(Internes) Manual Abläufe und Methoden

Teil 2

Ludwig Boltzmann Institut
Health Technology Assessment

HTA-Projektbericht Nr. 006
ISSN 1992-0488
ISSN-online 1992-0496

(Internes) Manual
Abläufe und Methoden

Teil 2

Ludwig Boltzmann Institut
Health Technology Assessment

Wien, März 2007

INSTITUT FÜR HEALTH TECHNOLOGY ASSESSMENT
DER LUDWIG BOLTZMANN GESELLSCHAFT

Projektleitung:
Dr. Gerald Gartlehner, MPH

Interne Begutachtung:
Dr. Claudia Wild
Mag. Rosemarie Felder-Puig, MSc

Externe Begutachtung:
Dr. Stefan Lange, IQWiG
Nathalie McGauran, IQWIG
Dr. Gerd Antes, Cochrane Collaboration

1. ÜBERARBEITUNG: DEZEMBER 2008

Danksagung: Herzlichen Dank an Smiljana Blagojevic, Stacey Williams und Laura Morgan für Hilfe bei der Formatierung des Berichts und Erstellung der Bibliographie.

IMPRESSUM

Medieninhaber und Herausgeber:
Ludwig Boltzmann Gesellschaft GmbH
Operngasse 6/5. Stock, A-1010 Wien
http://www.lbg.ac.at/gesellschaft/impressum.php

Für den Inhalt verantwortlich:

Ludwig Boltzmann Institut für Health Technology Assessment (LBI-HTA)
Garnisongasse 7/20, A-1090 Wien
http://hta.lbg.ac.at/

Die LBI-HTA-Projektberichte erscheinen unregelmäßig und dienen der Veröffentlichung der Forschungsergebnisse des Ludwig Boltzmann Instituts für Health Technology Assessment.

Die Berichte erscheinen in geringer Auflage im Druck und werden über das Internetportal „http://eprints.hta.lbg.ac.at" der Öffentlichkeit zur Verfügung gestellt:

HTA-Projektbericht Nr. 006
ISSN 1992-0488
ISSN-online 1992-0496

© 2007 LBI-HTA – Alle Rechte vorbehalten

Inhalt

Abbildungsverzeichnis

Zielsetzung des „internen" Manuals

Ziel dieses Manuals ist es, die Arbeitsmethoden des Ludwig Boltzmann Instituts für Health Technology Assessment (LBI für HTA) in Bezug auf systematische Übersichtsarbeiten und Rapid Assessments darzustellen und eine praktische Anleitung und Erklärung für einzelne Arbeitsschritte zu bieten. Dieses Manual setzt voraus, dass eine systematische Literaturdurchsicht als jene wissenschaftliche Methode bestimmt wurde, die eine gegebene Fragestellung am besten beantworten kann. Andere Methoden wie Datenbankanalysen, Anwendungsbeobachtungen, Fokus-Gruppen, Delphi-Prozesse etc. werden in diesem Manual nicht behandelt, können aber dennoch für bestimmte Fragen die relevante Methode der Wahl sein. Die systematische Literaturdurchsicht ist jedoch das international am häufigsten eingesetzte Verfahren bei Health Technology Assessments (HTAs).[1]

Ziel: Darstellung der Arbeitsmethoden ....

... & praktische Anleitung für Arbeitsschritte

Die Methodik des LBI für HTA orientiert sich am Vorgehen anderer Institutionen, dem Selbstverständnis des Institutes und der methodologischen Erfahrung der MitarbeiterInnen. Insbesondere wurden Richtlinien von NICE (National Institute for Health and Clinical Excellence)[2], AHRQ (Agency for Healthcare Research and Quality)[3 4], IQWiG (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen)[5] und der Cochrane Collaboration[6] in Betracht gezogen.

Grundlage: Selbstverständnis des Instituts .....

& international anerkannte Methodik

Die Struktur des Manuals folgt dem Ablauf eines HTAs von der Entwicklung der HTA-Fragestellung bis zur Publikation. Dieses Manual erläutert nicht im Detail, wie die Auswahl und Priorisierung von HTA-Themen auf gesundheitspolitischer Ebene erfolgt.

1 Einleitung

HTA: „Best Practice"-Aktivität zur Förderung von Effizienz und Qualität

Die Europäische Kommission hat Health Technology Assessments (HTAs) gemeinsam mit Evidenzbasierter Medizin (EBM) und klinischen Leitlinien als eine der „Best Practice"-Aktivitäten klassifiziert, die zur Förderung der Effizienz und Qualität der Gesundheitsversorgung ausschlaggebend sind.[7] Diese Aktivitäten sind durch eine systematische Arbeitsweise gekennzeichnet, die durch Evaluierung und Synthese vorhandener wissenschaftlicher Evidenz relevante gesundheitspolitische oder klinische Fragen zu beantworten versuchen.

HTA ist eine Wissenschaftsmethode

HTA ist ein Instrument, das systematisch und transparent vorhandenes Wissen zur klinischen Wirksamkeit und zu ökonomischen und organisatorischen Auswirkungen (neuer, aber auch etablierter) medizinischer Verfahren analysiert und auf die Vorbereitung von administrativen wie klinischen Entscheidungen abzielt.[8] Der Technologiebegriff ist dabei ein weiter und umfasst medizinisch-technische Großgeräte ebenso wie Arzneimittel, medizinische Verfahren und Methoden und auch sozialmedizinische oder psychologische Interventionen.[9] Das Europäische Netzwerk für Health Technology Assessment (EUnetHTA) fasst HTA wie folgt zusammen[10]:

internationale HTA-Netzwerke:

„Health technology assessment is a multidisciplinary process that summarises information about the medical, social, economic and ethical issues related to the use of a health technology in a systematic, transparent, unbiased, robust manner. Its aim is to inform the formulation of safe, effective health policies that are patient focused and seek to achieve best value: Despite its policy goals, HTA must always be firmly rooted in research and the scientific method."[11]

EUnetHTA-Definition

INAHTA-Definition

Das International Network for Agencies in Health Technology Assessment (INAHTA) definiert HTA folgendermaßen: *„Technology assessment in health care is a multidisciplinary field of policy analysis. It studies medical, social, ethical, and economic implications of development, diffusion, and use of health technology."*[12]

Unterstützung politischer Entscheidungen

HTAs werden erstellt, um die Entscheidungsfindung bei gesundheitspolitischen und versorgungstechnischen Fragen zu erleichtern.[13-17] Auftraggeber identifizieren relevante Themen, die von gesundheitspolitischer Bedeutung sind. In enger Zusammenarbeit mit den zuständigen WissenschaftlerInnen erfolgen Prioritätensetzung und exakte Definierung der wissenschaftlichen Fragestellung. Die Durchführung von HTAs erfolgt anhand standardisierter und transparenter Methoden.[18] Vorhandene wissenschaftliche Evidenz wird identifiziert und synthetisiert, um den Auftraggebern eine verständliche Zusammenfassung der wissenschaftlichen Datenlage zu präsentieren.[16 19]

standardisierte Arbeitsweise: methodisch valides Vorgehen ist Voraussetzung,.....

Eine standardisierte Arbeitsweise, Nachvollziehbarkeit, Transparenz und eine methodologisch gefestigte Vorgangsweise sind Voraussetzung für eine hohe Qualität eines HTAs.[16 19] Da HTAs zu kontroversen Ergebnissen führen können, muss davon ausgegangen werden, dass jeder Bericht einer sehr genauen Überprüfung durch einzelne Interessensgruppen unterzogen wird. Es ist daher wichtig, dass methodologische Entscheidungen wissenschaftlich gerechtfertigt werden können.

....um kontroverse Ergebnisse verteidigen zu können

Die folgenden Kapitel des internen Manuals setzen voraus, dass eine Identifizierung der zu evaluierenden Technologien und eine Prioritätensetzung durchgeführt worden sind. Sie bieten eine Anleitung zu einer standardisierten, methodologisch validen Arbeitsweise. Dies bietet die Voraussetzung,

dass ein HTA-Bericht erfolgreich in der Praxis eingesetzt werden kann und einer Evaluierung standhält.

Handbuch bietet Arbeitshilfen

Das interne Manual ist unterteilt in ein Kapitel über systematische Übersichtsarbeiten und ein Kapitel über Rapid Assessments. Im Appendix befinden sich Formulare, die zur Beurteilung der internen und externen Qualität verwendet werden können. Des Weiteren enthält der Anhang Beispiele von Dokumenten, die in abgewandelter Form bei einzelnen Arbeitsschritten eingesetzt werden können.

2 Systematische Übersichtsarbeiten

systematisches Arbeiten bedeutet zunächst Vereinheitlichung des Arbeitsprozesses

Systematische Übersichtsarbeiten durchlaufen einen einheitlichen Prozess von der Verfassung der eigentlichen Fragestellung bis zur externen Begutachtung (Peer Review). Dies gilt nicht nur für klinische Fragestellungen, sondern auch für Fragen der Versorgungsforschung, ökonomische Assessments oder für sozial-organisatorische Fragestellungen. Der Prozess der systematischen Übersichtsarbeit ist für alle Themenbereiche identisch, wenngleich sich Auswahlkriterien und Zielvariablen - abhängig von der Fragestellung - unterscheiden. Insbesondere für sozial-organisatorische oder gesundheitssystembezogene Fragen existieren derzeit noch wenig standardisierte Evaluierungsinstrumente.[20,21] In breiten sozioökonomischen Fragestellungen will das systematische Herangehen zumeist nur mögliche Perspektiven, Denkschulen und „Best Practice"-Modelle identifizieren und weniger zu einer finalen Ergebnisbeurteilung kommen.

interne Validität/ Qualität der eigenen Arbeit & Arbeitseffizienz stärken

Abbildung 2-1 zeigt eine Übersicht über einzelne Phasen einer systematischen Übersichtsarbeit. Abbildung 2-2 fasst einzelne Arbeitschritte zusammen. Nicht alle Arbeitschritte, die in diesem Manual beschrieben werden, sind obligat, um die Validität der Übersichtsarbeit zu gewährleisten. Manche Schritte dienen in erster Linie dazu, die Effizienz des Ablaufes zu verbessern und können mitunter weggelassen oder modifiziert werden. Andere müssen durchgeführt werden, um die interne Validität der Übersichtsarbeit sicherzustellen.

Themenpriorisierung in LBI-HTA Board,

Themenvergabe an Projektverantwortlichen,

Orientierung:

Präsierung der Policy-Fragestellung,

Ableitung der wissenschaftlichen Fragestellung

= Scoping

Bevor die eigentliche wissenschaftliche Arbeit beginnt, werden in den jährlichen Board/Kuratoriumssitzungen des LBI für HTA die Themen gesammelt und von den Board-Mitgliedern priorisiert (vgl. Externes Manual: Selbstverständnis & Arbeitsweise. http://eprints.hta.lbg.ac.at/714/1/HTA-Projektbericht_003.pdf). Diesem Prozess folgt eine LBI für HTA interne Themenvergabe an Projektverantwortliche, die interne Fach-Expertise wie Interdisziplinarität der Projektbearbeitung berücksichtigt. In einer darauf folgenden Orientierungsphase im Thema wird zunächst die Politik-Fragestellung entsprechend dem Entscheidungsunterstützungsbedarf präzisiert, um in der folge die wissenschaftliche(n) Fragestellung(en) (HTA-Fragestellung) heraus zu arbeiten und auch einen ev. Bedarf nach externer Expertise zu erkennen. Dieser Prozess der Orientierung im Thema kann in einer oder mehreren Feedback-Schleifen mit dem „Auftraggeber" (einzelnen oder mehreren Boardmitgliedern) und externer Fachexpertise erfolgen und wird international „Scoping" genannt.

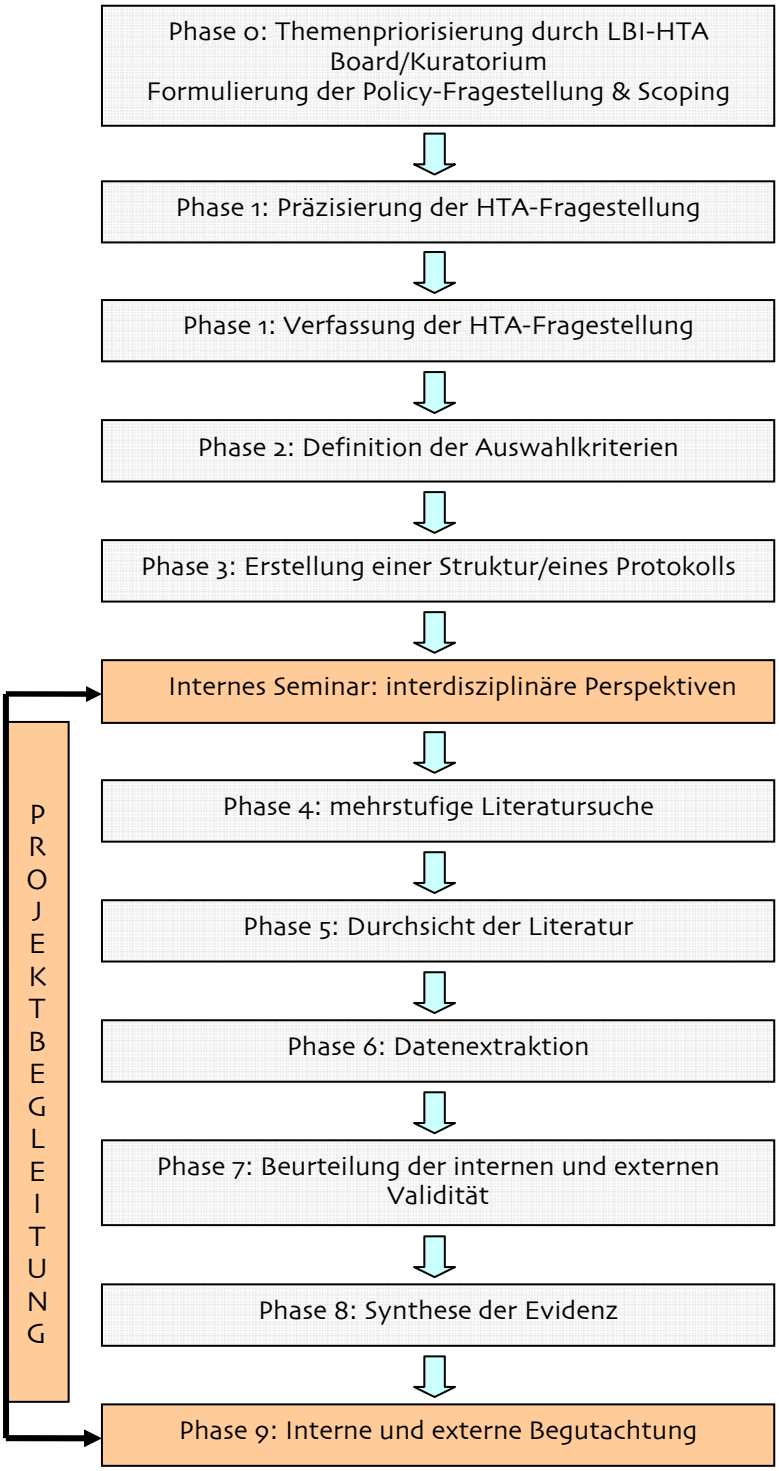

Abbildung 2-1: Arbeitsphasen einer systematischen Übersichtsarbeit

Arbeitsschritt	Obligat*?
Verfassung der HTA-Fragestellung	**Ja**
PIKO-System	Nein
Definition der Auswahlkriterien	**Ja**
Grobe Literatursuche	Nein
Erstellung eines Protokolls	**Ja**
Elektronische Literatursuche in mehreren Datenbanken	**Ja**
Manuelle Literatursuche	**Ja**
Suche nach grauer Literatur	Nein
Identifikation von Studien durch externe ExpertInnen	Nein
Duale Durchsicht der Abstracts	**Ja**
Verwendung eines Formulars zur Durchsicht der Abstracts	Nein
Duale Durchsicht der Volltext-Artikel	**Ja**
Verwendung eines Formulars zur Durchsicht der Volltext-Artikel	Nein
Kodierung der ausgeschlossenen Artikel im Literaturverwaltungsprogramm	Nein
Erstellung eines Flussdiagramms des Auswahlprozesses	**Ja**
Beurteilung der internen Validität	**Ja**
Beurteilung der externen Validität	Nein
Extraktion der Daten durch zwei Reviewer	**Ja**
Erstellung einer Evidenztabelle	Nein
Beurteilung der Stärke der Evidenz	**Ja**
Interne Begutachtung des Reports	**Ja**
Externe Begutachtung des Reports	**Ja**

*obligat, um die interne Validität der Übersichtsarbeit zu gewährleisten

Abbildung 2-2: Arbeitsschritte während einer systematischen Übersichtsarbeit

2.1 Verfassung der HTA-Fragestellung

HAUPTPUNKTE:

- Eine exakte Definition der Fragestellung ist Voraussetzung für die Literatursuche und für eine fokussierte Synthese der Evidenz.
- Die Fragestellung muss in Zusammenarbeit mit dem Auftraggeber überarbeitet und definiert werden, um für Entscheidungsträger relevant zu bleiben.
- Bei klinischen oder ökonomischen Fragestellungen sollen die Population, Intervention, Kontrollintervention und Zielvariablen in der Fragestellung definiert werden.
- Bei sozio-organisatorischen Fragestellungen sind die Einheiten (z.B. Klinikabteilungen oder ganze Institutionen) und die Perspektiven der Analyse (z.B. „Best Practice"-Modelle) zu definieren.

Unklarheiten früh beseitigen, ...

Jedem Thema eines HTAs liegt eine gesundheitspolitische Fragestellung zugrunde. Die Fragestellung wird vom Auftraggeber meist unzureichend und zu breit verfasst. Es ist Aufgabe der zuständigen WissenschaftlerInnen, eine exakte Fragestellung zu definieren und Änderungen mit dem Auftraggeber zu diskutieren. In diesen Prozess sollen sowohl Auftraggeber als auch bei Bedarf klinische, ökonomische oder methodologische ExpertInnen eingebunden werden, um eine wissenschaftliche Fragestellung zu formulieren, die

1. eine für den Auftraggeber relevante Frage beantwortet;
2. klinische, ökonomische oder andere FachexpertInnen für beantwortbar halten;
3. WissenschaftlerInnen des LBI für HTA innerhalb eines vorgegebenen Zeitrahmens beantworten können.

..... um Beantwortung der eigentlichen Frage zu gewährleisten

Bei der Einbindung von externen ExpertInnen müssen diese auf mögliche Interessenskonflikte befragt werden. Eine exakte Definition der Fragestellung verhindert Unklarheiten, welche Evidenz in den Bericht aufgenommen werden soll und wie diese Evidenz zusammengefasst wird.[22,23] Unzureichend formulierte, unpräzise Fragestellungen können dazu führen, dass der Umfang der vorhandenen Literatur ein Ausmaß annimmt, das innerhalb eines gegebenen Zeitrahmens nicht bewältigbar ist.[24]

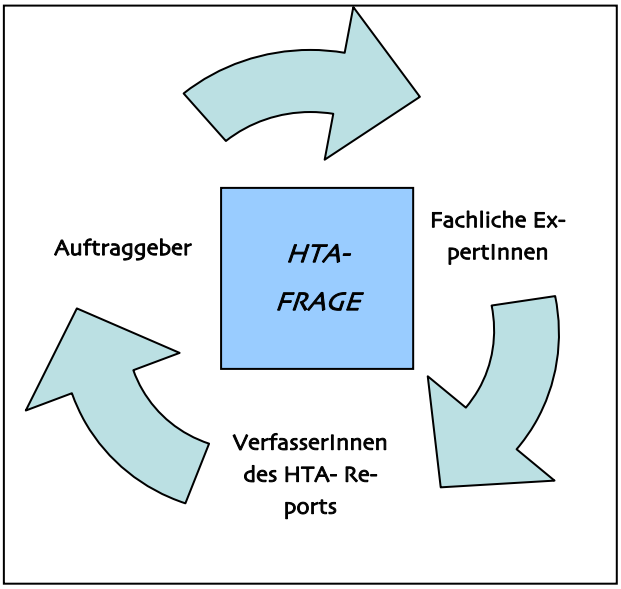

Abbildung 2.1-1: Prozess der Definition einer HTA-Frage

Für Fragestellungen, denen eine Intervention (z.B. diagnostische oder therapeutische Intervention, Screening, Präventionsmaßnahme) oder Exposition zugrunde liegt, soll eine gezielte Formulierung folgende Komponenten beinhalten[25]:

Definition der Fragestellung dient zur Abklärung & Eingrenzung…..

- **P**opulation
- **I**ntervention
- **K**ontrollintervention
- **O**utcome (Zielvariable)

Diese vier Säulen werden unter „PIKO" zusammengefasst.

Eine Fragestellung zu einer therapeutischen Intervention könnte zum Beispiel lauten:

„Ist Akupunktur genauso wirksam wie Nikotin-Substitutionstherapie, um bei erwachsenen RaucherInnen eine Langzeit-Nikotinabstinenz zu erreichen?"

Eine Fragestellung zu einer diagnostischen Intervention könnte zum Beispiel lauten:

… und ist erster Schritt zur Literatursuche

„Ist ambulante Spirometrie im Vergleich zur Ganzkörperpletysmographie ein adäquates diagnostisches Instrument, um eine Abnahme der Lungenfunktion in erwachsenen RaucherInnen mit COPD (chronic obstructive pulmonary disease) festzustellen?"

