Evidenzberichte mit Fragestellungen zur diagnostischen Güte

Version: 1.0 Stand: 23.05.2024

Impressum

Herausgeber

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Thema

Evidenzberichte mit Fragestellungen zur diagnostischen Güte

Anschrift des Herausgebers

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

Tel.: +49 221 35685-0
Fax: +49 221 35685-1
E-Mail: 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
KI	Konfidenzintervall
NICE	National Institute for Health and Care Excellence
PICO	Population, Intervention, Comparison, Outcome
PICo	Population, Phenomena of Interest, Context, Other/Outcomes
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
ROC	Receiver Operating Characteristic
SÜ	systematische Übersicht

1 Hintergrund

Das vorliegende Dokument beschreibt die Methodik bei der Erstellung eines Evidenzberichts für die Bearbeitung von Fragestellungen, in denen ≥ 1 diagnostische Maßnahmen (Indextest[s]) hinsichtlich ihrer diagnostischen Güte mit einem Referenztest verglichen werden.

Die Methodik für die Bearbeitung von Fragestellungen zu Interventionen wie präventive Maßnahmen (darunter auch Screeningmaßnahmen) und therapeutische Maßnahmen, zu Diagnoseverfahren, die sich nicht auf die Aufbereitung der Evidenz zur diagnostischen Güte beschränken oder zur Bearbeitung von qualitativen Fragestellungen wird in anderen Dokumenten beschrieben.

Auf Basis des Digitale-Versorgung-Gesetzes (DVG) kann die Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e. V. (AWMF) dem Bundesministerium für Gesundheit (BMG) Leitlinien vorschlagen, deren Entwicklung oder Aktualisierung das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) gemäß 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)- 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 zur Unterstützung der Leitliniengruppe mit einer Evidenzrecherche. Zur Auftragsbearbeitung finden ein oder mehrere Kick-off-Treffen zwischen den Leitlinien-koordinierenden, ggf. einer Ansprechperson der AWMF und Ansprechpersonen des IQWiG statt, in denen die Fragestellungen finalisiert werden. Zu jeder PICO- bzw. PICo-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 PICO-Fragestellung wird durch die Darstellung von Ergebnissen zu Vierfeldertafel-Daten 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 konkretisiert [3]. Das PICO-Schema wird von der Leitliniengruppe zur entsprechenden Leitlinie festgelegt. Hierfür ist die Rolle des fraglichen Tests im diagnostischen Algorithmus zentral: das zu prüfende diagnostische Verfahren (Indextest) kann den Referenztest entweder (zumindest teilweise) ersetzen oder derart ergänzen, dass der neue Test diesem vor- oder nachgeschaltet oder mit diesem kombiniert wird [4].

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	Indextest(s)
E3	Referenztest
E4	Zielgrößen: auf die Beobachtungseinheit Person bezogene Vierfeldertafel-Daten zur Berechnung der diagnostischen Güte
E5	Studientyp: Querschnitts- und Kohortenstudie[a]
E6	Publikationssprache: regelhaft Deutsch oder Englisch
E7	Vollpublikation verfügbar[b]
E8	Publikationszeitraum (sofern möglich)
E9	Setting (sofern relevant)
Ausschlusskriterien: Projektspezifisch können in Ausnahmefällen Ausschlusskriterien formuliert werden.	
a. Zentrale Aspekte der Ergebnissicherheit bei Studien zur diagnostischen Güte sind der konsekutive Einschluss der Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer, die Dokumentation fehlender Werte sowie die prospektive Festlegung des Trennkriteriums (z. B ein Grenzwert). Bei ausreichender Anzahl und Qualität werden nur die Studien eingeschlossen, die diese Voraussetzungen erfüllen. b. Als Vollpublikation gilt in diesem Zusammenhang auch ein Bericht über die Studie, der den Kriterien des STARD- [5] oder STROBE-Statements [6] genügt und eine Bewertung der Studie ermöglicht, sofern die in diesen Dokumenten enthaltenen Informationen zur Studienmethodik und zu den Studienergebnissen nicht vertraulich sind. STARD: Standards for the Reporting of Diagnostic Accuracy; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology	

