Evidenzberichte mit Fragestellungen zu Interventionen

Version: 2.0 Stand: 30.06.2025

Impressum

Herausgeber

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema

Evidenzberichte mit Fragestellungen zu Interventionen

Anschrift des Herausgebers

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Siegburger Str. 237
50679 Köln

Tel.: +49 221 35685-0
Fax: +49 221 35685-1
E-Mail: berichte@iqwig.de
Internet: www.iqwig.de

Inhaltsverzeichnis

Tabellenverzeichnis

Seite

Abkürzungsverzeichnis

Abkürzung	Bedeutung
AHRQ	Agency for Healthcare Research and Quality
AMSTAR	A Measurement Tool to Assess Systematic Reviews
AWMF	Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V.
BMG	Bundesministerium für Gesundheit
DVG	Digitale-Versorgung-Gesetz
G-BA	Gemeinsamer Bundesausschuss
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HTA	Health Technology Assessment
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ITT	Intention to treat
KI	Konfidenzintervall
MWD	Mittelwertdifferenz
NICE	National Institute for Health and Care Excellence
OR	Odds Ratio
PICO	Population, Intervention, Comparison, Outcome
PICo	Population, Phenomena of Interest, Context, Other/Outcomes
PIRD	Population, Index Test, Reference Standard, Diagnosis of Interest
RCT	Randomized controlled Trial (randomisierte kontrollierte Studie)
RD	absolute Risikodifferenz
ROBINS-I	Risk Of Bias In Non-randomised Studies – of Interventions
SGB	Sozialgesetzbuch
SMD	Standardized Mean Difference (standardisierte Mittelwertdifferenz)
SÜ	systematische Übersicht

1 Hintergrund

Das vorliegende Dokument beschreibt die Methodik bei der Erstellung eines Evidenzberichts für die Bearbeitung von Fragestellungen zu präventiven (darunter auch von Screeningmaßnahmen) und therapeutischen Maßnahmen sowie zu Diagnoseverfahren, sofern deren Evaluation sich nicht allein auf die diagnostische Güte beschränken soll.

Die Methodik für die Bearbeitung von Fragestellungen zur diagnostischen Güte oder zur Bearbeitung von qualitativen Fragestellungen wird jeweils in anderen Dokumenten beschrieben.

Die Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF) kann dem Bundesministerium für Gesundheit (BMG) Themen zur Entwicklung oder Weiterentwicklung von Leitlinien vorschlagen, die das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) gemäß Sozialgesetzbuch (SGB) V (§§ 139a Abs. 3 Nr. 3, 139b Abs. 6) mit Evidenzrecherchen unterstützen soll [1]. Hierfür formuliert die Leitliniengruppe in Abstimmung mit Patientenvertreterinnen und -vertretern und mit Beratung durch das IQWiG und die AWMF Vorschläge für 1 oder mehrere Population-Intervention-Comparison-Outcome(PICO)-, Population-Index-Test-Reference-Standard-Diagnosis-of-Interest(PIRD)- und / oder Population-Phenomena-of-Interest-Context-Other/Outcomes(PICo)-Fragestellungen. Im Anschluss schlägt die AWMF dem BMG das entsprechende Leitlinienthema mit den konkretisierten Fragestellungen vor. Nach Prüfung des Antrags und einem Auftakttreffen zwischen den Leitlinienkoordinierenden, der AWMF, dem BMG und dem IQWiG beauftragt das BMG das IQWiG mit einer Evidenzrecherche zur Unterstützung der Leitliniengruppe. Zur Auftragsbearbeitung findet 1 Kick-off-Treffen zwischen den Leitlinien-koordinierenden, ggf. 1 Ansprechperson der AWMF und Ansprechpersonen des IQWiG statt, in denen die Fragestellungen finalisiert werden. Zu jeder Fragestellung erstellt das IQWiG 1 Evidenzbericht, der nach Fertigstellung an die Leitlinienkoordinierenden, an die zuständige Ansprechperson für die Leitlinie bei der AWMF sowie an das BMG übermittelt wird.

Nach Abschluss aller Evidenzberichte für einen gesamten Auftrag werden diese zusammen an die Gremien des IQWiG und das BMG übermittelt sowie 4 Wochen später auf der Website des IQWiG veröffentlicht.

2 Methoden

Jede Fragestellung wird durch die Darstellung von Ergebnissen zu Endpunkten in Evidenzprofilen beantwortet. Die Erstellung der Evidenzprofile erfolgt auf Grundlage der methodischen Vorgaben von Grading of Recommendations Assessment, Development and Evaluation (GRADE) [2] und wird durch die IQWiG-Methodik [3] konkretisiert. Das PICO-Schema sowie die Einteilung der Endpunkte hinsichtlich ihrer Relevanz zur Entscheidungsfindung in kritisch und wichtig werden von der Leitliniengruppe zur entsprechenden Leitlinie festgelegt.

2.1 Kriterien für den Studieneinschluss

Folgende Kriterien für den Studieneinschluss werden in Absprache mit der Leitliniengruppe festgelegt:

Tabelle 1: Übersicht über die Kriterien für den Studieneinschluss

Einschlusskriterien	
E1	Population
E2	Prüfintervention[a, b]
E3	Vergleichsintervention[a, b]
E4	Endpunkte: (Optionale Einteilung in kritische und wichtige Endpunkte: Die Leitliniengruppe kann festlegen, ob die Endpunkte hinsichtlich ihrer Relevanz für die Empfehlungsbildung in „kritisch“ (für die Entscheidung) und „wichtig“ (aber nicht kritisch für die Entscheidung) eingeteilt werden. Hierbei wird sich an einer Gesamtzahl von 7 kritischen und / oder wichtigen Endpunkten orientiert. Bei mehr als 7 Endpunkten werden die kritischen Endpunkte mit verwertbaren Ergebnissen den wichtigen Endpunkten vorgezogen. Bei Vorliegen von weniger als 7 kritischen Endpunkten werden die wichtigen Endpunkte in der benannten Rangfolge untersucht, bis die Gesamtzahl von 7 Endpunkten mit verwertbaren Ergebnissen erreicht ist.)
E5	Studientyp: Für den zu erstellenden Evidenzbericht werden in erster Linie RCTs als relevante wissenschaftliche Informationsquelle einfließen. Sind Ergebnisse aus RCTs nicht in ausreichender Zahl und mit ausreichender Vertrauenswürdigkeit der Evidenz vorhanden, werden neben RCTs ggf. nicht randomisierte vergleichende Studien eingeschlossen (quasirandomisierte kontrollierte Studien, prospektive vergleichende Kohortenstudien, retrospektive vergleichende Kohortenstudien mit zeitlich paralleler Kontrollgruppe, retrospektive vergleichende Kohortenstudien mit nicht zeitlich paralleler Kontrollgruppe).
E6	Studiendauer (sofern relevant)
E7	Publikationssprache: regelhaft Deutsch oder Englisch
E8	Vollpublikation verfügbar[c]
E9	Publikationszeitraum / Zeitraum der Studiendurchführung (optional): Projektspezifisch können Eingrenzungen hinsichtlich Publikationszeitraum bzw. Zeitraum der Studiendurchführung sinnvoll sein.
E10	Setting (sofern relevant)
Ausschlusskriterien: Projektspezifisch können in Ausnahmefällen Ausschlusskriterien formuliert werden.	