[Diagnostische Studien sollen die Intervention immer mit einem Goldstandard vergleichen.]

Eine Fragestellung zu Screening könnte zum Beispiel lauten:

therapeutische, diagnostische, Screening-Fragestellungen

„Kann ambulante Spirometrie bei beschwerdefreien erwachsenen RaucherInnen beginnende COPD diagnostizieren und dadurch den weiteren Verlauf der Erkrankung positiv beeinflussen?"

[Die Kontrollgruppe wäre hier „kein Screening".]

Eine Fragestellung zu Präventionsmaßnahmen könnte zum Beispiel lauten:

„Kann die regelmäßige Einnahme von niedrig dosiertem Aspirin das Risiko von kardiovaskulären Erkrankungen in erwachsenen Rauchern reduzieren?"

[Die Kontrollgruppe wäre „kein Aspirin".]

ökonomische & organisatorische Fragestellungen

Eine gesundheitsökonomische Fragestellung könnte zum Beispiel lauten:

„Wie verhält sich die Kosten-Effektivität einer medikamentösen Therapie mit Statinen (cholesterinsenkenden Medikamenten) bei PatientInnen mit diagnostizierter Herz-Kreislauferkrankung (Sekundärprävention) im Vergleich zur nicht-medikamentösen Behandlung eines erhöhten Serum-Cholesterin-Spiegels?"

[Outcome: Kosten-Effektivität.]

Eine organisatorische Fragestellung könnte zum Beispiel lauten:

„*Welche organisatorischen Modelle und Ablaufstrukturen führen zu einer Verbesserung der klinischen und ökonomischen Ergebnisse in Intensivstationen (gemessen an ICU-Mortalität, iatrogene Morbidität, Wiedereinweisung und direktem Ressourceneinsatz, Liegedauer etc.)?"*

[Keine Kontrollgruppe.]

PIKO aber nicht immer anwendbar

Komponenten von PIKO werden bei Interventionen klar definiert oder können abgeleitet werden (z.B. „keine Intervention" als Kontrollgruppe). Bei Fragestellungen, die sich nicht auf Personen und Interventionen oder Expositionen beziehen, können PIKO-Kriterien nur begrenzt verwendet werden. Dies bezieht sich vor allem auf Fragen der Versorgungsforschung, sozioökonomische Fragestellungen und ethische oder methodologische Fragen.

2.1.1 Population

Analyseeinheit:

Patient, aber auch organisatorische Einheit

Eine genaue Definition der Population ist für die Generalisierbarkeit der Resultate wichtig. Eine Limitierung der Population sollte nicht willkürlich erfolgen, sondern nachvollziehbar sein (z.B. Raucher mit koronarer Herzerkrankung). Gründe für Limitierungen können Unterschiede in Prognose, Motivation usw. sein.

Die Definition einer Population kann sich auch auf ein Gesundheitssystem beziehen. Bei sozio-organisatorischen Fragestellungen sind die Einheiten (z.B. Klinikabteilungen oder ganze Institutionen) und die Perspektiven der Analyse (z.B. „Best Practice"-Modelle) zu definieren.

2.1.2 Intervention

Intervention muss definiert werden

Die Intervention sollte in der Fragestellung definiert werden. Eine genaue Definition muss jedoch in den Auswahlkriterien erfolgen. Zu viele Details in der wissenschaftlichen Fragestellung limitieren die Lesbarkeit. In unserem obigen Beispiel ist „Akupunktur" die Intervention. Nachdem diese nicht näher definiert ist, könnten alle Formen der Akupunktur beinhaltet sein (z.B. Körperakupunktur, Ohrakupunktur, Laserakupunktur, Akupunktur mit/ohne Moxabustion). Interventionen können jedoch nicht nur auf individueller PatientInnen-Ebene erfolgen, sondern ebenso auf der Ebene einer organisatorischen Einheit (z.B. klinische Leitlinien).

2.1.3 Kontrollintervention

Kontrollintervention muss definiert werden

Ebenso wie die Intervention muss die Kontrollintervention in der Fragestellung definiert werden. Die Kontrollintervention kann eine aktive (z.B. die bestehende Goldstandard-Intervention bei neuen Technologien) oder auch eine Scheinintervention sein. Bei Screening oder Präventionsmaßnahmen ist „keine Intervention" oftmals die relevante Kontrolle. Im Fall einer Exposition ist die Kontrollmaßnahme häufig „keine Exposition".

2.1.4 Outcome (Zielvariable)

patientenrelevante Zielvariablen sind zu bevorzugen

In der Fragestellung muss definiert werden, welche Zielvariablen in der systematischen Übersichtsarbeit berücksichtigt werden. Grundsätzlich sollen Zielvariablen für PatientInnen relevant sein. Eine genauere Definition der Zielvariablen erfolgt in den Auswahlkriterien.

2.2 Definition der Auswahlkriterien für Literatur

HAUPTPUNKTE:

- Die Auswahlkriterien für Literatur legen fest, welche Studien in den Bericht aufgenommen werden.
- Im Verlauf einer systematischen Übersichtsarbeit kann es notwendig sein, Auswahlkriterien zu adaptieren.

Studiendesign ist Auswahlkriterium für medizinische Fragestellungen

Auswahlkriterien legen fest, welche Studien in den Bericht aufgenommen werden.[25] Die Definition der Auswahlkriterien sollte anhand der Fragestellung erfolgen. Grundsätzlich sollten jene Studiendesigns eingeschlossen werden, welche die größtmögliche methodologische Validität besitzen. Für die Effektivität von Interventionen zum Beispiel wären dies etwa randomisierte kontrollierte Studien (randomized controlled trials; RCTs). Andere Studiendesigns können je nach Fragestellung zusätzlich wichtige Informationen liefern.

breite Perspektive soll nicht aus den Augen gelassen werden

Bei ethischen Fragestellungen zum Beispiel können qualitative Studien wichtige Information liefern. Für eine umfassende Beurteilung von Nebenwirkungen sollten immer zusätzlich Beobachtungsstudien eingeschlossen werden. Die Definition der Auswahlkriterien geht jedoch über das Studiendesign hinaus. PIKO-Kategorien sollen in den Auswahlkriterien genauer definiert werden:

- Population
  - Altersgruppen
  - Krankheit, Stadium der Krankheit
- Intervention (Exposition)
  - Exakte Definition der Intervention
  - Co-Interventionen
- Kontrollintervention (Kontrollexposition)

Zielvariablen
- Patientenrelevante Zielvariablen
- Surrogat-Zielvariablen

- Studiendesigns
  - Für die Beantwortung der HTA-Frage relevante Designs
  - Mindeststudiendauer
  - Eventuell Mindeststudiengröße

Auswahlkriterien bei Fülle/Mangel an Publikationen adaptieren

Auswahlkriterien sind ein gutes Instrument, um die Literatur in einem bewältigbaren Ausmaß zu halten. Zu restriktive Auswahlkriterien können allerdings dazu führen, dass keine Evidenz identifiziert werden kann. Nachdem ein erster Entwurf von Auswahlkriterien erstellt wurde, sollte eine grobe Literatursuche durchgeführt werden, um einen Eindruck über das Ausmaß der vorhandenen Literatur zu gewinnen. In Ausnahmefällen, wenn das Ausmaß der Literatur zu groß oder zu klein erscheint, müssen die Auswahlkriterien entsprechend überdacht werden.

2.3 Verfassung eines Protokolls

HAUPTPUNKTE:

- Ein systematischer Ablauf erfordert ein klar definiertes, a priori erstelltes Protokoll und eine Projektstruktur.
- Projektleitung, HTA-Fragestellung, Auswahlkriterien, Breite der Perspektive, Interventionen, Zielvariablen, Verantwortlichkeiten und Zeitplan sollen im Protokoll definiert werden.
- Das Protokoll dient als Referenzdokument während der Durchführung des Reviews.

Protokoll ist Leitfaden...

Die Durchführung einer systematischen Übersichtsarbeit ist ein komplexer Prozess, der während des gesamten Verlaufes vielfache Entscheidungen erfordert. Um einen systematischen Ablauf zu garantieren, müssen Prozesse standardisiert und wesentliche Punkte definiert werden.[26] Wie bei jeder wissenschaftlichen Arbeit sollen Definitionen vor Beginn des Projekts festgelegt werden. Die Erstellung eines Protokolls ist daher notwendig. Folgende Punkte sollen enthalten sein:

- Projektleitung
- HTA-Fragestellung
- Auswahlkriterien für Studien
- Quellen für Literatursuche
- Festlegung der Breite der Perspektiven
- Interventionen und Zielvariablen
- Zeitplan

... und wird von der Projektleitung erstellt

Üblicherweise wird das Protokoll vom/von der Projektleiter/in erstellt. Das Protokoll dient auch als Referenzdokument, auf das MitarbeiterInnen bei Unsicherheiten zurückgreifen können. Appendix A bietet einen Grundriss eines Protokolls.

Änderungen sind möglich, müssen aber festgehalten werden

Im Verlauf eines Projekts kann es sein, dass das Protokoll geändert werden muss. Diese Situation kann zum Beispiel eintreten, wenn im Zuge der Literaturdurchsicht klar wird, dass zu viele/zu wenige Studien zu den definierten Fragestellungen und Auswahlkriterien zur Verfügung stehen, oder wenn externe Hinweise auf wesentliche praxisrelevante Aspekte hinzukommen. Jede Adaption sollte im Protokoll festgehalten werden. Änderungen des Protokolls sollten keinesfalls post hoc, das heißt gegen Ende des Projekts, durchgeführt werden, wenn es absehbar ist, welchen Einfluss der Ausschluss oder die Einbeziehung zusätzlicher Studien auf das Resultat haben. Jegliche Änderungen des Protokolls sollten später im Methodenteil des Berichts dargestellt werden.

2.4 Literatursuche

HAUPTPUNKTE:

- Die Literatursuche muss systematisch in mehreren elektronischen Datenbanken durchgeführt werden.
- Fehlende Systematik führt zu Bias.
- Eine manuelle Suche sollte die elektronische Suche ergänzen.

systematische statt selektive Suche

Eine adäquate, systematische Literatursuche ist die Voraussetzung für die Validität einer systematischen Übersichtsarbeit.[6,25-28] Eine systematische Literatursuche unterscheidet systematische Übersichtsarbeiten von konventionellen, nicht-systematischen Übersichtsarbeiten (narrative reviews), die hauptsächlich von Expertenmeinungen getragen werden.[26,27] Publikationsbias kann nicht vermieden werden, während Retrieval- Bias mit einer sorgfältigen, systematischen Literatursuche minimiert werden kann.[29-31] Fehlende Systematik führt zusätzlich zu Selektionsbias.[31 32]

Biasminimierung

In dem vom LBI für HTA verwendeten bibliografischen Literaturverwaltungsprogramm werden die Suchergebnisse verwaltet und Duplikate eliminiert.

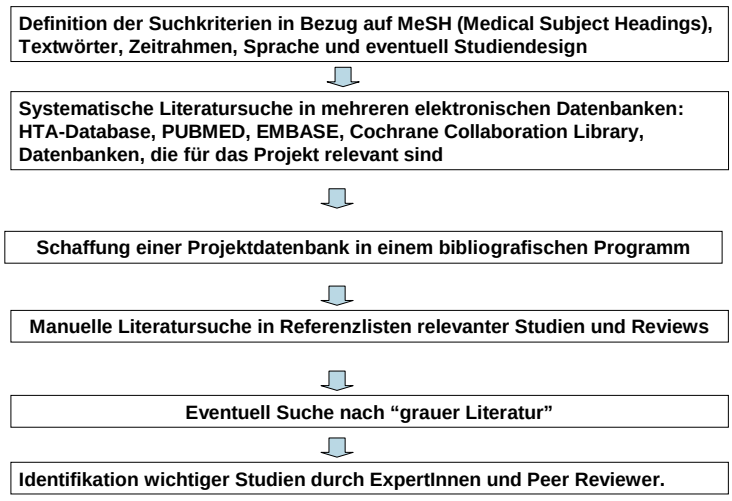

Abbildung 2.4-1: Literatursuche

mehrstufig

Eine systematische Literatursuche sollte mehrstufig durchgeführt werden. Abbildung 2.4-1 fasst die einzelnen Stufen zusammen. Am LBI für HTA wird die Literatursuche immer in Zusammenarbeit mit einem/einer Informationsspezialisten/in durchgeführt. Das detaillierte Vorgehen bei der Literatursuche in elektronischen Datenbanken wird daher nicht erläutert.

2.4.1 Suche in elektronischen Datenbanken

zuerst: HTA-Datenbank

Generell müssen immer mehrere elektronische Datenbanken durchsucht werden, um die Sensitivität der Suche zu erhöhen.[30,31,33] Zusätzlich geht jeglicher Suche eine Überprüfung der HTA-Datenbank voraus, ob eventuell bereits ein Assessment zum Thema vorliegt. Medline (National Library of Medicine) und Embase (Elsevier) sind die beiden größten elektronischen Datenbanken. Standardisiertes Vokabular (Medical Subject Headings [MeSH]) hilft, die Suche in diesen Datenbanken spezifischer und präziser durchzuführen. Zusätzlich zu Medline und Embase sollte auch immer eine Suche in der Cochrane Library durchgeführt werden. Wenn die Suche auf RCTs limitiert ist, kann eine kostenpflichtige Embase-Suche durch eine Suche in Cochrane CENTRAL (Central Register of Controlled Trials) ersetzt werden.[32]

MeSH-Terms

Medline, Embase, Cochrane CENTRAL

Spezialdatenbanken

Zusätzlich zu Medline und Embase gibt es eine Vielzahl anderer Datenbanken (z.B. CINAHL, AIDSLINE), die je nach Thema und Bedarf einbezogen werden müssen.[30] Die Spezifität der Suche kann durch eine Limitierung der Zeitperiode, spezifisches Suchvokabular und andere mögliche Limitierungen in Bezug auf Studiendesign, Population und Publikationssprache erhöht werden.

Die elektronische Literatursuche soll von dem/der Informationsspezialisten/in gemeinsam mit dem/der Projektleiter/in durchgeführt werden.

2.4.2 Manuelle Literatursuche

Schlüsselpublikationen identifizieren.....

Die manuelle Literatursuche ist ein obligater Bestandteil einer systematischen Literaturdurchsicht. Eine manuelle Suche ist notwendig, um publizierte Studien zu identifizieren, die bei der elektronischen Suche nicht identifiziert wurden.[34,35] Eine manuelle Literatursuche kann zum Beispiel anhand von 15 bis 20 relevanten Artikeln erfolgen, die während der letzten 3 bis 5 Jahre publiziert wurden. Dies können systematische Übersichtsarbeiten oder andere relevante Zusammenfassungen des Themas sein oder auch wesentliche Studien. In den Referenzlisten dieser Publikationen wird dabei nach relevanten Studien gesucht. Dieser Prozess der Handsuche wird durch die Zitationsdatenbank Scopus unterstützt. In der Projekt-Datenbank wird das Vorhandensein dieser Studien anschließend überprüft.

..... und Referenzlisten durchsehen,

Unterstützung durch Zitationsdatenbank Scopus

2.4.3 Suche nach nicht-publizierten Studien (grauer Literatur)

Konferenzabstracts etc.

Publikationsbias ist eines der wesentlichen Probleme von systematischen Übersichtsarbeiten.[36-39] Nur circa 45% aller publizierten Abstracts werden später als Artikel veröffentlicht.[40] Die Publikation einer Studie korreliert dabei stark mit dem Vorhandensein eines statistisch signifikanten Ergebnisses.[36] Publikationsbias ist ein Problem für jede systematische Übersichtsar-

beit, die Durchsicht von Konferenzabstracts (z.B. in SCOPUS)[41], Studienregistern (z.B. www.clinicaltrials.gov) oder Datenbanken wie die der CDER (Center for Drug Evaluation and Research) und der amerikanischen FDA (Food and Drug Administration)[42] kann jedoch Hinweise auf nicht-publizierte Studien geben. Auch das Kontaktieren von Haupt-AutorInnen und Studiensponsoren kann weitere nicht-publizierte Studien hervorbringen.

zeitaufwändig, daher Nutzenabwägung

Die Suche nach nicht-publizierten Studien ist jedoch sehr zeitaufwändig und fällt oft unergiebig aus. Die Suche nach nicht-publizierter Literatur ist nicht obligat. Der/die Projektleiter/in muss von Fall zu Fall entscheiden, ob der potentielle Ertrag den Aufwand rechtfertigt.

2.4.4 Identifizierung von Literatur durch externe GutachterInnen und InteressensvertreterInnen

auch bei FachexpertInnen nachfragen

Externe GutachterInnen und eventuell auch InteressensvertreterInnen können eingeladen werden, Studien zu nennen, die für ein Thema wesentlich sind.[43] Dies ist jedoch kein obligater Bestandteil der Literatursuche.[44] In der Projekt-Datenbank wird das Vorhandensein dieser Studien überprüft. Studien, die nicht bei der elektronischen Suche gefunden wurden, werden in die Datenbank importiert.

2.5 Durchsicht der Literatur

HAUPTPUNKTE:

- Die Durchsicht der Literatur sollte immer von zwei Personen unabhängig voneinander durchgeführt werden.
- Die Spezifität der Literatursuche wird schrittweise erhöht.
- Diskrepanzen werden durch Konsens gelöst.
- Gründe für den Ausschluss von Volltext-Artikeln müssen dokumentiert werden.
- Wird keine relevante Literatur gefunden, können Auswahlkriterien revidiert werden.
- Der gesamte Literatur-Selektionsprozess wird grafisch in einem Flussdiagramm dargestellt.

duale Durchsicht erhöht Spezifität der Literaturauswahl

Die Durchsicht der Literatur sollte immer von zwei Personen unabhängig voneinander durchgeführt, um die Präzision der Literaturauswahl zu erhöhen und falsch-negative Entscheidungen zu korrigieren.[6] Differenzen werden durch Diskussion und Konsens gelöst. Für die Durchsicht der Abstracts und der Volltext-Artikel können standardisierte Formulare verwendet werden, die eine Konsistenz der Beurteilung gewährleisten sollen. Diese Formulare basieren auf a priori definierten Ein- und Ausschlusskriterien. Studien, die nicht den Auswahlkriterien entsprechen, werden ausgeschlossen und entsprechend kodiert. Während des gesamten Prozesses wird die Spezifität (wie präzise identifizieren ReviewerInnen Studien, die die Auswahlkriterien *nicht* erfüllen) dabei schrittweise erhöht. Für Rapid Assessments kann dieser Vor-

Spezifität: Ausschluss irrelevanter Studien

gang modifiziert werden.

Darstellung in einem Flussdiagramm

Das Quality of Reporting of Meta-analyses (QUOROM) Statement empfiehlt, dass der gesamte Selektionsprozess grafisch dargestellt wird.[45] Gründe für den Ausschluss von Studien sollen zumindest auf Volltext-Ebene angeführt werden. Eine effiziente Möglichkeit, ein QUORUM-Flussdiagramm zu erstellen, wäre zum Beispiel die Kodierung von Ausschlüssen in einer bibliografischen Datenbank. Andere Vorgehensweisen können aber genauso zielführend sein. Ein Beispiel eines QUORUM-Flussdiagramms ist in Abbildung 2.5-1 dargestellt.

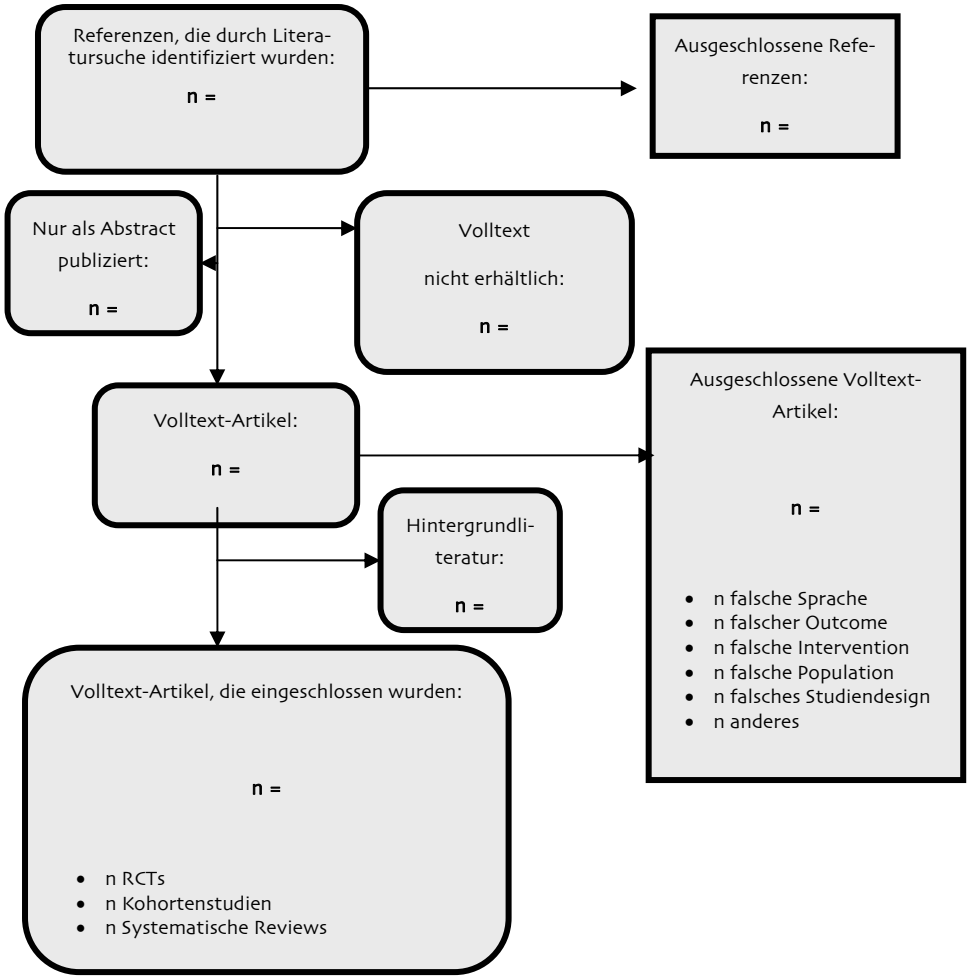

Abbildung 2.5-1: Beispiel eines Flussdiagramms des Auswahlprozesses (QUOROM tree)

2.5.1 Durchsicht der Abstracts und der Volltext-Publikationen

Abstracts:

Einschluss möglicherweise relevanter Zitate

Ziel der Durchsicht der Abstracts ist es, Studien zu identifizieren, die für den Review mit Sicherheit *nicht* relevant sind. Solche Studien werden ausgeschlossen. Bei der Durchsicht der Abstracts kann die Spezifität der Literaturauswahl zunächst noch niedrig sein. Dies bedeutet, dass ReviewerInnen in diesem Stadium jegliche Literatur, die *möglicherweise* relevant sein könnte, einschließen (hohe Sensitivität). In der Praxis bedeutet dies, dass es aus-

reicht, wenn einer der beiden ReviewerInnen befindet, dass eine Studie relevant sein könnte, um die Studie einzuschließen.

