %) | n = 4 (24 %) |
| Anteil (%) | | DE: n = 3 (18 %) | | | | KO: n = 1 (6 %) | QGB: n = 2 (12 %) | | | | | ◐ n = 4 (24 %) | | |

CASP = Critical Appraisal Skills Programme. CL = Checkliste. DE = Deutsch. EN = Englisch. GSWG = German Scientific Working Group. KO = Komponentensystem. QGB = Qualitative Gesamt-bewertung. QKB = Qualitative Komponentenbewertung. QUADAS = Quality assessment of diagnostic accuracy studies. SC = Skala.

Bei Skalen. Ausfüllhinweise: ◐ = Kurz. ● = Ausführlich.

## 8.15 Inhaltliche Charakteristika der QBI (Effektivität)

**Tabelle 43: Inhaltliche Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen, Teil 1**

| | Studienfrage | Ein- und Ausschluss-kriterien | | Literaturrecherche und -auswahl | | | | | | | | | Datenextraktion | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fragestellung | Kriterien a priori | Kriterien angemessen | Relevante Datenbanken | Weitere Datenquellen | Schlagworte/Freitext | Synonyme | Restriktionen | Suchstrategie | ≥ 2 Reviewer | Ausschlüsse | Ausreichend detailliert | Interventionen/ Expositionen | Outcomes | ≥ 2 Reviewer | Verblindung | Messung der Übereinstimmung | Ausreichend detailliert |
| | BQ | IV | BQ | IV | IV | IV | IV | IV | BQ | IV | BQ | BQ | IV | IV | IV | IV | BQ | BQ |
| ARIF[220] | ● | | ● | ● | ● | | | ● | ● | | | ● | | | | | | ● |
| Assendelft et al.[9] | | ● | | ● | ● | | | | ● | | | | ● | ● | | | | |
| Barnes & Bero[13] | ● | | ● | | | | | | ● | | | ● | | | | | | |
| CEBM[219] | ● | ● | ● | ● | ● | ● | | ● | | | | | | | | | | |
| CEBMH[36] | ● | | | ● | ● | | | ● | | | | | | | | | | |
| Ekkernkamp et al.[69] | ● | ● | | | | | | | ● | | ● | | | | ● | | | ● |
| Ekkernkamp et al.[69] | ● | ● | | | | | | | ● | | ● | | | | ● | | | ● |
| LBI[133] | ● | ● | ● | ● | ● | | | ● | | | | | | | | | | |
| Oxman et al.[170] | | | | ● | ● | | | | ● | | | | | | | | | |
| PHRU[159] | ● | | | ● | ● | | | ● | | | | | | ● | | | | |
| Rychetnik & Frommer[185] | ● | ● | | ● | ● | | | ● | | | ● | | | | | | | |
| Sackett[187] | ● | ● | | ● | ● | | | | ● | ● | ● | ● | | | ● | | ● | |
| Shea et al.[196] | ● | ● | | ● | ● | | | | ● | ● | ● | | ● | ● | ● | | | |
| SIGN 50[194] | ● | | | ● | ● | | | | | | | | | | | | | |
| Vigna-Taglianti et al.[230] | ● | | | | ● | | | ● | ● | | ● | | | | | | | |
| | | | | | | | | | | | | | | | | | | |
| Summe [156] | 13 | 8 | 4 | 11 | 12 | 1 | 0 | 7 | 9 | 2 | 6 | 3 | 2 | 3 | 4 | 0 | 1 | 3 |
| Summe (%) | 87 % | 53 % | 27 % | 73 % | 80 % | 7 % | 0 % | 47 % | 60 % | 13 % | 40 % | 20 % | 13 % | 20 % | 27 % | 0 % | 7 % | 20 % |

BQ = Berichtsqualität. HTA = Health Technology Assessment. IV = Interne Validität.

**Tabelle 44: Inhaltliche Charakteristika von Instrumenten für systematische Reviews, HTA und Metaanalysen, Teil 2**

| | Qualitätsbewertung | | | | | Datensynthese und -analyse | | | Ergebnisse | Diskussion | | Auftraggeber |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bewertungsmethode | ≥ 2 Reviewer | Verblindung | Bewertung der Übereinstimmung | Integration der Ergebnisse | Angemessene Synthese | Robustheit/ Heterogenität | Schlüsselelemente von Primärstudien | Zusammenfassung und Präzision | Schlussfolgerung basieren auf Ergebnissen | Berücksichtigung von Bias und Limitationen | Finanzierung |
| | IV | IV | IV | BQ | IV | IV | IV | BQ | BQ | BQ | BQ | IV |
| ARIF[220] | ● | ● | | | | ● | ● | | | | ● | |
| Assendelft et al.[9] | ● | ● | ● | ● | | ● | ● | ● | | ● | ● | |
| Barnes & Bero[13] | ● | | | | | ● | | | ● | ● | ● | |
| CEBM[219] | ● | | | | | | ● | | ● | | | |
| CEBMH[36] | ● | | | | | | ● | | ● | | | |
| Ekkernkamp et al.[69] | | ● | | | | | | | | ● | ● | |
| Ekkernkamp et al.[69] | | ● | | | | | ● | | | ● | ● | |
| LBI[133] | ● | ● | | | ● | | ● | | | | | |
| Oxman et al.[170] | ● | | | | | ● | | | | ● | | |
| PHRU[159] | ● | ● | | | | ● | ● | | ● | | | |
| Rychetnik & Frommer[185] | ● | | | ● | | ● | ● | | | ● | | |
| Sackett[187] | ● | ● | | ● | | ● | ● | ● | | ● | | |
| Shea et al.[196] | | | | | ● | | ● | | | | ● | ● |
| SIGN 50[194] | ● | | | | ● | | ● | | | | | |
| Vigna-Taglianti et al.[230] | ● | ● | | | | | ● | | | | ● | |
| | | | | | | | | | | | | |
| Summe[156] | 12 | 8 | 1 | 3 | 3 | 7 | 12 | 2 | 4 | 7 | 7 | 1 |
| Summe (%) | 80 % | 53 % | 7 % | 20 % | 20 % | 47 % | 80 % | 13 % | 27 % | 47 % | 47 % | 7 % |

BQ = Berichtsqualität. HTA = Health Technology Assessment. IV = Interne Validität.

**Tabelle 45: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 1**

| | Studienfrage | Studienpopulation | | | Randomisierung | | | Verblindung | | | | Interventionen | | | | | | | | Outcomes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fragestellung | Beschreibung | Ein-/Ausschlusskriterien | Stichprobengröße | Methode der Randomisierung | Gruppenzuweisung geheim | Vergleichbarkeit der Gruppen | Studienteilnehmer | Erheber des Outcomes | Übriges Studienpersonal | Ausreichend | Beschreibung | Kontrollgruppe | Behandlungsgleichheit | Placebo vergleichbar mit Verum | Co-Interventionen vermieden | Co-Interventionen beschrieben | Kontamination | Compliance | Outcomes präzise definiert | Valide Methoden | Reliable Methoden | Follow-up-Länge | Follow-up zeitgleich |
| | BQ | BQ | BQ | IV | IV | IV | IV | IV | IV | IV | IV | BQ | IV | IV | IV | IV | BQ | IV | IV | BQ | IV | IV | IV | IV |
| Ah-See & Molony[3] | ● | | ● | ● | ● | | | ● | ● | | | | | | | | | | | ● | | | | |
| Andrew et al.[7] | ● | ● | ● | | ● | | | ● | ● | | | ● | | | | | | | | ● | | | | |
| Assendelft et al.[8] | | | ● | | ● | ● | ● | ● | ● | | ● | ● | | | | ● | ● | | | | | | ● | ● |
| Balas et al.[11] | | | ● | ● | ● | | ● | ● | ● | ● | | ● | | | | | | | | ● | | | | |
| Balk et al.[12] | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | | | |
| Bath et al.[15] | | | | ● | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | | | |
| Bizzini et al.[18] | | | ● | ● | ● | | ● | | ● | | | ● | | | | ● | | | | ● | ● | ● | ● | |
| Borsody & Yamada[20] | | ● | | | ● | ● | ● | ● | ● | ● | ● | | | | ● | | | ● | | | | | | |
| Brown[26] | | ● | ● | | | | | ● | | | | ● | | | | ● | ● | | | ● | | | | |
| CEBM[217] | | | | | ● | | ● | ● | ● | | | | | ● | | | | | | | | | ● | |
| CEBMH[37] | | | | | | ● | ● | ● | ● | | | | | ● | | | | | | | | | | |
| Chalmers et al.[38] | | | | | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | | | |
| Chalmers et al.[39] | | ● | ● | ● | | ● | ● | ● | ● | ● | ● | ● | | | ● | | ● | | ● | | | | | |
| Cho & Bero[44] | ● | | ● | ● | ● | | | ● | ● | | | | | | | | | | | | | | | |
| Colditz et al.[49] | | | ● | | | | | ● | ● | | | | | | | | | | | | | | | |

