Evaluation diagnostischer Technologien

Hintergrund, Probleme, Methoden

Ludwig Boltzmann Institut
Health Technology Assessment

HTA-Projektbericht Nr.: 036
ISSN 1992-0488
ISSN online 1992-0496

Evaluation diagnostischer Technologien

Hintergrund, Probleme, Methoden

Ludwig Boltzmann Institut
Health Technology Assessment

Wien, August 2010

Projektleitung & Projektbearbeitung
Dr. med. Anna Nachtnebel, MSc

Projektbeteiligung

Systematische Literatursuche:	Tarquin Mittermayr, BA Hons
Externe Begutachtung:	Dr. med. Dagmar Lühmann Dr. Petra Schnell-Inderst, MPH
Interne Begutachtung:	Dr. rer. soc. oec. Ingrid Zechmeister, MA

Dieser Bericht soll folgendermaßen zitiert werden

Nachtnebel, A. Evaluation diagnostischer Technologien - Hintergrund, Probleme, Methoden. HTA-Projektbericht 2010; Nummer 36.

IMPRESSUM

Medieninhaber und Herausgeber:
Ludwig Boltzmann Gesellschaft GmbH
Nußdorferstr. 64, 6 Stock, A-1090 Wien
http://www.lbg.ac.at/de/lbg/impressum

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

Die HTA-Projektberichte erscheinen in geringer Auflage im Druck und werden über den Dokumentenserver „http://eprints.hta.lbg.ac.at" der Öffentlichkeit zur Verfügung gestellt:

HTA-Projektbericht Nr.: 036

ISSN 1992-0488
ISSN online 1992-0496

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

AUC = Area under curve

DOR = diagnostische Odds Ratio

EUnetHTA = European network for Health Technology Assessment

IQWiG = Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

LR = Likelihood Ratio

MRT = Magnetresonanztomographie

MSAC = Medical Services Advisory Committee

NHS = National Health Service

NICE = National Institute for Health and Clinical Excellence

NPV = negativer Vorhersagewert

PV = Vorhersagewert

PPV = positiver Vorhersagewert

QUADAS = Quality assessment tool for diagnostic accuracy studies

QUALY = qualitätskorrigiertes Lebensjahr

RCT = randomisiert kontrollierte Studien

ROC – Kurve = Receiver-Operating Characteristics Kurve

STARD = Standards for Reporting of Diagnostic Accuracy

Zusammenfassung

Hintergrund

keine einheitlichen Kriterien zur Bewertung von diagnostischen Verfahren

Auch aufgrund von steigenden Ausgaben im Gesundheitssektor kommt der evidenzbasierten Bewertung von medizinischen Technologien eine immer größere Bedeutung zu, um eine effiziente Allokation limitierter finanzieller Ressourcen zu gewährleisten. Während die Bewertung von Arzneimitteln vor Kostenerstattungsentscheidungen durch Versicherungsträger bereits Routine, und das damit verbundene methodische Vorgehen weitgehend standardisiert ist, stellt sich die Situation bei diagnostischen Maßnahmen anders dar.

dadurch häufig unkontrollierte Aufnahme in klinischen Alltag

Durch die Abwesenheit einheitlich formulierter Kriterien, die bei der Bewertung von diagnostischen Maßnahmen zu beachten sind, sowie durch das Fehlen von regulatorischen Standards, die vor der Marktzulassung zu erfüllen sind, wurden diese Verfahren häufig unkontrolliert in die klinische Praxis übernommen.

Evaluation aber wichtig, da Tests mit Kosten und Nebenwirkungen verbunden sind

Da mit dem Einsatz von „unwirksamen" Tests nicht nur beträchtliche Kosten assoziiert sind, sondern auch teilweise schwere Nebenwirkungen einhergehen können, stellt sich nun die Frage, wie diagnostische Verfahren aus einer entscheidungsträgerrelevanten Sichtweise heraus evaluiert werden können.

Zielsetzungen und Fragestellungen

Ziel: Problemstellungen und Methoden ausgewählter Institutionen bei der Evaluation darzustellen

Ziel dieses Berichts ist es daher

1. Eine Übersicht über spezifische Problemstellungen, die mit der Evaluierung von diagnostischen Technologien verbunden sind, zu geben.
2. Methoden ausgewählter Institutionen darzustellen, die bei der Nutzenbewertung von diagnostischen Technologien zur Anwendung gelangen.
3. Darauf aufbauend eine methodische Vorgehensweise zur Nutzenbewertung von diagnostischen Technologien in einem entscheidungsträgerrelevanten Kontext abzuleiten.

Relevante Fragestellungen umfassen:

Fragestellungen:

Ebenen der Nutzenbewertung?

Parameter der diagnostischen Genauigkeit?

Problemstellungen?

Methoden?

a. anhand welcher Ebenen kann der Nutzen von diagnostischen Verfahren bewertet werden kann?

b. welche Parameter können zur Bewertung der diagnostischen Genauigkeit herangezogen werden?

c. welche methodischen Probleme sind mit der Evidenz-basierten Nutzenbewertung von diagnostischen Verfahren verbunden?

d. welche Methoden werden von ausgewählten Institutionen angewandt; welche allgemeinen Vorgehensweisen lassen sich daraus für entscheidungsträgerrelevante Bewertungen ableiten?

Methodik

unsystematische Suche, Methodenhandbücher

Die Fragestellungen wurden im Rahmen einer narrativen Review bearbeitet, deren Grundlage eine unsystematische Handsuche, ergänzt durch eine systematische Suche in mehreren Datenbanken (Cochrane Library, EMBASE, Ovid MEDLINE und der Datenbank des Centre for Reviews and Dissemination), bildet. Zusätzlich wurden Methodenhandbücher ausgewählter Institutionen, die sich mit der Evaluation von diagnostischen Verfahren befassen, (MSAC, IQWiG, NICE, EUnetHTA) herangezogen.

Darstellung anhand von Evidenzmodell…

Die Beschreibung der Vorgehensweisen der genannten Institutionen als auch die Übersicht der Problemstellungen, die mit der Evaluation von diagnostischen Verfahren verbunden sind, wurden anhand eines von Fryback und Thornbury entwickelten Evidenzmodells dargestellt.

Ergebnisse

…welches 6 Level unterscheidet

Laut des von Fryback und Thornbury entwickelten Evidenzmodells können diagnostische Verfahren anhand von sechs Ebenen beurteilt werden. An unterster Stelle ist dabei die technische Qualität (Level 1) zu nennen und dann in aufsteigender Reihenfolge diagnostische Genauigkeit (Level 2), der diagnostische (Level 3) und therapeutische Impakt (Level 4) eines Testergebnisses, der patientenrelevante Nutzen (Level 5) und an letzter Stelle der Nutzen aus gesellschaftlicher Sicht (Level 6). Für EnscheidungsträgerInnen letztlich entscheidend ist aber der patientenrelevante Nutzen, als auch die mit einem Test vergesellschaftete Kosten.

häufigste Studien zu technischer Qualität und diagnostischer Genauigkeit, für Entscheidungsträger-Innen ist aber patientenrelevanter Nutzen und Kosten wichtig

Da die meisten Studien über diagnostische Verfahren allerdings Studien zu Level 1 und Level 2 sind, kann der für PatientInnen resultierende Nutzen nur in den seltensten Fällen direkt erhoben werden. Um dennoch Aussagen bezüglich des Nutzens zu treffen, können mittels „linked Evidence“ Ergebnisse von Studien zu diagnostischer Genauigkeit mit den Ergebnissen von Wirksamkeitsstudien verknüpft werden.

diagnostische Genauigkeitsstudien vergleichen Indextest mit Referenztest

Bei diagnostischen Genauigkeitsstudien wird das zu evaluierende diagnostische Verfahren (=Indextest) mit dem derzeit besten diagnostischen Verfahren (= Referenzstandard) verglichen und so Kenngrößen der diagnostischen Genauigkeit wie etwa Sensitivität und Spezifität oder Likelihood Ratios berechnet. Da mit diesen Studiendesigns zahlreiche methodische Probleme vergesellschaftet sein können, die einerseits zu einer Verzerrung der Resultate führen und andererseits die Übertragbarkeit der Ergebnisse auf den klinischen Alltag und auf andere klinische Settings kompromittieren können, sind spezielle Instrumente, wie etwa das QUADAS- Tool, zur Bewertung von diagnostischen Genauigkeitsstudien nötig.

„linked Evidence“: indirekt Aussage zu patientenrelevanten Nutzen durch Verknüpfung von diagnostischer Genauigkeitsstudie und Wirksamkeitsstudie

Um diagnostische Genauigkeitsstudien mit Ergebnissen von Wirksamkeitsstudien im Rahmen von „linked Evidence“ zu verknüpfen, müssen etliche Voraussetzungen erfüllt sein. Bedingungen sind etwa, dass die Populationen von diagnostischer Genauigkeitsstudie und Wirksamkeitsstudie miteinander vergleichbar sind und dass ein valider Referenzstandard existiert.

etliche Voraussetzungen müssen erfüllt sein

Bedingt durch diese Voraussetzungen kann „linked Evidence“ in der Praxis häufig nicht angewandt werden und soll daher, auch aufgrund von noch ungeklärten methodischen Fragen, nur nach eingehender Erörterung und Rechtfertigung, ob alle Kriterien auch tatsächlich erfüllt sind, angewandt werden.

Methoden der Institutionen ähnlich: Bewertung durch systematische Reviews

Die methodischen Vorgehensweisen, der in diesem Bericht untersuchten Institutionen, gleichen sich trotz einiger Unterschiede über weite Strecken. Als relevante Zielgröße für die Bewertung von diagnostischen Verfahren wird von allen Institutionen, der aus einem diagnostischen Verfahren resultierende Patientennutzen genannt, wobei dieser im Rahmen von systematischen Reviews erhoben wird. Auch „linked Evidence“ wird dabei von drei der Organisationen als Möglichkeit genannt, den damit vergesellschafteten Nutzen zu etablieren.

Fragenkatalog als Hilfestellung für EntscheidungsträgerInnen bei Bewertung von Tests

Basierend auf den Gemeinsamkeiten der Institutionen und der dargestellten spezifischen Herausforderungen, die mit der Evaluation von diagnostischen Verfahren einhergehen, wurde ein Fragenkatalog erstellt, der EntscheidungsträgerInnen eine Hilfestellung für die Bewertung von diagnostischen Verfahren bieten soll.