a. Grundlage für die Evidenzdarstellung von Arzneimitteln ist in der Regel die in Deutschland bestehende Arzneimittelzulassung. Es wird daher geprüft, ob die Anwendung der in den Studien eingesetzten Prüf- und Vergleichsinterventionen im Rahmen des für Deutschland gültigen Zulassungsstatus erfolgte. Eine weitere Voraussetzung ist die Marktverfügbarkeit in Deutschland.

b. Eine Prüfung der für Deutschland gültigen Zertifizierung eines Medizinprodukts erfolgt, wenn dieses explizit als solches vorab benannt und zur Bewertung eingebracht wurde sowie in Deutschland marktverfügbar ist. Alle anderen Interventionen, die auch als Medizinprodukt im weiteren Sinne gelten können (z. B. Apps), werden bezüglich der Zertifizierung und des entsprechenden Einsatzes in den Studien nicht beurteilt.

c. Abhängig von den einzuschließenden Studientypen gilt als Vollpublikation in diesem Zusammenhang auch ein Studienbericht gemäß ICH E3 [4] oder ein Bericht über die Studie, der den Kriterien des CONSORT- [5], TREND- [6] oder STROBE-Statements [7] genügt und eine Bewertung der Studie ermöglicht, sofern die in diesen Dokumenten enthaltenen Informationen zur Studienmethodik und zu den Studienergebnissen nicht vertraulich sind.

CONSORT: Consolidated Standards of Reporting Trials; ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use; RCT: randomisierte kontrollierte Studie; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology; TREND: Transparent Reporting of Evaluations with Nonrandomized Designs

Berücksichtigung von nicht randomisierten vergleichenden Studien

Nicht randomisierte vergleichende Studien werden schrittweise berücksichtigt, sofern Ergebnisse in der nächsthöheren Evidenzstufe nicht in ausreichender Zahl vorliegen: quasirandomisierte kontrollierte Studie, prospektive vergleichende Kohortenstudie, retrospektive vergleichende Kohortenstudie mit zeitlich paralleler Kontrollgruppe, retrospektive vergleichende Kohortenstudie mit nicht zeitlich paralleler Kontrollgruppe. Sie bleiben jedoch für die Evidenzdarstellung unberücksichtigt, wenn sie ein kritisches Verzerrungspotenzial aufweisen, bei dem keine sinnvolle Interpretation der vorliegenden Studienergebnisse möglich ist. Die Bewertung des Verzerrungspotenzials von relevanten nicht randomisierten vergleichenden Studien orientiert sich an den Domänen des ROBINS-I(Risk Of Bias In Non-randomised Studies – of Interventions)-Instruments [8].

2.2 Informationsbeschaffung

2.2.1 Fokussierte Informationsbeschaffung von systematischen Übersichten

Zunächst erfolgt eine systematische Recherche nach systematischen Übersichten (SÜs) in MEDLINE (umfasst auch die Cochrane Database of Systematic Reviews), der International Health Technology Assessment (HTA) Database sowie auf den Websites des National Institute for Health and Care Excellence (NICE) und der Agency for Healthcare Research and Quality (AHRQ).

Die Selektion erfolgt in der Regel durch 1 Person und wird anschließend von einer 2. Person überprüft. Diskrepanzen werden durch Diskussion zwischen den beiden aufgelöst. Wird mindestens 1 hochwertige und aktuelle SÜ identifiziert, die die beauftragte Fragestellung ausreichend abdeckt, wird geprüft, ob deren Informationsbeschaffung als Grundlage für die Evidenzdarstellung verwendet werden kann. Die SÜ wird dann als Basis-SÜ bezeichnet. Zur Überprüfung der Eignung als Basis-SÜ erfolgt eine Bewertung der Qualität der Informationsbeschaffung dieser SÜ(s) mit den entsprechenden Items aus A Measurement Tool to Assess Systematic Reviews 2 (AMSTAR 2) [9]. Die Bewertung erfolgt durch 1 Person und wird von einer 2. Person überprüft. Kann mindestens 1 diesbezüglich hochwertige und aktuelle Basis-SÜ identifiziert werden, werden die zugrunde liegenden Studien bzw. Dokumente von 1 Person auf deren Relevanz für die vorliegende Evidenzdarstellung geprüft und das Ergebnis von einer 2. Person überprüft. Die weitere Prüfung einer Basis-SÜ hinsichtlich einer möglichen Berücksichtigung von Ergebnissen wird in Abschnitt 2.3 beschrieben.

In jedem Fall werden die Referenzlisten der identifizierten SÜ(s) hinsichtlich relevanter Primärstudien gesichtet (siehe Abschnitt 2.2.2).

2.2.2 Fokussierte Informationsbeschaffung von Studien

Für die fokussierte Informationsbeschaffung wird eine systematische Recherche nach relevanten Studien beziehungsweise Dokumenten durchgeführt. Für den Fall, dass mindestens 1 SÜ als Basis-SÜ für die Informationsbeschaffung des Evidenzberichts verwendet werden kann (siehe Abschnitt 2.2.1), wird diese für die Informationsbeschaffung von Studien für den von der Basis-SÜ abgedeckten Zeitraum herangezogen. Dieser Teil der Informationsbeschaffung wird ergänzt um eine systematische Recherche nach relevanten Studien bzw. Dokumenten für den nicht von der Übersicht abgedeckten Zeitraum. Ggf. wird auf eine ergänzende fokussierte Informationsbeschaffung von Studien ganz verzichtet (z. B. aufgrund ausreichender Aktualität der Basis-SÜ).