Studien, die keinen Abstract haben bzw. deren Abstract zu wenig Aufschluss über die Studie gibt, werden immer eingeschlossen. Studien, die von beiden ReviewerInnen als nicht relevant identifiziert wurden, werden ausgeschlossen. Zu allen Abstracts, die von mindestens einem/einer Reviewer/in als relevant beurteilt werden oder die insuffizient für eine solche Beurteilung sind, wird die Volltext-Publikation erworben.

Folgende Fragen sollen bei der Durchsicht der Abstracts beantwortet werden:

1. Ist der Artikel relevant für das Thema?
2. Sind die Resultate relevant für die Beantwortung der HTA-Frage?
3. Erfüllt die Studie die Auswahlkriterien?

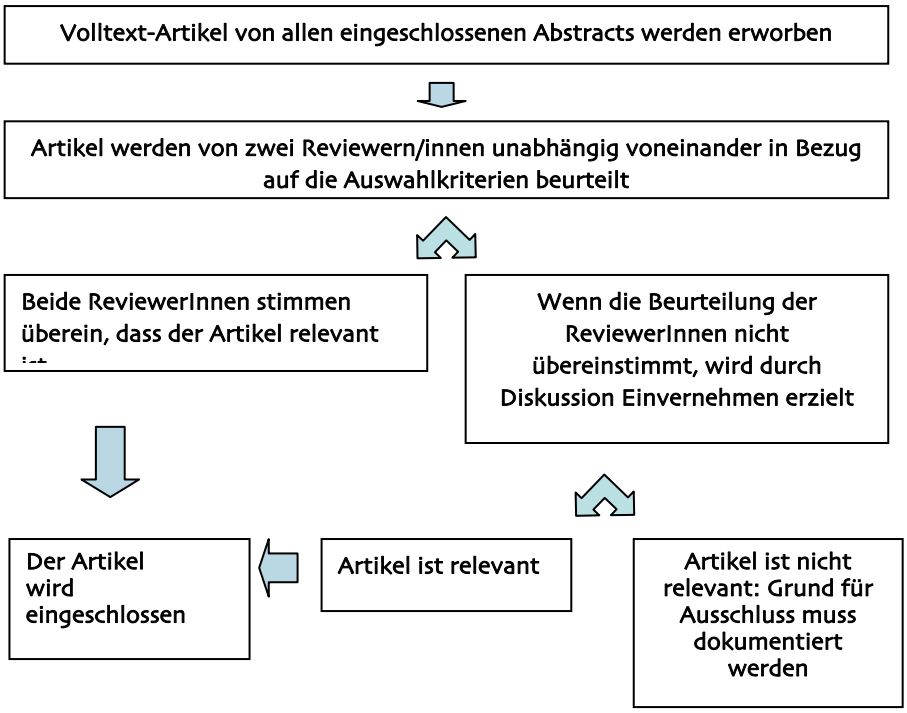

Abbildung 2.5.1-1: Durchsicht der Volltext-Artikel

Volltext:

Identifizierung wahrscheinlich relevanter Artikel

Ziel der Durchsicht der Volltext-Artikel ist es, Studien zu identifizieren, die für den Review mit großer Wahrscheinlichkeit relevant sind. In diesem Stadium wird die Spezifität daher erhöht, um Studien, welche die Auswahlkriterien nicht erfüllen, auszuschließen. ReviewerInnen müssen einstimmig über den Ein- bzw. Ausschluss von Studien entscheiden. In der Praxis evaluieren beide ReviewerInnen unabhängig voneinander die Studien und vergleichen anschließend die Resultate. Appendix B bietet einen Grundriss eines Formulars für die Durchsicht von Volltext-Artikeln.

Gründe für den Ausschluss von Artikeln müssen dokumentiert werden. Diese Information wird später für die Erstellung eines Flussdiagramms des Auswahlprozesses (QUOROM tree) benötigt. Abstracts, die nicht als Voll-

text-Publikation erhältlich sind (z.B. Abstracts von Posterpräsentationen oder Konferenzbeiträgen) sollten ebenfalls dementsprechend dokumentiert werden. Exakte Dokumentation erleichtert auch, Fragen über das Fehlen einzelner Studien während der internen und externen Begutachtung zu klären. Die Kodierung von ausgeschlossenen Studien kann auf verschiedene Art und Weise erfolgen. Eine Kodierung in einer bibliografischen Datenbank ist effizient und später leicht reproduzierbar. Appendix C bietet ein Beispiel für Kodierungen von Artikeln.

2.5.2 Klassifizierung von Studien

gleiches Verständnis bei Klassifizierung wichtig

Für die Validität der Beurteilung von HTA-Fragestellungen ist das Studiendesign von großer Bedeutung. Es gibt keine „offizielle" Terminologie für Studiendesigns, und verwendete Klassifikationen variieren stark zwischen Medizin und Sozialwissenschaften. Auch innerhalb der Epidemiologie gibt es Differenzen. Für einen systematischen Review ist es jedoch wichtig, dass von allen ReviewerInnen exakte, identische Definitionen verwendet werden. Die folgende Klassifizierung benutzt klinisch-epidemiologische Definitionen, die auf Schemata der CDC (Center for Disease Control) Task Force on Community Preventive Services[46] und des NICE beruhen.[2]

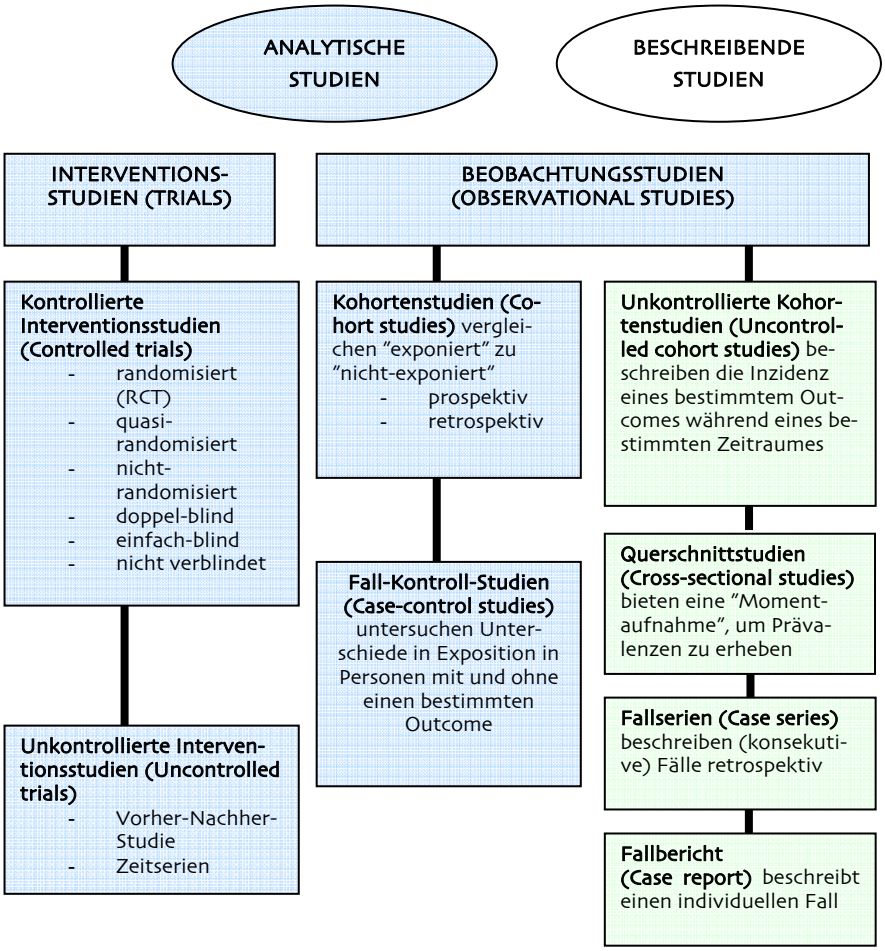

Abbildung 2.5.2-1: Klassifikationsschema von Studien

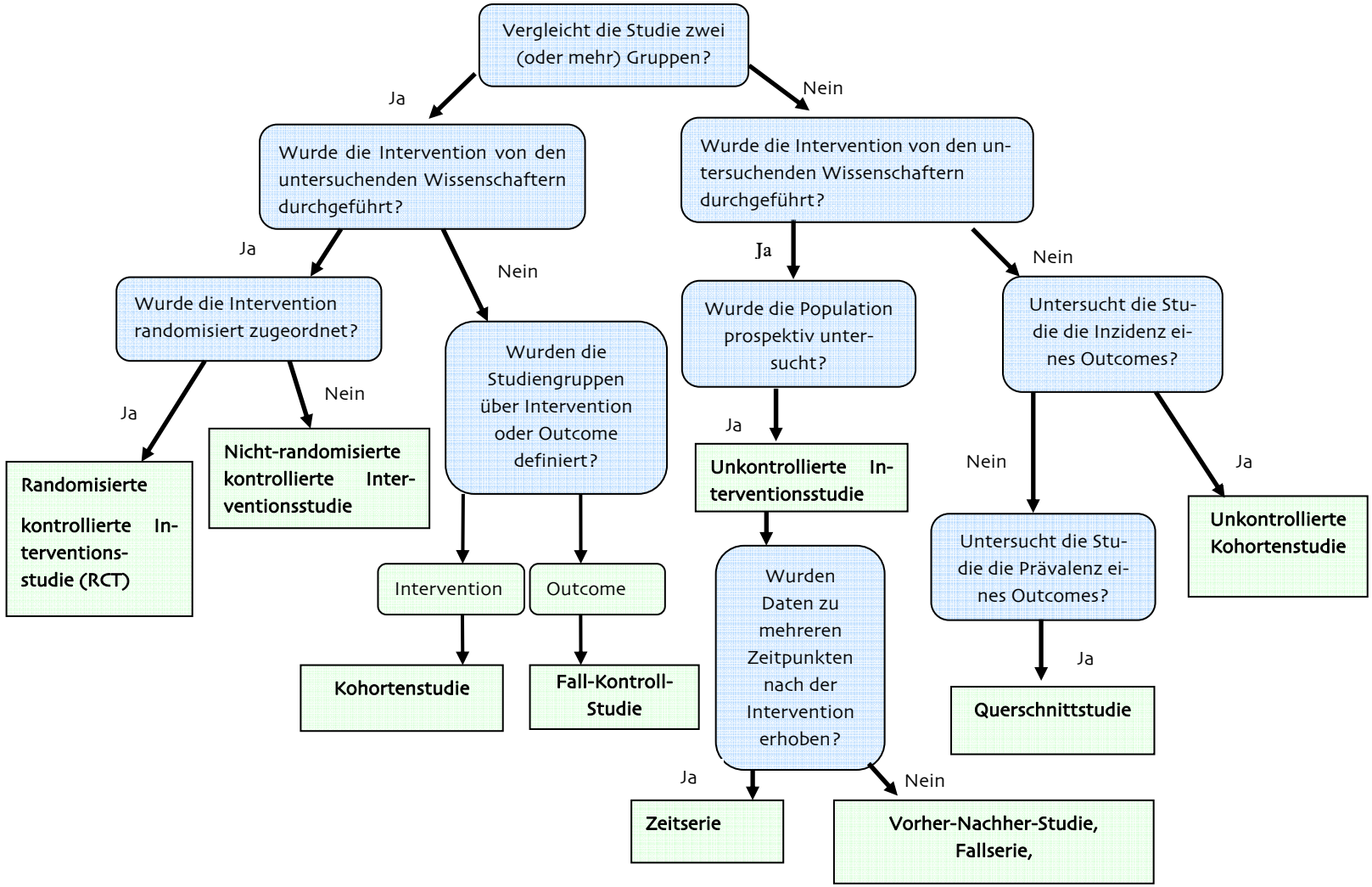

Abbildung 2.5.2-2: Algorithmus zur Klassifizierung von Studien

Studienglossar

Analytische Studien (Analytic studies)

untersuchen kausale Zusammenhänge & Kontrollgruppen

Analytische Studien untersuchen, ob ein kausaler Zusammenhang zwischen einer Intervention oder Exposition und einem definierten Outcome besteht. Grundsätzlich geht es in analytischen Studien darum, Ursache und Wirkung zu untersuchen. Analytische Studien haben meist eine Kontrollgruppe, die der Interventionsgruppe möglichst ähnlich sein soll, die aber keiner oder einer anderen Intervention ausgesetzt wird. Manche analytische Studien haben keine Kontrollgruppe, sondern beobachten eine Gruppe vor und nach einer Intervention. Analytische Studien können in zwei Gruppen unterteilt werden:

- Experimentelle Studien (Experimental studies)
- Beobachtungsstudien (Observational studies)

Beobachtungsstudien (Observational studies)

prospektiv oder retrospektiv

WissenschaftlerInnen setzen keine Intervention, sondern beobachten Gruppen mit und ohne Intervention. Beobachtungsstudien können pro- oder retrospektiv und abhängig vom Design analytisch oder beschreibend sein. Eine analytische Beobachtungsstudie wäre zum Beispiel eine Kohortenstudie oder eine Fall-Kontroll-Studie. Beispiele beschreibender Beobachtungsstudien wären Fallserien, Fallberichte, Querschnittstudien oder unkontrollierte Kohortenstudien zur Erhebung von Inzidenzdaten. Der Übergang zwischen analytischen und beschreibenden Beobachtungsstudien ist oft fließend.

Beschreibende Studien (Descriptive studies)

unterstützen Verständnis von Krankheit und Behandlung

Beschreibende Studien helfen ein besseres Verständnis für Krankheits- oder Behandlungscharakteristika zu entwickeln. Sie sind wichtig, um neue Hypothesen zu generieren, und sind häufig die erste verfügbare Evidenz. Beschreibende Studien haben keine Kontrollgruppen und erlauben daher keine Beurteilung von kausalen Zusammenhängen oder Assoziationen. Die wichtigsten beschreibenden Studien sind Querschnittstudien, Fallserien und unkontrollierte Kohortenstudien.

Experimentelle Studien (Interventionsstudien, Trials)

immer prospektiv

In experimentellen Studien werden Interventionen durchgeführt und Studiengruppen prospektiv beobachtet, um Unterschiede in Zielvariablen zu erheben. Experimentelle Studien sind RCTs, nicht-randomisierte kontrollierte Studien, Vorher-Nachher-Studien und Zeitserien. Experimentelle Studien sind immer prospektiv. Der Übergang zwischen experimentellen Studien und Beobachtungsstudien ist jedoch auch oft fließend.

Fall-Kontroll-Studien (Case-control studies)

immer retrospektiv

Fall-Kontroll-Studien identifizieren Populationen anhand eines bestimmten Outcomes. „Fälle" haben diesen Outcome (z.B. bestimmte Erkrankung), Personen in der Kontrollgruppe haben den Outcome nicht. Fall- und Kontrollgruppe sollen in Bezug auf Patientencharakteristika und Störgrößen (Confounder) möglichst ähnlich sein. Fall-Kontroll-Studien untersuchen Unterschiede in Expositionen zwischen Fall- und Kontrollgruppen, um die Ursache für einen bestimmten Outcome zu erklären. Fall-Kontroll-Studien sind immer retrospektiv.

Fallberichte (Case reports)

Beschreibung des Krankheitsverlaufs

Ein Fallbericht beschreibt den Krankheitsverlauf eines/einer bestimmten Patienten/in (Krankheit auf individueller Ebene). Fallberichte beschreiben häufig neue oder seltene Krankheiten. Fallberichte können auch wichtig für die Darstellung von seltenen Nebenwirkungen von Medikamenten oder anderen Interventionen sein.

Fallserien (Case series)

keine Kontrollgruppe, keine kausalen Zusammenhänge

Eine Fallserie synthetisiert den Verlauf einer Gruppe von Personen mit gleicher Diagnose, Intervention oder Exposition. Fallserien werden retrospektiv zusammengestellt. Idealerweise berichtet eine Fallserie über aufeinander folgende Fälle einer genau beschriebenen Population mit einer genau definierten Intervention oder Exposition. Prospektive Fallserien entsprechen Vorher-Nachher-Studien. Fallserien sind nützlich, um neue Hypothesen zu erstellen, neue Krankheiten zu definieren oder seltene Nebenwirkungen zu beschreiben. Fallserien haben keine Kontrollgruppen und können keinen kausalen Zusammenhang testen. Gelegentlich werden Fallserien in Zusammenhang mit historischen Kontrollgruppen präsentiert. Historische Kontrollgruppen unterscheiden sich jedoch oft wesentlich in Co-Interventionen und prognostischen Faktoren und haben daher eine sehr begrenzte Vergleichbarkeit.

Kohortenstudien (Cohort studies)

Gruppen mit ähnlichen Prognosen werden beobachtet

Analytische Kohortenstudien vergleichen Gruppen von Personen mit Exposition zu jenen ohne Exposition. Idealerweise sollten alle Gruppen möglichst ähnlich in prognostischen Faktoren sein. Selektionsbias ist das Hauptproblem bei Kohortenstudien. Kohortenstudien können prospektiv oder retrospektiv sein. In prospektiven Studien werden Gruppen über bestimmte Zeiträume beobachtet. Wenn zu Beginn der Studie Daten sowohl über Exposition als auch über Zielvariablen vorhanden sind, handelt es sich immer um eine retrospektive Studie.

Unkontrollierte Kohortenstudien werden zur Erhebung von Inzidenzdaten verwendet.

Unkontrollierte Interventionsstudien (Uncontrolled trials)

Interventionen ohne Kontrollgruppe

Interventionen werden durchgeführt, es gibt jedoch keine Kontrollgruppe. Abhängig davon, wie oft ein Outcome beurteilt wird, können unkontrollierte Interventionsstudien in Zeitserien oder Vorher-Nachher-Studien unterteilt werden.

Nicht-randomisierte Interventionsstudien (Nonrandomized trials)

mit Kontrollgruppe, aber ohne Zufall-Zuordnung

Interventionen werden durchgeführt, die Zuordnung der StudienteilnehmerInnen zu den Behandlungsgruppen wird jedoch nicht per Zufall bestimmt. Selektionsbias ist ein häufiges Problem in solchen Studien.

Randomisierte kontrollierte Interventionsstudien (Randomized controlled trials [RCTs])

mit Kontrollgruppe in Zufall-Zuordnung, aber oft geringer externer Validität

StudienteilnehmerInnen werden per Zufall (at random) einer Interventions- oder Kontrollgruppe zugeteilt. RCTs sind immer prospektiv. Randomisierung und die Unvorhersehbarkeit der Gruppenzuteilung (allocation concealment) führen dazu, dass prognostische Faktoren gleichmäßig über alle Gruppen verteilt werden. RCTs sind der „Goldstandard" der klinischen Prüfung von Interventionen. Ein Nachteil von RCTs ist, dass die Studienpopulation häufig stark selektiert ist und Resultate wenig externe Validität besitzen.

Querschnittstudien (Cross-sectional studies)

Prävalenzerhebung

Querschnittstudien bieten eine „Momentaufnahme" eines Outcomes. Querschnittstudien werden hauptsächlich zur Erhebung von Prävalenzdaten verwendet. Umfragen sind ein häufiges Format von Querschnittstudien.

Vorher-Nachher-Studien (Before-after studies)

Zeitserien:

anfällig für Störfaktoren & Verzerrungen

PatientInnen (oder Populationen, Gesundheitssysteme) werden vor und nach einer Intervention untersucht und beschrieben. Es gibt keine externe Kontrollgruppe. Der Status vor der Intervention wird mit dem nach der Intervention verglichen. Die Interventionsgruppe ist daher gleichzeitig Kontrollgruppe. Vorher-Nachher-Studien, die Daten über die Zielvariable zu mehreren Zeitpunkten erheben, werden oft als Zeitserien bezeichnet. Die Aussagekraft von Vorher-Nachher-Studien wird durch Regression zum Mittelwert und unspezifische Therapieeffekten (z.B. Placeboeffekt) eingeschränkt.