Summary

Background

no standardised methods for evaluation of diagnostics

Due to rising health care expenditures, the importance of evidence-based evaluations of medical technologies has increased in order to guarantee an efficient allocation of scarce financial resources.

uncontrolled introduction into clinical practice

Despite standardised methods for the evaluation of drugs prior to reimbursement decisions by social health insurers, commonly accepted methodological approaches for diagnostic technologies are missing. The absence of consistently defined criteria, according to which a test is assessed and the lack of regulatory standards for market authorization have repeatedly led to an uncontrolled introduction of diagnostics into clinical practice.

ineffective tests are waste of financial resources, associated with risks

Because "ineffective" tests not only represent a waste of financial resources but also potentially expose patients to unnecessary risks, the question posed is how to effectively evaluate diagnostic technologies from a decision-makers perspective.

Objective

objective: to identify methodological challenges and to formulate recommendations for evaluating tests

Objectives of this report were:

1. To summarize specific problems associated with the evaluation of diagnostic technologies.
2. To describe methods of selected institutions for the evaluation of diagnostics.
3. Based on these findings, to formulate recommendations for the evaluation of tests from a decision-maker's perspective.

Research questions were:

a. In what levels can diagnostics be evaluated?
b. What parameters are used for diagnostic accuracy?
c. What are the methodological challenges associated with the evidence-based evaluation of diagnostic tests?
d. What methods are used by selected institutions for the evaluation of diagnostic tests; can generally valid approaches for decision-makers be inferred from these findings?

Methods

unsystematic search, guidelines of selected institutions

The research questions were answered by performing a narrative review based on the methodological guidelines of selected institutions (MSAC, IQWiG, NICE, EUnetHTA), and an unsystematic hand-search which was complemented by a systematic literature search in various databases (Cochrane Library, EMBASE, Ovid MEDLINE, Database of the Centre for Reviews and Dissemination).

synthesis of findings based on evidence hierarchy model…

The description of the methodological approaches of the selected institutions, as well as the synthesis of the identified methodological challenges associated with the evaluation of diagnostics, followed an evidence hierarchy model developed by Fryback and Thornbury

Results

…which distinguishes 6 levels

According to the model developed by Fryback and Thornbury, diagnostic technologies can be evaluated according to six levels. Level 1 describes the technical efficacy, level 2 the diagnostic accuracy, level 3 and level 4 correspond to diagnostic and to therapeutic efficacy, level 5 equates to patient outcome efficacy, whereas the last level, that is level 6, defines the societal efficacy. Relevant for decision-makers, however, are the benefits for patients and the costs associated with a diagnostic test.

most of available literature on technical efficacy and diagnostic accuracy

Because most of the published literature on diagnostic tests consists of studies targeting only level 1 or level 2, it is often difficult to establish patient benefits directly. However, linked evidence offers a means to infer patient relevant outcomes indirectly by linking studies on diagnostic accuracy with studies on treatment efficacy.

diagnostic accuracy studies compare indextest with reference standard

Studies on diagnostic accuracy studies compare the new test (=indextest) with the best available diagnostic test (=reference standard) and therefore allow the calculation of parameters for diagnostic accuracy, such as sensitivity, specificity or likelihood ratios. Given that these studies can suffer from peculiar methodological flaws, opening up the possibility of bias or limited generalizability to clinical practice or other settings, specific tools, such as the QUADAS checklist, are required to assess the quality of these studies.

linked evidence: to infer patient benefits indirectly by linking diagnostic accuracy studies with treatment effectiveness

In order to link studies on diagnostic accuracy with studies on treatment effectiveness, certain prerequisites have to be met for "linked evidence". Requirements include that the population of both, the diagnostic accuracy study and the treatment effectiveness study, are comparable or that a validated reference standard exists. Due to these requirements linked evidence is often not an option to assess the effectiveness of diagnostic tests and even if so, its use must be reasonably justified.

methods of selected institutions similar

Regarding the institutions included in this report, many similarities of the methods for the evaluation of diagnostics exist: for example, all agencies demand evidence at least at the level of patient outcome efficacy. Moreover, linked evidence is mentioned in the majority of guidelines and the preferred method for the evaluation of diagnostics is a systematic review.

checklist for decision-makers to facilitate assessment of diagnostic tests

Finally, a checklist was composed as an aid for decision-makers to assess diagnostic tests, based on these similarities and on the specific challenges associated with the evaluation of diagnostic technologies.

1 Hintergrund

Evaluation von Nutzen, Risiken und Kosten einer Technologie erlaubt effiziente Allokation von Ressourcen

Vor allem steigende Kosten im Gesundheitswesen haben dazu geführt, dass wissenschaftliche Bewertungen neuer, oder bereits in Verwendung befindlicher medizinischer Technologien zunehmend Voraussetzungen für Entscheidungen über Kostenerstattung bilden. Durch die Evaluierung von Nutzen, Risiken, aber auch der mit einer Technologie vergesellschafteten Kosten, soll es EntscheidungsträgerInnen ermöglicht werden, limitierte finanzielle Ressourcen nur für (kosten-)effektive Verfahren bereitzustellen.

kein standardisiertes Vorgehen bei Evaluation von Diagnostika und fehlende gesetzliche Vorgaben für Marktzulassung

Während die Bewertung von Arzneimitteln vor Kostenerstattungsentscheidungen durch Versicherungsträger bereits Routine, und das damit verbundene methodische Vorgehen weitgehend standardisiert ist, stellt sich die Situation bei diagnostischen Maßnahmen anders dar. Zusätzlich fehlen bei Diagnostika - im Gegensatz zu Arzneimitteln bei denen der Wirksamkeitsnachweis anhand klar definierter gesetzlicher Vorgaben erfolgen muss - regulatorische Standards, die vor der Marktzulassung zu erfüllen sind [1].

diagnostische Verfahren

Generell werden unter dem Begriff „diagnostische Verfahren“ zahlreiche Technologien subsumiert, die im Rahmen der medizinischen Versorgung mit unterschiedlichen Zielsetzungen eingesetzt werden können. Dazu gehören:

- Unsicherheiten in Bezug auf den Gesundheitsstatus (krank vs gesund) zu reduzieren.
- Informationen zu gewinnen, um Entscheidungen in Bezug auf das weitere therapeutische/diagnostische Vorgehen zu erleichtern.
- prognostische Informationen bezüglich des weiteren Krankheitsverlaufes zu gewinnen.
- den Krankheitsverlauf während oder nach einer Therapie zu überwachen [2].

um Unsicherheiten bezüglich Gesundheitsstatus zu reduzieren

Beeinflussung von diagnostischem/therapeutischem Vorgehen

Prognose

Überwachung des Krankheitsverlaufs

Ziel: Verbesserung patientenrelevanter Endpunkte

Das Ziel diagnostischer Untersuchung ist also, diagnostische und/oder therapeutische Unsicherheiten zu reduzieren, um Entscheidungen bezüglich des weiteren diagnostischen/therapeutischen Managements zu erleichtern, wodurch letztlich patientenrelevante Endpunkte verbessert werden sollen [3]. Auch in einem von Fryback und Thornbury [4, 5] zur Evaluation von Diagnoseverfahren entwickelten Modell, werden als höchste Evidenz jene Studien angesehen, die diagnostische Maßnahmen in Hinsicht auf patientenrelevante Endpunkte und auf gesellschaftliche Konsequenzen bewerten.

Evidenzmodell: gesellschaftliche Konsequenzen als höchste Stufe

aber meisten Studien befassen sich mit technischer Qualität und diagnostischer Genauigkeit

häufig unkontrollierte Aufnahme in klinischen Alltag

Trotz allem befasst sich der Großteil der wissenschaftlichen Evaluationsstudien zu diagnostischen Verfahren lediglich mit der Bewertung der Testgenauigkeit und der technischen Qualität [6]. Zusätzlich sind mit der Bewertung von Tests spezifische methodische Herausforderungen verbunden, wodurch die Beurteilung der Konsequenzen für PatientInnen in erster, und der Konsequenzen für die Gesellschaft in zweiter Instanz erheblich erschwert wird. Auch das Fehlen von einheitlich formulierten Kriterien anhand derer Tests bewertet werden können und fehlende regulatorische Bedingungen vor der Marktzulassung, haben dazu geführt, dass diagnostische Maßnahmen rasch und unkontrolliert in den klinischen Alltag aufgenommen wurden [1, 2].

„unwirksame" Test mit beträchtlichen Kosten und mit Nebenwirkungen verbunden

Da mit dem Einsatz von „unwirksamen" Tests nicht nur beträchtliche Kosten assoziiert sind, sondern auch teilweise schwere Nebenwirkungen einhergehen können, wobei direkt durch einen Test entstehende (z.B. Nebenwirkungen durch Strahlenexposition, Komplikationen durch Testverfahren selbst) von indirekten (z.B. invasive Nachfolgeuntersuchungen, Nebenwirkungen der Therapie) unterschieden werden [7, 8], stellt sich nun die Frage, wie diagnostische Verfahren aus einer entscheidungsträgerrelevanten Sichtweise heraus evaluiert werden können.

2 Ziel und Forschungsfrage

Ziel des vorliegenden Berichtes ist es:

Ziele

1. Eine Übersicht über spezifische Problemstellungen, die mit der Evaluierung von diagnostischen Technologien verbunden sind, zu geben.

2. Methoden ausgewählter Institutionen darzustellen, die bei der Nutzenbewertung von diagnostischen Technologien zur Anwendung gelangen.

3. Darauf aufbauend eine methodische Vorgehensweise zur Nutzenbewertung von diagnostischen Technologien in einem entscheidungsträgerrelevanten Kontext abzuleiten.

Fragestellungen

Aus den angeführten Zielsetzungen ergeben sich daher folgende Fragestellungen:

a. In welchen Ebenen kann der Nutzen von diagnostischen Verfahren bewertet werden (z.B. Testgenauigkeit, PatientInnenrelevanz, systemische Implikationen)?

b. Welche Parameter werden zur Bewertung der diagnostischen Genauigkeit herangezogen? (z.B. Sensitivität, Spezifität, Vorhersagewerte).

c. Welche methodischen Probleme sind mit der Evidenz-basierten Nutzenbewertung von diagnostischen Verfahren verbunden (z.B. Bewertung der Qualität diagnostischer Studien, Bewertung der diagnostischen Kette, fehlender Referenztest)?

d. Welche Methoden werden von ausgewählten Institutionen zur Nutzenbewertung von Diagnostika angewandt; welche allgemeinen Vorgehensweisen lassen sich daraus für entscheidungsträgerrelevante Bewertungen ableiten?

3 Methodik

Entsprechend der oben angeführten Fragestellungen wurden mehrere methodische Herangehensweisen gewählt.