Für den Fall, dass keine Basis-SÜ identifiziert werden kann, findet eine systematische Recherche für den gesamten relevanten Zeitraum statt.

Folgende primäre und weitere Informationsquellen sowie Suchtechniken werden dabei berücksichtigt:

Primäre Informationsquellen

- bibliografische Datenbanken
  - MEDLINE
  - Cochrane Central Register of Controlled Trials
- Studienregister (es erfolgt eine Einschränkung auf Einträge mit Ergebnissen)
  - U.S. National Institutes of Health. ClinicalTrials.gov
  - World Health Organization. International Clinical Trials Registry Platform Search Portal (nur bei nicht medikamentösen Verfahren)

Weitere Informationsquellen und Suchtechniken

- Gemeinsamer Bundesausschuss(G-BA)-Website und IQWiG-Website
- Anwendung weiterer Suchtechniken
  - Sichten von Referenzlisten identifizierter SÜs
  - Identifizieren von Studienregistereinträgen zu eingeschlossenen Studien
- Autorenanfragen bei Bedarf

2.2.3 Orientierende Recherche zu Reporting Bias

Folgende Informationsquelle wird berücksichtigt:

- U.S. National Institutes of Health. ClinicalTrials.gov

Die Suche wird mit den Einschränkungen bezüglich des Studienstatus „abgeschlossen“, „vorzeitig abgebrochen“ sowie mit der Einschränkung auf Einträge ohne Ergebnisse durchgeführt.

2.2.4 Anwendung von Limitierungen auf Datenbankebene

Fokussierte Informationsbeschaffung von SÜs

Die Suchen können auf einen Publikationszeitraum beschränkt werden. Die MEDLINE-Suchstrategie enthält Limitierungen auf deutsch- und englischsprachige Publikationen [3] sowie auf Humanstudien.

Fokussierte Informationsbeschaffung von Studien

Die Suchen können auf einen Publikationszeitraum (siehe Einschlusskriterium E9) beschränkt werden.

Sollte die Informationsbeschaffung auf Grundlage einer Basis-SÜ erfolgen, wird eine entsprechende zeitliche Einschränkung in Betracht gezogen (siehe Abschnitt 2.2.2).

Mit der MEDLINE-Suchstrategie werden folgende Publikationstypen ausgeschlossen: Kommentare und Editorials, da diese in der Regel keine Studien enthalten (siehe Einschlusskriterium E8) [10]. Außerdem enthalten die Suchstrategien Limitierungen auf deutsch- und englischsprachige Publikationen (siehe Einschlusskriterium E7) sowie auf Humanstudien (MEDLINE). In der Cochrane Central Register of Controlled Trials Suche werden Einträge aus Studienregistern ausgeschlossen.

2.2.5 Selektion relevanter Studien

Selektion relevanter Studien bzw. Dokumente aus den Ergebnissen der bibliografischen Recherche

Duplikate werden mit Hilfe des Literaturverwaltungsprogramms EndNote entfernt. Die in bibliografischen Datenbanken identifizierten Treffer werden in einem 1. Schritt anhand ihres Titels und, sofern vorhanden, Abstracts in Bezug auf ihre potenzielle Relevanz bezüglich der Einschlusskriterien (siehe Tabelle 1) bewertet. Als potenziell relevant erachtete Dokumente werden in einem 2. Schritt anhand ihres Volltextes auf Relevanz geprüft. Beide Schritte erfolgen durch 2 Personen unabhängig voneinander. Diskrepanzen werden durch Diskussion zwischen den beiden aufgelöst.

Selektion relevanter Studien bzw. Dokumente aus weiteren Informationsquellen

Die Rechercheergebnisse aus den folgenden Informationsquellen werden von 1 Person auf Studien gesichtet:

- Studienregister

- Referenzliste(n) identifizierter SÜ(s)

Die identifizierten Studien werden auf ihre Relevanz geprüft. Der gesamte Prozess wird anschließend von einer 2. Person überprüft. Diskrepanzen werden durch Diskussion zwischen den beiden aufgelöst.

2.3 Informationsdarstellung und Synthese

Für den Fall, dass 1 oder mehrere Basis-SÜ(s) vorliegen (siehe Abschnitt 2.2.1), wird geprüft, ob diese auch dafür infrage kommt, deren Ergebnisse als Grundlage für die Evidenzdarstellung zu verwenden. Eine wesentliche Voraussetzung dafür ist, dass die Basis-SÜ über eine ausreichend hohe methodische Qualität verfügt. Die Prüfung der Qualität erfolgt in der Regel durch eine vollständige Bewertung der Qualität dieser SÜ mit AMSTAR 2 [9]. Die Bewertung erfolgt durch 1 Person und wird von einer 2. Person überprüft. Diskrepanzen werden durch Diskussion zwischen beiden aufgelöst. Sofern 1 methodisch hochwertige Basis-SÜ identifiziert wird, können aus dieser die relevanten Informationen für den Evidenzbericht herangezogen werden.

2.3.1 Darstellung der Studien

Alle für den Evidenzbericht notwendigen Informationen werden aus den Unterlagen zu den für die Evidenzdarstellung berücksichtigten Studien bzw. Basis-SÜ(s) in standardisierte Tabellen extrahiert. Ergeben sich im Abgleich der Informationen aus unterschiedlichen Dokumenten zu einer Studie (aber auch aus multiplen Angaben zu einem Aspekt innerhalb eines Dokumentes selbst) Diskrepanzen, die auf die Interpretation der Ergebnisse erheblichen Einfluss haben könnten, wird dies an den entsprechenden Stellen des Berichts dargestellt. Die Extraktion erfolgt durch 1 Person und wird von einer 2. Person auf Grundlage der Studien bzw. der Basis-SÜ kontrolliert.

Die Ergebnisse zu den von der Leitliniengruppe festgelegten und in den Studien bzw. Basis-SÜ(s) berichteten Endpunkten werden im Bericht vergleichend dargestellt. Anhand inhaltlicher und methodischer Kriterien kann eine Auswahl der dargestellten Operationalisierungen zu den Endpunkten, für die verwertbare Daten vorliegen, erfolgen. Die Auswahl erfolgt unter Berücksichtigung der Datenlage für den gesamten Evidenzbericht. Operationalisierungen aus qualitativ guten Studien und / oder Studien mit größeren Stichproben werden bevorzugt dargestellt. Werden in der überwiegenden Zahl der Studien Ergebnisse zu 1 spezifischen Operationalisierung dargestellt, kann die Darstellung im Bericht darauf begrenzt werden. Werden in den Studien Auswertungen mittels verschiedener Modelle berichtet, die nicht metaanalytisch zusammengefasst werden können, kann die Darstellung auf 1 Auswertungsart beschränkt werden. Es wird geprüft, ob die Ergebnisse, die nicht dargestellt werden, zu denen der ausgewählten Operationalisierung(en) bzw. Auswertung(en) passen. Abweichungen werden entsprechend eingeordnet.