**Tabelle 45: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 1 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cook et al.[50] | | | | | ● | ● | ● | | ● | ● | | | | | | | | | | ● | | | | |
| Dardennes et al.[53] | | ● | | ● | | | | | | | | | | | | ● | | | ● | | | | ● | |
| de Vet et al.[54] | | | ● | | ● | ● | ● | ● | ● | ● | ● | ● | | | | ● | ● | | ● | | | | ● | |
| Delfini Group[57] | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | | ● | ● | | ● | | ● | ● | ● | | | | ● |
| Downs & Black[60] | ● | ● | ● | ● | ● | ● | | ● | ● | | | ● | | | | | | ● | ● | ● | ● | ● | | ● |
| Earle & Hébert[66] | | ● | | ● | ● | | | | | | | ● | | | | | | | | ● | | | | |
| Ekkernkamp et al.[69] | | | ● | | ● | ● | ● | | ● | | | ● | | ● | | | | | | | ● | ● | | |
| Forsetlund & Rainer[72] | | | | | ● | ● | ● | | ● | ● | | ● | | | | | | | | ● | | ● | | |
| Geng[75] | ● | | ● | ● | ● | | ● | | ● | | | ● | | | | | | | | | | | | |
| Graf et al.[77] | | | ● | ● | ● | ● | | ● | ● | ● | | ● | | | | | ● | | | ● | | | ● | |
| Gupta et al.[79] | ● | ● | ● | ● | ● | | ● | ● | ● | | | ● | | | | | | | ● | ● | | | | |
| Hadorn et al.[82] | | | ● | | ● | | ● | ● | ● | | | ● | | | | ● | | ● | | | | | | |
| Hammerschlag & Morris[83] | | ● | ● | ● | ● | | | | ● | | | ● | | | | | ● | | ● | ● | | | | |
| Heneghan et al.[88] | | | ● | | ● | ● | ● | | ● | | | ● | | | | | | ● | | ● | | | | |
| Hettinga et al.[90] | | | | | | ● | ● | ● | ● | ● | | | | | | ● | ● | | ● | | | | | ● |
| Heyland et al.[91] | | | | | ● | | ● | ● | ● | | | ● | | | | | ● | ● | | | | | | |
| Higgins & Green[92] | | | | | ● | ● | ● | ● | ● | ● | ● | | | | | | | | | ● | | | | |
| Hill et al.[93] | | | | | ● | ● | | ● | ● | | | | | | ● | | | | | | | | | |
| Huwiler-Müntener et al.[99] | ● | ● | ● | ● | ● | ● | | | | | | ● | | | | | | | | ● | | | | |
| Imperiale & McCullough[100] | | | ● | | | | ● | | | | | | | | | | ● | | | | | | | |
| IQWiQ[103] | | | | ● | ● | ● | | ● | ● | | | | | | | | | | | | | | | |
| Jadad et al. (3 Items)[106] | | | | | ● | | | ● | ● | | | | | | ● | | | | | | | | | |
| Jadad et al.(6 Items)[106] | | | ● | | ● | | | ● | ● | | | | | | ● | | | | | | | | | |
| Jonas & Linde[107] | | | ● | ● | | ● | ● | ● | ● | ● | ● | ● | ● | | | | ● | ● | ● | ● | | ● | | ● |
| Kleijnen et al.[113] | | ● | | | ● | | | ● | ● | | | ● | | | ● | | | | | ● | | | | |
| Kmet et al.[114] | ● | ● | ● | ● | ● | | ● | ● | ● | | | | | | | | | | | ● | | | | |

**Tabelle 45: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 1 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Koes et al.[115] | | | ● | | ● | | ● | ● | ● | | ● | ● | | | | ● | | | | ● | | | | |
| Kwakkel et al.[120] | | | | | ● | ● | ● | ● | ● | | | ● | | | | ● | ● | | | | | | ● | ● |
| Lamont[121] | | | ● | ● | ● | ● | ● | ● | ● | ● | | ● | | | | | ● | | | ● | | | | |
| LBI[133] | | | | ● | ● | ● | ● | ● | ● | ● | | | | | | | | | | | | | | |
| Levine[123] | ● | | ● | ● | | | | ● | ● | ● | | ● | | | ● | ● | ● | ● | ● | | | | | |
| Linde et al.[127] | | ● | | | | | | | | | | | | | | | ● | | | | | | | |
| MacLehose et al.[134] | ● | ● | ● | ● | ● | ● | | ● | ● | ● | | ● | ● | | | | | | ● | ● | ● | ● | ● | |
| MacMillan et al.[135] | | | ● | | ● | | ● | | ● | | | | | | | | | | | | | | | |
| Moncrieff & Drummond[149] | | ● | ● | ● | ● | | | ● | ● | | ● | ● | | | | | ● | | | ● | | | | |
| Moncrieff et al.[148] | | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | | | | | ● | | ● | | | | | |
| Morley et al.[150] | | ● | | | | | | | ● | | | ● | | | | | ● | | | | | | | |
| Moseley et al.[151] | | | ● | | | ● | ● | ● | ● | ● | | | | | | | | | | | | | | |
| Ogilvie-Harris & Gilbart[167] | | | | | ● | | | ● | ● | | | | ● | | | | | | | | | | | |
| Onghena & Van Houdenhove[169] | | | | ● | | | | | | | | ● | | | | | | | ● | ● | | | | |
| Petrak et al.[171] | | ● | | ● | ● | | ● | ● | | | | ● | | | | | | | | | | | ● | |
| PHRU[158] | ● | | | ● | ● | | ● | ● | ● | ● | | | | ● | ● | | | | | | | | | |
| Prendiville et al.[177] | | | | | ● | ● | | ● | ● | | | | | | ● | | | | | ● | | | | ● |
| Pua et al.[178] | | | ● | ● | ● | | | | ● | | | | | | | | | | | | ● | ● | | |
| Reisch et al.[181] | ● | ● | ● | ● | ● | | ● | ● | ● | ● | | ● | | ● | | | | | ● | ● | ● | ● | | ● |
| Robeer et al.[183] | | | | | ● | | ● | ● | ● | | | ● | | | | ● | | | | | | | | |
| Rochon[184] | | | | ● | ● | ● | ● | ● | ● | | | | | | ● | | | | ● | ● | | | | |
| Rychetnik & Frommer[185] | | | | | ● | ● | ● | | | | | | | ● | | ● | | | | | | | | ● |
| Sackett[187] | | | | | ● | | | | | | | | | | | | | | | | ● | | | |
| SIGN 50[194] | ● | | ● | | ● | ● | ● | ● | ● | | | | | ● | | | ● | | | ● | | ● | | |
| Sindhu et al.[198] | ● | ● | | ● | ● | ● | ● | ● | ● | ● | | | | ● | | | | ● | | | | | | |
| Slim et al.[199] | ● | | ● | ● | | | ● | | ● | | | ● | | | | | | | | ● | | | ● | |