Methodenpapiere ausgewählter Institutionen (NICE, MSAC, IQWiG, EUnetHTA)

Einerseits wurden Methodenpapiere von vier ausgewählten Institutionen, das sind das „National Institute for Health and Clinical Excellence“ (NICE), das „Medical Services Advisory Committee“ (MSAC), das „Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen“ (IQWiG) und das vom „European Network for Health Technology Assessment“ (EUnetHTA) veröffentlichte „Core Model“ als Grundlage für den Bericht herangezogen. Die Auswahl der vier Institutionen erfolgte aufgrund der Vergleichbarkeit zum österreichischen Gesundheitssystem (IQWiG), wegen der strikten methodischen Vorgehensweisen (NICE) und der Relevanz für den gesamteuropäischen Kontext (EUnetHTA). Zusätzlich wurde das bereits 2005 publizierte Manual von MSAC berücksichtigt, da es zu einem der ersten und ausführlichsten Methodenpapiere über die Evaluation von Diagnostika gehört und auch in die Vorgehensweisen von zwei der ausgewählten Institutionen (EUnetHTA, IQWiG) einfließt.

unsystematische Literatursuche

Aufgrund des Umfangs des Themas und der damit verbundenen zeitlichen und finanziellen Implikationen wurden die Fragestellungen a) bis c) (siehe Kapitel 2) in erster Linie durch eine unsystematische Suche bearbeitet. Grundlage bildeten in erster Linie relevante Publikationen aus den Referenzlisten der Methodenmanuale, aber auch durch eine Handsuche identifizierte Schlüsselpublikationen. Ergänzend wurde am 4. und 5. November 2009 eine systematische Literatursuche ohne Einschränkungen bezüglich Publikationsjahr oder Sprache in vier Datenbanken durchgeführt (Cochrane Library, EMBASE, Ovid MEDLINE, Datenbank des Centre for Reviews and Dissemination) (siehe Appendix 7). Die Auswahl der Artikel erfolgte unsystematisch, wobei nur generelle Einschlusskriterien definiert worden waren (englische oder deutsche Sprache, Methodenartikel zu diagnostischer Genauigkeit, evidenzbasierter Evaluation von Diagnostika, relevanten Studiendesigns zur Bewertung von Diagnostika (primäre und sekundäre Studiendesigns), Instrumente zur Bewertung der methodischen Qualität von diagnostischen Genauigkeitsstudien).

ergänzend systematische Suche mit unsystematischer Auswahl von relevanten Publikationen

Obwohl eine unsystematische Suche zur Identifikation relevanter Publikationen herangezogen wurde, scheint diese Methode für die Bearbeitung der in diesem Bericht formulierten Fragestellungen gerechtfertigt zu sein. Darüber hinaus ist es unwahrscheinlich, dass eine systematische Literatursuche die Aussagen dieses Berichts verändert hätte.

Probleme bei Evaluation von Diagnostika und Vorgehensweisen der Institutionen wurden anhand Evidenzmodells im Rahmen einer narrativen Review bearbeitet

Anhand der gefundenen Publikationen wurden die in Kapitel 2 definierten Zielsetzungen im Rahmen einer narrativen Review bearbeitet. Die Darstellung der Problematik bei der Evaluation von Diagnostika erfolgte qualitativ anhand des von Fryback und Thornbury entwickelten hierarchischen Evidenzmodells zur Evaluierung von diagnostischen Verfahren. Der Fokus des Berichts wurde auf Verfahren gelegt, die zur Diagnosefindung eingesetzt werden (i.e. um das Vorliegen einer Erkrankung zu bestätigen, oder auszuschließen).

Die methodischen Vorgehensweisen der ausgewählten Institutionen, die zur Bewertung von diagnostischen Technologien verwendet werden, wurden den Methodenpapieren der jeweiligen Institutionen entnommen und in narrati-

basierend auf Ergebnissen, Ableitung eines Fragenkatalogs um Nutzenbewertung für EntscheidungsträgerInnen zu erleichtern

ver Form dargestellt. Anschließend wurden die Vorgehensweisen tabellarisch zusammengefasst und Gemeinsamkeiten und Unterschiede der einzelnen Institute ebenfalls anhand des Evidenzmodells dargestellt.

Als letzter Schritt wurde, basierend auf den identifizierten methodischen Herausforderungen und in Zusammenschau mit den Vorgehensweisen der untersuchten Institutionen, ein für EntscheidungsträgerInnen relevanter Fragenkatalog abgeleitet, der die Nutzenbewertung diagnostischer Verfahren erleichtern soll.

4 Evidenzhierarchie zur Evaluation diagnostischer Verfahren

Evidenzmodell nach Fryback unterscheidet sechs Level

Ein häufig genanntes Modell zur Klassifikation der zur Evaluierung von diagnostischen Verfahren identifizierten Evidenz, wurde 1990 von Fryback und Thornbury entwickelt [5]. Das Modell unterscheidet insgesamt sechs Level (siehe Abbildung 4.1-1) wobei die einzelnen Levels hierarchisch angeordnet sind, sodass die unteren Stufen theoretisch Voraussetzungen für die jeweils oberen Ebenen sind.

Level 1: technische Qualität

Level 2: diagnostische Genauigkeit

Als unterste Ebene gilt der Nachweis der technischen Qualität (z.B. Auflösung, Kontrast) eines diagnostischen Verfahrens, wobei Studien auf dieser Stufe aber nur beschränkte Relevanz für EntscheidungsträgerInnen haben. Level 2, die diagnostische Genauigkeit, beinhaltet die Bestimmung von Parametern wie Sensitivität, Spezifität oder der Vorhersagewerte und erlaubt daher eine Einschätzung inwieweit ein Test zwischen gesund und krank unterscheiden kann (siehe Kapitel 4.2).

Level 3: diagnostischer Impakt

Level 4: therapeutischer Impakt

Level 5: patientenrelevante Endpunkte

Ob ein Testergebnis aber auch tatsächlich Änderungen des diagnostischen und therapeutischen Vorgehens bewirken kann, wird mit Studien der Level 3 und 4 bewertet, die zum Beispiel Therapiepläne *vor* einem Test mit denen *nach* dem Vorliegen des Testergebnisses vergleichen (siehe Kapitel 4.3). Wurde nun das diagnostische/therapeutische Management durch ein Testresultat verändert, ist von Interesse, ob dadurch aber auch patientenrelevante Endpunkte verbessert werden (Level 5) (siehe Kapitel 4.4). Möglich Endpunkte dieser Evidenzstufe sind daher etwa Mortalität, Lebensqualität, unerwünschte Wirkungen oder positive Konsequenzen durch die Vermeidung weiterer, invasiver Tests.

Level 6: Nutzen aus gesellschaftlicher Sicht

Abbildung 4.1-1: Hierarchisches Evidenzmodell zur Evaluierung diagnostischer Untersuchungsmethoden (Quelle: [5])

Als oberste Stufe in der Evidenzhierarchie sind nun Studien angesiedelt, die diagnostische Verfahren nicht auf der individuellen Patientenebene bewerten, sondern die Nutzen und Kosten aus gesellschaftlicher Sicht evaluieren (siehe Kapitel 4.5).

für EntscheidungsträgerInnen Level 5 und 6 entscheidend

Aus Sicht von EntscheidungsträgerInnen sind daher Studien der Level 5 und 6 entscheidend, um beurteilen zu können, ob ein diagnostisches Verfahren einen sinnvollen Einsatz limitierter finanzieller Mittel darstellt. Diskutiert wird allerdings, ob für den Nutzennachweis eines Tests tatsächlich jeder einzelne Level durchlaufen werden muss, oder ob, zum Beispiel, vor der Bewertung eines patientenrelevanten Nutzens, unter allen Umständen auch die diagnostische Genauigkeit erhoben werden muss [1].

4.1 Level 1 – technische Qualität

nicht relevant für EntscheidungsträgerInnen

Studien, die die technische Qualität von diagnostischen Verfahren bewerten sind mit Abstand die am häufigsten publizierten [6]. Endpunkte die im Rahmen dieser Studien erhoben werden, befassen sich mit der Reproduzierbarkeit der Ergebnisse oder mit technischen Charakteristika wie etwa Bildkontrast. Daher sind Studien dieses Evidenzniveaus für EntscheidungsträgerInnen nicht relevant und werden in diesem Bericht daher nicht weiter berücksichtigt.

4.2 Level 2 – Diagnostische Genauigkeit

positives Testergebnis = krank

negatives Testergebnis = gesund

Diagnostische Verfahren werden eingesetzt, um bestimmte Eigenschaften zu messen, die Aufschluss geben sollen, ob Personen an einer bestimmten Krankheit leiden, oder nicht. Anhand des Testergebnisses wird auf das Vorliegen von Erkrankungen geschlossen, wobei ein positiver Test mit „erkrankt" gleichgesetzt wird und ein negatives Testergebnis mit „gesund". Allerdings kann ein positives Testergebnis nicht nur bei tatsächlich erkrankten PatientInnen (=richtig positiv), sondern auch bei tatsächlich Gesunden (=falsch positiv) möglich ist. Umgekehrt können negative Befunde richtigerweise nicht nur bei gesunden PatientInnen vorliegen (= richtig negativ), sondern auch bei tatsächlich Kranken (=falsch negativ) [9].

allerdings auch falsch positive/negative Befunde möglich

Diskriminationsfähigkeit: wie gut kann ein Test zwischen krank und gesund unterscheiden

Die Diskriminationsfähigkeit eines Tests beschreibt das Vermögen zwischen gesunden und kranken (im Sinne des Tests) Personen zu unterscheiden. Problem dabei ist aber eben, dass kaum ein Test fehlerfrei zwischen krank und gesund unterscheiden kann, sodass die Konsequenzen von falsch positiven und falsch negativen Befunden berücksichtigt werden müssen. Falsch positive Ergebnisse können etwa aufgrund der damit verbundenen psychischen Belastung zu einer erheblichen Verminderung der Lebensqualität führen, weitere zur Abklärung nötige Tests können mit gesundheitlichen Risiken verbunden sein und auch die, für das Gesundheitssystem entstehende Kosten für Nachfolgeuntersuchungen, können beträchtlich sein. Unerwünschte Folgen von falsch negativen Resultaten liegen vor allem im, wenn überhaupt, verzögerten Beginn einer effektiven Therapie, was im schlimmsten Fall zum vorzeitigen Tod führen kann [9].

Konsequenzen von falschen Befunden z.B. psychische Belastung, gesundheitliche Risiken durch anschließende Tests, …

Testergebnisse können prinzipiell in unterschiedlichen Formen vorliegen, wobei man qualitative und quantitative Daten unterscheidet:

Testergebnisse als quantitative …

- Quantitative Variable sind numerische Variablen die durch Messungen zustande kommen:
  - Stetige (=kontinuierliche): messen Merkmale und können alle Werte innerhalb eines bestimmten Intervalls annehmen, z.B. Gewicht, Größe, mmHG.

oder

  - Diskrete: zählen Merkmale und können daher nur abzählbar viele Werte annehmen, z.B. Anzahl der Personen in einem Haushalt, Anzahl der pro Tag konsumierten Zigaretten.