Das Vorgehen zur Bewertung des Einflusses des Verzerrungspotenzials auf die berichtsrelevanten Ergebnisse wird in Abschnitt 2.3.2 endpunktspezifisch pro Studie beschrieben. Diese Einschätzung fließt in die Bewertung der studienübergreifenden Vertrauenswürdigkeit der Evidenz ein. Diese und weitere Faktoren der Vertrauenswürdigkeit der Evidenz werden gemeinsam mit den Ergebnissen endpunktspezifisch in Evidenzprofilen zusammengeführt und dargestellt (siehe Abschnitt 2.3.5) [11]. Wenn möglich, werden die in den Abschnitten 2.3.3 bis 2.3.4 beschriebenen Verfahren durchgeführt.

Für binäre Daten wird als relative Effektschätzung primär das Odds Ratio (OR) herangezogen. Dabei wird im Fall von 0 Ereignissen in einem Studienarm bei der Berechnung von Effekt und Konfidenzintervall (KI) der Korrekturfaktor 0,5 in beiden Studienarmen verwendet. Als absolutes Effektmaß wird vorrangig die absolute Risikodifferenz (RD) verwendet. Das Basisrisiko wird in der Regel durch den Median des Risikos der Vergleichsgruppe(n) der berücksichtigten Einzelstudie(n) bestimmt. Auf Grundlage dieses angenommenen Basisrisikos werden mithilfe der relativen Gesamtschätzung der Metaanalyse das absolute Risiko in der Interventionsgruppe und die RD errechnet. Zur Bewertung der Vertrauenswürdigkeit der Evidenz werden ergänzend die obere und untere KI-Grenze der RD basierend auf den KI-Grenzen der relativen Gesamtschätzung der Metaanalyse berechnet. Wird keine Metaanalyse durchgeführt, erfolgt eine Berechnung der RD pro Studie durch die Risiken in den beiden Behandlungsgruppen; als zugehöriges KI wird standardmäßig dasjenige nach der Wilson-Score-Methode [12] angegeben. Sollte das Ergebnis des CSZ-Tests [13] qualitativ nicht zum Ergebnis des KI nach der Wilson-Score-Methode, sondern zum Wald-KI passen, wird dieses angegeben.

Für stetige Daten wird primär die Mittelwertdifferenz (MWD) herangezogen. Falls notwendig (z. B., wenn verschiedene Skalen gepoolt werden sollen oder dies zur Bewertung der Vertrauenswürdigkeit der Evidenz benötigt wird), wird außerdem eine standardisierte Mittelwertdifferenz (SMD, z. B. Hedges' g) angegeben. Ergebnisse können ggf. im Evidenzbericht unberücksichtigt bleiben, wenn ein großer Anteil der in die Auswertung eigentlich einzuschließenden Personen nicht in der Auswertung berücksichtigt worden ist. Für die Entscheidung hierüber wird sich an einem Anteil von ca. 70 % orientiert, der in der Auswertung mindestens berücksichtigt sein sollte.

Die Ergebnisse können ggf. auch dann unberücksichtigt bleiben, wenn der Unterschied der Anteile nicht berücksichtigter Personen zwischen den Gruppen sehr groß ist. Als Orientierung für die Entscheidung dient ein Unterschied von mehr als 15 Prozentpunkten.

2.3.2 Kriterien des Verzerrungspotenzials

Das Verzerrungspotenzial wird endpunktspezifisch pro Studie insbesondere anhand der nachfolgend aufgeführten Kriterien beurteilt. Dazu erfolgt jeweils eine Bewertung mit „ja“,

„unklar“ oder „nein“. Die Voraussetzungen für eine Bewertung mit „ja“ werden im Folgenden erläutert. Eine Bewertung mit „unklar“ erfolgt grundsätzlich dann, wenn keine bzw. keine ausreichenden Angaben zur adäquaten Bewertung Verfügung stehen. Die Bewertung erfolgt durch 1 Person und wird von einer 2. Person auf Grundlage der Studien kontrolliert.

Folgende Kriterien werden bei einer Beschränkung auf randomisierte kontrollierte Studien (RCTs) bewertet:

- adäquate Erzeugung der Randomisierungssequenz

  Voraussetzung für eine Bewertung mit „ja“ ist, dass aus der Beschreibung der Studie hervorgeht, dass die Gruppenzuteilung rein zufällig erfolgte und die Erzeugung der Zuteilungssequenz beschrieben und geeignet ist. Die Antwortmöglichkeit „nein“ besteht im Falle eines Evidenzberichts auf Basis von RCTs nicht, da die Studie in diesem Fall als nicht randomisierte Studie aufgefasst und ausgeschlossen wird.

- Verdeckung der Gruppenzuteilung

  Voraussetzung für eine Bewertung mit „ja“ ist, dass aus der Beschreibung der Studie hervorgeht, dass die Zuteilung der Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer in die verschiedenen Studiengruppen den Personen, die die Gruppenzuteilung veranlassen oder über den Einschluss der Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer entscheiden, nicht bekannt ist. Die Antwortmöglichkeit „nein“ besteht im Falle eines Evidenzberichts auf Basis von RCTs nicht, da die Studie in diesem Fall als nicht randomisierte Studie aufgefasst und ausgeschlossen wird.

- Verblindung

  Voraussetzung für eine Bewertung mit „ja“ ist, dass alle zu verblindenden Personengruppen (Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer, behandelnde Personen, endpunkterhebende Personen) jeweils explizit benannt werden und die Methode, die eine Verblindung gewährleistet, beschrieben und geeignet ist.

- Intention-to-treat(ITT)-Prinzip adäquat umgesetzt

  Voraussetzung für eine Bewertung mit „ja“ ist, dass randomisierte Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer entsprechend ihrer Gruppenzuteilung ausgewertet und der Anteil der fehlenden Werte gering ist oder sie in geeigneter Weise berücksichtigt wurden. Regelhaft erfolgt eine Bewertung mit „ja“, wenn der Anteil der nicht in der Auswertung berücksichtigten Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer insgesamt kleiner als 10 % oder der Unterschied der Anteile zwischen den Gruppen kleiner als 5 Prozentpunkte ist. Wurden fehlende Werte in geeigneter Weise berücksichtigt, hängt die Bewertung des Ergebnisses von weiteren Aspekten ab: u. a. der Häufigkeit fehlender Werte, der Art des Auswertungsverfahrens oder der Gründe für das Auftreten fehlender Werte.