**Tabelle 45: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 1 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Smeenk et al.[201] | | ● | ● | | ● | | ● | | ● | | | ● | ● | | | ● | | | | ● | | | | |
| Smith et al.[202] | | | | | | | ● | ● | ● | | | | ● | ● | | | | | ● | | | | | |
| Smith et al.[203] | | | | | | | | ● | ● | | | | | | | | | | | ● | | | ● | |
| Spitzer et al.[204] | | | ● | ● | ● | | ● | | ● | | | ● | | | | | | | | ● | | | | |
| Staiger et al.[206] | | | ● | | | ● | ● | ● | ● | ● | | ● | | | | ● | | | ● | ● | | | | |
| Stieb et al.[208] | | | | | ● | | ● | | ● | ● | | | | | | | | | | | | | | ● |
| ter Riet et al.[211] | | | | | | | ● | ● | ● | | | ● | | | | | | | | | | | | |
| Thomas et al.[213] | | | | | ● | | ● | ● | ● | | | | | | | ● | | ● | ● | ● | | | | |
| van Nieuwen-hoven et al.[224] | | | | | ● | ● | ● | | ● | ● | | | | | | | | | | | | | | |
| Verhagen et al.[227] | | | ● | | | ● | ● | ● | ● | ● | | | | | | | | | | ● | | | | |
| Yates et al.[243] | | ● | ● | ● | ● | ● | ● | | ● | | | ● | | | | | | | | | | | | |
| Yuen & Pope[244] | | | | ● | | | | ● | ● | | | | | | | | | | | | | | | |
| Zaza et al.[245] | | ● | ● | | | | ● | | | | | ● | | | | | | | | ● | | ● | | ● |
| | | | | | | | | | | | | | | | | | | | | | | | | |
| Summe | 17 | 25 | 43 | 36 | 59 | 36 | 52 | 55 | 69 | 26 | 9 | 40 | 6 | 10 | 10 | 16 | 18 | 10 | 18 | 36 | 7 | 10 | 11 | 11 |
| Summe (%) | 21 % | 31 % | 54 % | 45 % | 74 % | 45 % | 65 % | 69 % | 86 % | 33 % | 11 % | 50 % | 8 % | 13 % | 13 % | 20 % | 23 % | 13 % | 23 % | 45 % | 9 % | 13 % | 14 % | 14 % |

BQ = Berichtsqualität. EV = Externe Validität. IV = Interne Validität.

**Tabelle 46: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 2**

| | Statistische Analyse | | | | | | | Ergebnisse | | | | | | Diskussion | | Externe Validität | | | Finanzierung |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Angemessene Analyse | Multiples Testen | Intention-to-treat Analyse | Fehlenden Werte | Berücksichtigung von Confounding | Bewertung von Confounding | Bewertung von Heterogenität | Punktschätzer und Präzision | Studienabbrecher | Unterschiede Teilnehmer/ Abbrecher | Ursachen für Studienabbruch | Selektives Berichten | Vorzeitiger Abbruch der Studie | Schlussfolgerung basieren auf Ergebnissen | Einfluss von Confounding und Bias | Anteil Nichtteilnehmer | Unterschiede Teilnehmer/ Nichtteilnehmer | Studienpopulation repräsentativ | Art und Quelle der Förderung |
| | IV | IV | IV | IV | IV | IV | IV | BQ | IV | IV | BQ | IV | IV | BQ | BQ | EV | EV | EV | IV |
| Ah-See & Molony[3] | ● | | | | | | | ● | | | | | | ● | | | | | |
| Andrew et al.[7] | ● | | | | | | | | | | ● | | | ● | | | | | |
| Assendelft et al.[8] | | | ● | | | | | ● | ● | | ● | | | | | | | | |
| Balas et al.[11] | ● | | ● | | | | | ● | ● | | | | | | | | | ● | |
| Balk et al.[12] | ● | | ● | | ● | | | | | | ● | | | ● | | | | | |
| Bath et al.[15] | | | ● | | | | | | | | | | | | | | | | |
| Bizzini et al.[18] | ● | | ● | | | | | | | | ● | | | | | | | | |
| Borsody & Yamada[20] | | | ● | | | | | | ● | | ● | | | | | | | | |
| Brown[26] | | | | | | | | | ● | | | | | | | | | | |
| CEBM[217] | ● | | | | ● | | | | ● | | | | | | | | | | |
| CEBMH[37] | | | ● | | | | | | | | | | | | | | | | |
| Chalmers et al.[38] | | | ● | | | | | | ● | | | | | | | | | | |
| Chalmers et al.[39] | ● | | ● | | | | | ● | ● | | ● | | ● | | | | | | |
| Cho & Bero[44] | ● | | | | ● | | | ● | | | ● | ● | | ● | | | | | |
| Colditz et al.[49] | ● | | | | | | | ● | | | ● | | | | | | | | |

**Tabelle 46: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 2 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cook et al.[50] | | | | | | | | | | | | | | | | | | | |
| Dardennes et al.[53] | ● | | | | ● | | | | ● | | | | | | | | | | |
| de Vet et al.[54] | ● | | ● | | ● | | | | ● | | ● | | | | | | | | |
| Delfini Group[57] | ● | ● | ● | | | | | | ● | ● | | | | | ● | | | | |
| Downs & Black[60] | ● | | ● | | ● | | | ● | ● | | | ● | | | | | | ● | |
| Earle & Hébert[66] | | | ● | | | | | | ● | | | | | | | | | | |
| Ekkernkamp et al.[69] | ● | | ● | | | | | ● | ● | ● | ● | | | | | | | ● | |
| Forsetlund & Rainer[72] | ● | | ● | | | | | | ● | | | | | | | | | | |
| Geng[75] | ● | | ● | | ● | | | ● | ● | | | | | ● | | | | ● | |
| Graf et al.[77] | | | ● | | ● | | | ● | ● | | ● | | | | | ● | | | |
| Gupta et al.[79] | ● | | ● | | | | | | | | | | | | | | | | |
| Hadorn et al.[82] | ● | | | | ● | | | ● | ● | | ● | ● | | | | | | ● | |
| Hammerschlag & Morris[83] | | | | | | | | | ● | | | | | | | | | | ● |
| Heneghan et al.[88] | | | ● | | | | | | | | | | | | | | | | |
| Hettinga et al.[90] | | | ● | | | | | | ● | | | | | | | | | | |
| Heyland et al.[91] | | | ● | | | | | | ● | | | | | | | | | | |
| Higgins & Green[92] | | | ● | ● | | | | | ● | | ● | ● | ● | | | | | | |
| Hill et al.[93] | | | ● | | | | | | | | | | | | | | | | ● |
| Huwiler-Müntener et al.[99] | ● | | | | | | | ● | | | ● | | ● | | | | | | |
| Imperiale & McCullough[100] | | | | | | | | | | | | | | | | | | | |
| IQWiQ | | | ● | | | | | | ● | ● | | | | | | | | | |
| Jadad et al. (3 Items)[106] | | | | | | | | | | | ● | | | | | | | | |
| Jadad et al.(6 Items)[106] | | | | | | | | | ● | | ● | | | | | | | | |
| Jonas & Linde[107] | ● | ● | ● | | | | | ● | ● | | | ● | | ● | | ● | | ● | |
| Kleijnen et al.[113] | ● | | | | | | | ● | | | | | | ● | | | | | |
| Kmet et al.[114] | ● | | | | | | | ● | | | | | | ● | | | | | |
| Koes et al.[115] | ● | | | | | | | | | | ● | | | ● | | | | | |
| Kwakkel et al.[120] | | | ● | | | | | ● | ● | | ● | | | | | | | | |
| Lamont[121] | ● | | ● | | | | | ● | ● | | | | | | | | | ● | |
| LBI[133] | | | ● | | | | | | ● | ● | | | | | | | | | |