Zeitserien (Time series)

PatientInnen (oder Populationen, Gesundheitssysteme) werden vor und nach einer Intervention untersucht und beschrieben. Es gibt keine externe Kontrollgruppe. Der Status vor der Intervention wird mit dem nach der Intervention verglichen. Die Interventionsgruppe ist daher gleichzeitig Kontrollgruppe. Die Zielvariable wird prospektiv zu mehreren Zeitpunkten erhoben. Interventionen werden durchgeführt, es gibt jedoch keine Kontrollgruppe. Zeitserien, die Daten nur zu Beginn und am Ende der Studie erheben, werden als Vorher-Nachher-Studie bezeichnet. Die Aussagekraft von Zeitserien wird durch Regression zum Mittelwert und unspezifische Therapieeffekte (z.B. Placeboeffekt) eingeschränkt.

2.5.3 Studienhierarchie

Unterschiedliche Studiendesigns sind zu unterschiedlichen Graden anfällig für Bias und Confounding. [47] Es besteht kein internationaler Konsens bezüglich einer Rangordnung unterschiedlicher Studientypen. [6,48-52] RCTs werden generell als jenes Design gesehen, das systematische Fehler am besten verhindern kann. Die folgende Hierarchie für therapeutische Studien basiert auf einem Schema des Centre for Reviews and Dissemination, University of York[52]:

1. Systematische Übersichtsarbeiten und Meta-Analysen von RCTs
2. RCTs
3. Nicht-randomisierte Interventionsstudien

Kontrollierte Beobachtungsstudien
   a. Kohortenstudien (prospektiv, retrospektiv)
   b. Fall-Kontroll-Studien
4. Beobachtungsstudien ohne Kontrollgruppe
5. ExpertInnenmeinung

RCTs oft,

aber keinesfalls immer am aussagekräftigsten

Dieses Stufensystem sollte jedoch mit Bedacht und unter Einbeziehung anderer Faktoren (z.B. interne und externe Validität, Studiengröße, Finanzierung) verwendet werden.[53,54] Für manche Fragestellungen sind RCTs nicht durchführbar[55,56] oder sie sind durch Studiengröße und Länge limitiert. In solchen Situationen sind Beobachtungsstudien aufschlussreicher.[55] Für therapeutische Studien sind RCTs daher meistens, aber nicht immer das ideale Studiendesign.[47] Eine gute prospektive Beobachtungsstudie kann aufschlussreicher sein als ein mittelmäßiger RCT.[57] Für diagnostische, prognostische, ökonomische oder Inzidenz/Prävalenz-Fragestellungen muss die Hierarchie der Studientypen anders gewichtet werden.[58]

2.5.4 Einschluss von Studien

breiten, offenen Zugang zu Publikationen bewahren, ...

Auch wenn Auswahlkriterien explizit definiert werden, bleibt ein Teil der Entscheidungsfindung subjektiv. Der endgültige Einschluss von Studien in einen systematischen Review erfolgt nach der Beurteilung der internen Validität (methodologische Qualität) und sobald ein Überblick über die Gesamtheit der Evidenz gewonnen werden konnte. Zu diesem Zeitpunkt ist es manchmal nötig, Auswahlkriterien zu revidieren, um die bestmögliche Evidenz finden zu können (Best Evidence Approach). Wenn zum Beispiel eine therapeutische Fragestellung mit RCTs nicht beantwortet werden kann, sollte man auf nicht-randomisierte Studien oder kontrollierte Beobachtungsstudien zurückgreifen.

... um Nebenaspekte nicht zu verlieren

Je nach Breite der Perspektive sind zur Klärung von Umfeldfragestellungen und bei sozial-organisatorischen Aspekten der Anwendung (Praxisvarianzen in der Indikationsstellung, Fragen professioneller Qualitätssicherung etc.) jedenfalls auch andere Studien und Quellen einzubeziehen.

2.6 Datenextraktion

HAUPTPUNKTE:

- Evidenztabellen sollen studienspezifisch erstellt werden und für die Fragestellung relevante Informationen beinhalten.
- Formulare für die Datenextraktion sollten getestet werden.
- Der Inhalt des Extraktionsformulars muss genau definiert werden, um Genauigkeit und Konsistenz zu gewährleisten.
- Extrahierte Daten sollen von einer zweiten Person unabhängig verifiziert werden.

2.6.1 Design einer Evidenztabelle

kompakte Zusammenfassungen...

Evidenztabellen sollen nur jene Daten präsentieren, die für die Fragestellung relevant sind. Evidenztabellen können sich daher von Projekt zu Projekt unterscheiden. Die Appendices D bis F bieten Grundrisse von Formularen, die für die Datenextraktion von therapeutischen, diagnostischen und ökonomischen[59] Studien verwendet werden können. Diese Formulare können auch für die Erstellung von Evidenztabellen eingesetzt werden. Evidenztabellen erfüllen zwei wesentliche Aufgaben:

1. Sie bieten eine kompakte Zusammenfassung einzelner Studien und können beim Schreiben des Berichts verwendet werden.
2. Leser finden in den Evidenztabellen genauere Details zu einzelnen Studien.

...mit relevanten Detailinformationen

Folgende Punkte sollten in Evidenztabellen enthalten sein:

- Allgemeine Information über Publikation und Studie
  - AutorIn, Jahr
  - Finanzierung
  - Staat/Gesundheitssystem, in dem die Studie durchgeführt wurde
  - Zielsetzung der Studie
- Spezifische Information über die Studie
  - Studiendesign
  - Studiendauer
  - Studiengröße
  - Auswahlkriterien der Population
  - Charakteristika der Studienpopulation
  - Intervention/Exposition
  - Zielvariablen
- Resultate
  - Zielvariablen, die für die Fragestellung relevant sind; die Information sollte Effektgröße, Konfidenzintervall und P – Wert beinhalten und nicht nur auf statistisch signifikante Resultate beschränkt sein.
- Beurteilung der internen Validität (Qualität)
- Kommentare

2.6.2 Extraktion der Daten

Trend: elektronische Datenextraktion

Datenextraktion kann auf verschiedene Art und Weise erfolgen. Meistens werden Formulare in Word, Access oder Excel erstellt. Für die Extraktion können Papier- oder elektronische Formulare verwendet werden. Der Trend geht wahrscheinlich in Richtung webbasierte Datenextraktion (z.B. TrialStat®).

Wesentlich sind klare Anleitungen und Definitionen, welche Daten wie extrahiert werden müssen. Dies verhindert, dass Evidenztabellen zu unterschiedlich werden und später mit viel Aufwand korrigiert und editiert werden müssen.

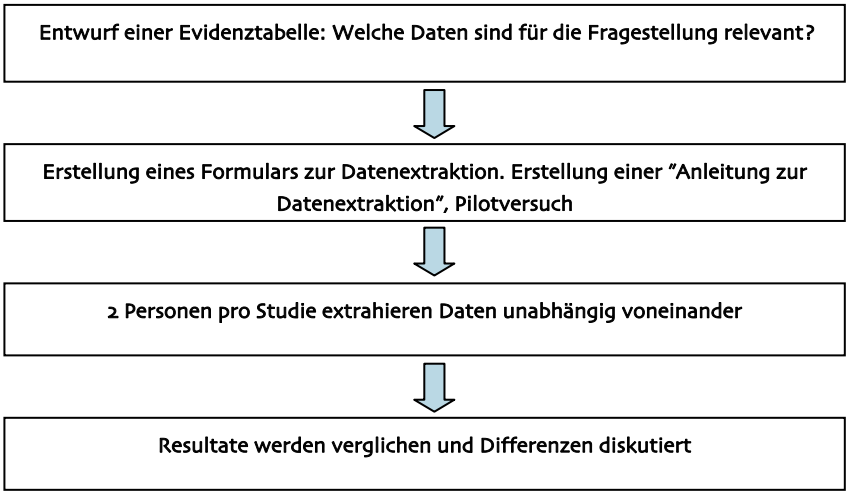

Abbildung 2.6.2-1: Datenextraktion

Meistens werden mehr Daten extrahiert als später in den Evidenztabellen präsentiert werden können. Daten für Meta-Analysen werden zum Beispiel selten in Evidenztabellen dargestellt. Es ist jedoch effizient, ein grobes Konzept der Evidenztabelle zur Datenextraktion zu benutzen.

Datenextraktionen sind fehleranfällig

Datenextraktionen sind extrem fehleranfällig und erfordern gelegentlich subjektive Entscheidungen. Methodikstudien haben gezeigt, dass Daten, die von nur einer Person extrahiert wurden, eine um 21% höhere Fehlerrate haben als Daten, die von zwei Personen extrahiert wurden.[60] Datenextraktion sollte daher immer durch zwei Personen unabhängig voneinander erfolgen. Dies kann durch Parallel-Extraktion erfolgen, das heißt, beide ReviewerInnen extrahieren alle Daten einer Studie und vergleichen die Ergebnisse. Eine andere, effizientere, aber ungenauere Methode ist, dass extrahierte Daten des ersten Reviewers von einem zweiten Reviewer kontrolliert werden. Auf jeden Fall sollten Unterschiede immer diskutiert und per Konsens gelöst werden.

Pilotversuch für Extraktionsformulare

Zu Beginn jedes Projekts ist es unbedingt nötig, Datenformulare in einem Pilotversuch zu testen. Alle ReviewerInnen extrahieren dabei Daten der gleichen Studie(n) und vergleichen die Ergebnisse bzw. diskutieren Probleme mit dem Formular. Mehrere Pilotversuche können notwendig sein, um ein Formular zu entwickeln, das Genauigkeit und Konsistenz garantiert.

Mehrfachpublikationen ausschließen,

Im Zuge der Datenextraktion wird es immer wieder vorkommen, dass Reviewer-Innen befinden, dass Studien doch nicht die Auswahlkriterien erfüllen. Der Grund für den Ausschluss muss dann dokumentiert werden. Zusätzlich sollte vermieden werden, dass Mehrfachpublikationen, die sich auf eine Studie beziehen, als unterschiedliche Studien extrahiert werden. Eine genaue Durchsicht und Gruppierung der Artikel nach identischen Studien ist daher extrem wichtig. Sekundärartikel zitieren den Primärartikel oft unzureichend oder werden vor einem Primärartikel publiziert. Eine exakte Durchsicht der einzelnen Artikel verhindert, dass ein Sekundärartikel fälschlicherweise als Primärstudie extrahiert und präsentiert wird.

Primär- und Sekundärartikel unterscheiden

fehlende Daten nicht schätzen

Publizierte Studien beinhalten nicht immer alle Daten, die für einen systematischen Review von Interesse sind. Fehlende Daten sind immer problematisch und können zu Bias führen.[38] Manche Daten können errechnet oder von anderen Daten abgeleitet werden (z.B. Dropout-Raten). Häufig ist dies jedoch nicht möglich. In einem solchen Fall sollten StudienautorInnen kontaktiert werden. Daten sollen keinesfalls von Abbildungen „geschätzt" werden.

2.7 Beurteilung der internen und externen Validität

HAUPTPUNKTE:

- Kriterien zur Beurteilung der Studienqualität sollen auf einem validierten System beruhen.
- Punktsysteme sollten vermieden werden.
- 2 Personen sollen unabhängig voneinander die Qualität der Studien beurteilen.
- Es muss vor Beginn des Reviews entschieden werden, wie mit Studien, die schwere methodologische Mängel haben, verfahren wird.

2.7.1 Was ist interne Validität (methodologische Qualität)?

Studien, die den Auswahlkriterien entsprechen, werden einer kritischen methodologischen Evaluierung unterzogen. Ziel dieser Evaluierung ist es, die beste, verfügbare Evidenz zu identifizieren.[53,61] 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.[61, 62] In anderen Worten, interne Validität ist die Wahrscheinlichkeit, dass Resultate möglichst nahe an die „Wahrheit" herankommen. Jedes Resultat einer Studie setzt sich aus folgenden Komponenten zusammen.

Resultat = Wahrheit + Bias + Zufallsfehler (random error)

Das grundsätzliche Problem bei der Beurteilung der internen Validität ist, dass wir zwar das Resultat kennen, nicht aber das Gewicht der einzelnen Komponenten. „Wahrheit" bleibt immer eine unbekannte Variable, „Zufall" kann durch Studiengröße entsprechend minimiert werden. Bei der Beurteilung der internen Validität untersucht man daher die Wahrscheinlichkeit, ob Bias das Resultat wesentlich beeinflusst.

Biasminimierung

Der Einfluss von Bias und methodologischer Qualität auf Resultate lässt sich jedoch nicht quantifizieren. Wir müssen uns daher bei der Beurteilung der Studienqualität auf einzelne Aspekte des Studiendesigns beschränken, die Bias minimieren können.[63] Zum Beispiel: War die Randomisierungsmethode bei RCTs adäquat? Wurde eine Intention-to-Treat-Analyse (ITT-Analyse) korrekt durchgeführt? Wie hoch war die Dropout-Rate? Wenn diese methodologischen Aspekte zufrieden stellend ausgeführt wurden, können wir davon ausgehen, dass Bias die Studienergebnisse nicht wesentlich verzerrt.

Die Information, die man durch kritische Beurteilung der Methodik gewinnt, ist wesentlich für die Beurteilung der „Stärke der Evidenz". Abbildung 2.7.1-1 fasst die Vorgangsweise bei der Beurteilung der internen und externen Validität zusammen.

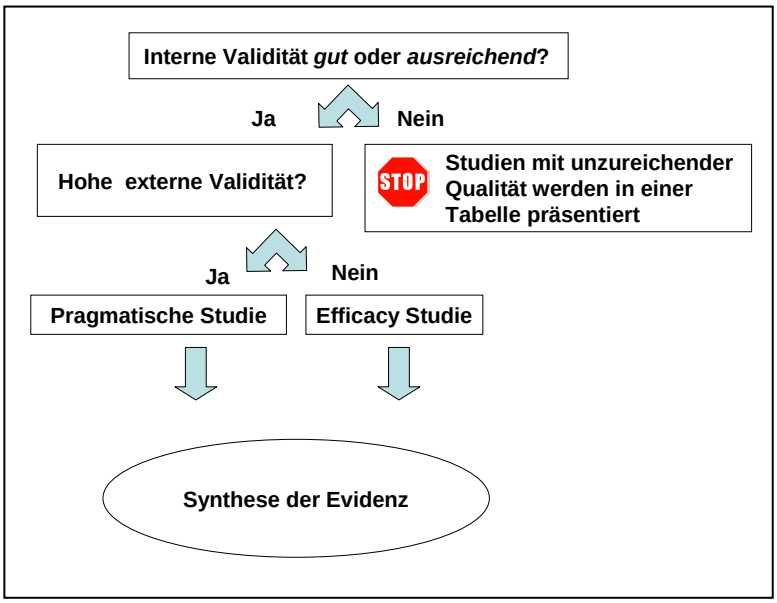

Abbildung 2.7.1-1: Beurteilung der internen und externen Validität

2.7.2 Kriterien zur Beurteilung der internen Validität

Komponenten des Studiendesigns werden evaluiert

Es gibt eine Vielzahl an Instrumenten zur Beurteilung der Qualität von Studien.[50,64] Methodikstudien haben jedoch gezeigt, dass Systeme, die ein Punkte-system verwenden, zu wenig verlässlichen Ergebnissen führen.[65] Eine bessere Methode ist, elementare Komponenten des Designs und der Ausführung von Studien zu beurteilen.[62] Solche Komponenten unterscheiden sich für unterschiedliche Studientypen. Für RCTs wären dies zum Beispiel Randomisierung, Unvorhersehbarkeit der Gruppenzuordnung (allocation concealment), Verblindung und Dropout-Rate.

standardisierte Instrumente & Kriterienlisten liegen vor

Bei der Beurteilung der internen Validität wird eine Studie in Bezug auf diese Komponenten evaluiert und diesbezüglich beurteilt. Die Qualität einzelner Studien wird immer von zwei Personen unabhängig beurteilt. Differenzen werden durch Diskussion und Konsens oder durch die Einbringung einer anderen Person gelöst. Standardisierte Formulare, die zur Beurteilung der internen Validität verschiedener Studientypen verwendet werden können, sind in Appendix G zusammengefasst. Die Instrumente zur Beurteilung der internen Validität von RCTs und systematischen Reviews basieren auf Systemen, die von der U.S. Preventive Services Task Force[51] und dem Centre for Reviews and Dissemination, University of York[52] entwickelt wurden. Beobachtungsstudien werden nach Kriterien beurteilt, die von Deeks et al. beschrieben wurden.[66] Die interne Validität diagnostischer Studien wird unter Berücksichtigung von Kriterien beurteilt, die im QUADAS (Quality Assessment of Diagnostic Accuracy Studies) dargelegt wurden.[67] Kriterien zur Beurteilung der internen Validität von ökonnomischen und entscheidungsanalytischen Studien basieren auf Arbeiten von Siebert et al.[68] und Philips et al.[69]

Kriterien zur Beurteilung von RCTs

RCTs:

Wie bei allen kontrollierten Studien geht es auch bei RCTs darum, wie ähnlich die einzelnen Studiengruppen sind. Studiengruppen, die sich in keinen wesentlichen Charakteristika voneinander unterscheiden, sind Voraussetzung für interpretierbare Resultate.[70,71] Im folgenden Abschnitt werden Fragen präsentiert, die für die Evaluierung kritischer Aspekte der internen Validität unterschiedlicher Studiendesigns verwendet werden können.

Folgende Fragen sollten bei der kritischen Beurteilung von RCTs beantwortet werden:

adäquate Randomisierung

- War die Randomisierung adäquat?

  Eine adäquate Randomisierung wäre:
  - Computer-generierte Randomisierung
  - Randomisierung durch unbeteiligte Dritte (zentralisierte Randomisierung)

  Eine inadäquate Randomisierung wäre:
  - Geburtsdaten
  - Ambulanznummer
  - Abwechselnde Zuordnung

Unvorhersehbarkeit der Gruppenzuordnung

- War die Unvorhersehbarkeit der Gruppenzuordnung adäquat (allocation concealment)?

  Ein adäquater Prozess wäre:
  - Zentralisierte Randomisierung
  - Seriell nummerierte, identische Container
  - Jeder andere Prozess, der es unmöglich macht, die Sequenz der Gruppenzuordnung vorauszusehen

  Ein inadäquater Prozess wäre:
  - Briefumschläge, die nicht verschlossen, blickdicht und durchlaufend nummeriert sind
  - Offene Randomisierungslisten
  - Ambulanznummer, Geburtsdaten etc.

Ähnlichkeit der Gruppen in klinisch relevanten Aspekten

- Waren wesentliche Charakteristika der Studiengruppen ähnlich?

  Eine adäquate Randomisierung führt zu Studiengruppen, die sich in allen Aspekten ähnlich sind (bekannte und unbekannte Confounder). Bei kleinen Studien können Zufallsfehler (random errors) allerdings dazu führen, dass sich einzelne Charakteristika zwischen den Studiengruppen auch bei gut durchgeführter Randomisierung unterscheiden. Eine Beurteilung der Charakteristika der Studiengruppen sollte daher in erster Linie die klinische Relevanz von Unterschieden beurteilen und sich nicht auf statistische Resultate verlassen.

Studiengröße

- Basiert die Studiengröße auf einer adäquaten Berechnung, die Power und einen kleinsten wesentlichen Unterschied einbezieht (minimal important difference)?

  Studien, die nicht genug Power haben, führen zu statistisch nichtsignifikanten Resultaten, denen jedoch ein klinisch wesentlicher Unterschied zugrunde liegen könnte.

Verblindung bei der Outcome-Beurteilung

- Wurde die Verblindung adäquat durchgeführt?

  Idealerweise sollten alle beteiligten Parteien verblindet sein. Manchmal ist dies allerdings nicht durchführbar (z.B. chirurgische Studien). Jene Personen, die Outcomes beurteilen, sollten allerdings bis auf wenige Ausnahmen immer verblindet sein (evaluator blinding).

Dropout-Rate & ...

- Gab es eine hohe Dropout-Rate?

  Die Dropout-Rate (attrition, loss to follow-up) setzt sich aus jenen Personen zusammen, die randomisiert wurden, aber nicht bis zum Studienende verblieben sind. Eine erhöhte Dropout-Rate kann zu Selektionsbias führen. Es gibt keine einheitlichen Empfehlungen, wie hoch die Dropout-Rate sein darf. Eine Dropout-Rate von 20% wird häufig als Grenze verwendet. In manchen Populationen (z.B. psychiatrische Populationen) muss die Grenze aus praktischen Gründen höher angesetzt werden.

...differentielle Dropout-Rate

- Gab es eine hohe differentielle Dropout-Rate?

  Dropouts können insbesondere dann zu Problemen führen, wenn die Dropout-Rate in einer Gruppe wesentlich größer ist als in einer anderen. Es gibt keine einheitlichen Empfehlungen, wie hoch eine differentielle Dropout-Rate sein kann. Ein Unterschied von 15 Prozentpunkten wird häufig als Grenze verwendet.