…qualitative Variable

- Qualitative (=kategoriale) Variable entstehen durch Kategorisierung von Merkmalen:
  - Ordinale: die Merkmale werden anhand einer Rangordnung bewertet: Tumorstadium, Schweregrade einer Erkrankung.
  - Nominale: benennen Variable ohne Rangordnung: hierzu gehören dichotome Variablen, die nur in zwei Ausprägungen vorliegen (z.B. Geschlecht, krank/gesund) und polytome, bei denen es mehr als zwei Ausprägungen gibt (z.B. Blutgruppe, Augenfarbe) [10, 11].

Ergibt ein Test nun nicht nur negative und positive Ergebnisse, sondern wird zum Beispiel der Blutzucker gemessen, dann muss ein Grenzwert definiert werden, anhand dessen zwischen krank und gesund unterschieden wird [9] (siehe Kapitel „Zusammengesetzte Kenngrößen“).

Bestimmung der diagnostischen Genauigkeit durch Vergleich von Indextest mit Referenzstandard

Um die Fähigkeit eines Tests zu beschreiben, inwieweit er tatsächlich zwischen gesund und krank differenzieren kann, werden Kenngrößen der diagnostischen Genauigkeit ermittelt, indem die Ergebnisse des zu evaluierenden Tests (=Indextest) mit den Ergebnissen eines anderen Tests verglichen werden. Zum Vergleich sollten im besten Falle die Ergebnisse des Goldstandards herangezogen werden. Als Goldstandard wird ein Verfahren bezeichnet, das sicher zwischen Gesunden und Kranken unterscheiden kann. Da aber, wie erwähnt, kaum ein Test zu hundert Prozent zwischen Erkrankten und Nicht-Erkrankten differenzieren kann, existieren Goldstandards *in realiter* fast nicht [2]. Deshalb sollte der Indextest dann mit dem Referenzstandard verglichen werden, wobei unter Referenzstandard das derzeit beste diagnostische Verfahren verstanden wird.

Wie aus der Evidenzhierarchie nach Fryback ersichtlich wird (siehe Abbildung 4.1-1) sind Studien zur diagnostischen Genauigkeit lediglich auf Level 2 anzusiedeln und erlauben in der Regel keine Verfahrensbewertung hinsichtlich patienten- oder auf systemrelevanter Zielgrößen. Unter bestimmten Voraussetzungen können Testgenauigkeitsstudien aber ausreichen, um Rückschlüsse auf patientenrelevante Ergebnisse ziehen zu können (siehe Kapitel 4.4.2).

4.2.1 Parameter zur Messung der diagnostischen Genauigkeit

Einfache Kenngrößen

bei Tests mit nur zwei möglichen Ergebnissen (krank/gesund) Berechnung der Parameter mittels Vierfelder-Tafel

Die diagnostische Genauigkeit kann anhand von mehreren Parametern gemessen werden. Zu den Gebräuchlichsten gehören jene, die auf dichotomen Testergebnissen (Test ist positiv oder negativ) (siehe Kapitel 4.2) beruhen. Wird lediglich unterschieden ob ein Test negativ oder positiv ist, dann können die wichtigsten Parameter mittels Vierfelder-Tafeln berechnet werden (siehe Tabelle 4.2-1).

In der Vierfelder-Tafel werden in Abhängigkeit vom tatsächlichen Erkrankungszustand und dem Testergebnis richtig positive, falsch positive, richtig negative und falsch negative Ergebnisse eingetragen. Der tatsächliche Erkrankungszustand wird, wie erwähnt, am besten mittels Goldstandard, bei Fehlen eines solchen, mittels Referenzstandards erhoben.

Tabelle 4.2-1: Vierfelder-Tafel

Testergebnis	Krankheit	
	krank	gesund
positiv	Richtig positiv A	Falsch positiv B
negativ	Falsch negativ C	Richtig negativ D

Sensitivität und Spezifität

Sensitivität und Spezifität am gebräuchlichsten

Die am häufigsten angegebenen Kenngrößen zur Beschreibung der Diskriminationsfähigkeit eines Tests sind Sensitivität und Spezifität. Da sie die Testgenauigkeit nur gemeinsam vollständig beschreiben, sollten sie daher immer paarweise angegeben werden.

Sensitivität = Wahrscheinlichkeit, dass ein tatsächlich Kranker als krank diagnostiziert wird

- Die Sensitivität ist die Wahrscheinlichkeit, dass eine tatsächlich erkrankte Person als krank diagnostiziert wird, beziehungsweise entspricht dem Anteil erkrankter Personen in einer Population, der richtigerweise als krank diagnostiziert wurde. Eine Sensitivität von 90% bedeutet daher, dass von 100 Kranken 90 erkannt wurden (richtig positiv), während 10 Personen fälschlicherweise als gesund (falsch negativ) eingestuft worden waren. Die Berechnung ist wie folgt:

  Sensitivität = a/(a + c).

Spezifität = Wahrscheinlichkeit, dass ein tatsächlich Gesunder als gesund diagnostiziert wird

- Die Spezifität ist die Wahrscheinlichkeit, dass eine tatsächlich gesunde Person als gesund diagnostiziert wird, beziehungsweise entspricht dem Anteil gesunder Personen in einer Population, der richtigerweise als gesund identifiziert wurde. Eine Spezifität von 75% bedeutet, dass von 100 Gesunden 75 richtig als gesund (richtig negativ) identifiziert wurden, 25 aber fälschlicherweise als krank (falsch positiv) eingestuft wurden. Die Berechnung ist wie folgt:

  Spezifität = d/(b +d) [12, 13].

hohe Spezifität wichtig, um Krankheit zu bestätigen

hohe Sensitivität wichtig, um Krankheit auszuschließen

Stehen mehrere Tests zur Verfügung, die sich in Bezug auf Sensitivität und Spezifität unterscheiden, dann wird die Auswahl eines geeigneten Tests davon abhängen, ob mehr Wert auf den Ausschluss oder den Einschluss der Erkrankung gelegt wird. Ist ein Testergebnis eines sehr spezifischen Tests positiv, dann ist die Person zu hoher Wahrscheinlichkeit auch tatsächlich krank (Rule-In). Soll dagegen eine Krankheit mit hoher Sicherheit ausgeschlossen werden (Rule-Out), dann eignen sich Test mit hoher Sensitivität, da ein negatives Ergebnis dann ziemlich sicher mit dem Befund „gesund" einhergeht [14, 15].

Sensitivität/Spezifität unabhängig von Prävalenz

Vorteile: Sensitivität und Spezifität sind von der Häufigkeit (Prävalenz) der Zielerkrankung in der untersuchten Population unabhängig.

aber Unterschiede können in Subgruppen auftreten

Nachteile: Diese Kenngrößen können in verschiedenen Subgruppen, wie zum Beispiel bei unterschiedlichen Krankheitsstadien, variieren [16, 17], sodass zwei Tests, wenn schon nicht in der gleichen Patientenpopulation, zumindest in sehr ähnlichen Gruppen verglichen werden sollten (siehe auch Kapitel 4.2.5) [18].

Vorhersagewerte

Vorhersagewerte = Wahrscheinlichkeit mit der eine Erkrankung bei bestimmtem Testergebnis vorliegt

Vorhersagewerte (Englisch: predictive values (PV)) drücken die Wahrscheinlichkeit aus, mit der eine Erkrankung bei gegebenem Testergebnis vorliegt [19], besitzen damit also vor allem klinische Relevanz.

- Der positive Vorhersagewert (PPV) ist ein Maß für die Wahrscheinlichkeit einer Erkrankung bei Vorliegen eines positiven Testergebnisses. Beträgt der PPV etwa 0.62, bedeutet dies, dass 62% der PatientInnen mit positivem Testergebnis auch tatsächlich krank sind. Die Berechnung ist wie folgt:

  $PPV = a/(a + b)$.

- Der negative Vorhersagewerte (NPV) ist ein Maß für die Wahrscheinlichkeit gesund zu sein, wenn ein negatives Testergebnis vorliegt. Ein NPV von 0.98 bedeutet daher, dass 98% aller testnegativen PatientInnen tatsächlich gesund sind. Die Berechnung ist wie folgt:

  $NPV = d/ (c + d)$.

klinisch relevante Parameter

Vorteil: Der PPV ergibt die Wahrscheinlichkeit, dass ein Testpositiver auch tatsächlich erkrankt ist, unterstützt also den/die KlinkerIn um Aussagen bezüglich der Erkrankungswahrscheinlichkeit bei positivem Testergebnis zu machen und erhöht so die Wahrscheinlichkeit einer richtigen Diagnose. Er ist somit der wichtigste Parameter in der klinischen Praxis.

aber prävalenzabhängig

Nachteil: Vorhersagewerte werden auch häufig Nachtestwahrscheinlichkeiten (= *a posteriori* Wahrscheinlichkeit) genannt, da sie von der Prävalenz (= *a priori* Wahrscheinlichkeit) der Erkrankung in der Studienpopulation abhängig sind [17, 19, 20]. Geht man davon aus, dass die diskriminatorische Fähigkeit eines Tests (Sensitivität, Spezifität) konstant ist, wird klar, dass mit demselben Test in einer Gruppe mit höherer Ausgangswahrscheinlichkeit für eine Erkrankung höhere PPVs erzielt werden können, als in einer Gruppe mit niedriger Ausgangswahrscheinlichkeit. Das bedeutet, dass sich unterschiedliche PV für Tests in einer z.B. asymptomatischen Patientengruppe und in einer symptomatischen Gruppe [3], oder auch im niedergelassenen Bereich und in einem Schwerpunktkrankenhaus ergeben würden

daher unterschiedliche Ergebnisse in Patientengruppen mit unterschiedlicher Prävalenz

(siehe Kapitel 4.2.5) [17, 20]. Die Abhängigkeit von der Prävalenz bedeutet daher auch, dass PVs nicht auf andere Populationen umgelegt werden können und damit Studienergebnisse von PVs nicht generalisiert werden können. Ausnahme ist, wenn die Prävalenzen der Studienpopulation und der Population, in der der Test angewandt werden soll, vergleichbar (d.h. gleich) sind [3].

Likelihood Ratios

Likelihood Ratios (LR) geben an, wie viel Mal wahrscheinlicher ein Testergebnis bei Erkrankten, als bei Gesunden ist [21]. Zwei LR werden unterschieden [22]:

Likelihood Ratios = wie viel Mal wahrscheinlicher ein Testergebnis bei Erkrankten als bei Gesunden ist

- die positive LR ist die Wahrscheinlichkeit, dass ein erkrankter PatientIn ein positives Testergebnis erhält, dividiert durch die Wahrscheinlichkeit, dass ein gesunder PatientIn ein positives Testergebnis erhält [23] und wird wie folgt berechnet:

  LR+ = (a/(a+c))/(b/(b+d)) = Sensitivität/(1-Spezifität).