**Tabelle 46: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 2 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Levine[123] | ● | | ● | | ● | | | | | | ● | | | ● | | | | | |
| Linde et al.[127] | | | ● | | | | | | | | | | | | | | | | |
| MacLehose et al.[134] | ● | | ● | | | | | | | | ● | | | | | | | | |
| MacMillan et al.[135] | | | ● | | | | | | ● | | ● | | | | | | | | |
| Moncrieff & Drummond[149] | | | | | | | | | ● | | | | | | | | | | |
| Moncrieff et al.[148] | ● | | ● | | ● | | | | | | | | | | | | | | |
| Morley et al.[150] | | | ● | | | | | | ● | | | | | | | | | | |
| Moseley et al.[151] | ● | | ● | | | | | ● | ● | | ● | | ● | | | | | | |
| Ogilvie-Harris & Gilbart[167] | ● | | | | | | | ● | | | ● | | | | | | | | |
| Onghena & Van Houdenhove[169] | | | | | | | | | | | | | | | | | | | |
| Petrak et al.[171] | ● | | ● | | ● | | | | ● | | ● | | | | | | | | |
| PHRU[158] | ● | | | | ● | | | ● | | | ● | ● | | ● | | | | | |
| Prendiville et al.[177] | ● | ● | ● | | | | | | ● | ● | | | | | ● | | | | |
| Pua et al.[178] | ● | | ● | | | | | ● | ● | ● | ● | | | | | | | ● | |
| Reisch et al.[181] | ● | | ● | | ● | | | ● | ● | | | ● | | | | | | ● | |
| Robeer et al.[183] | ● | | | | | | | ● | ● | | ● | | | | | | | | |
| Rochon[184] | | | ● | | | | | | ● | | | | | | | | | | |
| Rychetnik & Frommer[185] | | | ● | | | | | | | | ● | | | | | | | | |
| Sackett[187] | | | | | | | | | ● | | | | | | | | | | |
| SIGN 50[194] | ● | | ● | | | | | | ● | | | | | | | | | | |
| Sindhu et al.[198] | ● | | ● | | ● | | | ● | ● | | | | | ● | | | | ● | |
| Slim et al.[199] | | | ● | | ● | | | ● | ● | | ● | | | | | ● | | | |
| Smeenk et al.[201] | ● | | ● | | | | | | | | | | | | | | | | |
| Smith et al.[202] | ● | | | | ● | | | ● | ● | | ● | ● | | | | | | ● | |
| Smith et al.[203] | ● | | | | | | | | ● | | | ● | | ● | | | | | |
| Spitzer et al.[204] | | | | | | | | | ● | | | | | | | | | | ● |
| Staiger et al.[206] | | | ● | | | | | | | | | | | | | | | | |
| Stieb et al.[208] | | | ● | | | | | | ● | | | | | | | | | | |
| ter Riet et al.[211] | | | ● | | | | | | ● | | | | | | | | | | |

**Tabelle 46: Inhaltliche Charakteristika von Instrumenten für Interventionsstudien, Teil 2 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Thomas et al.[213] | | | ● | ● | | | | | ● | | ● | ● | ● | | | | | | |
| van Nieuwenhoven et al.[224] | | | ● | | | | | | | | | | | | | | | | ● |
| Verhagen et al.[227] | ● | | | | | | | ● | | | ● | | ● | | | | | | |
| Yates et al.[243] | | | | | | | | | | | | | | | | | | | |
| Yuen & Pope[244] | | | | | | | | | ● | | ● | | | | | | | | |
| Zaza et al.[245] | ● | ● | ● | | | | | ● | ● | | | ● | | ● | | ● | | ● | |
| | | | | | | | | | | | | | | | | | | | |
| Summe | 42 | 4 | 51 | 2 | 17 | 0 | 0 | 28 | 49 | 6 | 33 | 11 | 6 | 14 | 2 | 4 | 0 | 12 | 4 |
| | 53 % | 5 % | 64 % | 3 % | 21 % | 0 % | 0 % | 35 % | 61 % | 8 % | 41 % | 14 % | 8 % | 18 % | 3 % | 5 % | 0 % | 15 % | 5 % |

BQ = Berichtsqualität. EV = Externe Validität. IV = Interne Validität.

**Tabelle 47: Inhaltliche Charakteristika von Instrumenten für Beobachtungsstudien, Teil 1**

| | Studien-frage | Studienpopulation | | | | | | | | Exposition | | | | | Outcome | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fragestellung präzise und angemessen | Beschreibung | Ein- und Ausschlusskriterien | Identische Ein- und Ausschlusskriterie | Stichprobengröße | Kontrollgruppe | Vergleichbarkeit der Gruppen | FKS: Falldefinition | FKS: Gleichheit von Fällen und Kontrollen | Definition | Valide Methoden | Reliable Methoden | Messung gleich in allen Studiengruppen | FKS: verblindete Erhebung | Definition | Valide Methoden | Reliable Methoden | FKS: verblindete Erhebung | Verblindete Erhebung | Messung gleich in allen Studiengruppen | Follow-up-Länge angemessen | Länge des Follow-up gleich |
| | BQ | BQ | BQ | IV | IV | IV | IV | BQ | IV | BQ | IV | IV | IV | IV | BQ | IV | IV | IV | IV | IV | IV | IV |
| Andrew et al.[7] | ● | ● | ● | | | | | | | ● | | | | | | | | | | | | |
| Brown[26] | | ● | ● | | | | | | | ● | | | | | ● | ● | | | | | | |
| Carson et al.[30] | | | ● | | | | | | | | | | | | | | | | | | | |
| CEBMH[35] | | | | | | | ● | | | | | | | | | ● | | | ● | | ● | |
| Cho & Bero[44] | ● | | ● | | ● | | | | | | | | | | | | | ● | ● | | | |
| Colditz et al.[49] | | | ● | | | | | | | | | | | | | | | ● | ● | | | |
| Downs & Black[60] | ● | ● | | | ● | ● | | | | ● | | | | | ● | ● | ● | ● | ● | | | ● |
| Earle & Hébert[66] | | ● | | | ● | | | | | ● | | | | | ● | | | | | | | |
| Ekkernkamp et al.[69] | | | ● | | | ● | ● | ● | ● | | ● | ● | ● | | | ● | ● | | ● | | | |
| Geng[75] | ● | | ● | | ● | | ● | | | ● | | | ● | ● | | | | ● | ● | | | |
| Hadorn et al.[82] | | | ● | | | ● | ● | | | ● | | | | | | | | ● | ● | | | |
| Jonas & Linde[107] | | | ● | | ● | ● | ● | | | ● | | | | ● | ● | | ● | ● | ● | | | |
| Kmet et al.[114] | ● | ● | ● | | ● | | ● | | | ● | ● | ● | | | ● | ● | ● | ● | ● | | | |
| LBI[133] | | | | | ● | ● | ● | | | | ● | | ● | | | ● | ● | | ● | ● | ● | ● |
| Linde et al.[127] | | ● | | | | | | | | | | | | | | | | | | | ● | |
| MacLehose et al.[134] | ● | ● | ● | | ● | ● | | ● | | ● | | | | | ● | ● | ● | ● | ● | | | ● |
| Nguyen et al.[162] | ● | ● | ● | | ● | | | | | | | | | ● | | | | ● | ● | | | |

**Tabelle 47: Inhaltliche Charakteristika von Instrumenten für Beobachtungsstudien, Teil 1 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PHRU (FKS)[160] | ● | | | | ● | | | ● | | ● | | | ● | ● | | ● | | | | | | |
| PHRU (KS)[157] | ● | | | | | | | | | | ● | | ● | | | ● | ● | | ● | ● | ● | |
| Rychetnik & Frommer[185] | | ● | | | | | ● | ● | ● | | | ● | ● | | | | ● | ● | ● | | ● | |
| SIGN 50 (FKS)[194] | ● | | ● | | | | ● | | | | | ● | | | ● | ● | ● | | ● | ● | | |
| SIGN 50 (KS)[194] | ● | | | ● | | | | ● | | | ● | ● | | ● | ● | | | ● | ● | | | |
| Slim et al.[200] | ● | | | | ● | ● | ● | | | | | | | | ● | | | ● | ● | | ● | |
| Spitzer et al.[204] | | | ● | | ● | | ● | | | ● | | | ● | | ● | | | ● | ● | ● | | |
| Spooner et al.[205] | | | | | | | | | | | | | | | | | | | ● | | | |
| Stieb et al.[208] | | | | | | | ● | ● | | | ● | | | ● | | | | | | | | |
| Thomas et al.[213] | | | | | | | ● | | | | ● | ● | | | | ● | ● | ● | ● | | | |
| Wells et al., FKS[234] | | | | | | | ● | ● | | | | | ● | | | | | | | | | |
| Wells et al., KS[234] | | | | | | | ● | | | | ● | | | | | ● | | | ● | | ● | |
| Wong et al.[241] | | | | | | | | | | | | | | | | ● | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | |
| Summe | 12 | 9 | 14 | 1 | 12 | 7 | 15 | 7 | 2 | 11 | 8 | 6 | 8 | 6 | 10 | 12 | 10 | 14 | 21 | 4 | 7 | 3 |
| Prozent | 41% | 31% | 48% | 3% | 41% | 24% | 52% | 24% | 7% | 38% | 28% | 21% | 28% | 21% | 34% | 41% | 34% | 48% | 72% | 14% | 24% | 10% |

BQ = Berichtsqualität. FKS = Formular für Fall-Kontrollstudien. IV = Interne Validität. KS = Formular für Kohortenstudien.