- ergebnisunabhängige Berichterstattung

  Voraussetzung für eine Bewertung mit „ja“ ist, dass die Endpunkte gemäß einer prospektiven Planung (in der Regel im Studienregistereintrag) operationalisiert, erfasst und ausgewertet werden oder relevante Abweichungen von den Studienautorinnen und Studienautoren plausibel erklärt werden.

Folgende Kriterien werden bei der Verwendung von nicht randomisierten vergleichenden Studien bewertet:

- zeitliche Parallelität der Interventionsgruppen

  Voraussetzung für eine Bewertung mit „ja“ ist, dass die Interventionen in beiden Gruppen im gleichen Zeitraum starten oder (bei retrospektiven Studien) beobachtet werden.

- Vergleichbarkeit der Interventionsgruppen bzw. adäquate Berücksichtigung prognostisch relevanter Faktoren

  Voraussetzung für eine Bewertung mit „ja“ ist, dass alle wesentlichen Confounder adäquat berücksichtigt wurden oder eine Homogenität der Gruppen bezüglich aller prognostisch relevanten Faktoren besteht.

- Verblindung

  Voraussetzung für eine Bewertung mit „ja“ ist, dass alle zu verblindenden Personengruppen (Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer, behandelnde Personen, endpunkterhebende Personen) jeweils explizit benannt werden und die Methode, die eine Verblindung gewährleistet, beschrieben und geeignet ist.

- ITT-Prinzip adäquat umgesetzt bzw. Daten vollständig

  Regelhaft erfolgt eine Bewertung mit „ja“, wenn der Anteil der nicht in der Auswertung berücksichtigten Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer insgesamt kleiner als 10 % oder der Unterschied der Anteile zwischen den Gruppen kleiner als 5 Prozentpunkte ist. Wurden fehlende Werte in geeigneter Weise berücksichtigt, hängt die Bewertung des Ergebnisses von weiteren Aspekten ab: u. a. der Häufigkeit fehlender Werte, der Art des Auswertungsverfahrens oder der Gründe für das Auftreten fehlender Werte.

- ergebnisunabhängige Berichterstattung

  Voraussetzung für eine Bewertung mit „ja“ ist, dass der Endpunkt gemäß einer prospektiven Planung operationalisiert, erfasst und ausgewertet wird oder relevante Abweichungen von den Studienautorinnen und Studienautoren thematisiert und adäquat eingeordnet werden.

Pro Endpunkt wird für jede Studie nach den oben genannten Kriterien das Verzerrungspotenzial bewertet. Dieses Verzerrungspotenzial wird in einer studienübergreifenden Bewertung der Studienlimitationen für die Evidenzprofile zusammengefasst (siehe Abschnitt 2.3.5.1).

2.3.3 Metaanalysen

Liegen mehrere Studien zu einem Endpunkt vor, werden die Ergebnisse in Metaanalysen zusammengefasst und die gepoolte Schätzung wird im Evidenzprofil dargestellt. Wird in der Metaanalyse Heterogenität beobachtet, fließt dies zudem in die Bewertung der Inkonsistenz ein (siehe Abschnitt 2.3.5).

Die geschätzten Effekte und KIs aus den Studien werden mittels Forest Plots dargestellt. Die Heterogenität zwischen den Studien wird mithilfe des statistischen Tests auf Vorliegen von Heterogenität [14] untersucht. Ergibt der Heterogenitätstest ein statistisch nicht signifikantes Ergebnis ($p \geq 0{,}05$), wird davon ausgegangen, dass die Schätzung eines gemeinsamen (gepoolten) Effekts sinnvoll ist.

In Abhängigkeit von der Anzahl der Studien wird zur Durchführung von Metaanalysen folgendes Standardvorgehen gewählt, sofern keine klaren Gründe dagegensprechen:

- 2 Studien: Anwendung des Modells mit festem Effekt, und zwar mithilfe der inversen Varianzmethode bei stetigen Daten bzw. der Mantel-Haenszel-Methode bei binären Daten [15].
- 3 bis 4 Studien für die Effektmaße SMD, OR, relatives Risiko und Hazard Ratio: Anwendung des Modells mit zufälligen Effekten, und zwar – mithilfe einer bayesschen Metaanalyse mit nicht informativen A-priori-Verteilungen für den Behandlungseffekt und informativen A-priori-Verteilungen für den Heterogenitätsparameter τ gemäß Lilienthal et al. [16].
- 3 bis 4 Studien für sonstige Effektmaße: Vorzugsweise Anwendung des Modells mit zufälligen Effekten, und zwar mithilfe der Knapp-Hartung-Methode. Zunächst werden gepoolte Effekte nach der Methode von Knapp-Hartung – mit und ohne Ad-hoc-Varianzkorrektur – sowie der Paule-Mandel-Methode zur Schätzung des Heterogenitätsparameters τ [17] und gepoolte Effekte nach der Methode von DerSimonian-Laird berechnet. Es wird geprüft, ob das Konfidenzintervall nach Knapp-Hartung (ohne Ad-hoc-Varianzkorrektur) schmaler ist als das nach DerSimonian-Laird. In diesem Fall wird die Effektschätzung nach Knapp-Hartung mit Ad-hoc-Varianzkorrektur, ansonsten ohne Ad-hoc-Varianzkorrektur weiterverwendet. Im Anschluss wird geprüft, ob diese Effektschätzung im Widerspruch zu einer qualitativen Zusammenfassung steht. In diesem Fall wird ein Modell mit festem Effekt verwendet.