- die negative LR gibt die Wahrscheinlichkeit an, mit der ein erkrankter PatientIn ein negatives Testergebnis erhält, dividiert durch die Wahrscheinlichkeit, dass ein gesunder PatientIn ein negatives Testergebnis erhält [23] und wird wie folgt berechnet:

  LR- = (c/(a+c)) / (d/(b+d))= (1-Sensitivität) / Spezifität.

Die Werte, die LR einnehmen können rangieren zwischen 0 und ∞ (siehe Tabelle 4.2-2). 1 bedeutet also, dass der Test keine zusätzliche Information bringt, da die Wahrscheinlichkeit für ein bestimmtes Testergebnis bei Gesunden und Kranken gleich groß ist. Bei Werten darunter (LR-) oder darüber (LR+) ist die Wahrscheinlichkeit für das Vorliegen einer Erkrankung erniedrigt/erhöht [17, 24].

Tabelle 4.2-2: Überblick über Effizienz eines diagnostischen Test (Quelle: [22])

LR +	LR -	Testeffizienz
>10	< 0,10	sehr gut
5 – 10	0,1 – 0,2	gut
2 -5	0,2 – 0,5	mäßig
1 - 2	0,5 – 1,0	schlecht

Wenn mehrere voneinander unabhängige Tests (das Testergebnis des einen Tests beeinflusst nicht die Wahrscheinlichkeiten der Ergebnisse des anderen Tests) durchgeführt werden, so entspricht deren gemeinsame LR, dem Produkt der LR der einzelnen Tests [24].

Vorteil: LR werden einerseits verwendet, um abschätzen zu können, wie gut ein diagnostischer Test ist und sind daher Kenngrößen für den Informationsgewinn, den ein Test bringt. Andererseits eignen sie sich auch, um geeignete Testmethoden, beziehungsweise die Reihenfolge mehrerer Tests festzulegen [15]. Zusätzlich sind sie unabhängig von der Prävalenz einer Erkrankung [15] und können im Gegensatz zu Sensitivität, Spezifität und PV

um abschätzen zu können, wie gut ein Test ist

prävalenzunabhängig

nicht nur für dichotome Variablen berechnet werden, sodass die Festlegung eines (arbiträren) Grenzwertes unnötig wird (siehe Kapitel "Zusammengesetzte Kenngrößen") [17, 20, 25].

mittels Bayes-Theorem Berechnung von PV aus LR

Mit den LRs können auch unter Zuhilfenahme des Bayes-Theorem, das Rückschlüsse von *a priori* - Wahrscheinlichkeiten auf *a posteriori* - Wahrscheinlichkeit (= PVs) erlaubt, die PVs berechnet werden [3, 10, 12, 26, 27]. Der gebräuchlichste Weg ist jener der Verwendung von „Odds", was so viel wie „Chance" bedeutet. Odds beschreiben in einer Gruppe das Verhältnis zwischen der Anzahl von TeilnehmerInnen mit einem bestimmten Endpunkt und der Anzahl von TeilnehmerInnen ohne diesen Endpunkt [10, 28]. Odds können leicht in Wahrscheinlichkeiten umgerechnet werden, da gilt:

a posteriori Wahrscheinlichkeit = *a posteriori* Odds/ *(a posteriori* Odds + *1)*

a posteriori - Odds = *a priori* - Odds* LR [12, 15, 20, 21, 25, 29],

oder anders ausgedrückt:

Informationsgewinn nach Test = Information vor Test * Information durch den Test [3, 30].

Die *a priori* - Odds wird wiederum mittels folgender Formel berechnet:

a priori - Odds = Prävalenz/(1- Prävalenz)

Eine einfache Möglichkeit Nachtestwahrscheinlichkeiten aus LR und Vortestwahrscheinlichkeiten zu berechnen, bietet das FAGAN - Nomogramm (siehe Abbildung 4.2-1) [15].

einfache Möglichkeit: Fagan Nomogramm

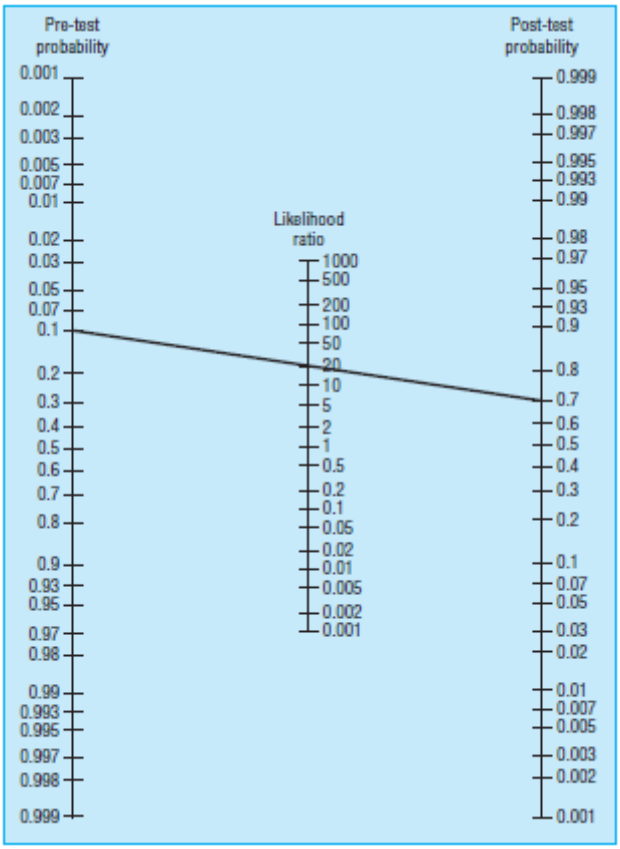


Abbildung 4.2-1: Fagan Nomogramm zur Berechnung von Nachtestwahrscheinlichkeiten (Quelle: [31])

Zusammengesetzte Kenngrößen

Liegen Testergebnisse nicht nur als positive oder negative Resultate vor, sondern sind sie zum Beispiel stetig oder polytom (siehe Kapitel 4.2), dann muss ein Grenzwert festgelegt werden, anhand dessen zwischen „krank“ (d.h. positiv) und „gesund“ (d.h. negativ) unterschieden wird [13, 32]. Dies ist kritisch, da sich Sensitivität und Spezifität in Abhängigkeit von diesem Grenzwert ändern.

bei nicht-dichotomen Variablen muss anhand eines Grenzwertes zwischen gesund und krank unterschieden werden

Wird zum Beispiel der Blutglucosewert, anhand dessen zwischen Diabetikern und Nicht-Diabetikern unterschieden werden soll, hoch angesetzt (z.B. 200 mg/dl), dann werden Nicht-Diabetiker zu einem Großteil richtigerweise als gesund diagnostiziert (= hohe Spezifität). Im Gegensatz dazu werden aber etliche Zuckerkranke fälschlicherweise als gesund klassifiziert (=niedrige Sensitivität). Wird der Grenzwert dagegen niedrig angesetzt, zum Beispiel bei 90 mg/dl, dann werden tatsächlich Erkrankte richtigerweise als erkrankt diagnostiziert (=hohe Sensitivität), aber es werden auch zahlreiche Nicht-Diabetiker als zuckerkrank diagnostiziert (=niedrige Spezifität)[9, 20, 32]. Das bedeutet, dass ein Trade-off zwischen Sensitivität und Spezifität gemacht werden muss: bei Erhöhung der Sensitivität durch einen

Wahl des Grenzwertes beeinflusst Sensitivität/Spezifität

niedrigen Grenzwert wird die Spezifität erniedrigt, bei Erhöhung der Spezifität durch einen hohen Grenzwert wird die Sensitivität erniedrigt [9].

Wahl des Grenzwertes abhängig ob richtiges Erkennen von Gesunden oder Kranken wichtiger ist

Dies bedeutet, dass vor Festlegung eines Grenzwertes bestimmt werden muss, ob größerer Wert auf das richtige Erkennen von tatsächlich Erkrankten oder auf das Erkennen von tatsächlich Gesunden gelegt wird. Eine hohe Sensitivität ist vor allem dann wichtig, wenn mit dem Nichterkennen (=falsch negativ) gravierende, fatale Folgen verbunden sind, da eine relativ sichere und wirksame Therapie zur Verfügung gestanden hätte [9]. Eine hohe Sensitivität bedeutet nämlich, dass nahezu alle Erkrankten diagnostiziert werden, d.h. es gibt fast keine falsch-negativen Testergebnisse. Dadurch können etwa differenzialdiagnostische Überlegungen am Anfang der Diagnosestellung deutlich eingeschränkt werden [10, 14, 15, 32].

Bestimmung durch Konsequenzen von falsch positiven/falsch negativen Befunden

hohe Sensitivität wenn wirksame Therapie zur Verfügung steht und fatale Folgen mit Nicht-Erkennen verbunden sind

Eine hohe Spezifität und damit möglichst wenig falsch-positive Befunde sind besonders dann relevant, wenn eine Diagnose bestätigt werden soll, da ein Therapieregime mit gravierenden Nebenwirkungen oder hohen Kosten verbunden ist, oder wenn die weitere diagnostische Abklärung mit hohen Risiken oder psychischen Beanspruchungen verbunden ist [10, 14, 32]. Eine hohe Spezifität bedeutet nämlich, dass nahezu alle Gesunden richtig diagnostiziert werden, d.h. es gibt kaum falsch-positive Ergebnisse.

hohe Spezifität wenn Therapie mit starken Nebenwirkungen oder hohen Kosten verbunden ist

Zusammengesetzte Kenngrößen eignen sich nun, um die Gesamtgenauigkeit eines Tests mit einem Parameter festzustellen, da sie das Verhältnis von richtigen zu falschen Testergebnissen reflektieren.

Receiver-Operating Characteristics Kurve & Area under the curve

ROC stellt Sensitivität und Spezifität in Abhängigkeit der Grenzwerte dar

In der Receiver-Operating Characteristics Kurve (ROC) werden nun Kombinationen richtig positiver (Sensitivität) und falsch positiver Ergebnisse („1-Spezifität) in Abhängigkeit von unterschiedlichen Grenzwerte dargestellt (siehe Kapitel „Zusammengesetzte Kenngrößen“) [13, 33].

daher zur Bestimmung des optimalen Grenzwertes geeignet

Je nachdem ob nun besonderes Augenmerk auf Sensitivität oder Spezifität gelegt wird, kann mittels der ROC-Kurve der optimale Grenzwert gefunden werden. Wird beiden Parametern der gleiche Stellenwert eingeräumt, dann wäre der optimale Schwellenwert der Punkt der Kurve, der am nächsten zum Punkt (0, 1) des Koordinatensystems liegen würden, also jener Punkt der am nächsten zur oberen linken Ecke ist (siehe Abbildung 4.2-2)[10, 12, 13].