**Tabelle 48: Inhaltliche Charakteristika von Instrumenten für Beobachtungsstudien, Teil 2**

| | Statistische Analyse | | | | | | | Ergebnisse | | | | | Diskussion | | Externe Validität | | | Finanzierung |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Angemessene Analyse | Multiples Testen | Confounderkontrolle | Fehlende Werte | Bewertung von Confounding | Heterogenität | Dosiswirkungsbeziehung | Punktschätzer und Präzision | Studienabbrecher | Unterschiede Teilnehmer/Abbrecher | Ursachen für Studienabbruch | Selektives Berichten | Schlussfolgerung basieren auf Ergebnissen | Einfluss von Confounding/ Bias | Anteil Nichtteilnehmer | Unterschiede Teilnehmer/ Nichtteilnehmer | Studienpopulation repräsentativ | Art und Quelle der Förderung |
| | IV | IV | IV | IV | IV | IV | BQ | BQ | IV | IV | BQ | IV | BQ | BQ | EV | EV | EV | IV |
| Andrew et al.[7] | ● | | | | | | | | | | ● | | ● | | | | | |
| Brown[26] | | | | | | | | | ● | | | | | | | | | |
| Carson et al.[30] | | | ● | | | | | | | | | | | | | | | |
| CEBMH[35] | | | ● | | | | | | ● | | | | | | | | ● | |
| Cho & Bero[44] | ● | | ● | | | | | ● | | | ● | ● | ● | | | | | |
| Colditz et al.[49] | ● | | | | | | | ● | | | ● | | | | ● | | | |
| Downs & Black[60] | ● | | ● | | | | | ● | ● | | | ● | | | | ● | ● | |
| Earle & Hébert[66] | | | | | | | | | ● | | | | | | | | | |
| Ekkernkamp et al.[69] | ● | | | | | | | ● | ● | ● | ● | | | | ● | | ● | |
| Geng[75] | ● | | | | | | | ● | | | | | ● | ● | | | ● | |
| Hadorn et al.[82] | ● | | | | | | | ● | ● | | ● | ● | | | | | ● | |
| Jonas & Linde[107] | ● | ● | | | | | | ● | ● | | | ● | ● | | ● | | | |
| Kmet et al.[114] | ● | ● | ● | | | | | ● | | | | ● | ● | | | | | |
| LBI[133] | ● | | ● | | | | | | ● | ● | ● | | | | | ● | | |
| Linde et al.[127] | | | | | | | | | ● | | | | | | | | | |
| MacLehose et al.[134] | ● | | ● | | | | | ● | ● | | | ● | | | | | | |

**Tabelle 48: Inhaltliche Charakteristika von Instrumenten für Beobachtungsstudien, Teil 2 – Fortsetzung**

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nguyen et al.[162] | ● | | ● | | | | | | | | | | ● | | | | | |
| PHRU (FKS)[160] | ● | | ● | | | | ● | ● | | | | | | | ● | ● | ● | |
| PHRU (KS)[157] | | | ● | | | | ● | ● | | ● | | | | | | | ● | |
| Rychetnik & Frommer[185] | ● | | ● | | | | ● | | | ● | ● | | | | | ● | ● | |
| SIGN 50 (FKS)[194] | | | ● | | | | | ● | ● | ● | ● | | | | | | ● | |
| SIGN 50 (KS)[194] | | | ● | | | | | ● | | | | | | ● | ● | ● | ● | |
| Slim et al.[200] | ● | | | | | | | ● | ● | | | | | | | | | |
| Spitzer et al.[204] | ● | ● | ● | | | | | | ● | | | | ● | | | | ● | |
| Spooner et al.[205] | | | | | | | | | | | | | | | | | | |
| Stieb et al.[208] | | | ● | | | | | | | | | | | | | | | |
| Thomas et al.[213] | ● | | ● | | | | | | ● | | ● | | | | ● | | ● | |
| Wells et al., FKS[234] | | | ● | | | | | | | | | | | | | | ● | |
| Wells et al., KS[234] | | | ● | | | | | | ● | | | | | | | | ● | |
| Wong et al.[241] | | | ● | | | | | | | | | | | | ● | | | |
| | | | | | | | | | | | | | | | | | | |
| Summe | 17 | 3 | 18 | 0 | 0 | 0 | 3 | 14 | 15 | 5 | 9 | 6 | 7 | 2 | 6 | 5 | 14 | 0 |
| Prozent | 59 % | 10 % | 62 % | 0 % | 0 % | 0 % | 10 % | 48% | 52 % | 17 % | 31 % | 21 % | 24 % | 7 % | 21 % | 17 % | 48 % | 0 % |

BQ = Berichtsqualität. EV = Externe Validität. FKS = Formular für Fall-Kontrollstudien. IV = Interne Validität. KS = Formular für Kohortenstudien.

**Tabelle 49: Inhaltliche Charakteristika von Instrumenten für Diagnosestudien**

| | Interne Validität | | | | | | | | | | Externe Validität | | | | Studiendurchführung | | | | Berichtsqualität | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Referenzstandard | Bias durch Krankheitsprogression | Verifikationsbias | Bias durch nicht-unabhängige Tests (Incorporation bias) | Behandlungsparadox | Review bias Referenztest | Review bias Referenztest | Klinischer Review bias | Beobachter-/ Instrumentenvariabilität | Umgang mit nicht bewert-baren Testergebnissen | Spektrum der Teilnehmer (spectrum composition) | Rekrutierung | Krankheitsprävalenz/ -schwere | Methodenwechsel beim Indextest | Subgruppenanalysen | Stichprobengröße | Studienziele | Protokoll | Einschlusskriterien | Indextestdurchführung | Referenztestdurchführung | Definition von „normalem" Testergebnis | Angemessene Ergebnisse | Genauigkeit der Ergebnisse | Studienabbrecher | Datentabelle | Nützlichkeit des Tests |
| | IV | IV | IV | IV | IV | IV | IV | IV | IV | IV | EV | EV | EV | EV | IV | IV | BQ | IV | BQ | BQ | BQ | BQ | BQ | BQ | IV | BQ | BQ |
| CEBM [218] | ● | | ● | ● | | ● | ● | | | | ● | | ● | | | | | | | ● | ● | | ● | | | ● | |
| CEBMH[33] | ● | | ● | ● | | ● | ● | | | | | ● | | ● | | | | | | | | | | | | | |
| de Vet et al.[55] | ● | | ● | | | ● | ● | | | ● | ● | ● | | | | ● | | | ● | ● | | ● | ● | | | | |
| Ekkernkamp et al.[69] | ● | | ● | | | ● | ● | | | | ● | | ● | | | | | | | | | ● | ● | | | ● | |
| Hoffmann et al.[95] | ● | | ● | | | ● | ● | | | | | ● | | | | ● | | | | | | | | | | | |
| Huebner et al.[98] | | | ● | | ● | | | | ● | | | | | | | | | | ● | ● | ● | ● | ● | ● | ● | | ● |
| IQWiQ[102] | ● | ● | ● | ● | | ● | ● | | | ● | ● | | ● | | | | | | ● | ● | ● | | | | | | |
| LBI[133] | ● | ● | ● | ● | | ● | ● | | | ● | | ● | | ● | | | | ● | | | | ● | | | | | |
| Mullins et al.[154] | | | ● | | | ● | ● | ● | | | | | | | | | | | | ● | | ● | | | | | |
| Mulrow et al.[155] | ● | | ● | | | ● | ● | | | ● | ● | | ● | | | ● | | | ● | | | ● | ● | | | | |
| PHRU[161] | ● | | ● | ● | ● | ● | | | | | ● | | ● | | | | | | | ● | | | ● | ● | | | ● |
| Sackett[187] | ● | | | | | ● | | | ● | | | | ● | | | | | | | ● | | | | | | | |
| SIGN 50[194] | ● | ● | ● | ● | | ● | ● | ● | | ● | ● | ● | | | | | | | ● | ● | ● | | | | | | |
| van den Hoogen et al.[221] | ● | | ● | | | ● | ● | | ● | | ● | ● | ● | | | ● | | | ● | | | ● | ● | | | | |
| Varela-Lema & Ruano-Ravina[226] | | ● | ● | | | ● | ● | ● | | ● | ● | | ● | | | | | | | | | | | | | | |
| Whiting et al.[238] | ● | ● | ● | ● | | ● | ● | ● | | ● | ● | ● | ● | | | | | | ● | ● | ● | | | | ● | | |
| Wurff et al.[223] | | | | | | ● | ● | | | | | | | | | | | | ● | | | | ● | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Summe | 13 | 5 | 15 | 7 | 2 | 16 | 14 | 4 | 3 | 7 | 10 | 7 | 9 | 2 | 0 | 4 | 0 | 1 | 8 | 9 | 5 | 7 | 8 | 2 | 2 | 2 | 2 |
| Prozent | 76 % | 29 % | 88 % | 41 % | 12 % | 94 % | 82 % | 24 % | 18 % | 41 % | 59 % | 41 % | 53 % | 12 % | 0 % | 24 % | 0 % | 6 % | 47 % | 53 % | 29 % | 41 % | 47 % | 12 % | 12 % | 12 % | 12 % |