- 5 Studien und mehr: Anwendung des Modells mit zufälligen Effekten, und zwar mithilfe der Knapp-Hartung-Methode. Zunächst werden gepoolte Effekte nach der Methode von Knapp-Hartung – mit und ohne Ad-hoc-Varianzkorrektur – sowie der Paule-Mandel-Methode zur Schätzung des Heterogenitätsparameters τ [17] und gepoolte Effekte nach der Methode von DerSimonian-Laird berechnet. Es wird geprüft, ob das Konfidenzintervall nach Knapp-Hartung (ohne Ad-hoc-Varianzkorrektur) schmaler ist als das nach DerSimonian-Laird. In diesem Fall wird die Effektschätzung nach Knapp-Hartung mit Ad-hoc-Varianzkorrektur, ansonsten ohne Ad-hoc-Varianzkorrektur weiterverwendet. Im Anschluss wird geprüft, ob diese Effektschätzung informativ ist. Als informativ wird die Schätzung dann bezeichnet, falls das Konfidenzintervall (des gemeinsamen Effekts) in der Vereinigung der Konfidenzintervalle der Einzelstudien enthalten ist. In diesem Fall wird diese Effektschätzung (nach Knapp-Hartung) zur finalen Bewertung herangezogen. Ansonsten wird projektspezifisch entschieden, welches Modell infrage kommt.
- Ergibt der Heterogenitätstest ein statistisch signifikantes Ergebnis ($p < 0,05$), wird untersucht, welche Faktoren eine vorhandene Heterogenität möglicherweise verursachen. Dazu zählen methodische und klinische Faktoren, sogenannte Effektmodifikatoren. Falls vorhandene Heterogenität durch solche Faktoren zumindest zum Teil erklärt werden kann, so wird der Studienpool nach diesen Faktoren aufgespaltet und die weiteren Berechnungen erfolgen in den getrennten Studienpools. Kann die Heterogenität nicht erklärt werden, so wird, sofern möglich, ebenfalls ein gemeinsamer (gepoolter) Effekt berechnet. Dabei ist zu beachten, dass ein gepoolter Effekt bei bedeutsamer Heterogenität möglicherweise nicht sinnvoll interpretierbar ist. Von der Berechnung eines gepoolten Effekts wird abgesehen, falls sich die KIs der eingehenden Studien nur gering oder gar nicht überlappen und gleichzeitig die Effektschätzungen der Studien in unterschiedliche Richtungen weisen.

2.3.4 Subgruppenmerkmale und andere Effektmodifikatoren

Optional können die Ergebnisse hinsichtlich potenzieller Effektmodifikatoren, das heißt klinischer Faktoren, die die Effekte beeinflussen, untersucht werden. Ziel ist es, mögliche Effektunterschiede zwischen Patienten- bzw. Teilnehmergruppen und Behandlungsspezifika aufzudecken. Für einen Nachweis unterschiedlicher Effekte ist die auf einem Homogenitäts- bzw. Interaktionstest basierende statistische Signifikanz Voraussetzung. In den Evidenzbericht werden die vorliegenden Ergebnisse aus Regressionsanalysen, die Interaktionsterme beinhalten, und aus Subgruppenanalysen einbezogen. Außerdem erfolgen ggf. eigene Analysen in Form von Metaregressionen oder Metaanalysen unter Kategorisierung der Studien bezüglich der möglichen Effektmodifikatoren. Subgruppenanalysen werden nur durchgeführt, falls jede Subgruppe mindestens 10 Personen umfasst und bei binären Daten

mindestens 10 Ereignisse in einer der Subgruppen aufgetreten sind. Folgende Faktoren werden bezüglich einer möglichen Effektmodifikation regelhaft in die Analysen einbezogen:

- Geschlecht,
- Alter.

Sollten sich aus den verfügbaren Informationen weitere mögliche Effektmodifikatoren ergeben, können diese ebenfalls begründet einbezogen werden.

2.3.5 Bewertung der Vertrauenswürdigkeit der Evidenz

Alle für den Evidenzbericht relevanten Ergebnisse werden hinsichtlich einer Beeinflussung durch Faktoren, die zu einer Ab- oder ggf. Aufwertung der Vertrauenswürdigkeit der Evidenz führen können, überprüft. Für jeden Endpunkt wird eine studienübergreifende Aussage zur Vertrauenswürdigkeit der Evidenz bezüglich des jeweiligen Ausmaßes des Vertrauens in die Effektschätzung getroffen. Hierzu erfolgt eine Einteilung der Vertrauenswürdigkeit der Evidenz entsprechend der 4 Stufen der GRADE-Guideline in „hoch“, „moderat“, „niedrig“ und „sehr niedrig“ [18-20]:

- Eine hohe Vertrauenswürdigkeit der Evidenz bedeutet, dass der wahre Effekt sehr sicher nahe bei der Effektschätzung liegt.
- Eine moderate Vertrauenswürdigkeit der Evidenz bedeutet, dass der wahre Effekt wahrscheinlich nahe bei der Effektschätzung liegt, aber die Möglichkeit besteht, dass er relevant verschieden ist.
- Eine niedrige Vertrauenswürdigkeit der Evidenz bedeutet, dass der wahre Effekt durchaus relevant verschieden zur Effektschätzung sein kann.
- Eine sehr niedrige Vertrauenswürdigkeit der Evidenz bedeutet, dass der wahre Effekt durchaus relevant sehr verschieden von der Effektschätzung sein kann.

Die Bewertung erfolgt durch 1 Person und wird von einer 2. Person überprüft. Diskrepanzen werden durch Diskussion zwischen den beiden aufgelöst.

In der Regel wird Ergebnissen aus mindestens 2 RCTs im 1. Bewertungsschritt eine hohe, Ergebnissen aus 1 RCT wird aufgrund einer fehlenden Replikation eine moderate und Ergebnissen aus nicht randomisierten vergleichenden Studien eine niedrige Vertrauenswürdigkeit der Evidenz attestiert (angelehnt an Balshem et al. 2011 [19]). Von dieser Einschätzung ausgehend kann die Vertrauenswürdigkeit der Evidenz ab- oder aufgewertet werden. Die Faktoren Studienlimitationen, Inkonsistenz, Indirektheit oder fehlende Genauigkeit der Effekte können mit „nicht schwerwiegend“, „schwerwiegend“ oder „sehr schwerwiegend“ bewertet werden. Der Faktor Reporting Bias kann mit „keiner entdeckt“ oder

„anzunehmen“ bewertet werden. Die Gründe für eine Bewertung mit „anzunehmen“ bzw. „schwerwiegend“ oder „sehr schwerwiegend“ werden durch Fußnoten in den Evidenzprofilen erläutert. Je nach Einschätzung der Faktoren kann die übergreifende Vertrauenswürdigkeit der Evidenz um bis zu 3 Stufen abgewertet werden. Bei großen Effekten, einer Dosis-Wirkungs-Beziehung oder, wenn die Berücksichtigung aller potenziellen Confounder zu einer Effekterhöhung / Verstärkung des beobachteten Ergebnisses führen würde, kann die Vertrauenswürdigkeit der Evidenz ggf. aufgewertet werden. Unterstützend für diese Bewertungsschritte können Sensitivitätsanalysen durchgeführt werden.