AUC für Gesamtgenauigkeit eines Tests

Mittels der Area under the curve (AUC - die Fläche unter der ROC-Kurve) kann die Gesamtgenauigkeit eines Tests bestimmt werden. Hat diese Fläche einen Wert von O,5, ist der Test nicht besser als eine rein zufällige Zuordnung zu gesund oder krank (in Abbildung 4.2-2 die strichlierte Linie). Je größer dieser Wert wird, desto besser ist die diagnostische Genauigkeit und bei einem Wert von 1, sind falsche Befunde ausgeschlossen, was allerdings nur möglich wäre, wenn der Test unabhängig vom Grenzwert eine Sensitivität und Spezifität von 100% haben würde [13, 33].

Vergleich mehrerer Tests möglich

Vorteil: Durch die Darstellung von Sensitivität und Spezifität in Abhängigkeit von unterschiedlichen Schwellenwerten kann eine Einschätzung der gesamten Testgenauigkeit erfolgen. Durch die graphische Darstellung des Trade-offs von richtig positiven und falsch positiven Testergebnissen, lässt sich die ROC Kurve also zur Bestimmung des optimalen Schwellenwertes heranziehen [10, 12] und anhand der AUC können mehrere Tests miteinander verglichen werden [10, 32], wobei die AUC Sensitivität und Spezifität gleich gewichtet, was nur in Ausnahmefällen sinnvoll erscheinen mag.

Weiters ist das Ergebnis unabhängig von der Prävalenz der Erkrankung [33].

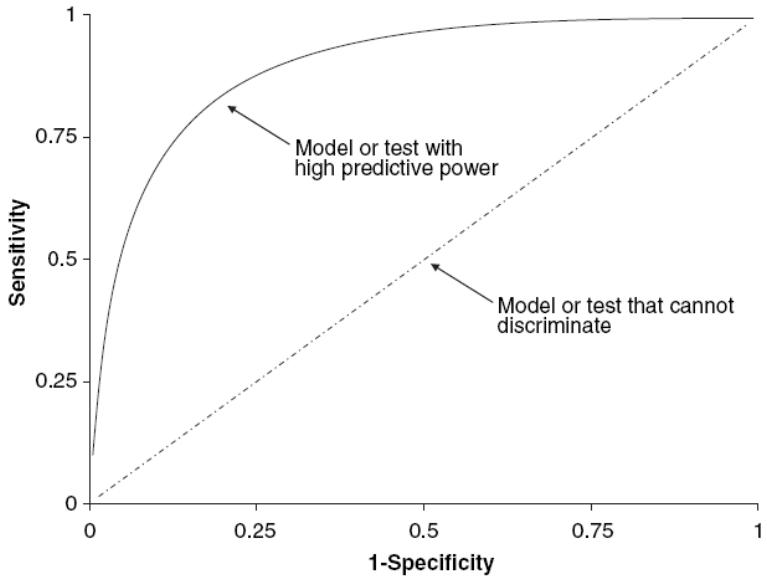


Abbildung 4.2-2: Zwei ROC- Kurven mit unterschiedlicher Diskriminationsfähigkeit. Die gebogene Linie entspricht einem Test mit guter diagnostischer Genauigkeit, die Gerade einem Test, der nicht zwischen gesund und krank unterscheiden kann (Quelle: [13])

Diagnostische Odds Ratio

DOR zusammengesetztes Maß für Testgenauigkeit

Die diagnostische Odds Ratio (DOR) ist eine zusammengesetzte Kenngröße für die Testgenauigkeit und eignet sich im Rahmen von Meta-Analysen zur kombinierten Darstellung der Sensitivitäten und Spezifitäten, oder LR aus mehreren Studien. Sie beschreibt die Chance eines positiven Testergebnisses einer erkrankten Person in Relation zu der Chance eines positiven Testergebnisses einer gesunder Person [34]. Bei einer DOR = 1 besitzt der Test keinerlei Diskriminationsfähigkeit. Je weiter die DOR aber von 1 entfernt ist, desto besser ist die Aussagekraft [35].

DOR = (a x d)/(c x b)= LR+/LR- =

= (Sensitivität/ (1-Sensitivität))/((1-Spezifität)/Spezifität) [34].

erlaubt Darstellung der Genauigkeit in einer einzelnen Zahl

Vorteil: Im Unterschied zu LRs oder Sensitivität/Spezifität, erlaubt die DOR die Darstellung der Testgüte in einer einzelnen Zahl [17]. Zusätzlich ist sie sowohl weitgehend vom Grenzwert, als auch gänzlich von der Prävalenz der Erkrankung unabhängig [21].

Informationsverlust von Trade-Off zwischen Sensitivität und Spezifität

Nachteil: DORs sind einerseits nicht geeignet, um direkt klinische Entscheidungen treffen zu können [21], andererseits sind sie durch den Verlust der Information zu dem Trade-off, der zwischen Sensitivität und Spezifität gemacht werden muss, nicht geeignet, zwei Tests miteinander zu vergleichen [21, 36].

Zusammenfassung

Tabelle 4.2-3: Testgütekriterien (adaptiert nach: [6])

Maßzahl	Definition	Vorteile	Nachteile
Sensitivität	a/(a + c)	Unabhängig von Prävalenz	Abhängig von Schwellenwert
Spezifität	d/(b + d)	Unabhängig von Prävalenz	Abhängig von Schwellenwert
PPV	a/(a + b)	Klinisch relevant	Abhängig von Prävalenz
NPV	d/(c + d)	Klinisch relevant	Abhängig von Prävalenz
LR +	(a/(a + c)) / (b/(b + d))	Unabhängig von Prävalenz	Nur bei positiven Testergebnis anwendbar
LR -	(c/(a + c)) / (d/(b + d))	Unabhängig von Prävalenz	Nur bei negativem Testergebnis anwendbar
Odds Ratio	(a x d)/ (c x b)	Unabhängig von Prävalenz, kombiniert Sensitivität und Spezifität	nicht selbsterklärend, nicht relevant für klinische Entscheidungen
AUC	Area under curve	Unabhängig von Prävalenz, kombiniert Sensitivität und Spezifität	Nicht direkt relevant für klinische Entscheidungen

4.2.2 Studiendesigns

diagnostische Genauigkeitsstudie: Querschnittstudie mit konsekutiv eingeschlossenen PatientenInnen mit Vergleich von Index- und Referenztest unter Verblindung

Als bestes Studiendesign zur Bestimmung der diagnostischen Genauigkeit gilt eine Querschnittsstudie, die PatientInnen konsekutiv einschließt und die Ergebnisse des Indextests in der *gesamten* Population mit einem validen Referenzstandard unter Verblindung vergleicht (=Verifizierung) [6, 34, 37-41]. Ist der gewählte Vergleichstest nicht mit dem Referenztest identisch, dann sollte idealerweise ein- und dieselbe Patientengruppe sowohl mit Index-, Vergleichs-, aber auch Referenztest untersucht werden [18].

Modifikationen möglich

Je nach geplantem Verwendungszweck eines neuen Tests, kann dieses Basisdesign aber modifziert werden (siehe Kapitel 4.2.4). Ist der Indextest etwa zu invasiv um in allen PatientInnen durchgeführt zu werden, dann sind randomisierte Studien, die Index- und Vergleichstest zufällig auf die Studienpopulation aufteilen, zu bevorzugen [18].

andere Designs minderer Qualität

Von geringerer methodischer Qualität sind Studien, die PatientInnen *nicht* konsekutiv einschließen, nicht verblindet sind, diagnostische Fall-Kontroll-Studien, oder Studien, die den Indextest entweder gar nicht, oder mit einem nicht relevanten Referenzstandard vergleichen [41](siehe Kapitel 4.2.5).

Mehrere Institutionen haben Evidenzhierarchien für diagnostische Studien entwickelt [41, 42] (siehe Appendix 9), wobei an oberster Stelle jeweils auf hochwertigen Diagnosestudien basierende systematische Reviews stehen.

4.2.3 Mutiple Testergebnisse: Serielles vs paralleles Testen

mehrere Tests können nacheinander (seriell) oder gleichzeitig (parallel) durchgeführt werden

Gelangen im Rahmen eines Diagnosepfades mehrere diagnostische Verfahren zum Einsatz, sollte auch überlegt werden, in welcher Abfolge die Tests eingesetzt werden und wie mehrere Testergebnisse insgesamt zu interpretieren sind. Prinzipiell können Untersuchungen nacheinander (***serielles*** Testen) oder gleichzeitig (***paralleles*** Testen) durchgeführt werden (siehe Abbildung 4.2-3).

Die Entscheidung in welcher Abfolge einzelne Tests gemacht werden sollen, ist abhängig von den zu erwartenden positiven, als auch negativen Konsequenzen für PatientInnen [27, 43].

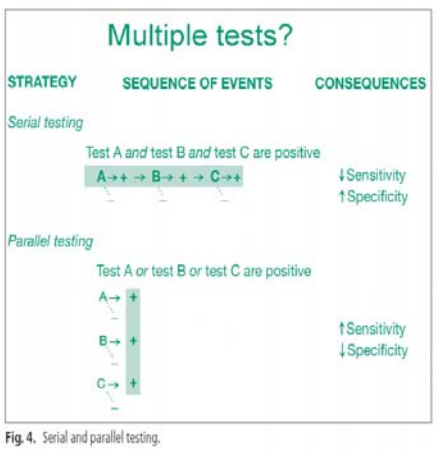


Abbildung 4.2-3: Serielles und paralleles Testen
(Quelle: [43])

paralleles Testen bei nicht-invasiven, kostengünstigen Tests, bei Zeitdruck

Parallelem Testen sollte der Vorzug gegeben werden, wenn Untersuchungsmethoden einfach, nicht-invasiv und kostengünstig sind und/oder wenn Zeitdruck (z.B. schwere Erkrankung, hohe Kosten bei langen Spitalsaufenthalten) besteht [44]. In Abhängigkeit von der Interpretation der unterschiedlichen Testergebnisse, kann ein PatientIn bereits als krank eingestuft werden, wenn einer der Tests positiv ist und als gesund nur dann, wenn alle Tests negativ sind (ODER-Regel). Bei der UND-Regel ist ein PatientIn nur dann erkrankt, wenn alle Ergebnisse auf das Vorliegen einer Erkrankung hindeuten (UND-Regel), ist also bereits „gesund", wenn einer der Tests negativ ist [27, 43]. Bei der ODER-Regel ist die kombinierte Sensitivität hö-

her, die kombinierte Spezifität dafür aber niedriger, als die der Einzelergebnisse. Bei der UND-Regel gilt wiederum das Gegenteil [9, 27] (siehe Tabelle 4.2-4.).

serielles Testen bei teuren, invasiven Untersuchungen, Zeit kein limitierenden Faktor ist

Im Gegensatz dazu wird *serielles* Testen bei eher teuren und invasiven Untersuchungen empfohlen, Zeit kein limitierender Faktor ist, oder nachfolgende Untersuchungen von den Ergebnissen vorhergehender Tests abhängig sind [44]. Obwohl für serielles Testen zwar mehr Zeit benötigt wird, ist der Vorteil, dass unnötige Tests vermieden werden können [27]. Wird die ODER-Regel bei seriellem Testen angewandt, dann wird die Diagnose bei Vorliegen *eines* positiven Testergebnisses, sei es das erste, zweite oder irgendeines beliebigen in einer Testserie, gestellt. Die UND-Regel hingegen bedingt, dass sowohl der erste, als auch alle weiteren Tests positiv sein müssen. Die ODER-Regel erhöht die Sensitivität, die UND-Regel die Spezifität (siehe Tabelle 4.2-4)[27, 43].