Einschluss von Studien, die die vorgenannten Kriterien nicht vollständig erfüllen

Für die Einschlusskriterien E1 (Population), E2 (Indextest[s]) und E3 (Referenztest) reicht es aus, wenn bei ca. 80 % der eingeschlossenen Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer diese Kriterien erfüllt sind. Liegen für solche Studien Subgruppenanalysen für Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer vor, die die Einschlusskriterien erfüllen, wird auf diese Analysen zurückgegriffen. Studien, bei denen die Einschlusskriterien E1, E2 und E3 bei weniger als ca. 80 % erfüllt sind, können ggf. für die Evidenzdarstellung unberücksichtigt bleiben.

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 (im Folgenden: Basis-SÜ). Zur Überprüfung der Eignung als Basis-SÜ erfolgt eine Bewertung der Qualität der Informationsbeschaffung dieser SÜ(s) mit den entsprechenden Items aus AMSTAR 2 (A Measurement Tool to Assess Systematic Reviews 2) [7]. 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 Übersicht 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 Studienergebnisse)
  - 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 Publikationsbias

Folgende Informationsquelle wird berücksichtigt:

- U.S. National Institutes of Health. ClinicalTrials.gov (Studienstatus: “completed “ und “terminated”)

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 E8) 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 E7) [8]. Außerdem enthalten die Suchstrategien Limitierungen auf deutsch- und englischsprachige Publikationen (siehe Einschlusskriterium E6) [3] 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
- Referenzlisten identifizierter SÜs

Die identifizierten Studien werden dann 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(s) mit AMSTAR 2 [7]. 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 kontrolliert.

Die Ergebnisse zu den von der Leitliniengruppe festgelegten und in den Studien bzw. Basis-SÜ(s) berichteten Vierfeldertafel-Daten werden im Bericht in Forest Plots dargestellt, sofern mindestens 2 Studien vorliegen. Als Zielgrößen für die diagnostische Güte werden primär Sensitivität und Spezifität herangezogen. Mithilfe der Gesamtschätzung von Sensitivität und Spezifität werden unter Annahme verschiedener Prävalenz-Szenarien Schätzungen für die Einträge einer Vierfeldertafel basierend auf 1000 untersuchten Personen angegeben. Zur Abschätzung der quantitativen Ergebnisunsicherheit werden ergänzend die oberen und unteren Konfidenzintervall(KI)-Grenzen der Vierfeldertafel-Einträge basierend auf den KI-Grenzen der Gesamtschätzungen von Sensitivität und Spezifität und der angenommenen Prävalenz berechnet.

Das Vorgehen zur Bewertung des Einflusses des Verzerrungspotenzials auf die berichtsrelevanten Ergebnisse wird in Abschnitt 2.3.2 zielgrößenspezifisch 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 zielgrößenspezifisch in Evidenzprofilen zusammengeführt und dargestellt (siehe Abschnitt 2.3.4) [9]. Wenn möglich werden über die Darstellung der Ergebnisse der Einzelstudien hinaus die im Abschnitt 2.3.3 beschriebenen Verfahren durchgeführt.

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.

2.3.2 Kriterien des Verzerrungspotenzials

Das Verzerrungspotenzial wird zielgrößenspezifisch pro Studie insbesondere anhand der nachfolgend aufgeführten Quality-Assessment-of-Diagnostic-Accuracy-Studies(QUADAS)-2-Kriterien [10,11] beurteilt. Dazu erfolgt jeweils eine Bewertung mit „niedrig“, „unklar“ oder „hoch“. Voraussetzungen für eine Bewertung mit „niedrig“ werden im Folgenden erläutert. Eine Bewertung mit „unklar“ erfolgt grundsätzlich dann, wenn keine bzw. keine ausreichenden Angaben zur Bewertung zur Verfügung stehen. Die Bewertung erfolgt durch 1 Person und wird von einer 2. Person auf Grundlage der Studien kontrolliert.