BQ = Berichtsqualität. EV = Externe Validität. IV = Interne Validität.

# 8.16 Darstellung verschiedener Instrumente (Effektivität)

## 8.16.1 Beispiele für Checklisten ohne Komponenten- oder Gesamtbewertung

**Checkliste 1: QBI der German Scientific Working Group (2003) für systematische Reviews/Metaanalysen[69]**

| **Checkliste 1b: Systematische Reviews und Meta-Analysen** | | | | |
|---|---|---|---|---|
| **Bericht-Nr.:** | | | | |
| **Referenz-Nr.:**<br>**Titel:**<br>**Autoren:** | | | | |
| Das vorliegende Dokument enthält:<br>qualitative Informationssynthesen quantitative Informationssynthese | | | | |
| **Klass** | **A Fragestellung** | **Ja** | **Nein** | **unklar** |
| QA | 1. Ist die Forschungsfrage relevant für die eigene Fragestellung? | | | |
| | **B Informationsgewinnung** | | | |
| | 1. Dokumentation der Literaturrecherche: | | | |
| QA | a) Wurden die genutzten Quellen dokumentiert? | | | |
| QB | b) Wurden die Suchstrategien dokumentiert? | | | |
| QB | 2. Wurden Einschlusskriterien definiert? | | | |
| QB | 3. Wurden Ausschlusskriterien definiert? | | | |
| | **C Bewertung der Informationen** | | | |
| | 1. Dokumentation der Studienbewertung: | | | |
| QA | a) Wurden Validitätskriterien berücksichtigt? | | | |
| QB | b) Wurde die Bewertung unabhängig von mehreren Personen durchgeführt? | | | |
| QC | c) Sind ausgeschlossene Studien mit ihren Ausschlussgründen dokumentiert? | | | |
| QC | 2. Ist die Datenextraktion nachvollziehbar dokumentiert? | | | |
| QC | 3. Erfolgte die Datenextraktion von mehreren Personen unabhängig? | | | |
| | **D Infomationssynthese** | | | |
| | 1. Quantitative Informationssynthesen: | | | |
| QA | a) Wurde das Meta-Analyse-Verfahren angegeben? | | | |
| QB | b) Wurden Heterogenitätstestungen durchgeführt? | | | |
| QC | c) Sind die Ergebnisse in einer Sensitivitätsanalyse auf Robustheit überprüft? | | | |
| | 2. Qualitative Informationssynthesen: | | | |
| QA | a) Ist die Informationssynthese nachvollziehbar dokumentiert? | | | |
| QB | b) Gibt es eine Bewertung der bestehenden Evidenz? | | | |
| | **E Schlussfolgerungen** | | | |
| QB | 1. Wird die Forschungsfrage beantwortet? | | | |
| QB | 2. Wird die bestehende Evidenz in den Schlussfolgerungen konsequent umgesetzt? | | | |
| QA | 3. Werden methodisch bedingte Limitationen der Aussagekraft kritisch diskutiert? | | | |
| I | 4. Werden Handlungsempfehlungen ausgesprochen? | | | |
| I | 5. Gibt es ein Grading der Empfehlungen? | | | |
| I | 6. Wird weiterer Forschungsbedarf identifiziert? | | | |
| I | 7. Ist ein „Update“ des Reviews eingeplant? | | | |
| **Beurteilung:**<br>Die vorliegende Publikation wird berücksichtigt ausgeschlossen<br>**Legende:**<br>Klass. Klassifikation der Frage<br>Q Frage, die Aspekte der methodischen Qualität erfasst; in absteigender Relevanz mit A, B oder C bewertet<br>I Frage mit reinem Informationsgehalt, irrelevant für Qualitätsbeurteilung | | | | |

NR = Nummer. QBI = Qualitätsbewertungsinstrument.

**Checkliste 2: QBI der German Scientific Working Group (2003) für Primärstudien[69]**

| **Checkliste 2a: Primärstudien (RCTs, Fall-Kontrollstudien, Kohortenstudien, Längsschnittstudien, Fallserien)** | | | | |
|---|---|---|---|---|
| **Referenz-Nr.:**<br>**Titel:**<br>**Autoren:** | | | | |
| Dokumenttyp: RCT | | | | |
| **Klass.** | **A Auswahl der Studienteilnehmer** | **Ja** | **Nein** | **unklar** |
| QA | 1. Sind die Ein- und Ausschlusskriterien für Studienteilnehmer ausreichend/ eindeutig definiert? | | | |
| QA | 2. Wurden die Ein- und Ausschlusskriterien vor Beginn der Intervention festgelegt? | | | |
| QA | 3. Wurde der Erkrankungsstatus valide und reliabel erfasst? | | | |
| QBI | 4. Sind die diagnostischen Kriterien der Erkrankung beschrieben? | | | |
| QB | 5. Ist die Studienpopulation/exponierte Population repräsentativ für die Mehrheit der exponierten Population bzw. die „Standardnutzer" der Intervention? | | | |
| QA | 6. Bei Kohortenstudien: Wurden die Studiengruppen gleichzeitig betrachtet? | | | |
| | **B Zuordnung und Studienteilnahme** | | | |
| QA | 1. Entstammen die Exponierten/Fälle und Nicht Exponierten/Kontrollen einer ähnlichen Grundgesamtheit? | | | |
| QA | 2. Sind Interventions-/Exponierten und Kontroll-/Nicht-Exponiertengruppen zu Studienbeginn vergleichbar? | | | |
| QB | 3. Erfolgte die Auswahl randomisiert mit einem standardisierten Verfahren? | | | |
| QC | 4. Erfolgte die Randomisierung blind? | | | |
| QA | 5. Sind bekannte/mögliche Confounder zu Studienbeginn berücksichtigt worden? | | | |
| | **C Intervention/Exposition** | | | |
| QA | 1. Wurden Intervention bzw. Exposition valide, reliabel und gleichartig erfasst? | | | |
| QB | 2. Wurden Interventions-/Kontrollgruppen mit Ausnahme der Intervention gleichartig therapiert? | | | |
| QB | 3. Falls abweichende Therapien vorlagen, wurde diese valide und reliabel erfasst? | | | |
| QA | 4. Bei RCTs: Wurden für die Kontrollgruppen Placebos verwendet? | | | |
| QA | 5. Bei RCTs: Wurde dokumentiert wie die Placebos verabreicht wurden? | | | |
| | **D Studienadministration** | | | |
| QB | 1. Gibt es Anhaltspunkte für ein Overmatching? | | | |
| QB | 2. Waren bei Multicenterstudien die diagnostischen und therapeutischen Methoden sowie die Outcomemessung in den beteiligten Zentren identisch? | | | |
| | **E Outcome Messung** | | | |
| I | 1. Wurden patientennahe Outcome-Parameter verwendet? | | | |
| QA | 2. Wurden die valide und reliabel erfasst? | | | |
| QB | 3. Erfolgte die Outcomemessung verblindet? | | | |
| QC | 4. Bei Fallderien: Wurde die Verteilung prognostischer Faktoren ausreichend erfasst? | | | |
| | **F Drop Outs** | | | |
| QA | 4. War die Response-Rate bei Interventions-/Kontrollgruppen ausreichend hoch bzw. bei Kohortenstudien: konnte ein ausreichend großer Teil der Kohorte über die gesamte Studiendauer verfolgt werden? | | | |
| QA | 5. Wurde die Gründe für Ausscheiden von Studienteilnehmern aufgelistet? | | | |
| QB | 6. Wurden die Outcomes der Drop-Outs beschrieben und in der Auswertung berücksichtigt? | | | |
| QB | 7. Falls Differenzen gefunden wurden – sind diese signifikant? | | | |
| QB | 8. Falls Differenzen gefunden wurden – sind diese relevant? | | | |
| | **G Statistische Analyse** | | | |
| QA | 1. Sind die beschriebenen analytischen Verfahren korrekt und die Informationen für eine einwandfreie Analyse ausreichend? | | | |
| QB | 2. Wurden für Mittelwerte und Signifikanztests Konfidenzintervalle angegeben? | | | |
| I | 3. Sind die Ergebnisse in graphischer Form präsentiert und wurden die den Graphiken zugrunde liegenden Werte angegeben? | | | |