Tabelle 4.2-4: Berechnung der Sensitivität und Spezifität bei parallelem und seriellem Testen nach Weinstein [27]:

	Und - Regel		Oder - Regel	
	Berechnung	Auswirkung auf Sens/Spez	Berechnung	Auswirkung auf Sens/Spez
Parallele Tests	SEk = SEa x SEb SPk = SPa + SPb – (Spa x SPb)	SE ↓ SP ↑	SEk = SEa + SEb – (SEa x SEb) SPk = SPa x SPb	SE ↑ SP ↓
Serielle Tests	SEk = SEa x SEb SPk = SPa + (1 – SPa) x SPb	SE ↓ SP ↑	SEk = SEa + (1 – SEa) x SEb SPk = SPa x SPb	SE ↑ SP ↓

(SPk= kombinierte Spezifität, SPa = Spezifität Test A, SEk= kombinierte Sensitivität, SEa = Sensitivität Test A)

4.2.4 Diagnostische Pfade

Tests können an unterschiedlichen Stellen in existierende diagnostische Strategien eingebettet werden: Ersatz, Triage, Add-On

Diagnostische Verfahren können in unterschiedlichster Weise in bestehende diagnostische Strategien eingebettet werden. Vor der Bewertung eines Tests ist es daher sinnvoll zu überlegen, ob der Test *nach, anstatt* oder *vor* einem bereits bestehenden Verfahren (= in Abbildung 4.2-4 „inital Test“) eingesetzt werden soll. Neben der Bewertung der reinen diagnostischen Genauigkeit können bei Entscheidungen über den Verwendungszweck eines neuen Tests auch andere Testattribute, wie etwa niedrigere Kosten, geringere Nebenwirkungen oder eine schnellere Verfügbarkeit der Ergebnisse eine Rolle spielen [18]. Je nachdem wo der Indextest in eine bestehende diagnostische Strategie eingebettet werden soll, können sich auch die Studiendesigns, die zum Nachweis der diagnostischen Genauigkeit benötigt werden, verändern.

Ersatz

Bestimmung der diagnostischen Genauigkeit durch Vergleich von Index,- Referenztest und zu ersetzender Test in ein und derselben Population (=fully paired study)

Um herauszufinden, ob eine neue diagnostische Technologie eine bereits bestehende Untersuchungsmethode ersetzten kann, müssen deren diagnostische Genauigkeiten verglichen werden. Da die diagnostische Genauigkeit innerhalb von unterschiedlichen Patientengruppen variieren kann (siehe Kapitel Sensitivität und Spezifität und 4.2.5), sind die besten Studiendesigns dafür Studien, bei denen alle PatientInnen mit dem Indextest, dem zu ersetzenden Test und dem Referenztest untersucht werden („fully paired study“) [18] (siehe Kapitel 4.2.2). Die Vorteile des letztgenannten Designs sind, dass die PatientInnen absolut identisch sind und dass eine nur relativ kleine Patientenpopulation benötigt wird.

bei invasiven Test auch RCTs mit Zuteilung entweder zu Indextest oder zu ersetzenden Test

Eine andere Möglichkeit sind randomisiert kontrollierte Studien (RCT), bei denen PatientInnen zufällig, entweder dem Indextest oder zu dem Vergleichstest zugeordnet werden und anschließen mittels Referenztest verifiziert (= Feststellen des tatsächlichen Krankheitszustands) werden. RCTs sind dann angezeigt, wenn sich die zu vergleichenden Tests gegenseitig beeinflussen würden oder wenn die Tests zu invasiv sind, um in ein- und derselben Patientengruppe getestet zu werden [18].

Triage

Triage um nachfolgende teure oder invasive Tests zu reduzieren

Unter Triage versteht man, wenn der Indextest *vor* einem bereits bestehenden Test, oder einer Testsequenz, zur Anwendung gelangen soll (siehe Abbildung 4.2-4) und PatientInnen nur bei positivem Indextest weitere Untersuchungen erhalten [18]. Triagetests zielen generell nicht darauf ab, die diagnostische Genauigkeit zu verbessern und sollen bereits existierende Tests auch nicht ersetzten, sondern sollen mit geringeren Kosten oder eine einfache Handhabung, die Anzahl anschließender teurer oder invasiver Folgeuntersuchungen reduzieren.

Studiendesigns: fully paired study, oder Studien mit limitierter Verifizierung

Um die Triagestrategie mit einer bereits bestehenden Strategie zu vergleichen, eignen sich auch wieder „fully paired studies“ (siehe Kapitel „Ersatz“), wobei es aber auch ausreichend sein kann, nicht alle PatientInnen mittels Referenzstandards zu untersuchen, sondern nur Patientinnen mit negativen Indextestergebnis aber positivem Ergebnis des existierenden Test mittels Referenztest zu verifizieren (=limitierte Verifizierung). So können PatientInnen identifziert werden, die bei Verwendung des Indextests unerkannt geblieben wären, aber auch jene bei denen der existierende Test vermieden hätte werden können [18]. Dabei zu beachten ist aber, dass die Ergebnisse dieser Studien durch „Verfication Bias“ verfälscht werden können (siehe Kapitel 4.2.5).

Add-on Test

Add-On Tests erhöhen Sensitivität oder Spezifität

Add-on Tests erhöhen die Sensitivität oder die Spezifität und werden im Anschluss an eine bereits bestehende diagnostische Strategie durchgeführt (siehe Abbildung 4.2-4), sodass oft nur eine ausgewählte Patientengruppe für diese oft teuren oder invasiven, dafür aber genauen Untersuchungen in Frage kommt [45]. Mögliche Anwendungsbereiche sind entweder, wenn das Vorliegen einer Erkrankung bestätigt werden soll, oder aber das Gegenteil, wenn falsch positive Befunde ausgeschlossen werden sollen.

Studiendesigns: fully paired study, Studien mit limitierter Verifizierung

Als Studientyp eigenen sich zwar auch wiederum“ fully paired studies“ oder RCTs, allerdings kann es bei Add-On Tests ausreichend sein, nur jene PatientInnen einzuschließen, die nach der bisherigen Strategie ein negatives Testergebnis hatten, und dann nur jene, die ein positives Indextestergebnis haben, mittels Referenzstandard zu verifizieren [18]. Wiederum ist hier aber die Möglichkeit von Verfication Bias gegeben (siehe Kapitel 4.2.5).

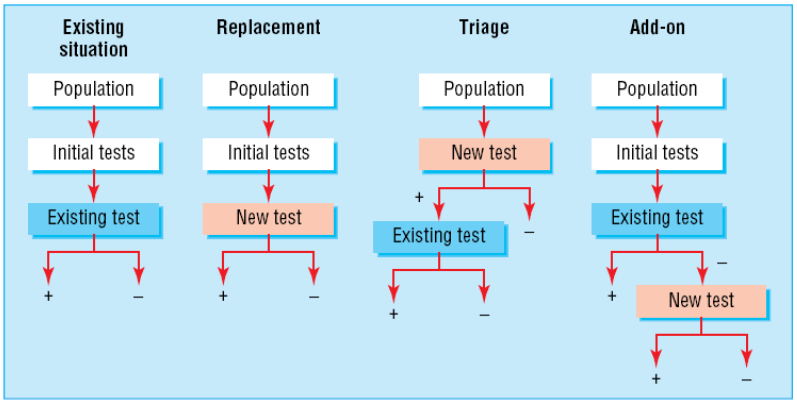


Abbildung 4.2-4: Verwendungsmöglichkeiten neuer Tests bei einer bereits bestehenden diagnostischen Strategie (Quelle: [18])

4.2.5 Bias & Variation in diagnostischen Genauigkeitsstudien

Validität von diagnostischen Genauigkeitsstudien wird bestimmt durch interne und externe Validität

Wie bereits erwähnt, bilden diagnostische Genauigkeitsstudien den Hauptanteil der zu diagnostischen Untersuchungsverfahren vorhandenen Evidenz. Da mit diesem Studiendesigns spezielle methodische Probleme vergesellschaftet sein können, soll in diesem Kapitel genauer auf mögliche Ursachen eingegangen werden, die die Validität der Ergebnisse kompromittieren können.

Wie auch bei Interventionsstudien wird bei diagnostischen Studien zwischen interner und externer Validität unterschieden.

interne Validität kann durch Bias,…

- die *interne Validität* bezeichnet das Ausmaß mit dem Studienergebnisse den wahren Effekt einer Intervention/Technologie/Exposition ausdrücken. Der wahre Effekt kann durch systematische Fehler (=Bias), bedingt etwa durch Mängel im Studiendesign oder in der Ausführung der Studie, verzerrt sein [28, 38] und so Studienergebnisse in eine bestimmte Richtung verfälschen [10, 38].

externe Validität durch Variation kompromittiert sein

- die *externe Validität* hingegen beschreibt die Generalisierbarkeit, also die Übertragbarkeit von Studienergebnissen auf den klinischen Alltag oder auf andere klinische Settings [28, 38] und daher möglicherweise auf PatientInnen, die nicht in der ursprünglichen Studienpopulation enthalten waren. Unter Variation wird nun verstanden, wenn sich Studien in Hinblick auf Charakteristika der Studienpopulation, dem Testprotokoll oder etwa dem klinischen

Settings unterscheiden und daher die Übertragbarkeit der Ergebnisse limitiert wird [38, 46].