Intention-to-Treat-Analyse

- Wurde eine ITT-Analyse adäquat durchgeführt?

  In einer idealen ITT-Analyse müssen ALLE Personen in jener Gruppe analysiert werden, in die sie randomisiert wurden, unabhängig von Cross-Over, Protokollverstößen und Dropouts.

- Gab es Ausschlüsse nach der Randomisierung (post-randomization exclusions)?

  In einer idealen ITT-Analyse sollen keine randomisierten PatientInnen von der Analyse ausgeschlossen werden. Eine kleine Anzahl wird jedoch aus praktischen Gründen meistens toleriert (z.B. PatientInnen, die nie eine Intervention erhielten).

Kriterien zur Beurteilung von Kohortenstudien

Problem Selektionsbias

Das methodologische Hauptproblem bei Kohortenstudien (wie bei allen Beobachtungsstudien) ist Selektionsbias.[72-74] Eine Ungleichverteilung von prognostischen Faktoren kann zu verzerrten Resultaten führen. Ob die Studiengruppen ausreichend ähnlich sind, lässt sich allerdings nie mit Sicherheit beurteilen.

Gleichverteilung der Prognosefaktoren

Folgende Fragen sollten bei der kritischen Beurteilung von Kohortenstudien beantwortet werden:

Studiengruppen aus derselben Population

- Wurden die Studiengruppen aus derselben Population rekrutiert?

  Kontroll- und Expositionsgruppen sollen aus derselben Population rekrutiert werden, um Unterschiede von Störgrößen (Confounder) zu minimieren. Historische oder externe Kontrollgruppen können sich zum Beispiel wesentlich in Begleiterkrankungen, zusätzlichen Risikofaktoren, prognostischen Faktoren usw. unterscheiden. Die Folge kann Selektionsbias sein.

Vergleich anhand von Outcomes

- Ist die Verteilung der prognostischen Faktoren zwischen den Gruppen ausreichend beschrieben?

  In vielen Publikationen vergleicht Tabelle 1 die Charakteristika der Studiengruppen. In dieser Tabelle sollen die wichtigsten prognostischen Faktoren der Studiengruppen in Bezug auf einen bestimmten Outcome verglichen werden.

- Ist die Verteilung der prognostischen Faktoren zwischen den Gruppen ähnlich?

  Um Selektionsbias zu vermeiden, sollen prognostische Faktoren zwischen den Studiengruppen zu Beginn der Studie möglichst ähnlich sein. Falls wesentliche Unterschiede auftreten, müssen diese bei der statistischen Analyse adjustiert werden.

ähnliches Basisrisiko

- Haben alle Gruppen dasselbe Risiko für den Outcome?

  Um die Auswirkung einer Exposition (Intervention) beurteilen zu können, müssen Studiengruppen zu Beginn der Studie ein ähnliches Basisrisiko haben, einen bestimmten Outcome zu erfahren. Unterschiede in Risiken führen zu Selektionsbias.

zeitgleiche Rekrutierung

- Wurden alle Gruppen während derselben Zeitperiode rekrutiert?

  Wenn Studiengruppen zu unterschiedlichen Zeitperioden rekrutiert werden (z.B. historische Kontrollgruppen), kann dies zu wesentlichen Unterschieden in prognostischen Faktoren (z.B. Fortgeschrittenheit einer Erkrankung, zusätzliche Behandlungen) und somit Selektionsbias führen., da sich Diagnosekriterien im Laufe der Zeit ändern können bzw. Begleitbehandlungen verbessert werden.

identische Outcome-Beurteilung

- Wurden Outcomes in allen Gruppen auf gleiche Art und Weise beurteilt?

  Die Beurteilung des Outcomes muss in beiden Gruppen auf identische, valide und verlässliche Art und Weise durchgeführt werden (Beobachtungsgleichheit). Unterschiede können zu einer Verzerrung der Ergebnisse durch Messfehler (measurement bias) und Missklassifikationen führen.

Verblindung bei der Outcome-Beurteilung

- Wurden Outcomes verblindet beurteilt?

  Idealerweise sollten Personen, die Outcomes beurteilen, verblindet sein, um Beobachtungsgleichheit zu erreichen, allerdings ist dies nicht immer möglich. Mangelnde Verblindung kann zu einer Verzerrung der Ergebnisse durch Messfehler (measurement bias und detection bias) führen.

gleiche Studienlaufzeit

- War die Studienlaufzeit für alle Gruppen identisch?

  Unterschiedliche Studienlaufzeiten zwischen den einzelnen Gruppen führen zu einem Missverhältnis in prognostischen Faktoren und dadurch zu Selektionsbias.

Dropout-Rate &

- Gab es eine hohe Dropout-Rate?

  Die Dropout-Rate (attrition, loss to follow-up) setzt sich aus jenen Personen zusammen, die zu Studienbeginn in die Kohorte aufgenommen wurden, aber nicht bis zum Studienende verblieben sind. Eine erhöhte Dropout-Rate führt zu Selektionsbias.

differentielle Dropout-Rate

- Gab es eine hohe differentielle Dropout-Rate?

  Dropouts können insbesondere dann zu Problemen führen, wenn die Dropout-Rate in einer Gruppe wesentlich größer ist als in einer anderen. Es gibt keine einheitlichen Empfehlungen, wie hoch eine differentielle

Dropout-Rate sein kann. Ein Unterschied von 15 Prozentpunkten wird häufig als Grenze verwendet.

Risiko-Adjustierung

- Wurden potentielle Störgrößen (Confounder) in der statistischen Auswertung berücksichtigt?

  Um Bias zu vermeiden, müssen Unterschiede in prognostischen Faktoren während der statistischen Analyse adjustiert werden. Multifaktorielle Methoden können solche Unterschiede berücksichtigen. Es bleibt jedoch immer das Risiko, dass unbekannte Confounder zu systematischen Unterschieden zwischen Studiengruppen führen.

Kriterien zur Beurteilung von Fall-Kontroll-Studien

Problem Selektionsbias

Wie bei Kohortenstudien ist Selektionsbias das wesentliche Problem bei Fall-Kontroll-Studien.[75] In Fall-Kontroll-Studien müssen Charakteristika der „Fall-Population" und der „Kontroll-Population" nicht unbedingt identisch sein. Wichtiger ist, dass jede Person in der Kontrollgruppe potentiell auch ein Fall sein könnte. Folgende Beurteilung ist spezifisch für das Design von Fall-Kontroll-Studien:

exakte „Fall"-Definition ist wichtig

- Wurde ein „Fall" exakt definiert?

  Ein „Fall" muss genau definiert werden, um später einen kausalen Zusammenhang mit einer Exposition herstellen zu können.

Kontrollgruppe aus derselben Population

- Wurde die Kontrollgruppe aus derselben Population wie die Fälle ausgewählt?

  Fall und Kontrollgruppen müssen aus derselben Population rekrutiert werden, um Unterschiede von Störgrößen zu minimieren. Historische oder externe Kontrollgruppen können sich zum Beispiel wesentlich in Begleiterkrankungen, zusätzlichen Risikofaktoren, prognostischen Faktoren usw. unterscheiden. Die Folge ist Selektionsbias.

- Ist die Verteilung der prognostischen Faktoren zwischen den Gruppen ausreichend beschrieben?

  In vielen Publikationen vergleicht Tabelle 1 die Charakteristika der Studiengruppen. In dieser Tabelle sollen die wichtigsten prognostische Faktoren der Studiengruppen in Bezug auf einen bestimmten Outcome verglichen werden.

Risiko-Adjustierung

- Ist die Verteilung der prognostischen Faktoren zwischen den Gruppen ähnlich?

  In Fall-Kontrollstudien ist es unwahrscheinlich, dass sich prognostische Faktoren zwischen den Studiengruppen ähnlich sind. Häufig werden die wichtigsten prognostische Faktoren bei der Auswahl der Kontrollpersonen ähnlich gehalten (matching). Wichtig ist, dass wesentliche Unterschiede bei der statistischen Analyse adjustiert werden.

identische Outcome-Beurteilung

- Wurde die Exposition (Intervention) in Fall und Kontrollgruppen auf gleiche Art und Weise beurteilt?

  Bei einer Fall-Kontrollstudie ist der Outcome zu Beginn der Studie bekannt. Die Beurteilung der Exposition muss in beiden Gruppen auf identische, valide und verlässliche Art und Weise durchgeführt werden (Beobachtungsgleichheit). Unterschiede können zu einer Verzerrung der Ergebnisse durch Messfehler (measurement bias) und Missklassifikationen führen.

- War die Rate der Personen, die eine Beteiligung an der Studie abgelehnt haben, in beiden Gruppen ähnlich?

  Personen mit Outcome (Fälle) und ohne Outcome (Kontrolle) müssen bei Fall-Kontrollstudien aktiv rekrutiert werden. Wenn es in einer Gruppe Schwierigkeiten gibt, Personen zu rekrutieren, kann dies zu ungleichen Studiengruppen führen.

- Ist „Over-matching" möglich?

  Over-matching:

  zu große Ähnlichkeit zwischen Fall- & Kontrollpersonen

  In Fall-Kontrollstudien müssen Fall und Kontrollpersonen *nicht* möglichst identisch sein. Für jeden „Fall" wird eine Kontrollperson rekrutiert, die in einigen wichtigen Charakteristika ähnlich ist. Individuelle Fall- und Kontrollpersonen können jedoch *zu* ähnlich sein, wenn sie sich in Faktoren ähneln, die kausal mit dem Outcome verbunden sind. Overmatching kann zu falsch-negativen Ergebnissen führen.

- Wurde eine adäquate statistische Analyse durchgeführt?

  Effektmaß: Odds ratio

  Um Bias zu vermeiden, müssen Unterschiede in prognostischen Faktoren während der statistischen Analyse adjustiert werden. Multifaktorielle Methoden können solche Unterschiede berücksichtigen. Es bleibt jedoch immer das Risiko, dass unbekannte Confounder zu systematischen Unterschieden zwischen Studiengruppen führen. Odds ratios sind jenes Effektmaß, das bei Fall-Kontroll-Studien verwendet wird.

Kriterien zur Beurteilung von systematischen Übersichts arbeiten

a priori definiertes Vorgehen:

Das wesentlichste Qualitätskriterium bei systematischen Reviews ist ein a priori definiertes, systematisches Vorgehen, das Subjektivität und nachfolgenden Selektionsbias minimieren kann.[26] Eine systematische, objektive Beurteilung der Evidenz unterscheidet systematische Reviews von konventionellen, nicht-systematischen Übersichtsarbeiten (narrative reviews), die hauptsächlich von Expertenmeinungen getragen werden.[26 27]

Folgende Fragen sollten bei der kritischen Beurteilung von systematischen Übersichtsarbeiten und Meta-Analysen beantwortet werden:

- Basiert der Review auf einer klar definierten Frage?

  wissenschaftliche Fragestellung

  Die wissenschaftliche Fragestellung, die ein systematischer Review zu beantworten versucht, muss a priori formuliert werden. Die Fragestellung darf nicht post hoc der vorhandenen Evidenz angepasst werden. Populationen, Interventionen, Kontrollinterventionen und Outcomes müssen in der Fragestellung definiert werden.

- Wurden Auswahlkriterien klar definiert?

  Auswahlkriterien statt Überblick über Gesamtliteratur

  Auswahlkriterien für Studien müssen ebenfalls a priori definiert werden und sollen klare Entscheidungen bezüglich den Ein- und Ausschluss von Studien ermöglichen. Eine systematische Übersichtsarbeit soll nicht eine Zusammenfassung der gesamten, jemals publizierten Evidenz sein, sondern soll die beste, verfügbare Evidenz zusammenfassen und synthetisieren.

- Wurde eine systematische Literatursuche in mehreren Datenbanken durchgeführt?

  systematische Literatursuche in mehreren Quellen

  Publikationsbias und Retrievalbias lassen sich nie gänzlich vermeiden, können aber mit einer sorgfältigen, systematischen Literatursuche minimiert werden. Fehlende Systematik führt zusätzlich zu Selektionsbias.

duale Beurteilung

- Haben zumindest 2 Personen die Studien beurteilt?

  Die Entscheidung, ob Studien die Auswahlkriterien erfüllen, ist immer auch von der subjektiven Einschätzung abhängig. Entscheidungen sollen daher von einer zweiten Person evaluiert werden, um falsch-positive oder falsch-negative Resultate zu korrigieren.

Qualitätsvalidierung

- Wurde die methodologische Qualität der Studien beurteilt?

  Unzureichende methodologische Qualität kann Resultate verzerren. Die Beurteilung der internen Validität soll mit Hilfe eines validierten und verlässlichen Schemas erfolgen.

- Wurde die methodologische Qualität der Studien bei der Evidenzsynthese berücksichtigt?

  Studien mit guter interner Validität sollen bei der Synthese der Evidenz mehr Gewicht haben als Studien mit methodologischen Problemen. Vor allem Studien mit unzureichender Qualität sollen nur in Ausnahmefällen bei der Evidenzsynthese berücksichtigt werden.

Kriterien zur Beurteilung von Meta-Analysen

Statistische Berücksichtigung von:

Publikationsbias

- Wurde Publikationsbias beurteilt?

  Publikationsbias ist eines der wesentlichen Probleme von systematischen Übersichtsarbeiten. Nur circa 45% aller publizierten Abstracts werden später als Artikel veröffentlicht. Publikationsbias soll bei Meta-Analysen statistisch beurteilt (z.B. durch funnel plots) und bewertet werden.

Heterogenität beurteilt & analysiert

- Wurde Heterogenität beurteilt?

  Heterogenität muss bei Meta-Analysen statistisch beurteilt werden (z.B. I^2 -Statistik). Das Vorhandensein von Heterogenität kann mitunter bedeuten, dass Studien *zu verschieden* sind, um meta-analytisch verwendet werden zu können.

- Wurde Heterogenität adäquat analysiert?

  Wenn Heterogenität festgestellt wurde, sollte diese untersucht und erklärt werden. Bei hoher Heterogenität, die nicht erklärt werden kann, soll von einer Meta-Analyse Abstand genommen werden.

Gewichtung & adäquate Zusammenlegung

- Waren Studien die Einheit der statistischen Analyse?

  In Meta-Analysen müssen die einzelnen Komponentenstudien als solche erhalten werden, um die Vorteile der Randomisierung, zumindest teilweise zu erhalten. Resultate der individuellen Studien werden gewichtet und anschließend statistisch zusammengeführt. Daten einzelner Studien dürfen nicht ohne Gewichtung einfach zusammengelegt und gemittelt werden. Alle gängigen statistischen Modelle erhalten die einzelnen Studien als die Einheit der Analyse.

Kriterien zur Beurteilung von diagnostischen Studien

wesentliche Unterschiede zu Interventionsstudien

Diagnostische Studien und Screening-Studien unterscheiden sich wesentlich von Interventionsstudien.[76] [77-79] Folgende Fragen sollten bei der kritischen Beurteilung von diagnostischen Studien beantwortet werden:

- Repräsentiert die Studie jene PatientInnen, die den Test in der Praxis erhalten werden?

  Sensitivität & Spezifität eines Tests sind von Population abhängig

  Diagnostische Studien können in Populationen mit unterschiedlichem Risiko zu unterschiedlichen Ergebnissen führen. Resultate von diagnostischen Studien, die in PatientInnen mit hohem Risiko durchgeführt wurden, sind mitunter nicht auf PatientInnen mit niedrigem Risiko übertragbar. Dies bedeutet, dass Sensitivität und Spezifität eines diagnostischen Tests von der Population abhängig sind und daher variieren. Diagnostische Tests sollen überdies in konsekutiven PatientInnen durchgeführt werden.

- Wurden Auswahlkriterien exakt definiert?

  Aufgrund der variierenden diagnostischen Eigenschaften eines Tests müssen die Auswahlkriterien der Studienpopulation genau definiert und beschrieben sein.

- Wurde ein Referenztest verwendet, der als „Goldstandard" angesehen werden kann?

  Referenztest

  Diagnostische Studien sind nur dann anwendbar und interpretierbar, wenn das Ergebnis mit einem validierten Referenztest verglichen wird. Dies kann der beste erhältliche Test sein, ein neuer Test kann auch mit dem gängigen Standardtest verglichen werden.

- Ist die Zeitperiode zwischen Durchführung des Indextests und des Referenztests entsprechend kurz?

  Vergleich von Index- und Referenztest in Zeitnähe

  Um Unterschiede in Risiko und Prognose zwischen den Studiengruppen zu vermeiden, sollen Index- und Referenztests innerhalb eines kurzen Zeitraumes durchgeführt werden. Die Auswirkungen, die unterschiedliche Zeitpunkte auf das Ergebnis haben können, sind jedoch abhängig vom klinischen Verlauf einer Erkrankung.

- Wurde der Referenztest an der gesamten Studienpopulation ausgeführt?

  Idealerweise sollte der Referenztest an der gesamten Studienpopulation durchgeführt werden. Wenn dies nicht möglich ist, soll eine Teilpopulation, an der der Referenztest durchgeführt wird, per Randomisierung ermittelt werden. Unterschiede bei Risikofaktoren und prognostischen Faktoren zwischen Gesamt- und Teilpopulation müssen dann genau untersucht werden.

- Wurde der Referenztest unabhängig vom Indextestresultat durchgeführt?

  Referenztest wird unabhängig vom Ergebnis des Indextests..

  Die Durchführung des Referenztests darf nicht vom Ergebnis des Indextests abhängen. Bei der Durchführung des Referenztests müssen Ergebnisse des Indextests unbekannt sein und umgekehrt. Kenntnis des Ergebnisses kann zu Detection und Measurement Bias führen.

- Wurden beide Testresultate unabhängig voneinander ausgewertet?

  ...durchgeführt und..

  ..ausgewertet

  Die Auswertung des Referenztests darf nicht vom Ergebnis des Indextests abhängen. Bei der Auswertung des Referenztests müssen Ergebnisse des Indextests unbekannt sein und umgekehrt. Kenntnis des Ergebnisses kann zu Detection und Measurement bias führen.

Berichterstattung nicht-interpretierbarer Resultate ..

- Wurde der Anteil an nicht-interpretierbaren Resultaten berichtet?

  Für die Anwendbarkeit des Tests und für die Glaubhaftigkeit des Resultates ist es nötig, die Rate an nicht-interpretierbaren Testresultaten zu kennen. Es besteht zurzeit kein Konsens, ab welchem Prozentsatz die Rate an nicht-interpretierbaren Resultaten die Validität der Ergebnisse negativ beeinflusst.

..und Dropouts

- Wurde die Höhe der Dropout-Rate während der Tests genannt?

  Ähnlich wie bei der Dropout-Rate in RCTs kann eine zu hohe Dropout-Rate zu Selektionsbias führen. Es gibt keine einheitlichen Empfehlungen, wie hoch die Dropout-Rate sein darf.

Kriterien zur Beurteilung von ökonomischen Studien und entscheidungsanalytischen Studien

Kriterien zur Beurteilung von ökonomischen[68] und entscheidungsanalytischen[69] Studien werden hier nicht im Detail beschrieben. Appendix H und I bietet Formulare zur kritischen Beurteilung dieser Studientypen.

2.7.3 Einstufung der internen Validität

Die interne Validität einzelner Studien wird in Anlehnung an ein System der U.S. Preventive Services Task Force[51] in drei Stufen dargestellt.

Einstufung auf Wahrscheinlichkeit für Bias & systematische Fehler

1. **Gut:** Die interne Validität ist hoch. Alle Komponenten des Qualitätsinstrumentes konnten zufrieden stellend beantwortet werden. Die Wahrscheinlichkeit, dass systematische Fehler die Resultate wesentlich verzerren, ist gering.
2. **Ausreichend:** Die interne Validität ist moderat. Nicht alle Komponenten des Qualitätsinstrumentes konnten zufrieden stellend beantwortet werden. Es besteht jedoch kein kritisches methodologisches Problem. Die Wahrscheinlichkeit, dass systematische Fehler die Resultate wesentlich verzerren, ist gegeben.
3. **Unzureichend:** Die interne Validität ist schlecht. Es besteht ein offensichtliches kritisches methodologisches Problem, oder ein kritisches methodologisches Problem kann nicht mit Sicherheit ausgeschlossen werden. Die Wahrscheinlichkeit, dass systematische Fehler die Resultate wesentlich verzerren, ist hoch.

situationsbedingte Entscheidungen

Wie mit *unzureichenden* Studien verfahren wird, muss situationsbedingt entschieden werden. Methodikstudien sind widersprüchlich, inwieweit unzureichende methodologische Qualität die Resultate verzerren kann.[28,63,80-84] Generell ist es wahrscheinlich nicht zielführend, *unzureichende* Studien in die Synthese der Evidenz einzubeziehen, wenn *gute* oder *ausreichende* Studien vorhanden sind. Meistens genügt es, *unzureichende* Studien in einer Tabelle zu präsentieren und den Grund für diese Beurteilung zu präsentieren. In Ausnahmefällen kann es notwendig sein, im Sinne einer „Strategie der besten Evidenz" (Best Evidence Approach), Studien mit unzureichender Qualität in die Analyse mit einzubeziehen. Ein Beispiel wäre eine Situation, in der nur eine einzige Studie einen direkten Vergleich zweier Interventionen bietet. Auch wenn die interne Validität einer solchen Studie unzureichend ist, kön-

nen Resultate mitunter mehr Aussagekraft haben als der indirekte Vergleich dieser Interventionen.