**Checkliste 2: QBI der German Scientific Working Group (2003) für Primärstudien[69] – Fortsetzung**

| | |
|---|---|
| **Beurteilung:** | |
| Die vorliegende Publikation wird | berücksichtigt ausgeschlossen |
| **Legende:** | |
| Klass. | Klassifikation der Frage |
| Q | Frage, die Aspekte der methodischen Qualität erfasst; in absteigender Relevanz mit A, B oder C bewertet |
| I | Frage mit reinem Informationsgehalt, irrelevant für Qualitätsbeurteilung |

RCT = Randomisierte kontrollierte Studie. NR = Nummer. QBI = Qualitätsbewertungsinstrument.

**Checkliste 3: QBI der German Scientific Working Group (2003) für Diagnosestudien[69]**

| **Checkliste 2b: Diagnosestudie** | | | | |
|---|---|---|---|---|
| **Bericht-Nr.:** | | | | |
| **Titel:**<br>**Autoren:**<br>**Quelle:** | | | | |
| **Klass.** | **A Beschreibung der Ausgangssituation** | **Ja** | **Nein** | **unklar** |
| QA | 1. Gibt es eine klar formulierte Fragestellung vor Beginn der Studie? | | | |
| QA | 2. Wurde die Zielkrankheit eindeutig definiert? | | | |
| QA | 3. Erfolgte die Festlegung der Trenngröße vor Beginn? | | | |
| QA | 4. Wurde ein „Goldstandard“ festgelegt Angaben über seine Zuverlässigkeit gemacht? | | | |
| | **B Durchführung der Prüfung** | | | |
| QB | 1. Ausreichende Beschreibungen für Nachprüfungen? | | | |
| QA | 2. Erfolgte die Auswertung des Testergebnisses ohne Kenntnis der Diagnose und umgekehrt die Diagnose ohne Kenntnis des Testergebnisses? | | | |
| QB | 3. Wurde die Zusammensetzung der Versuchskollektive in Bezug auf die Übertragbarkeit der Ergebnisse berücksichtigt? | | | |
| QA | 4. Wurde die zu untersuchende Technik und der „Goldstandard“ bei allen Patienten angewendet? | | | |
| | **C Ergebnisinterpretation** | | | |
| QA | 1. Ist eine Vierfeldertafel vorhanden bzw. ist eine Erstellung aus den gegebenen Daten möglich? | | | |
| QB | 2. Erfolgte eine Verwendung von eindeutig definierten Parametern zur Beschreibung der Ergebnisse? | | | |
| | **D Diskussion** | | | |
| QA | 1. Wurde die Abhängigkeit des prädiktiven Wertes von der a priori Wahrscheinlichkeit ausreichend diskutiert? | | | |
| QB | 2. Wurde die Definition des pathologischen Testergebnisses als Erkrankung selbst vermieden? | | | |
| QC | 3. Wurde eine Nutzen-Schaden-Abwägung für die vier Ergebnisgruppen durchgeführt? | | | |
| QA | 4. Wurden die Folgen von Fehlklassifikationen ausreichend diskutiert? | | | |
| **Beurteilung:** | | | | |
| Die vorliegende Publikation wird | berücksichtigt ausgeschlossen | | | |
| **Legende:** | | | | |
| Klass. | Klassifikation der Frage | | | |
| Q | Frage, die Aspekte der methodischen Qualität erfasst; in absteigender Relevanz mit A, B oder C bewertet | | | |
| I | Frage mit reinem Informationsgehalt, irrelevant für Qualitätsbeurteilung | | | |

QBI = Qualitätsbewertungsinstrument.

**Checkliste 4: QUADAS von Whiting et al. (2003), QBI für diagnostische Studien[238]**

| Item | Yes | No | Unclear |
|---|---|---|---|
| 1. Was the spectrum of patients representative of the patients who will receive the test in practice? | | | |
| 2. Were selection criteria clearly described? | | | |
| 3. Is the reference standard likely to correctly classify the target condition? | | | |
| 4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? | | | |
| 5. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis? | | | |
| 6. Did patients receive the same reference standard regardless of the index test result? | | | |
| 7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? | | | |
| 8. Was the execution of the index test described in sufficient detail to permit replication of the test? | | | |
| 9. Was the execution of the reference standard described in sufficient detail to permit its replication? | | | |
| 10. Were the index test results interpreted without knowledge of the results of the reference standard? | | | |
| 11. Were the reference standard results interpreted without knowledge of the results of the index test? | | | |
| 12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? | | | |
| 13. Were uninterpretable/ intermediate test results reported? | | | |
| 14. Were withdrawals from the study explained? | | | |

QBI = Qualitätsbewertungsinstrument. QUADAS = Quality assessment of diagnostic accuracy studies.

## 8.16.2 Beispiele für Checklisten mit qualitativer Gesamtbewertung

**Checkliste 5: QBI des Ludwig-Boltzmann-Instituts (2007) für RCT, Kohortenstudien, systematische Reviews/Metaanalysen und diagnostische Studien[133]**

| Kriterien zur Beurteilung von RCTs | Ja | Nein | Nicht enthalten | Nicht anwendbar |
|---|---|---|---|---|
| War die Randomisierung adäquat? | | | | |
| War die Unvorhersehbarkeit der Gruppenzuweisung adäquat (allocation concealment)? | | | | |
| Waren wesentliche Charakteristika der Studiengruppe ähnlich? | | | | |
| Basiert die Studiengröße auf einer adäquaten Berechnung, die Power und einen kleinsten wesentlichen Unterschied einbezieht (minimal important difference)? | | | | |
| Wurde die Verblindung adäquat durchgeführt? | | | | |
| Gab es eine hohe Drop-out-Rate? (> 20%) | | | | |
| Gab es eine hohe differentielle Drop-out-Rate? (>15%) | | | | |
| Wurde eine Intention-to-Treat-Analyse (ITT-Analyse) adäquat durchgeführt? | | | | |
| Gab es Ausschlüsse nach der Randomisierung (post randomization exclusions)? | | | | |
| Beurteilung der internen Validität | **Gut** | **Aus-reichend** | **Unzu-reichend** | |
| Kommentare | | | | |