2.3.5.1 Abwertung der Vertrauenswürdigkeit der Evidenz

A: Studienlimitationen

Die Vertrauenswürdigkeit der Evidenz wird aufgrund von schwerwiegenden oder sehr schwerwiegenden Studienlimitationen in der Regel um 1 oder 2 Stufen abgewertet, wenn starke Limitierungen in 1 oder mehreren der in Abschnitt 2.3.2 genannten Kriterien des Verzerrungspotenzials das Vertrauen in die Effektschätzung beeinträchtigen. Die endpunktbezogene studienübergreifende Bewertung der Studienlimitationen erfolgt unter Berücksichtigung des Einflusses der einzelnen Studien auf die Effektschätzung für jeden Endpunkt [21].

B: Inkonsistente (heterogene) Effektschätzungen

Die Vertrauenswürdigkeit der Evidenz wird bei schwerwiegender oder sehr schwerwiegender Inkonsistenz (unerklärter Heterogenität) zwischen Studienergebnissen in der Regel um 1 oder 2 Stufen abgewertet. Die Einschätzung einer möglichen Heterogenität erfolgt anhand der in Abschnitt 2.3.3 beschriebenen Kriterien.

Falls für binäre Daten Metaanalysen für 2 Effektmaße durchgeführt wurden und 1 der beiden einen statistisch signifikanten Gruppenunterschied anzeigt, wird die Inkonsistenz anhand dieser Metaanalyse bewertet. Ist der Gruppenunterschied für beide Effektmaße nicht statistisch signifikant, wird die Bewertung in der Regel anhand des relativen Effektmaßes vorgenommen.

Bei stetigen Daten wird die Bewertung in der Regel anhand der Metaanalyse der Mittelwert-differenz vorgenommen. Falls die fehlende Genauigkeit anhand der standardisierten Mittelwertdifferenz bewertet wurde, wird die Inkonsistenz auch anhand dieser bewertet.

Eine Abwertung um 1 Stufe erfolgt in der Regel, wenn der Heterogenitätstest ein signifikantes Ergebnis liefert. Um 2 Stufen kann z. B. abgewertet werden, wenn aufgrund der Heterogenität der Studienergebnisse keine gepoolte Effektschätzung berechnet wird.

Sofern für einen definierten Endpunkt Ergebnisse nur aus 1 Studie vorliegen, kommt die Bewertung der Inkonsistenz für diesen Endpunkt nicht zur Anwendung.

C: Indirektheit

Die Vertrauenswürdigkeit der Evidenz wird bei schwerwiegender oder sehr schwerwiegender Indirektheit (eingeschränkter Übertragbarkeit) in der Regel um 1 oder 2 Stufen abgewertet. Indirektheit kann z. B. auf Abweichungen zwischen dem PICO und den eingeschlossenen Studienpopulationen, der durchgeführten Prüf- und / oder Vergleichsinterventionen oder der eingesetzten Operationalisierungen der Endpunkte der Studien basieren [22].

D: Reporting Bias

Die Vertrauenswürdigkeit der Evidenz wird in der Regel um 1 Stufe abgewertet, wenn Reporting Bias anzunehmen ist (angelehnt an Page et al. 2023 [23]).

Es besteht Grund zur Annahme eines Reporting Bias, wenn Kenntnisse über unpublizierte Daten vorliegen und es entweder unklar ist, ob diese zu einer substanziellen Änderung der Effektschätzung (z. B. eine Änderung, die einen vorhandenen statisch signifikanten Effekt infrage stellt) führen würden, oder sie (wahrscheinlich) zu einer substanziellen Änderung der Effektschätzung führen würden und damit verzerrt sein könnte. Zur Beurteilung werden z. B. der Anteil und die Informationen zur (möglichen) Lage der Effekte der unpublizierten Daten betrachtet. Dies wird mit der Lage der entsprechenden gemeinsamen Effektschätzung aus den verfügbaren und verwertbaren Daten verglichen. Aus der Gesamtheit der herangezogenen Informationen wird abgeschätzt, ob und inwieweit die gemeinsame Effektschätzung aus den verwertbaren Daten aufgrund der fehlenden Daten verzerrt sein könnte.

E: Fehlende Genauigkeit der Effektschätzung

Die Vertrauenswürdigkeit der Evidenz wird wegen schwerwiegender oder sehr schwerwiegender fehlender Genauigkeit der Effektschätzung in der Regel um 1 oder 2 Stufen abgewertet. Maßgeblich hierfür ist neben der Lage und Breite des 95 %-KI auch die Wahl des Metaanalysemodells (zufällige Effekte oder fester Effekt), da diese sich auf die Breite des KIs auswirken kann (siehe Abschnitt 2.3.3). Außerdem können sehr kleine Fallzahlen zu einer Abwertung wegen fehlender Genauigkeit führen [24]. Bei ausreichender Patienten- bzw. Teilnehmerzahl wird regelhaft nicht abgewertet, wenn das KI eines Effekts (relativ oder absolut im Fall binärer Daten bzw. MWD oder SMD im Fall stetiger Daten) einen signifikanten Unterschied anzeigt.

Andernfalls erfolgt regelhaft eine Abwertung wegen fehlender Präzision um 1 Stufe, falls das KI (des relativen Effekts bei binären Daten bzw. Hedges' g bei stetigen Daten) einen mittelgroßen Effekt und gleichzeitig auch den Nulleffekt überdeckt. Bei binären Daten wird unter einem mittelgroßen Effekt eine Halbierung oder eine Verdopplung (0,5 oder 2) des relativen Effekts verstanden. Bei stetigen Daten wird von einem mittelgroßen Effekt ausgegangen, wenn das KI für eine SMD die Grenze −0,5 oder 0,5 umfasst. Eine Abwertung

um 2 Stufen wird in Fällen in Betracht gezogen, in denen die zuvor genannten KI-Grenzen beide, d. h. z. B. im Fall binärer Daten sowohl 0,5 als auch 2, im KI enthalten sind.