Da die Beurteilung der internen Validität mit subjektiven Entscheidungen verbunden ist, sollten zumindest zwei Personen diese unabhängig voneinander durchführen.

2.7.4 Beurteilung der externen Validität (Generalisierbarkeit)

Pragmatische Studien oder „Effectiveness"-Studien

Die externe Validität hängt in erster Linie von der Population und vom Gesundheitssystem ab und ist daher eine subjektive Beurteilung (ist die Studie relevant für „meine" Population, für „mein" Gesundheitssystem?)[85 86] Bei Interventionsstudien gibt es allerdings wesentliche Aspekte des Studiendesigns, die die externe Validität mitbestimmen können.[87-89] Studien mit guter externer Validität werden in der Literatur häufig als „pragmatische Studien" (pragmatic studies oder effectiveness studies im Gegensatz zu explanatory oder efficacy studies) bezeichnet.[88,90-93] Folgende Kriterien zur Beurteilung der externen Validität von Interventionsstudien wurden vom RTI-UNC (Research Triangle Institute - University of North Carolina) Evidence-based Practice Center definiert[89]:

- Handelt es sich bei der Studienpopulation um eine Population in der Primärversorgung?

  Studienpopulation wie Primärversorgung

  Die meisten Studien und klinischen Prüfungen werden an Universitätskliniken durchgeführt. PatientInnen und behandelndes Personal unterscheiden sich jedoch wesentlich von jenen in der Primärversorgung. Zusätzlich sind Universitätskliniken meist technologisch besser ausgerüstet und haben hoch spezialisiertes Personal. Die Generalisierbarkeit von Resultaten, die in universitärer Umgebung gewonnen wurden, ist daher häufig limitiert. In seltenen Situationen können Universitätskliniken allerdings auch zur Primärversorgung dienen (z.B. Transplantationen).

- Waren die Auswahlkriterien der Studienpopulation wenig restriktiv?

  Population mit Co-Morbiditäten

  Auswahlkriterien von Studien sind häufig sehr restriktiv, um eine hochselektierte Studienpopulation mit wenig Begleiterkrankungen zu erhalten. Die Generalisierbarkeit der Resultate auf durchschnittliche PatientInnen wird dadurch reduziert. Pragmatische Studien verwenden wenig restriktive Auswahlkriterien, um eine repräsentative PatientInnenpopulation zu erhalten.

- Wurden patientenrelevante Outcomes (health outcomes) untersucht?

  Surrogat- oder patientenrelevante Outcomes

  Patientenrelevante Outcomes sind Outcomes, die PatientInnen erfahren oder spüren können. Surrogat-Outcomes wie zum Beispiel Laborwerte sind in pragmatischen Studien nicht ausschlaggebend: Es bestehen allerdings wesentliche Ausnahmen wie z.B. HbA1c bei Diabetes oder andere Outcomes, die einen starken kausalen Zusammenhang mit einem patientenrelevanten Outcome haben.

- Sind Studiendauer und Behandlungsmodalitäten klinisch relevant?

  Widerspiegeln klinischer Wirklichkeit

  Studiendauer und Behandlungsmodalitäten sollen in pragmatischen Studien der klinischen Wirklichkeit entsprechen. Bei Medikamentenstudien soll zum Beispiel die Dosierung flexibel und nicht fixiert sein. Die Studiendauer soll, wenn möglich, der Behandlungsdauer während eines klinischen Verlaufs entsprechen.

- Wurden Nebenwirkungen der Intervention adäquat erhoben?

Nebenwirkungen adäquat erhoben

Die Erhebung von Nebenwirkungen soll mittels eines validen Instrumentes erfolgen (z.B. WHO-Klassifikation). Zu erwartende Nebenwirkungen sollen außerdem a priori definiert und gezielt abgefragt werden.

- Ist die Studiengröße adäquat, um einen minimal-wesentlichen Unterschied aus Patientenperspektive erheben zu können?

Studiengröße: Signifikanz ≠ Relevanz

Statistische Signifikanz ist zum Großteil von der Studiengröße abhängig. Statistische Signifikanz ist jedoch nicht immer mit klinischer Relevanz gleichzusetzen. Statistisch signifikante Verbesserungen, die von PatientInnen nicht als solche wahrgenommen werden, sind in den meisten Fällen nicht klinisch relevant. In pragmatischen Studien sollte ein minimalwesentlicher Unterschied aus Patientenperspektive bei der Berechnung der Studiengröße jedoch immer in Betracht gezogen werden.

- Wurde eine ITT-Analyse durchgeführt?

ITT-Analyse: Non-Compliance

ITT-Analysen reflektieren Probleme wie Nicht-Befolgung der Medikamenteneinnahme, falsche oder unregelmäßige Durchführung von Interventionen usw. und sollten daher bei pragmatischen Studien immer die statistische Analyse der Wahl sein.

Wenn sechs dieser sieben Kriterien erfüllt werden, handelt es sich um eine pragmatische Studie.[89] Appendix H enthält ein Formular, das für die Beurteilung der externen Validität verwendet werden kann.

2.8 Synthese der Literatur

HAUPTPUNKTE:

- Die Zusammenfassung der Evidenz soll dem Leser helfen, diese zu interpretieren.
- „Katalogisieren" soll vermieden werden.
- Die Stärke der Evidenz muss beurteilt werden.
- Quantitative Analysen sollen in Betracht gezogen werden.

Zusammenfassen bedeutet nicht Katalogisieren

Das Ziel eines systematischen Reviews ist es, die Evidenz in Bezug auf die Fragestellungen zusammenzufassen.[94] Dies kann qualitativ oder quantitativ (Meta-Analyse) geschehen. Eine qualitative Synthese sollte eine Antwort auf die HTA-Fragestellung bieten und keine „Katalogisierung" der vorhandenen Studien sein.[94] Bei Interventionen muss das Verhältnis von Nutzen zu Risiko beurteilt werden. Interventionen sind nur dann sinnvoll, wenn mehr Nutzen als Risiko besteht (net-benefit). Die Betonung soll dabei auf Effektmaßen liegen, die leicht interpretierbar sind (z.B. number needed to treat [NNT], number needed to harm [NNH]). Diese Synthese der Evidenz muss in Bezug auf die Stärke der Evidenz erfolgen.[95] Die Stärke der Evidenz reflektiert das Vertrauen, dass die vorhandene Evidenz die richtige Antwort gibt.[96]

klinisch relevante Effektmaße betonen

2.8.1 Stärke der Evidenz

Eine systematische und explizite Methode zur Beurteilung der Stärke der Evidenz verhindert Fehler und erleichtert die Reproduzierbarkeit und kritische Evaluierung der Entscheidungen durch Außenstehende.[49] Das folgende System zur Beurteilung der Stärke der Evidenz wurde von der GRADE Working Group entwickelt. [96]

GRADE beurteilt folgende Aspekte der Evidenz:

GRADE unterstützt Vertrauen,

- Studiendesign
- Qualität der Studien
- Konsistenz
- Direktheit

dass vorliegendes Wissen, auch die *richtigen* Antworten gibt

GRADE benutzt folgende Klassifizierungen und Definitionen, um die Stärke der Evidenz zu beurteilen:

- Hoch: Es ist unwahrscheinlich, dass neue Studien einen wichtigen Einfluss auf die Einschätzung des Effektes haben werden.
- Mittel: Neue Studien werden möglicherweise einen wichtigen Einfluss auf die Einschätzung des Effektes haben.
- Niedrig: Neue Studien werden sehr wahrscheinlich einen wichtigen Einfluss auf die Einschätzung des Effektes haben.
- Sehr niedrig: Jegliche Einschätzung des Effektes ist sehr unsicher.

Wesentlich bei GRADE ist, dass die Stärke der Evidenz für jede Fragestellung und jeden primären Outcome getrennt beurteilt wird. Die genaue Anwendung des GRADE-Systems ist der entsprechenden Publikation zu entnehmen.[96]

2.8.2 Effektmaße

Die Wahl der richtigen Effektmaße ist ausschlaggebend, um Resultate entsprechend zusammenzufassen und zu kommunizieren.[97-100] Meistens bieten Artikel genug Information, um mit Hilfe einer Vierfeldertafel (Abbildung 2.8.2-1) Effektmaße oder diagnostische Parameter (Abbildung 2.8.2-2) zu berechnen. Im Folgenden eine kurze Zusammenfassung der am häufigsten verwendeten Effektmaße.

Effektmaße: Kommunikation der Resultate

	Intervention	*Keine Intervention*	*Gesamt*
Outcome vorhanden	*A*	*b*	*a+b*
Outcome nicht vorhanden	*C*	*d*	*c+d*
Gesamt	*n der Interventionsgruppe*	*n der Kontrollgruppe*	*N der Studie*

Abbildung 2.8.2-1: Vierfeldertafel zur Berechnung von Effektmaßen für kategorielle Zielvariablen

Relatives Risiko (RR)

RR beschreibt Wahrscheinlichkeit, mit oder ohne Intervention den Endpunkt zu erreichen

- Das RR beschreibt, um wie viel wahrscheinlicher es ist, für jemanden mit Exposition (Intervention), einen bestimmten Outcome zu haben, verglichen mit jemandem ohne Exposition.
- RR beschreibt die Stärke des Verhältnisses von Exposition zu einem bestimmten Outcome.
- Wenn das Konfidenzintervall "1" kreuzt, besteht kein statistisch signifikanter Unterschied.
- Formel: [a/(a+c)] / [b/ (b+d)]
  - RR = 1: kein Unterschied
  - RR = 2: Die Wahrscheinlichkeit, mit Exposition einen bestimmten Outcome zu haben, ist doppelt so groß (100%) wie ohne Exposition.
  - RR = 0,5: Die Wahrscheinlichkeit, mit Exposition einen bestimmten Outcome zu haben, ist halb so groß (50%) wie ohne Exposition.

	Erkrankung vorhanden	*Erkrankung NICHT vorhanden*	*Gesamt*
Diagnostisches Testergebnis ***POSITIV***	*a* richtig-positiv	*b* falsch-positiv	*a+b*
Diagnostisches Testergebnis ***NEGATIV***	*c* falsch-negativ	*d* richtig-negativ	*c+d*
Gesamt	*n der Interventior gruppe*	*n der Kontroll- gruppe*	*N der Studie*

Abbildung 2.8.2-2: Vierfeldertafel zur Berechnung von diagnostischen Parametern

Relative Risikoreduktion (RRR, relativer Effekt)

RRR: Prozentsatz beschreibt die relative Differenz des Effekts mit/ohne Intervention

- Die RRR beschreibt, um welchen Prozentsatz eine Intervention das Risiko relativ reduziert.
- Formel: RR – 1
  - RRR = 0,1: Die Intervention führt zu einer Verringerung des Risikos um relativ 10 %.
- Ohne Kenntnis des Basisrisikos können RR und RRR keine Auskunft über das absolute Risiko geben.

Absolute Risikoreduktion (ARR, Risiko-Differenz)

- Die ARR gibt an, um wie viel das Risiko durch eine Intervention reduziert werden kann.
- Formel: $R_1 - R_2$
- ARR ist die Grundlage zur Berechnung der NNT und der NNH.
- ARR ist klinisch aussagekräftiger als relative Effektmaße.
  - Interpretation: ARR = 0,1: Die Intervention führt zu einer Verringerung des Risikos um absolut 10 % (wie bei RRR; ARR kann jedoch in Bezug auf klinische Relevanz beurteilt werden, RRR nicht).

ARR:
Prozentsatz beschreibt die absolute/reale Differenz des Effekts, mit Intervention Endpunkt zu erreichen

Odds ratio (OR)

- OR werden meistens wie RR verwendet, um zu beschreiben, wie stark das Verhältnis einer Intervention (Exposition) zu einem bestimmten Outcome ist.
- OR und RR unterscheiden sich wesentlich, wenn die Ereignisrate hoch ist.
- Formel: (a/c) / (b/ d)
- OR beschreibt die Stärke des Verhältnisses von Exposition zu einem bestimmten Outcome.
- Wenn das Konfidenzintervall "1" kreuzt, besteht kein statistisch signifikanter Unterschied.

Wie RR,
beschreibt Stärke des Zusammenhangs von Intervention & Outcome

Number needed to treat (NNT) / Number needed to harm (NNH)

- RR und OR sind relative Effektmaße, die nur beschränkt Auskunft über die klinische Relevanz geben können.
- NNT und NNH geben an, wie viele Personen behandelt werden müssen, um bei einer Person einen bestimmten Outcome (Endpunkt) zu erzielen bzw. zu vermeiden.
- Formel: 1/ARR
- Interpretation: z.B. NNT = 10; 10 Personen müssen die Intervention erhalten, damit eine Person den Outcome (Endpunkt) erreicht (z.B. von der Intervention profitiert).

NNT & NNH machen Aussagen über klinische Relevanz

2.9 Interne/Externe Begutachtung

HAUPTPUNKT:

- Interne und externe Begutachtungen werden vor jeder Publikation durchgeführt.

zur Qualitätsbeurteilung: interne & externe Begutachtung

Eine kritische interne und externe Begutachtung ist ein wesentlicher Teil eines systematischen Reviews.[2,5,6] Zur Gewährleistung und Absicherung der Qualität werden alle Berichte des LBI für HTA einem internen Peer Review unterzogen. GutachterInnen (ReviewerInnen) sind WissenschaftlerInnen des Institutes, die nicht unmittelbar am Projekt gearbeitet haben, methodologische ExpertenInnen und jene fachlichen ExpertInnen, die in beratender Funktion am Projekt beteiligt waren.

Stakeholder-Involvement fallweise

Zusätzlich wird jeder Bericht von mindestens einem/r externen Experten/in begutachtet. Der/Die Projektleiter/in entscheidet, ob ein oder mehrere externe ReviewerInnen involviert werden. In Einzelfällen kann es auch ratsam sein, den Entwurf eines Berichtes öffentlich zugänglich zu machen, um Kommentare von Interessensgruppen, PatientInnen und deren Angehörigen oder der Industrie zu erhalten.

entsprechend einem peer-review:

Die externen Gutachten dienen zu Qualitätssicherung unserer wissenschaftlichen Arbeit. Die Evaluationen/Assessments haben das definierte Ziel,

- methodisch transparent und
- fachlich korrekt

Beurteilung von Methode, Originalität, fachliche Korrektheit, Relevanz. etc.

eine unabhängige Entscheidungsgrundlage für politische Entscheidungsträger zu liefern, die wissenschaftlichen Qualitätskriterien entspricht. Wir verstehen demnach die externe Begutachtung durch wissenschaftliche FachexpertInnen aus unterschiedlichen Disziplinen – in Anlehnung an einen „peer-review" Prozess in wissenschaftlichen Fachzeitschriften. Qualitätskriterien für wissenschaftlichen Arbeitens sind:

- „fachliche Korrektheit" (stimmen die Informationen),
- „Originalität" (ist die Fragestellung neu),
- „Adäquatheit und Transparenz der Methode" (wird die richtige Methode eingesetzt),
- „logischer Aufbau der Arbeit und Konsistenz in der Struktur" (ist das Ergebnis nachvollziehbar),
- „Relevanz für die nationale und internationale Fachöffentlichkeit" (haben die Ergebnisse eine Relevanz für Anwender),
- „Berücksichtigung des aktuellen Stands der Forschung" (am letzten Stand der Forschung).

Auswahl der GutachterInnen:

Die Auswahl der GutachterInnen erfolgt nach deren „Profilen" als ein/e MethodikerIn und ein/e FachexpertIn mit praktischen Erfahrungen. Die Gutachten selbst sind verfügbar und können von den AutorInnen des Berichts abgerufen werden.

MethodikerIn & praktische-klinische Erfahrung

GutachterInnen sind gebeten den Bericht anhand eines standardisierten Fragebogens (vgl. Appendix I) sowie zusätzlich detaillierten im Freitext zu beurteilen. Der Fragebogen erleichtert GutachterInnen, die wissenschaftliche Validität des Reportes, die Verständlichkeit des Textes sowie die Validität der Interpretation der Evidenz mittels einer Checkliste zu beurteilen. Der Fragebogen erhebt auch mögliche Interessenskonflikte der GutachterInnen.

offener Peer-Review

Um Transparenz zu gewährleisten, führt das LBI für HTA ein offenes Peer-Review-System durch. GutachterInnen werden dabei namentlich im Bericht genannt.

3 Rapid Assessments

HAUPTPUNKTE:

- Rapid Assessments werden eingesetzt, um innerhalb eines kurzen Zeitrahmens eine Zusammenfassung der Evidenz zu erstellen.
- Die Methodik von Rapid Assessments ist systematisch.
- Rapid Assessments haben eine geringere Detailtiefe als systematische Reviews.

3.1 Einsatzbereiche

keine Original-Kausalitätsbeurteilungen

Ein Rapid Assessment wird eingesetzt, um innerhalb eines kurzen Zeitrahmens eine Zusammenfassung der Evidenz zu erstellen. Die Durchführung erfolgt systematisch, Resultate haben jedoch nicht die Detailtiefe von systematischen Reviews. Aufgrund der vereinfachten Methodik besteht ein erhöhtes Risiko für Retrievalbias, Selektionsbias und Reviewerbias. Rapid Assessments sollen daher nicht für die Beurteilung von Kausalitäten herangezogen werden. Rapid Assessments können jedoch kostengünstig eine Übersicht über den Status der Evidenz bieten.

aber Berichterstattung zum Evidenz-Status

Zurzeit besteht kein internationaler Konsens über eine valide Methodik für Rapid Assessments. Die folgende Methodik für Rapid Assessments basiert auf Resultaten einer Studie zur Beurteilung der Validität von klinischen Leitlinien, die von der U.S. Preventive Services Task Force entwickelt und validiert wurde.[101,102] Abbildung 3.1-1 stellt das Prinzip dieser Methode dar

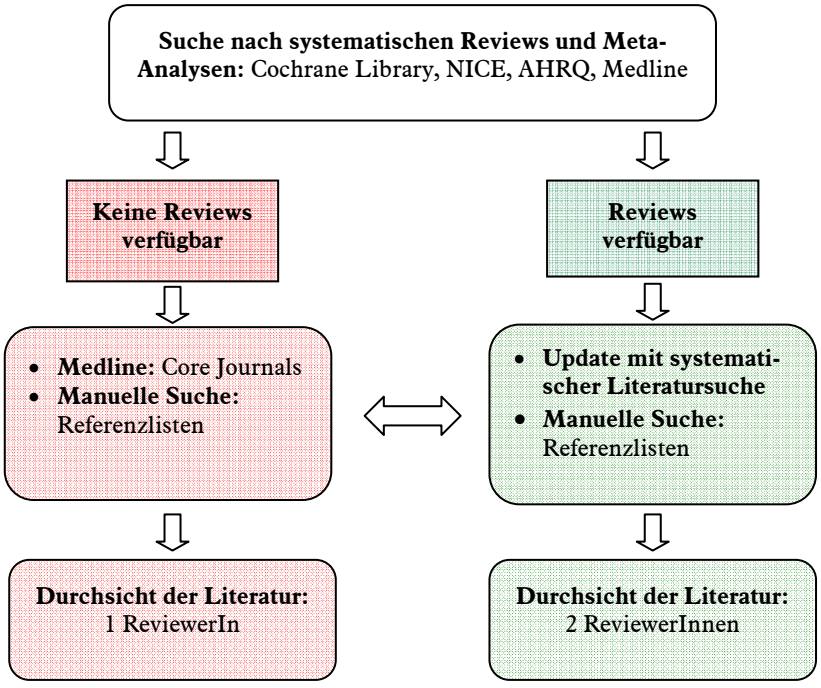

Abbildung 3.1-1: Schema eines Rapid Assessment

3.2 Literatursuche

Rapid Assessments bauen zunächst auf bestehenden Reviews auf

Die folgende Methode bedient sich schon bestehender systematischer Reviews und konzentriert sich auf Literatur von Zeitschriften mit hohem „Impact". Dieses Konzept beruht auf der Annahme, dass wichtige Studien in großen Journals mit hohem „Impact" publiziert werden. Eine manuelle Literatursuche kann zusätzlich kleinere Studien identifizieren. Jene Studien, die bei dieser Methode nicht gefunden werden, sind Arbeiten, die kaum in der Literatur zitiert werden. Es ist daher unwahrscheinlich, dass diese Studien Resultate und Schlussfolgerungen wesentlich verändern würden.

1. Suche nach systematischen Reviews und Meta-Analysen:

Suche in HTA-Datenbank etc.

In der ersten Stufe werden relevante Datenbanken nach systematischen Reviews durchsucht, welche die Fragestellung des Rapid Assessments (oder Teile davon) beantworten.[103] Mögliche Datenbanken wären Cochrane Library, NICE, AHRQ, Medline und andere themenspezifische Literatur-Datenbanken.

2. Bestehende systematische Reviews können verwendet werden:

Aktualisieren der Literatur

Wenn systematische Reviews vorhanden sind, die Teile oder die gesamte Fragestellung abdecken, sollte zuerst die interne Validität überprüft werden. Nur Reviews mit ausreichend guter methodologischer Qualität sollten als Basis für das Rapid Assessment benutzt werden. In einem weiteren Schritt werden diese Reviews durch eine gezielte Literatursuche aktualisiert. Die Literatursuche kann den Suchzeitraum mit Bezug auf den Review limitieren. Die Suchzeiträume sollten sich jedoch zumindest überschneiden. Wenn zum Beispiel der Suchzeitraum des systematischen Reviews von 1990 bis 2005 geht, sollte die aktualisierende Literatursuche 2004 beginnen.