Folgende Domänen werden bewertet:

- Selektion der Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer

  Falls aus der Beschreibung der Studie hervorgeht, dass eine konsekutive oder zufällige Stichprobe von Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmern eingeschlossen wurde und inadäquate Studienausschlüsse vermieden wurden, erfolgt eine Bewertung mit „niedrig“.

- Durchführung und Auswertung des Indextests

  Falls aus der Beschreibung der Studie hervorgeht, dass der Indextest ohne Wissen über die Ergebnisse des Referenztests ausgewertet wurde und das verwendete Trennkriterium prospektiv festgelegt worden war. Sofern mehr als ein Indextest untersucht wurde, müssen diese wechselseitig ohne Wissen über das jeweils andere Testergebnis erhoben worden sein um mit „niedrig“ bewertet zu werden.

- Auswahl, Durchführung und Auswertung des Referenztests

  Falls aus der Beschreibung der Studie hervorgeht, dass der Referenztest die Zielerkrankung korrekt klassifiziert und ohne Wissen über die Ergebnisse des Indextests ausgewertet wurde, erfolgt eine Bewertung mit „niedrig“.

- Patienten- bzw. Teilnehmerfluss und zeitlicher Ablauf

  Falls aus der Beschreibung der Studie hervorgeht, dass

- die Zeitspanne zwischen Indextest und Referenztest adäquat war,
- alle Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer mittels Referenztest untersucht wurden,
- alle Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer denselben oder einen gleichwertigen Referenztest erhalten haben und
- alle Patientinnen und Patienten bzw. Teilnehmerinnen und Teilnehmer in die Analyse einbezogen wurden, erfolgt eine Bewertung mit „niedrig.

Pro Zielgröße wird für jede Studie das Verzerrungspotenzial anhand der oben genannten Kriterien bewertet. Dieses Verzerrungspotenzial wird in einer studienübergreifenden Bewertung der Studienlimitationen für die Evidenzprofile zusammengefasst (siehe Abschnitt 2.3.4.1).

2.3.3 Metaanalysen

Liegen mehrere Studien zu einer Zielgröße vor, werden die Ergebnisse nach Möglichkeit mithilfe von 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 2.3.4.1).

Die Punktschätzungen und dazugehörigen univariaten 95 %-KIs [12] aus den Studien werden in 2 nebeneinander platzierten Forest Plots für die Sensitivität und die Spezifität gemeinsam mit den Einträgen der Vierfeldertafeln der Einzelstudien dargestellt. Außerdem werden, sofern die dafür nötigen Anforderungen erfüllt sind, für die Maße der diagnostischen Güte uni- oder bivariate Metaanalysen durchgeführt [13]. Die Schätzung der Sensitivität und der Spezifität erfolgt über ein generalisiertes lineares gemischtes Modell [14,15].

Die bivariaten Metaanalysen werden in Anlehnung an den Programm-Code der Cochrane Diagnostic Test Accuracy Working Group [16] mithilfe der Statistik-Software R erstellt [17]. Im Fall von weniger als 5 Studien erfolgt die Berechnung der KIs von Sensitivität und Spezifität mithilfe der Normalverteilungsannahme entsprechend des Cochrane-Programms. Bei 5 Studien oder mehr wird statt der Normalverteilung die t-Verteilung zur Berechnung der KIs verwendet. Der Algorithmus zum Schätzen der Parameter im bivariaten Modell kann zu unpräzisen Schätzungen führen, das heißt zu Schätzungen mit großen Standardfehlern. Auch kann der Algorithmus ggf. keine Schätzungen liefern, wenn das Maximum-Likelihood-Verfahren nicht konvergiert. In beiden Fällen fehlen brauchbare Schätzungen. Die Gründe hierfür können beispielsweise sein, dass zu wenige Studien vorliegen oder dass einzelne Studien extreme Werte aufweisen. Sind die resultierenden Schätzungen unpräzise, werden die Ergebnisse der bivariaten Metaanalysen in der Regel nicht dargestellt. In diesem Fall wird auf univariate Metaanalysen für Sensitivität und Spezifität auf Basis von generalisierten

gemischten Modellen zurückgegriffen. Von der Berechnung einer gepoolten Schätzung wird abgesehen, falls sich die 95 %-KIs der eingehenden Studien nur gering oder gar nicht überlappen und gleichzeitig sich die Schätzungen der Studien zu stark unterscheiden.