2.3.5.2 Aufwertung der Vertrauenswürdigkeit der Evidenz („Andere Faktoren“)

Methodisch hochwertige Studien können durch die sogenannten „Anderen Faktoren“ bezüglich ihrer Vertrauenswürdigkeit der Evidenz aufgewertet werden, sofern insbesondere folgende Aspekte nicht dagegensprechen: relevante Verzerrungsrisiken (z. B. durch fehlende Adjustierung plausibler Confounder), statistisch nicht signifikante oder unpräzise Effektschätzungen (mit nicht hinreichend schmalen KIs) [25]. Eine Aufwertung sollte in der Regel gut abgewogen werden und nicht zur Anwendung kommen, wenn die Vertrauenswürdigkeit der Evidenz zuvor bereits abgewertet wurde.

A: Große Effektschätzungen

Die Aufwertung der Vertrauenswürdigkeit der Evidenz ist in der Regel um 1 Stufe bei mittelgroßen Effektschätzungen möglich, wenn beispielsweise das geschätzte relative Risiko bei binären Daten zwischen 2 und 5 bzw. zwischen 0,2 und 0,5 liegt. Gleiches gilt für stetige Daten, wenn die geschätzte SMD zwischen 0,5 und 0,8 bzw. −0,8 und −0,5 liegt. Eine Aufwertung um 1 weitere Stufe kann bei großen Effektschätzungen erfolgen, wenn beispielsweise das geschätzte relative Risiko bei binären Daten über 5 bzw. unter 0,2 oder bei stetigen Daten die geschätzte SMD über 0,8 bzw. unter −0,8 liegt.

B: Dosis-Wirkungs-Beziehung

Die Vertrauenswürdigkeit der Evidenz wird bei Nachweis einer Dosis-Wirkungs-Beziehung regelhaft aufgewertet [25].

C: Plausibles Confounding

Die Vertrauenswürdigkeit der Evidenz wird bei einer möglichen Verstärkung des beobachteten Effektes unter Berücksichtigung von potenziellen Confoundern oder Verzerrungsquellen regelhaft aufgewertet [25].

3 Literatur

1. Bundestag. Gesetz für eine bessere Versorgung durch Digitalisierung und Innovation (Digitale-Versorgung-Gesetz – DVG). Bundesgesetzblatt Teil 1 2019; (49): 2562-2584.

2. Schünemann H, Brożek J, Guyatt G et al. GRADE Handbook [online]. 2013 [Zugriff: 02.09.2024]. URL: https://gdt.gradepro.org/app/handbook/handbook.html.

3. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden [online]. URL: https://www.iqwig.de/ueber-uns/methoden/methodenpapier/.

4. ICH Expert Working Group. ICH harmonised tripartite guideline: structure and content of clinical study reports; E3 [online]. 1995 [Zugriff: 02.09.2024]. URL: https://database.ich.org/sites/default/files/E3_Guideline.pdf.

5. Moher D, Hopewell S, Schulz KF et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869. https://doi.org/10.1136/bmj.c869.

6. Des Jarlais DC, Lyles C, Crepaz N et al. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004; 94(3): 361-366. https://doi.org/10.2105/ajph.94.3.361.

7. Von Elm E, Altman DG, Egger M et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007; 147(8): 573-577. https://doi.org/10.7326/0003-4819-147-8-200710160-00010.

8. Sterne JA, Hernan MA, Reeves BC et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355: i4919. https://doi.org/10.1136/bmj.i4919.

9. Shea BJ, Reeves BC, Wells G et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017; 358: j4008. https://doi.org/10.1136/bmj.j4008.

10. Waffenschmidt S, Navarro-Ruan T, Hobson N et al. Development and validation of study filters for identifying controlled non-randomized studies in PubMed and Ovid MEDLINE. Res Synth Methods 2020; 11(5): 617-626. https://doi.org/10.1002/jrsm.1425.

11. Guyatt GH, Oxman AD, Santesso N et al. GRADE guidelines: 12. Preparing summary of findings tables—binary outcomes. J Clin Epidemiol 2013; 66(2): 158-172. https://doi.org/10.1016/j.jclinepi.2012.01.012.

12. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 1998; 17(8): 873-890. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8%3c873::aid-sim779%3e3.0.co;2-i.

13. Martín Andrés A, Silva Mato A. Choosing the optimal unconditioned test for comparing two independent proportions. Comput Stat Data Anal 1994; 17(5): 555-574. https://doi.org/10.1016/0167-9473(94)90148-1.

14. Sutton AJ, Abrams KR, Jones DR et al. Methods for meta-analysis in medical research. Chichester: Wiley; 2000.

15. Schulz A, Schürmann C, Skipka G et al. Performing Meta-analyses with Very Few Studies. In: Evangelou V, Veroniki AA (Ed). Meta-Research; Methods and Protocols. New York: Humana; 2022. S. 91-102.

16. Lilienthal J, Sturtz S, Schürmann C et al. Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies. Res Synth Methods 2024; 15(2): 275-287. https://doi.org/10.1002/jrsm.1685.

17. Veroniki AA, Jackson D, Bender R et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. Res Synth Methods 2019; 10(1): 23-43. https://doi.org/10.1002/jrsm.1319.

18. Hultcrantz M, Rind D, Akl EA et al. The GRADE Working Group clarifies the construct of certainty of evidence. J Clin Epidemiol 2017; 87: 4-13. https://doi.org/10.1016/j.jclinepi.2017.05.006.

19. Balshem H, Helfand M, Schunemann HJ et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011; 64(4): 401-406. https://doi.org/10.1016/j.jclinepi.2010.07.015.

20. Meerpohl JJ, Langer G, Perleth M et al. GRADE-Leitlinien: 3. Bewertung der Qualität der Evidenz (Vertrauen in die Effektschätzer). Z Evid Fortbild Qual Gesundhwes 2012; 106(6): 449-456. https://doi.org/10.1016/j.zefq.2012.06.013.

21. Guyatt GH, Oxman AD, Vist G et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011; 64(4): 407-415. https://doi.org/10.1016/j.jclinepi.2010.07.017.

22. Guyatt GH, Oxman AD, Kunz R et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J Clin Epidemiol 2011; 64(12): 1303-1310. https://doi.org/10.1016/j.jclinepi.2011.04.014.

23. Page MJ, Sterne JAC, Boutron I et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. BMJ 2023; 383: e076754. https://doi.org/10.1136/bmj-2023-076754.

24. Guyatt GH, Oxman AD, Kunz R et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. J Clin Epidemiol 2011; 64(12): 1283-1293. https://doi.org/10.1016/j.jclinepi.2011.01.012.

25. Guyatt GH, Oxman AD, Sultan S et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011; 64(12): 1311-1316. https://doi.org/10.1016/j.jclinepi.2011.06.004.