**Checkliste 5: QBI des Ludwig-Boltzmann-Instituts (2007) für RCT, Kohortenstudien, systematische Reviews/Metaanalysen und diagnostische Studien[133] – Fortsetzung**

| **Kriterien zur Beurteilung von Kohortenstudien** | **Ja** | **Nein** | **Nicht enthalten** | **Nicht anwendbar** |
|---|---|---|---|---|
| Wurden die Studiengruppen aus derselben Population rekrutiert? | | | | |
| Ist die Verteilung der prognostischen Faktoren zwischen den Gruppen ausreichend beschrieben? | | | | |
| Ist die Verteilung der prognostischen Faktoren zwischen den Gruppen ähnlich? | | | | |
| Haben alle Gruppen dasselbe Risiko für den Outcome? | | | | |
| Wurden alle Gruppen während derselben Zeitperiode rekrutiert? | | | | |
| Wurden Outcomes in allen Gruppen auf gleiche Art und Weise beurteilt? | | | | |
| Wurden Outcomes verblindet beurteilt? | | | | |
| War die Studienlaufzeit für alle Gruppen identisch? | | | | |
| Gab es eine hohe Drop-out-Rate? (> 20%) | | | | |
| Gab es eine hohe differentielle Drop-out-Rate? (>15%) | | | | |
| Wurden potentielle Confounder (Störgrößen) in der statistischen Analyse berücksichtigt? | | | | |
| Beurteilung der internen Validität | **Gut** | **Ausreichend** | **Unzureichend** | |
| Kommentare | | | | |

| **Kriterien zur Beurteilung von systematischen Reviews und Meta-Analysen** | **Ja** | **Nein** | **Nicht enthalten** | **Nicht anwendbar** |
|---|---|---|---|---|
| Basiert der Review auf einer klar definierten Frage? | | | | |
| Wurden Auswahlkriterien klar definiert? | | | | |
| Wurde eine systematische Literatursuche durchgeführt? | | | | |
| Haben zumindest 2 Personen die Studie beurteilt? | | | | |
| Wurde die methodologische Qualität der Studien beurteilt? | | | | |
| Wurde die methodologische Qualität der Studien bei der Evidenzsynthese berücksichtigt? | | | | |
| **Für Meta-Analysen** | | | | |
| Wurde Publikationsbias beurteilt? | | | | |
| Wurde Heterogenität statistische beurteilt? | | | | |
| Wurde Heterogenität adäquat analysiert? | | | | |
| Waren Studien die Einheit der statitischen Analyse? | | | | |
| Beurteilung der internen Validität | **Gut** | **Ausreichend** | **Unzureichend** | |
| Kommentare | | | | |

| **Kriterien zur Beurteilung von diagnostischen Studien** | **Ja** | **Nein** | **Nicht enthalten** | **Nicht anwendbar** |
|---|---|---|---|---|
| Repräsentiert die Studie jene PatientInnen, die den Test in der Praxis erhalten werden? | | | | |
| Wurden Auswahlkriterien exakt formuliert? | | | | |
| Wurde ein Referenztest verwendet, der als „Goldstandard“ angesehen werden kann? | | | | |
| Ist die Zeitperiode zwischen Durchführung des (Index-)Tests und Referenztests entsprechend kurz? | | | | |
| Wurden beide Testresultate unabhängig voneinander ausgewertet? | | | | |
| Wurde die Höhe der Drop-out-Rate während des Tests genannt? | | | | |
| Wurde der Anteil an nicht-interpretierbaren Ergebnissen genannt? | | | | |
| Beurteilung der internen Validität | **Gut** | **Ausreichend** | **Unzureichend** | |
| Kommentare | | | | |

RCT = Randomisierte kontrollierte Studie.

## 8.16.3 Beispiel für ein Komponentensystem

**Checkliste 6: „Risk of bias tool" der Cochrane Collaboration (2008)[92]**

| | Yes | No | Unclear |
|---|---|---|---|
| Sequence generation: Was the allocation sequence adequately generated? | | | |
| Allocation concealment: Was allocation adequately concealed? | | | |
| Blinding of participants, personnel and outcome assessors: Was knowledge of the allocated interventions adequately prevented during the study? | | | |
| Incomplete outcome data: Were incomplete outcome data adequately addressed? | | | |
| Selective outcome reporting: Are reports of the study free of suggestion of selective outcome reporting? | | | |
| Other potential threats to validity: Was the study apparently free of other problems that could put it at a risk of blues? | | | |

## 8.16.4 Beispiele für Skalen

**Checkliste 7: QBI von Downs & Black (1998) für RCT und Beobachtungsstudien[60]**

| | Yes | No | Partially | Unable to determine |
|---|---|---|---|---|
| **Reporting** | | | | |
| 1. Is the hypothesis/aim/objective of the study clearly described? | | | | |
| 2. Are the main outcomes to be measured clearly described in the Introduction or Methods section? | | | | |
| 3. Are the characteristics of the patients included in the study clearly described? | | | | |
| 4. Are the interventions of interest clearly described? | | | | |
| 5. Are the distributions of principal confounders in each group of subjects to be compared clearly described? | | | | |
| 6. Are the main findings of the study clearly described? | | | | |
| 7. Does the study provide estimates of the random variability in the data for the main outcomes? | | | | |
| 8. Have all important adverse events that may be a consequence of the intervention been reported? | | | | |
| 9. Have the characteristics of patients lost to follow-up been described? | | | | |
| 10. Have actual probability values been reported (e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001? | | | | |
| **External validity** | | | | |
| 11. Were the subjects asked to participate in the study representative of the entire population from which they were recruited? | | | | |
| 12. Were those subjects who were prepared to participate representative of the entire population from which they were recruited? | | | | |
| 13. Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive? | | | | |
| **Internal validity – bias** | | | | |
| 14. Was an attempt made to blind study subjects to the intervention they have received? | | | | |
| 15. Was an attempt made to blind those measuring the main outcomes of the intervention? | | | | |
| 16. If any of the results of the study were based on "data dredging", was this made clear? | | | | |
| 17. In trials and cohort studies, do the analyses adjust for different lengths of follow-up of patients, or in case-control studies, is the time period between the intervention and outcome the same for cases and controls? | | | | |

**Checkliste 7: QBI von Downs & Black (1998) für RCT und Beobachtungsstudien[60] – Fortsetzung**

| | | | | |
|---|---|---|---|---|
| 18. Were the statistical tests used to assess the main outcomes appropriate? | | | | |
| 19. Was compliance with the intervention/s reliable? | | | | |
| 20. Were the main outcome measures used accurate (valid and reliable)? | | | | |
| **Internal validity – confounding (selection bias)** | | | | |
| 21. Were the patients in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited from the same population? | | | | |
| 22. Were study subjects in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited over the same period of time? | | | | |
| 23. Were study subjects randomised to intervention groups? | | | | |
| 24. Was the randomised intervention assignment concealed from both patients and health care staff until recruitment was complete and irrevocable? | | | | |
| 25. Was there adequate adjustment for confounding in the analyses from which the main findings were drawn? | | | | |
| 26. Were losses of patients to follow-up taken into account? | | | | |
| **Power** | | | | |
| 27. Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%? (Size of smallest intervention group) | | | | |

**Checkliste 8: 6 Item-Scale von Jadad et al. (1996) für RCT[106]**

| | Yes | No |
|---|---|---|
| Was the study described as randomized? | | |
| Was the study described as double-blind? | | |
| Was there a description of withdrawals and drop-outs? | | |
| Was there a clear description of the inclusion and exclusion criteria? | | |
| Was the method used to assess adverse effects described? | | |
| Were the methods of statistical analysis described? | | |

Die systematische Bewertung medizinischer Prozesse und Verfahren, *Health Technology Assessment* (HTA), ist mittlerweile integrierter Bestandteil der Gesundheitspolitik. HTA hat sich als wirksames Mittel zur Sicherung der Qualität und Wirtschaftlichkeit im deutschen Gesundheitswesen etabliert.

Seit Einrichtung der Deutschen Agentur für HTA des DIMDI (DAHTA) im Jahr 2000 gehören die Entwicklung und Bereitstellung von Informationssystemen, speziellen Datenbanken und HTA-Berichten zu den Aufgaben des DIMDI.

Im Rahmen der Forschungsförderung beauftragt das DIMDI qualifizierte Wissenschaftler mit der Erstellung von HTA-Berichten, die Aussagen machen zu Nutzen, Risiko, Kosten und Auswirkungen medizinischer Verfahren und Technologien mit Bezug zur gesundheitlichen Versorgung der Bevölkerung. Dabei fallen unter den Begriff Technologie sowohl Medikamente als auch Instrumente, Geräte, Prozeduren, Verfahren sowie Organisationsstrukturen. Vorrang haben dabei Themen, für die gesundheitspolitischer Entscheidungsbedarf besteht.

DIMDI
Deutsches Institut für Medizinische Dokumentation und Information