Wenn eine systematische Suche über den fehlenden Zeitraum durchgeführt wird, kann diese Methode als „systematisch" betrachtet werden und der weitere Ablauf kann wie bei einem systematischen Review erfolgen. Als Alternative kann die Suche aber auch auf die wesentlichsten Journals limitiert werden. Die Auswahl dieser Journals wird vom Thema abhängen, als Ausgangsbasis kann in Medline die Funktion „limit to core journals" verwendet werden. Zusätzlich sollte eine manuelle Suche von Referenzlisten von relevanten Artikeln (ist so nicht verständlich, Vorschlag BG: Durchsicht von Referenzlisten relevanter Artikel) durchgeführt werden.

3. Keine systematischen Reviews vorhanden

Prozess wie bei Übersichtsarbeiten,

aber engere Fragestellung & ohne duale Literatursichtung

Wenn keine systematischen Reviews mit ausreichender Validität vorhanden sind, wird eine systematische Literatursuche in Medline durchgeführt. Resultate werden mittels der Funktion „limit to core journals" auf Zeitschriften mit hohem „Impact" limitiert. Zusätzlich sollten Journals einbezogen werden, die im jeweiligen Fachbereich als wesentlich gelten, aber nicht in den „Medline core journals" gelistet sind. Expertenbefragung kann helfen, solche Zeitschriften zu identifizieren. Die weiteren Schritte entsprechen denen eines systematischen Reviews. Die Literaturdurchsicht und die manuelle Suche werden jedoch nur von einer Person durchgeführt.

3.3 Durchsicht der Literatur

Die Durchsicht der Literatur erfolgt nach a priori definierten Kriterien wie unter 3.5 „Durchsicht der Literatur" beschrieben. Die Durchsicht der Literatur wird bei Rapid Assessments jedoch nur von einer Person durchgeführt.

Unterschiede in einigen,

3.4 Klassifizierung der Studien

Die Klassifizierung der Studien erfolgt nach dem in 3.5.2 „Klassifizierung der Studien" beschriebenen Schema.

aber nicht allen Arbeitsschritten

3.5 Beurteilung der internen/externen Validität

Die Beurteilung der internen und externen Validität erfolgt nach den in 3.7 „Beurteilung der internen und externen Validität von Studien" dargelegten Algorithmen.

3.6 Datenextraktion

Bei Rapid Assessments werden Daten nicht in Evidenztabellen präsentiert.

3.7 Synthese der Evidenz

Die Synthese der Evidenz in Rapid Assessments wird eine kurze, prägnante, qualitative Zusammenfassung der vorhandenen Studien sein. Es werden keine quantitativen und ökonomischen Analysen durchgeführt.

kurze, prägnante Zusammenfassung

3.8 Interne und externe Begutachtung

Die interne und externe Begutachtung von Rapid Assessments erfolgt nach dem gleichen Schema wie unter 2.9 „Interne und externe Begutachtung" beschrieben.

Arbeitsschritte	Systematische Reviews	Rapid Assessments
Systematisches Vorgehen	Ja	Ja
Literatursuche in elektronischen Datenbanken	Ja	Ja
Manuelle Literatursuche	Ja	Ja
Suche nach nicht-publizierten Studien	Ja	Nein
Identifizierung von Literatur durch externe GutachterInnen	Ja	Nein
Duale Durchsicht der Abstracts	Ja	Nein
Duale Durchsicht der Volltext-Artikel	Ja	Nein
Beurteilung der internen Validität	Ja	Ja
Beurteilung der externen Validität	Ja	Ja
Daten Extraktion in Evidenztabellen	Ja	Nein
Qualitative Synthese der Evidenz	Ja	Ja
Umfassende Beurteilung relevanter Outcomes	Ja	Ja
Quantitative Analyse	Bei Bedarf	Nein
Ökonomische Analyse	Bei Bedarf	Bei Bedarf
Beurteilung der Stärke der Evidenz	Ja	Ja
Interner Peer Review	Ja	Ja
Externer Peer Review	Ja	Ja

Abbildung 3.8-1: Gegenüberstellung der Arbeitsschritte systematischer Übersichtsarbeiten und Rapid Assessments

4 Appendizes

4.1 Appendix A: Grundriss eines Protokolls

Thema des Reviews:

ProjektleiterIn:

Zeitperiode:

Methode:

HTA-Fragestellungen:

- x
- x
- x

Breite der Perspektiven:

Interventionen und Zielvariablen:

Auswahlkriterien:

Geplante quantitative Analysen:

Zeitplan:

4.2 Appendix B: Grundriss eines Formulars zur Durchsicht von Volltext-Artikeln

StudienautorIn und Jahr:

Referenznummer:

Name des/der Reviewers/in:

Handelt es sich bei dem Artikel um primäre wissenschaftliche Literatur (keine Editorials, Zusammenfassungen, Letters to the editor etc.)?

Beantwortet die Studie eine HTA-Fragestellung?

Ist das Gesundheitssystem, in dem die Studie durchgeführt wurde, relevant?

Entspricht das Studiendesign den Auswahlkriterien?

Entsprechen die Interventionen/Expositionen den Auswahlkriterien?

Entsprechen die Zielparameter den Auswahlkriterien?

Appendix C: Beispiel für die Kodierung von Artikeln in einem Literaturverwaltungsprogramm

Literaturverwaltungsprogramm Custom Field	Code*	Bedeutung des Codes	Erklärung & Kommentare
	H	Hintergrund	Artikel, die nicht den Auswahlkriterien entsprechen, die aber wertvolle Hintergrundinformation bieten können.
	N	Nicht verfügbar	Der Artikel konnte trotz mehrerer Versuche nicht erworben werden.
	A	Abstract only	Artikel erfüllt Auswahlkriterien, ist jedoch nur als Abstract publiziert.
	E1	Volltext ausgeschlossen: Keine primäre Studie	z. B. Editorials, Zusammenfassungen, Letters to the editor usw.
	E2	Volltext ausgeschlossen: Falsche Population	Population entspricht nicht den Auswahlkriterien.
	E3	Volltext ausgeschlossen: Falsches Design	Studiendesign entspricht nicht den Auswahlkriterien, z.B. zu kurze Studiendauer, falscher Studientyp usw.
	E4	Volltext ausgeschlossen: Falscher Outcome	Zielvariablen der Studie entsprechen nicht den Auswahlkriterien.
	E5	Volltext ausgeschlossen: studienspezifisch	
	E6	Volltext ausgeschlossen: studienspezifisch	

4.3 Appendix D: Grundriss eine Formulars zur Datenextraktion für therapeutische Studien

Autor, Jahr, Referenznummer	
Fragestellung/Zielsetzung	
Finanzierung der Studie	
Geografisches Setting der Studie	
Studiendesign	
Einschluss-/Ausschlusskriterien	
Intervention: Verum(n)/Kontrolle(n)	
Beobachtungsdauer	
Vergleichbarkeit der Charakteristika der Gruppen zu Beginn der Studie	
Primäre Zielvariablen (Outcomes)	
Sekundäre Zielvariablen (Outcomes)	
Resultate	
Dropout-Rate	
Anmerkungen	

4.4 Appendix E: Grundriss eines Formulars zur Datenextraktion für diagnostische Studien

Autor, Jahr, Referenznummer	
Fragestellung/Zielsetzung	
Finanzierung der Studie	
Geografisches Setting der Studie	
Studiendesign	
Einschluss-/Ausschlusskriterien	
Anwendungssituation	
Zu prüfendes Verfahren	
Referenzverfahren (Goldstandard)	
Vergleichbarkeit der Charakteristika der Gruppen zu Beginn der Studie	
Primäre Zielvariablen (Outcomes)	
Sekundäre Zielvariablen (Outcomes)	
Resultate	
Dropout-Rate	
Anmerkungen	

4.5 Appendix F: Grundriss eines Formulars zur Datenextraktion für ökonomische Studien

Basis: NHS Centre for Reviews and Dissemination[59]; eigene Modifikation

Allgemeine Informationen	
Datum der Datenextraktion	
Verfasser/Publikationsjahr der Studie	
Quelle der Finanzierung	
Methodologische Studiencharakteristika	
Allgemein	
Forschungsfrage	
Charakteristika der Studienpopulation und Setting der Intervention	
Studientyp (CMA, CEA, CUA, CBA)[1]	
Perspektive (Gesundheitssystem, Gesellschaft, Kostenträger, PatientIn)	
Land	
Untersuchte Intervention	
Komparator(en)	
Diskontierungsrate	
Zeithorizont	
Variablen in Sensitivitätsanalyse	
Outcome-Parameter (Kosten, Gesundheitseffekte)	
Datenquellen	
Quelle für klinische Daten (Meta-Analyse, Einzelstudie, Expertenmeinung etc)	
Quelle für Kostendaten (Primär/Sekundär)	
Währung und Jahr für Kostendaten	
Berechnung von Kosten und Gesundheitseffekten	
Verwendete Parameter für Gesundheitseffekte (z.B. LYG, QALY)[2]	
Bewertungsmethoden für klinische Outcomes (z.B. monetäre Bewertung, HrQoL-Bewertung)	
Eingeschlossene Kosten	
(Modelltyp)	
Beurteilung der methodologischen Qualität	
Ergebnisse	
Ergebnis Gesundheitseffekte	
Ergebnis Kosten	
Synthese von Kosten und Gesundheitseffekten (IKEV bei CEA, IKNV bei CUA, Nettonutzen bei CBA)[3]	
Ergebnis Sensitivitätsanalyse	
Synthese der Ergebnisse	
Esümee	
Sonstige Anmerkungen	

[1] CMA: Kosten-Minimierungsanalyse; CEA: Kosten-Effektivitätsanalyse; CUA: Kosten-Nutzwertanalyse; CBA: Kosten-Nutzenanalyse

[2] LYG: gewonnene Lebensjahre; QALY: qualitätsadjustierte Lebensjahre

HrQoL: Health related Quality of Life

[3] IKEV: Inkrementelles Kosten-Effektivitätsverhältnis; IKNV: Inkrementelles Kosten-Nutzwertverhältnis

4.6 Appendix G: Formulare zur Beurteilung der internen Validität

Kriterien zur Beurteilung von RCTs	Ja	Nein	Nicht enthalten	Nicht anwendbar
War die Randomisierung adäquat?				
War die Unvorhersehbarkeit der Gruppenzuordnung adäquat (allocation concealment)?				
Waren wesentliche Charakteristika der Studiengruppen 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	Ausreichend	Unzureichend

Kommentare	

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 Auswertung berücksichtigt?				

Beurteilung der internen Validität	Gut	Ausreichend	Unzureichend

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 Studien 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 statistisch beurteilt?				
Wurde Heterogenität adäquat analysiert?				
Waren Studien die Einheit der statistischen Analyse?				

Beurteilung der internen Validität	Gut	Ausreichend	Unzureichend

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 definiert?				
Wurde ein Referenztest verwendet, der als „Goldstandard" angesehen werden kann?				
Ist die Zeitperiode zwischen Durchführung des (Index-) Tests und des Referenztests entsprechend kurz?				
Wurde der Referenztest an der gesamten Studienpopulation ausgeführt?				
Wurde der Referenztest unabhängig vom (Index-) Testresultat durchgeführt?				
Wurden beide Testresultate unabhängig voneinander ausgewertet?				
Wurde die Höhe der Drop-out-Rate während der Tests genannt?				
Wurde der Anteil an nicht-interpretierbaren Resultaten berichtet?				

Beurteilung der internen Validität	Gut	Ausreichend	Unzureichend

4.7 Appendix H: Formular zur Beurteilung der Qualität ökonomischer Studien

Basis: Siebert et al.[68]; leicht modifiziert

	Formular zur Beurteilung der Qualität ökonomischer Studien	Ja	Nein	Nicht Enthal-ten
	Fragestellung			
1.	Wurde die Fragestellung präzise formuliert?			
2.	Wurde der medizinische und ökonomische Problemkontext ausreichend dargestellt?			
	Evaluationsrahmen			
3.	Wurden alle in die Studie einbezogenen Technologien hinreichend detailliert beschrieben?			
4.	Wurden alle im Rahmen der Fragestellung relevanten Technologien verglichen?			
5.	Wurde die Auswahl der Vergleichstechnologien schlüssig begründet?			
6.	Wurde die Zielpopulation klar beschrieben?			
7.	Wurde ein für die Fragestellung angemessener Zeithorizont für Kosten und Gesundheitseffekte gewählt und angegeben?			
8.	Wurde der Typ der gesundheitsökonomischen Evaluation explizit genannt?			
9.	Wurden sowohl Kosten als auch Gesundheitseffekte untersucht?			
10.	Wurde die Perspektive der Untersuchung eindeutig gewählt und explizit genannt?			
	Analysemethoden und Modellierung			
11.	Wurden adäquate statistische Tests/Modelle zur Analyse der Daten gewählt und hinreichend gründlich beschrieben?			
12.	Wurden in entscheidungsanalytischen Modellen die Modellstruktur und alle Parameter vollständig und nachvollziehbar dokumentiert (in der Publikation bzw. einem technischen Report)?			
13.	Wurden die relevanten Annahmen explizit formuliert?			
14.	Wurden in entscheidungsanalytischen Modellen adäquate Datenquellen für die Pfadwahrscheinlichkeiten gewählt und eindeutig genannt?			
	Gesundheitseffekte			
15.	Wurden alle für die gewählte Perspektive und den gewählten Zeithorizont relevanten Gesundheitszustände berücksichtigt und explizit aufgeführt?			
16.	Wurden adäquate Quellen für die Gesundheitseffektdaten gewählt und eindeutig genannt?			

17.	Wurden das epidemiologische Studiendesign und die Auswertungsmethoden adäquat gewählt und beschrieben und wurden die Ergebnisse detailliert dargestellt? (falls auf einer einzelnen Studie basierend)			
18.	Wurden angemessene Methoden zur Identifikation, Extraktion und Synthese der Effektparameter verwendet und wurden sie detailliert beschrieben? (falls auf einer Informationssynthese basierend)			
19.	Wurden die verschiedenen Gesundheitszustände mit Präferenzen bewertet und dafür geeignete Methoden und Messinstrumente gewählt und angegeben?			
20.	Wurden adäquate Quellen der Bewertungsdaten für die Gesundheitszustände gewählt und eindeutig genannt?			
21.	Wurde die Evidenz der Gesundheitseffekte ausreichend belegt? (s. ggf. entsprechende Kontextdokumente)			
	Kosten			
22.	Wurden die den Kosten zugrunde liegenden Mengengerüste hinreichend gründlich dargestellt?			
23.	Wurden adäquate Quellen und Methoden zur Ermittlung der Mengengerüste gewählt und eindeutig genannt?			
24.	Wurden die den Kosten zugrunde liegenden Preisgerüste hinreichend gründlich beschrieben?			
25.	Wurden adäquate Quellen und Methoden zur Ermittlung der Preise gewählt und eindeutig genannt?			
26.	Wurden die einbezogenen Kosten anhand der gewählten Perspektive und des gewählten Zeithorizontes schlüssig begründet und wurden alle relevanten Kosten berücksichtigt?			
27.	Wurden Daten zu Produktivitätsausfallskosten (falls berücksichtigt) getrennt aufgeführt und methodisch korrekt in die Analyse einbezogen?			
28.	Wurde die Währung genannt?			
29.	Wurden Währungskonversionen adäquat durchgeführt?			
30.	Wurden Preisanpassungen bei Inflation oder Deflation adäquat durchgeführt?			
	Diskontierung			
31.	Wurden zukünftige Gesundheitseffekte und Kosten adäquat diskontiert?			
32.	Wurde das Referenzjahr für die Diskontierung angegeben bzw. bei fehlender Diskontierung das Referenzjahr für die Kosten?			
33.	Wurden die Diskontraten angegeben?			
34.	Wurde die Wahl der Diskontraten bzw. der Verzicht auf eine Diskontierung plausibel begründet?			
	Ergebnispräsentation			
35.	Wurden Maßnahmen zur Modellvalidierung ergriffen und beschrieben?			
36.	Wurden absolute Gesundheitseffekte und absolute Kosten jeweils pro Kopf bestimmt und dargestellt?			
37.	Wurden inkrementelle Gesundheitseffekte und inkrementelle Kosten jeweils pro Kopf bestimmt und dargestellt?			

38.	Wurde eine für den Typ der gesundheitsökonomischen Evaluation sinnvolle Maßzahl für die Relation zwischen Kosten und Gesundheitseffekt angegeben?			
39.	Wurden reine (nicht lebensqualitätsadjustierte) klinische Effekte berichtet?			
40.	Wurden die relevanten Ergebnisse in disaggregierter Form dargestellt?			
41.	Wurden populationsaggregierte Kosten und Gesundheitseffekte dargestellt?			
	Behandlung von Unsicherheiten			
42.	Wurden Sensitivitätsanalysen für die relevanten Parameter durchgeführt?			
43.	Wurde Sensitivitätsanalysen für die relevanten strukturellen Elemente durchgeführt?			
44.	Wurden geeignete Methoden für Sensitivitätsanalysen angewandt? (univariat, multivariat, probabilistisch)			
45.	Wurden in den Sensitivitätsanalysen realistische Werte oder Wertebereiche bzw. Strukturvarianten berücksichtigt und angegeben?			
46.	Wurden die Ergebnisse der Sensitivitätsanalysen hinreichend dokumentiert?			
47.	Wurden adäquate statistische Inferenzmethoden (statistische Tests, Konfidenzintervalle) für stochastische Daten eingesetzt und die Ergebnisse berichtet?			
	Gerechtigkeit			
48.	Wurden die Gerechtigkeitsannahmen explizit formuliert? (z.B. gleicher Wert eines QALY für alle)			
49.	Wurden für Untergruppen relevante Gerechtigkeitscharakteristika identifiziert und beschrieben?			
	Diskussion			
50.	Wurde die Datenqualität kritisch beurteilt?			
51.	Wurden Richtung und Größe des Einflusses unsicherer oder verzerrter Parameterschätzung auf das Ergebnis konsistent diskutiert?			
52.	Wurde Richtung und Größe des Einflusses struktureller Modellannahmen auf das Ergebnis konsistent diskutiert?			
53.	Wurden die wesentlichen Einschränkungen und Schwächen der Studie diskutiert?			
54.	Wurden plausible Angaben zur Generalisierbarkeit der Ergebnisse gemacht?			
55.	Wurden wichtige ethische und Verteilungsfragen diskutiert?			
56.	Wurde das Ergebnis sinnvoll im Kontext mit unabhängigen Gesundheitsprogrammen diskutiert?			
	Schlussfolgerungen			
57.	Wurden in konsistenter Weise Schlussfolgerungen aus den berichteten Daten/Ergebnissen abgeleitet?			
58.	Wurde eine auf Wissensstand und Studienergebnissen basierende Antwort auf die Fragestellung gegeben?			

4.8 Appendix I: Formular zur Beurteilung der Qualität von entscheidungsanalytischen gesundheitsökonomischen Modellen

Basis: Philips et al.[69]; eigene Übersetzung und leichte Modifikation

Beurteilung der Qualität von entscheidungsanalytischen gesundheitsökonomischen Modellen	Ja	Nein	Nicht enthalten
Modellstruktur			
1. Beschreibung des Entscheidungsproblems			
1.1. Wird das Entscheidungsproblem klar beschrieben?			
1.2. Sind das Ziel der Evaluation und das beschriebene Modell mit dem Entscheidungsproblem konsistent?			
1.3. Wird der primäre Entscheidungsträger genannt?			
2. Beschreibung der Perspektive			
2.1. Wird die Perspektive der Untersuchung eindeutig gewählt und explizit genannt?			
2.2. Sind die Input-Parameter des Modells konsistent mit der genannten Perspektive?			
2.3. Wird der Gültigkeitsbereich des Modells angegeben und begründet?			
2.4. Sind die Modellergebnisse mit Perspektive, Gültigkeitsbereich und Gesamtziel der Modellierung konsistent?			
3. Logik der Modellstruktur			
3.1. Ist die Struktur des Modells mit einer kohärenten Theorie zum evaluierten Gesundheitszustand konsistent?			
3.2. Werden die Daten(quellen) zur Konstruktion der Modellstruktur beschrieben und schlüssig begründet?			
3.3. Werden die im Modell verwendeten kausalen Zusammenhänge beschrieben und schlüssig begründet?			
4. Strukturelle Annahmen			
4.1. Werden die strukturellen Annahmen transparent beschrieben und begründet?			
4.2. Sind die strukturellen Annahmen vor dem Hintergrund von Evaluierungsziel, Perspektive und Gültigkeit zweckvoll?			
5. Strategien/Vergleichstechnologien			
5.1. Werden alle im Rahmen der Fragestellung relevanten Technologien hinreichend detailliert beschrieben?			
5.2. Werden alle im Rahmen der Fragestellung relevanten Technologien verglichen?			
5.3. Wird der Ausschluss einer plausiblen Alternative ausreichend begründet?			