2.3.4 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 die jeweiligen diagnostischen Eigenschaften wird eine studienübergreifende Aussage zur Vertrauenswürdigkeit der Evidenz bezüglich des jeweiligen Ausmaßes des Vertrauens in die Schätzung der diagnostischen Güte 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 Wert sehr sicher nahe bei der Schätzung liegt.
- Eine moderate Vertrauenswürdigkeit der Evidenz bedeutet, dass der wahre Wert wahrscheinlich nahe bei der Schätzung liegt, aber die Möglichkeit besteht, dass er relevant verschieden ist.
- Eine niedrige Vertrauenswürdigkeit der Evidenz bedeutet, dass der wahre Wert durchaus relevant verschieden zur Schätzung sein kann.
- Eine sehr niedrige Vertrauenswürdigkeit der Evidenz bedeutet, dass der wahre Wert durchaus relevant sehr verschieden von der Schätzung sein kann.

Die Bewertung erfolgt durch 2 Personen unabhängig voneinander. Diskrepanzen werden durch Diskussion zwischen den beiden aufgelöst.

Ergebnissen aus Querschnitts- oder Kohortenstudien zur diagnostischen Güte mit einem adäquaten Referenztest wird im 1. Bewertungsschritt eine hohe Vertrauenswürdigkeit der Evidenz attestiert [21]. Von dieser rein designbedingten Einschätzung ausgehend kann die Vertrauenswürdigkeit der Evidenz abgewertet werden. Die Faktoren Studienlimitationen, Inkonsistenz, Indirektheit oder fehlende Genauigkeit der Zielgrößen können mit „nicht schwerwiegend“, „schwerwiegend“ oder „sehr schwerwiegend“ bewertet werden. Der Faktor Publikationsbias 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. Im Einzelfall kann ggf. eine mögliche Aufwertung erwogen werden (siehe Abschnitt 2.3.4.2).

2.3.4.1 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 einem oder mehreren der in Abschnitt 2.3.2 genannten Kriterien des Verzerrungspotenzials das Vertrauen in die Schätzung der diagnostischen Güte beeinträchtigen. Die studienübergreifende Bewertung der Studienlimitationen erfolgt unter Berücksichtigung des Einflusses der einzelnen Studien auf die Schätzung der diagnostischen Güte [21].

B: Inkonsistente (heterogene) Schätzer

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 unter Berücksichtigung der Ähnlichkeit der Punktschätzungen und der Überlappung der 95 %-KIs [22].

Sofern für eine definierte Zielgröße Ergebnisse nur aus 1 Studie vorliegen, kommt die Bewertung der Inkonsistenz für diese Zielgröße 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 den Einschlusskriterien E1 (Population), E2 (Indextest[s]) und E3 (Referenztest) in Tabelle 1 und den diesbezüglichen Einschlusskriterien der Studien basieren [21].

D: Publikationsbias

Die Vertrauenswürdigkeit der Evidenz wird in der Regel um 1 Stufe abgewertet, wenn ein Publikationsbias anzunehmen ist.

Kriterien für die Annahme eines Publikationsbias umfassen eine Evidenzgrundlage, welche vorrangig auf kleinen Studien und wenigen Ereignissen beruht, welche einen beobachteten Zusammenhang zwischen der Größe der Schätzung der diagnostischen Güte und der Studiengröße (oder Genauigkeit der Schätzung der diagnostischen Güte) zeigt oder welche Auffälligkeiten in der gesichteten Datenlage erkennen lässt [22].