6 Modelltyp Wurde ein dem Entscheidungsproblem angemessener Modelltyp gewählt?			
7. Zeithorizont 7.1. Ist der Zeithorizont des Modells ausreichend lang, um alle Unterschiede zwischen den gewählten Alternativen aufzuzeigen? 7.2. Werden der gewählte Zeithorizont des Modells sowie die Dauer von Behandlung und Behandlungseffekt beschrieben und schlüssig begründet?			
8. Krankheitszustände/Pfade Spiegeln die Krankheitszustände oder die Pfade (Entscheidungsbaum) den tatsächlichen biologischen Prozess der untersuchten Erkrankung und der Auswirkung der Intervention wider?			
9. Zykluslänge Wird die gewählte Zykluslänge beschrieben und auf Basis des natürlichen Krankheitsverlaufs ausreichend begründet?			
Daten			
1. Datenidentifikation 1.1. Werden die Methoden zur Identifikation der Daten transparent dargestellt und sind sie für das Evaluationsziel geeignet? 1.2. Wird die Entscheidung zur Datenauswahl ausreichend begründet? 1.3. Wird der Identifikation von Daten für zentrale Modellparameter entsprechende Aufmerksamkeit gewidmet? 1.4. Wird die Datenqualität angemessen evaluiert? 1.5. Wird bei der Nutzung von ExpertInnenmeinungen die Methode der Datengenerierung beschrieben und begründet?			
2. Modellierung 2.1. Basiert die Modellierungsmethode auf begründeten und anerkannten statistischen und epidemiologischen Methoden? *2a. Basisfalldaten* 2a.1.Wird die Datenauswahl für die Basisfallanalyse ausreichend beschrieben und begründet? 2a.2.Werden die Übergangswahrscheinlichkeiten richtig kalkuliert? 2a.3.Werden ‚half-cycle' Korrekturen für Kosten und Gesundheitseffekte durchgeführt oder die Unterlassung begründet? *2b. Behandlungseffekte* 2b.1.Wird die Datensynthese bei der Verwendung relativer Effekte aus klinischen Studien methodisch korrekt und transparent durchgeführt? 2b.2.Werden die Methoden und Annahmen zur Extrapolation von kurzfristigen auf langfristige Gesundheitseffekte beschrieben und begründet? 2b.3.Werden Annahmen bezüglich des Behandlungseffektes nach Behandlungsstop beschrieben und begründet? 2b.4.Werden Sensitivitätsanalysen zu alternativen Annahmen durchgeführt? 2b.5.Werden die Diskontierungsraten für Gesundheitseffekte beschrieben und begründet?			

2c. Kosten			
2c.1. Werden alle relevanten Kosten berücksichtigt?			
2c.2. Werden alle Quellen für die Kostendaten beschrieben?			
2c.3. Werden geeignete Methoden zur Ermittlung der Mengen- und Preisgerüste angewendet?			
2c.4. Werden die Diskontierungsraten für Kosten beschrieben und begründet?			
2d. Lebensqualitätsbewertung			
2d.1.Werden Nutzwerte korrekt in das Modell integriert?			
2d.2. Wird die Quelle für Nutzwerte beschrieben?			
2d.3. Ist die zur Gewinnung der Nutzwerte verwendete Methode geeignet und begründet?			
3. Datensynthese			
3.1. Werden alle Quellen für die verwendeten Daten ausreichend beschrieben?			
3.2. Werden die Datenauswahl und damit verbundene Annahmen ausreichend begründet?			
3.3. Ist der Prozess der Datensynthese transparent?			
3.4. Wird bei einer Verwendung von Wahrscheinlichkeitsverteilungen die gewählte Verteilung begründet?			
3.5. Wird bei einer Verwendung von Wahrscheinlichkeitsverteilungen die Unsicherheit zweiten Ranges berücksichtigt?			
4. Behandlung von Unsicherheit			
4.1. Werden die vier zentralen Unsicherheitstypen mittels Sensitivitätsanalysen behandelt?			
4.2. Wenn nicht, wird die Entscheidung begründet?			
4a. Methodologische Unsicherheit			
Wird die Modellierung mit mehreren methodologischen Annahmen zur Behandlung methodologischer Unsicherheit durchgeführt?			
4b. Strukturelle Unsicherheit			
Wird strukturelle Unsicherheit in Sensitivitätsanalysen behandelt?			
4c. Heterogenität			
Wird das Modell mit unterschiedlichen Subgruppen gerechnet, um Heterogenität zu berücksichtigen?			
4d.Parameterunsicherheit			
4d.1. Werden geeignete Methoden zur Analyse von Parameterunsicherheit angewandt?			
4d.2. Werden bei der Verwendung von Punktschätzern die für die Sensitivitätsanalyse herangezogenen Intervalle klar beschrieben und begründet?			
Validierung			
1. Interne Validierung			
Wurde die mathematische Logik des Modells getestet?			
2. Externe Validierung			
2.1. Werden kontra-intuitive Ergebnisse erklärt und begründet?			
2.2. Werden bei einer Validierung mit externen Daten auftretende Unterschiede erklärt und begründet?			
2.3. Werden die Modellergebnisse mit früheren Modellen verglichen und Abweichungen erklärt?			

4.9 Appendix J: Formular zur Beurteilung der externen Validität

Beurteilung der externen Validität (Generalisierbarkeit)	**Ja**	**Nein**	**Nicht enthalten**
Handelt es sich bei der Studienpopulation um eine Population in der Primärversorgung?			
Waren die Auswahlkriterien der Studienpopulation wenig restriktiv?			
Wurden patientenrelevante Outcomes (health outcomes) untersucht?			
Sind Studiendauer und Behandlungsmodalitäten klinisch relevant?			
Wurden Nebenwirkungen der Intervention adäquat erhoben?			
Ist die Studiengröße adäquat, um einen minimal-wesentlichen Unterschied aus Patientenperspektive erheben zu können?			
Wurde eine Intention-to-Treat-Analyse durchgeführt?			

Handelt es sich um eine pragmatische Studie (mindestens 6 Kriterien sind erfüllt)?	**Ja**	**Nein**

4.10 Appendix K: Checkliste für GutachterInnen und Darlegung von potentiellen Interessenskonflikten

Kriterien	Ja	Nein
Thema und analytisches Konzept		
1. Ist die HTA-Fragestellung klar definiert und ausreichend erklärt?		
2. Wurde die Relevanz des Themas ausreichend erklärt?		
3. Ist der methodologische Ablauf ausreichend erklärt und valide?		
Literatursuche		
4. Wurden adäquate Suchbegriffe und Datenbanken für die Literatursuche verwendet?		
5. Sind die Auswahlkriterien klar definiert und adäquat?		
6. Fehlen wesentliche Studien? Wenn ja, erstellen Sie bitte eine Liste dieser Literatur.		
7. Gibt es Studien, die der Bericht nicht beinhalten sollte? Wenn ja, erstellen Sie bitte eine Liste dieser Literatur.		
Durchsicht und Analyse der Literatur		
8. Wurde die wissenschaftliche Evidenz korrekt analysiert und dargestellt? a. für Frage 1?		
b. für Frage 2?		
c. für Frage 3?		
d. für Frage 4?		
9. Wurden Resultate und Schlussfolgerungen klar präsentiert: a. für Frage 1?		
b. für Frage 2?		
c. für Frage 3?		
d. für Frage 4?		
10. Tabellen und Abbildungen: a. Sind Tabellen und Abbildungen verständlich und interpretierbar?		
b. Sind Tabellen und Abbildungen vollständig?		
Interpretation der Evidenz		
11. Wurde die Evidenz korrekt interpretiert?		
12. Unterstützt die vorhandene Evidenz die Schlussfolgerungen?		
13. Wurden die Schlussfolgerungen ausreichend diskutiert?		
14. Wurde fehlende wissenschaftliche Evidenz klar verdeutlicht?		
Generelle Präsentation		
15. Ist der HTA-Bericht verständlich strukturiert?		
16. Ist der HTA-Bericht verständlich geschrieben?		
17. Ist der HTA-Bericht vollständig?		
18. Wie könnte der HTA-Bericht verbessert werden?		

Appendix K (fortgesetzt)

Darlegung potentieller Interessenskonflikte

(Adaption des IQWiG-Formulars):

1) Sind oder waren Sie innerhalb der letzten 3 Jahre bei einer Person, Institution oder Firma† abhängig (angestellt) beschäftigt, die von den Ergebnissen Ihrer wissenschaftlichen Arbeit für das Institut‡ finanziell profitieren könnte?

   O ja O nein

   Falls ja, wo und in welcher Position?

   _____

   _____

   _____

   _____

2) Beraten Sie oder haben Sie innerhalb der letzten 3 Jahre eine Person, Institution oder Firma direkt oder indirekt§ beraten, die von den Ergebnissen Ihrer wissenschaftlichen Arbeit für das Institut finanziell profitieren könnte?

   O ja O nein

   Falls ja, wen und wie hoch ist/war das Honorar?

   _____

   _____

   _____

   _____

5 Referenzliste

1. Draborg E, Gyrd-Hansen D, Poulsen PB, Horder M. International comparison of the definition and the practical application of health technology assessment. Int J Technol Assess Health Care 2005; 21(1):89-95.
2. National Institute for Health and Clinical Excellence (April 2006). The guidelines manual. London: National Institute for Health and Clinical Excellence. Verfügbar unter: www.nice.org.uk. Zitiert am 25.05.2006.
3. Agency for Healthcare Research and Quality. Verfügbar unter: www.ahrq.gov/clinic/epcindex.htm#methodology. Zitiert am 29.05.2006.
4. Atkins D, Fink K, Slutsky J. Better information for better health care: the Evidence-based Practice Center program and the Agency for Healthcare Research and Quality. Ann Intern Med 2005; 142(12 Pt 2):1035-41.
5. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Methoden. Verfügbar unter: www.iqwig.de/methoden.428.html. Zitiert am 23.05.2006.
6. Higgins JPT GSCHfSRoIe425. Verfügbar unter: www.cochrane.org/resources/handbook/hbook.htm Zitiert am 29.05.2006.
7. Jakubowski E, Perleth M, Busse R. European Commission, Directorate-General for Employment, Industrial Relations and Social Affairs (ed.). "Best Practice": State of the art and perspectives in the EU for improving the effectiveness and efficiency of European health systems. Luxembourg: Office for Official Publications of the European Communities, 1999.
8. Wild C. Health Technology Assessment. In: Qualitätssicherung in der Medizin, Fischer R. und H. Tragl (Hrsg), Wien: 249-264.
9. Busse R, Orvain J, Drummond M *et al*. Best Practice in undertaking and reporting HTA. ECHTA Working Group 4 Final Report. 2001. http://www.oeaw.ac.at/ita/ebene5/WG4_FinalReport_010719.pdf
10. Perleth M, Busse R. Health Technology Assessment (HTA) - Teil und Methode der Versorgungsforschung. Gesundh Okon Qual Mang 2004; 9: 172-176.
11. European Network for Health Technology Assessment. Verfügbar unter: www.eunethta.net/HTA/. Zitiert am 08.12.2006.
12. International Network for Agencies in Health Technology Assessment. Verfügbar unter: http://www.inahta.org. Zitiert am 08.07.2006.
13. Claxton K, Cohen JT, Neumann PJ. When is evidence sufficient? Health Aff (Millwood) 2005; 24(1):93-101.
14. Fox DM. Evidence of evidence-based health policy: the politics of systematic reviews in coverage decisions. Health Aff (Millwood) 2005; 24(1):114-22.
15. Clancy CM, Cronin K. Evidence-based decision making: global evidence, local decisions. Health Aff (Millwood) 2005; 24(1):151-62.

16. Helfand M. Using evidence reports: progress and challenges in evidence-based decision making. Health Aff (Millwood) 2005; 24(1):123-7.
17. Atkins D, Best D, Briss PA *et al*. Grading quality of evidence and strength of recommendations. BMJ 2004; 328(7454):1490.
18. Akobeng AK. Principles of evidence based medicine. Arch Dis Child 2005; 90(8):837-40.
19. Moynihan R. Evaluating Health Services: A Reporter Covers the Science of Research Synthesis. 2004.
20. Alborz A, McNally R. Developing methods for systematic reviewing in health services delivery and organization: an example from a review of access to health care for people with learning disabilities. Part 2. Evaluation of the literature - a practical guide. Health Info Libr J 2004; 21(4):227-36.
21. McNally R, Alborz A. Developing methods for systematic reviewing in health services delivery and organization: an example from a review of access to health care for people with learning disabilities. Part 1. Identifying the literature. Health Info Libr J 2004; 21(3):182-92.
22. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995; 123(3):A12-3.
23. Akobeng AK. Evidence in practice. Arch Dis Child 2005; 90(8):849-52.
24. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 1993; 270(17):2093-5.
25. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997; 127(5):380-7.
26. Mulrow CD. Rationale for systematic reviews. BMJ 1994; 309(6954):597-9.
27. Akobeng AK. Understanding systematic reviews and meta-analysis. Arch Dis Child 2005; 90(8):845-8.
28. Egger M JPBCHFSJ. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empiracal Study. Health Technology Assess 2003; 7(1).
29. Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. Int J Epidemiol 2002; 31(1):150-3.
30. Betran AP, Say L, Gulmezoglu AM, Allen T, Hampson L. Effectiveness of different databases in identifying studies for systematic reviews: experience from the WHO systematic review of maternal morbidity and mortality. BMC Med Res Methodol 2005; 5(1):6.
31. Savoie I, Helmer D, Green CJ, Kazanjian A. Beyond Medline: reducing bias through extended systematic review search. Int J Technol Assess Health Care 2003; 19(1):168-78.
32. Sampson M, McGowan J. Errors in search strategies were identified by type and frequency. J Clin Epidemiol 2006; 59(10):1057-63.
33. Bartkowiak BA. Searching for evidence-based medicine in the literature part 2: resources. Clin Med Res 2005; 3(1):39-40.

34. Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. Stat Med 2002; 21(11):1625-34.
35. Armstrong R, Jackson N, Doyle J, Waters E, Howes F. It's in your hands: the value of handsearching in conducting systematic reviews of public health interventions. J Public Health (Oxf) 2005; 27(4):388-91.
36. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997; 315(7109):640-5.
37. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. BMJ 2001; 323(7304):101-5.
38. Burdett S, Stewart LA, Tierney JF. Publication bias and meta-analyses: a practical example. Int J Technol Assess Health Care 2003; 19(1):129-34.
39. Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. BMJ 2000; 320(7249):1574-7.
40. Scherer R, Langenberg P, von Elm E. Full publication of results initially presented in abstracts (Review). The Cochrane Library, 2006.
41. Dundar Y, Dodd S, Williamson P, Dickson R, Walley T. Case study of the comparison of data from conference abstracts and full-text articles in health technology assessment of rapidly evolving technologies: does it make a difference? Int J Technol Assess Health Care 2006; 22(3):288-94.
42. MacLean CH, Morton SC, Ofman JJ, Roth EA, Shekelle PG. How useful are unpublished data from the Food and Drug Administration in meta-analysis? J Clin Epidemiol 2003; 56(1):44-51.
43. McManus RJ, Wilson S, Delaney BC *et al*. Review of the usefulness of contacting other experts when conducting a literature search for systematic reviews. BMJ 1998; 317(7172):1562-3.
44. Royle P, Milne R. Literature searching for randomized controlled trials used in Cochrane reviews: rapid versus exhaustive searches. Int J Technol Assess Health Care 2003; 19(4):591-603.
45. 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. Onkologie 2000; 23(6):597-602.
46. Center for Disease Control. The Guide to Community Preventive Services. Verfügbar unter: www.thecommunityguide.org/methods/abstractionform.pdf. Zitiert am 23.07.2006.
47. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet 2002; 359(9300):57-61.
48. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000; 342(25):1887-92.

49. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care 2004; 16(1):9-18.
50. West S, King V, Carey TS *et al*. Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ) 2002; (47):1-11.
51. Harris RP, Helfand M, Woolf SH *et al*. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001; 20(3 Suppl):21-35.
52. NHS Centre for Reviews and Dissemination (2001) *Understaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews. CRD Report Number 4*. 2nd edition York: NHS Centre for Reviews and Dissemination, University of New York.
53. Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ 2004; 328(7430):39-41.
54. Steinberg EP, Luce BR. Evidence based? Caveat emptor! Health Aff (Millwood) 2005; 24(1):80-92.
55. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996; 312(7040):1215-8.
56. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. Health Technol Assess 1998; 2(13):i-iv, 1-124.
57. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004; 363(9422):1728-31.
58. Oxford Centre for Evidence-based Medicine. Levels of Evidence. Verfügbar unter: http://www.cebm.net/levels_of_evidence.asp#levels. Zitiert am 13.09.2006.
59. NHS Centre for Reviews and Dissemination. Making cost-effectiveness information accessible. The NHS economic evaluation database project. CRD guidelines for reporting critical summaries of economic evaluations. York: University of York. NHS Centre for Reviews and Dissemination; 1996.
60. 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.
61. Lohr KN, Carey TS. Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv 1999; 25(9):470-9.
62. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001; 323(7303):42-6.
63. Balk EM, Bonis PA, Moskowitz H *et al*. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002; 287(22):2973-82.
64. Berlin JA, Rennie D. Measuring the quality of trials: the quality of quality scales. JAMA 1999; 282(11):1083-5.
65. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999; 282(11):1054-60.

66. Deeks JJ, Dinnes J, D'Amico R *et al*. Evaluating non-randomised intervention studies. Health Technol Assess 2003; 7(27):iii-x, 1-173.
67. 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 Med Res Methodol 2003; 3:25.
68. Siebert U, Behrend C, Mühlberger Neal. Entwicklung eines Kriterienkatalogs zur Beschreibung und Bewertung ökonomischer Evaluationsstudien in Deutschland. In: Leidl R, Graf von der Schulenburg JM, Wasem J, Hrsg. Ansätze und Methoden der ökonomischen Evaluation Eine internationale Perspektive. Baden-Baden: Nomos Verlag; 1999.
69. Philips Z, Ginnelly L, Sculpher M *et al*. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 2004; 8(36):iii-iv, ix-xi, 1-158.
70. Grossman J, Mackenzie FJ. The randomized controlled trial: gold standard, or merely standard? Perspect Biol Med 2005; 48(4):516-34.
71. Akobeng AK. Understanding randomised controlled trials. Arch Dis Child 2005; 90(8):840-4.
72. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002; 359(9302):248-52.
73. Bavry AA, Bhatt DL. Interpreting observational studies -- look before you leap. J Clin Epidemiol 2006; 59(8):763-4.
74. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet 2002; 359(9303):341-5.
75. Schulz KF, Grimes DA. Case-control studies: research in reverse. Lancet 2002; 359(9304):431-4.
76. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140(3):189-202.
77. Grimes DA, Schulz KF. Uses and abuses of screening tests. Lancet 2002; 359(9309):881-4.
78. 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 Technol Assess 2004; 8(25):iii, 1-234.
79. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006; 144(11):850-5.
80. Verhagen AP, de Vet HC, Vermeer F *et al*. The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 2002; 18(1):11-23.
81. Moher D, Pham B, Jones A *et al*. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998; 352(9128):609-13.

82. 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):408-12.
83. 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):125-8.
84. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998; 317(7167):1185-90.
85. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" Lancet 2005; 365(9453):82-93.
86. Green LW, Glasgow RE. Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. Eval Health Prof 2006; 29(1):126-53.
87. Califf RM, DeMets DL. Principles from clinical trials relevant to clinical practice: Part I. Circulation 2002; 106(8):1015-21.
88. Godwin M, Ruhland L, Casson I *et al*. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003; 3(1):28.
89. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol 2006; 59(10):1040-8.
90. Brook RH, Lohr KN. Efficacy, effectiveness, variations, and quality. Boundary-crossing research. Med Care 1985; 23(5):710-22.
91. Helms PJ. 'Real world' pragmatic clinical trials: what are they and what do they tell us? Pediatr Allergy Immunol 2002; 13(1):4-9.
92. MacRae KD. Pragmatic versus explanatory trials. Int J Technol Assess Health Care 1989; 5(3):333-9.
93. Roland M, Torgerson DJ. What are pragmatic trials? BMJ 1998; 316(7127):285.
94. Anonymous. Reviews: making sense of an often tangled skein of evidence. Ann Intern Med 2005; 142(12 Pt 1):1019-20.
95. Guyatt G, Vist G, Falck-Ytter Y, Kunz R, Magrini N, Schunemann H. An emerging consensus on grading recommendations? ACP J Club 2006; 144(1):A8-9.
96. Atkins D, Eccles M, Flottorp S *et al*. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res 2004; 4(1):38.
97. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995; 310(6977):452-4.
98. Anonymous. Measuring the Size of an Intervention. Verfügbar unter: http://gim.unmc.edu/dxtests/Effect1.htm. Zitiert am 12.08.2006.
99. Carney S, Doll H. Introduction to biostatistics: Part 2. Measures of association as used to address therapy, harm, and etiology questions. ACP J Club 2005; 143(2):A8.
100. Bewick V, Cheek L, Ball J. Statistics review 11: assessing risk. Crit Care 2004; 8(4):287-91.

101. Shekelle P, Eccles MP, Grimshaw JM, Woolf SH. When should clinical guidelines be updated? BMJ 2001; 323(7305):155-7.
102. Gartlehner G, West SL, Lohr KN *et al*. Assessing the need to update prevention guidelines: a comparison of two methods. Int J Qual Health Care 2004; 16(5):399-406.
103. Montori VM, Wilczynski NL, Morgan D, Haynes RB. Optimal search strategies for retrieving systematic reviews from Medline: analytical survey. BMJ 2005; 330(7482):68.