E: Fehlende Genauigkeit der Zielgrößen

Die Vertrauenswürdigkeit der Evidenz wird wegen schwerwiegender oder sehr schwerwiegender fehlender Genauigkeit der Schätzung in der Regel um 1 oder 2 Stufen

abgewertet. Maßgeblich hierfür ist die Lage und Breite des 95 %-KI. Zur Bewertung wird mittels zweier Entscheidungsgrenzen (empfehlenswert bzw. unbrauchbar für die klinische Praxis) der Wertebereich der Sensitivität bzw. Spezifität in 3 Sektoren aufgeteilt. Wegen fehlender Genauigkeit wird um 1 Stufe abgewertet, wenn das 95 %-KI der (gepoolten) Schätzung für die Sensitivität bzw. die Spezifität über mehr als 1 Sektor erstreckt (2 Sektoren: Abwertung um 1 Stufe; 3 Sektoren: Abwertung um 2 Stufen). Außerdem können sehr kleine Personenzahlen zu einer Abwertung wegen fehlender Genauigkeit führen [23].

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

In der Literatur gibt es bislang keine voll entwickelte Methodik zur Aufwertung der Vertrauenswürdigkeit der Evidenz bei Studien zur diagnostischen Güte und es wird auf die Notwendigkeit einer weiteren theoretischen und empirischen Erarbeitung von Kriterien hingewiesen. Diskutiert werden z. B. ein starker, konsistenter Zusammenhang von Sensitivität und Spezifität in der Receiver-Operating-Characteristic(ROC)-Kurve oder sehr hohe Schätzungen für die diagnostische Güte, die nicht durch übrige Faktoren infrage gestellt werden [21,22].

Die genannten Faktoren werden, soweit möglich, betrachtet und eine mögliche Aufwertung im Einzelfall erwogen.

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, Oxman A. GRADE Handbook [online]. 2013 [Zugriff: 06.10.2023]. 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. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332(7549): 1089-1092. https://doi.org/10.1136/bmj.332.7549.1089.

5. Bossuyt PM, Reitsma JB, Bruns DE et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138(1): W1-12. https://doi.org/10.7326/0003-4819-138-1-200301070-00012-w1.

6. 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.

7. 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.

8. 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.

9. 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.

10. Whiting PF, Rutjes AW, Westwood ME et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155(8): 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009.

11. Whiting PF, Rutjes AW, Westwood ME et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013; 66(10): 1093-1104. https://doi.org/10.1016/j.jclinepi.2013.05.014.

12. Leemis LM, Trivedi KS. A Comparison of Approximate Interval Estimators for the Bernoulli Parameter. Am Statistn 1996; 50(1): 63-68. https://doi.org/10.2307/2685046.

13. Reitsma JB, Glas AS, Rutjes AW et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005; 58(10): 982-990. https://doi.org/10.1016/j.jclinepi.2005.02.022.

14. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol 2006; 59(12): 1331-1332; author reply 1332-1333. https://doi.org/10.1016/j.jclinepi.2006.06.011.

15. Menke J. Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. Methods Inf Med 2010; 49(1): 54-62, 62-54. https://doi.org/10.3414/ME09-01-0001.

16. Partlett C, Takwoingi Y. Meta-analysis of test accuracy studies in R: A summary of user-written programs and step-by-step guide to using glmer; Version 2.0 [online]. 2021 [Zugriff: 06.10.2023]. URL: https://methods.cochrane.org/sdt/sites/methods.cochrane.org.sdt/files/public/uploads/r_dta_meta-analysis_v2.0.zip.

17. R Core Team. R: A Language and Environment for Statistical Computing [online]. 2022 [Zugriff: 06.10.2023]. URL: https://www.R-project.org/.

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. Schünemann HJ, Mustafa RA, Brozek J et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol 2020; 122: 129-141. https://doi.org/10.1016/j.jclinepi.2019.12.020.

22. Schünemann HJ, Mustafa RA, Brozek J et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol 2020; 122: 142-152. https://doi.org/10.1016/j.jclinepi.2019.12.021.

23. 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.