Variationen und Bias können heterogene Studienergebnisse bedingen

Variationen und Bias sind von besonderem Interesse, da sie zu Unterschieden in den Ergebnissen (= Heterogenität) einzelner Studien führen können. Eine Untersuchung des Potentials für Bias oder Variationen ist daher nötig, um die Qualität der Studienergebnisse, als auch die Relevanz für den eigenen Kontext bewerten zu können. Ein weiterer Aspekt ist, dass vergleichbare Studienpopulationen in Wirksamkeits- und Diagnosestudien eine der Voraussetzungen für „linked Evidence" (siehe Kapitel 4.4.2) sind und daher mögliche Variationen in den untersuchten Populationen, dem Schweregrad der Erkrankung, etc. berücksichtigt werden müssen.

mögliche Quellen für Bias und Variation durch:

Im Folgenden sollen in der Literatur beschriebene, mögliche Quellen für Bias und Variationen kurz dargestellt werden, wobei aber anzumerken ist, dass nicht alle davon empirisch belegt sind. Für eine umfassendere Darstellung wird auf das „Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy" verwiesen [38].

Bias

Limited challenge bias

nicht konsekutiv eingeschlossene PatientInnen

Wie in Kapitel 4.2.2 erwähnt, sollten in diagnostischen Studien die PatientInnen konsekutiv eingeschlossen werden. Ist dies nicht der Fall, könnten schwieriger zu diagnostizierende Individuen ausgeschlossen werden und zu einer Überschätzung der Testgenauigkeit führen [34, 47, 48]. Ähnliche Ergebnisse wären auch zu erwarten, wenn die diagnostische Genauigkeit eines Tests mittels einer Fall-Kontroll Studie erhoben worden wäre, in der schwerkranke PatientInnen *mit* Zielerkrankung mit *gesunden* PatientInnen verglichen werden [38, 47].

Disease progression bias (recovery bias) & Therapieparadoxon

unterschiedlicher Zeitpunkt wann Index- und Referenztest durchgeführt werden

Normalerweise sollten die Ergebnisse von Referenz- und Indextest für dieselben PatientInnen zum *gleichen* Zeitpunkt erhoben werden [38]. Ist dies nicht der Fall, dann kann durch die Veränderung des Krankheitszustandes (z.B. Heilung, Progression der Erkrankung) [46], das Ergebnis falsch eingeschätzt werden [38]. Disease progression bias ist dann sehr wahrscheinlich, wenn der Indextest mit deutlichem zeitlichen Abstand vor dem Referenzstandard durchgeführt worden war und die Erkrankung daher weiter fortgeschritten ist [46]. Wenn aufgrund eines positiven Ergebnisses des Indextests eine Behandlung eingeleitet wurde und daher der nachfolgende Referenztest negativ ausfällt, spricht man von einem „Therapieparadoxon" [38, 46].

Mangelhafter Referenzstandard

mangelhafter Referenzstandard kann zu Unter-, aber auch zu Überschätzung der diagnostischen Genauigkeit führen

Um die diagnostische Genauigkeit eines Tests festzustellen, wird der Indextest mit dem derzeitigen Referenzstandard (idealer Weise dem Goldstandard) verglichen. Kommen die beiden Tests zu abweichenden Ergebnissen, wird angenommen, dass der Indextest PatientInnen falsch klassifiziert hat [38]. Wird nun der Indextest mit einem ungeeigneten Referenztest verifiziert, kommt es zu einer Verzerrung der diagnostischen Genauigkeit: einerseits kann es sein, dass der Indextest PatientInnen richtig klassifiziert, der Referenztest aber selbst ungenau ist, wodurch es zu einer Unterschätzung der Genauigkeit des Indextests kommt. Auf der anderen Seite kann aber auch das Ergebnis des Indextests fälschlicherweise durch den Referenztest

bestätigt werden, wodurch die Testperformance überschätzt werden würde [46].

noch problematischer, wenn Referenzstandard fehlt

Neben der Auswahl eines ungeeigneten Referenzstandards, ist ein größeres Problem, wenn gar kein Referenztest oder Goldstandard zur Verifizierung zur Verfügung steht [2]. Mehrere Methoden, um mit diesem Problem umzugehen, werden beschrieben, allerdings ist dies nach wie vor Gegenstand wissenschaftlicher Forschung [49-52]. Eine Methode bei fehlendem Referenzstandard wäre, zum Beispiel, die Indextestergebnisse mit daraus resultierenden klinischen Konsequenzen zu validieren (z.B. ob der Indextest vorhersagen kann, welche Patientengruppe tatsächlich von einer Therapie profitieren wird)[52]. Bei ungenauem Referenzstandard können etwa Sensitivitätsanalysen eine Möglichkeit darstellen, Unterschiede in den Ergebnissen der diagnostischen Genauigkeit zu berücksichtigen [51].

Verificationsbias

Verificationsbias kann dann entstehen, wenn die Verifizierung mittels Referenzstandards abhängig von dem Ergebnis des Indextests ist [34]. Man unterscheidet:

unterschiedliche Test für Verifizierung

Differential Verificationsbias: in Abhängigkeit des Indextestergebnisses werden unterschiedliche Referenztests zur Verifizierung verwendet. Differential Verificationbias kann also etwa entstehen, wenn bei Verdacht auf Krebs nur Indextest-positive PatientInnen mittels des derzeitige Referenzstandards verifiziert werden (z.B. eine histologische Untersuchung), während der tatsächliche Gesundheitszustand von Indextest-negativen PatientInnen lediglich durch eine klinische Routineuntersuchung überprüft wird - die Genauigkeit der beiden Vergleichstest also unterschiedlich ist [34, 38, 46].

nur ein Teil der PatientInnen verifiziert wurden

Partial Verificationsbias: nur ein Teil der Indextestergebnisse werden mittels Referenzstandard überprüft [34, 35, 38, 46]. Problematisch wird dies vor allem dann, wenn PatientInnen für die Überprüfung nicht zufällig ausgewählt werden, sondern die Auswahl durch das Ergebnis des Indextest beeinflusst worden war. Dieser Bias kann zum Beispiel auftreten, wenn es sich bei dem Referenzstandard um einen invasiven Test handelt und daher nur Indextest-negative PatientInnen verizifiert werden [34].

Review bias

Interpretation der Testergebnisse ist nicht verblindet

Review Bias kann entstehen, wenn die Interpretation von Ergebnissen des Indextests oder des Referenzstandards durch das bekannte Resultat des jeweils anderen Tests beeinflusst wird [46, 53]. Dieser Bias ist besonders häufig bei Tests, die Gegenstand individueller Bewertung (z.B. Interpretation eines Röntgenbildes durch einen RadiologIn im Unterschied zur technischen Bewertung von Laborparametern) sind, könnten aber vermieden werden, indem alle Tests unter Verblindung und anhand eines standardisierten Protokolls durchgeführt werden [20, 38, 53].

Unter *Test Review Bias* wird die Verzerrung der Interpretation des Indextests bei bekanntem Ergebnis des Referenzstandards verstanden.

Unter *Diagnostischer Review Bias* wird die Verzerrung der Interpretation des Referenzstandards bei bekanntem Ergebnis des Indextests verstanden. [38].

Klinischer Review Bias

Interpretation des Testergebnisses wird durch klinische Daten beeinflusst

Die Interpretation eines Testergebnisses wird durch das Wissen um klinische Daten (z.B. Geschlecht, Alter, Ko-Morbidität) beeinflusst. Wird die diagnostische Genauigkeit eines Tests durch zusätzliche klinische Informationen verbessert, so sollten in Studien auch nur jene Patienteninformationen bekannt sein, die auch in der klinischen Praxis vorliegen würden [38, 46].

Inkorporations Bias

Indextestergebnis fließt in Referenzstandard ein

Inkorporationsbias entsteht, wenn das Indextestergebnis zur endgültigen Diagnosefindung herangezogen wird, dessen Ergebnis also in den Referenzstandard einfließt [38, 46, 53]. Da die angenommene „wahre" Diagnose, gegen die Indextestergebnisse verglichen werden, nicht mehr unabhängig von dem Indextest ist, kann es zu einer Verzerrung der diagnostischen Genauigkeit kommen [20, 53]. Ein Beispiel für Inkorporationsbias wäre, wenn die Genauigkeit von Magnetresonanztomographie (MRT) zur Diagnose von multipler Sklerose untersucht werden soll, der Referenzstandard aber durch die Zusammenschau sämtlicher Informationen definiert ist und somit auch das MRT Ergebnis ein Teil der Referenzstandards wäre.

Uneindeutige Testergebnisse

Testergebnisse sind nicht eindeutig, oder schlecht beurteilbar

Sowohl Indextest, als auch Referenzstandard, können manchmal Ergebnisse liefern, die nicht eindeutig sind, zum Beispiel wenn Computertomographien durch Bewegungsartefakte schlecht beurteilbar sind [46]. Die Häufigkeit mit der uneindeutige Ergebnisse in Studien auftreten, sollte unbedingt angegeben werden. Sind diese nicht-interpretierbaren Ergebnisse zufällig und damit unabhängig vom tatsächlichen Erkrankungsstatus, so ist keine systematische Verzerrung der Testgenauigkeit zu erwarten. Sind sie allerdings mit dem Vorliegen oder der Abwesenheit der Zielerkrankung korreliert, können die Ergebnisse verfälscht werden [38].

Tabelle 4.2-5: Überblick über mögliche Quellen für Bias (adaptiert nach: [46])

Quelle des Bias	Art des Bias
Testprotokoll: Material und Methoden	
Therapieparadoxon und Disease Progression Bias	Therapieparadoxon: die Therapie wird aufgrund des Ergebnisses des Indextests eingeleitet, der Referenztest wird erst nach Beginn der Behandlung angewandt.
	Disease progression bias: Indextest wird früher als der Referenztest angewandt; bei Verwendung des Referenztests ist die Krankheit bereits in einem anderen Stadium
Limited challenge bias	Einschluss der PatientInnen erfolgte nicht konsekutiv, schwieriger zu diagnostizierende Gruppen wurden ausgeschlossen
Referenzstandard und Verifzierung	
Mangelhafter Referenzstandard	Verzerrung der Genauigkeit des Indextest bei Verifizierung mittels eines ungeeigneten Referenzstandards
Differential Verification Bias	Ergebnisse des Indextests werden durch unterschiedliche Referenzstandards bewertet
Partial Verification Bias	Nur ausgewählte Indextestergebnisse werden mittels Referenzstandard verglichen
Interpretation	
Review bias	Interpretation des Index-/Referenztests werden durch bekannte Testergebnisse des jeweils anderen Verfahrens verfälscht
Klinischer Review Bias	Interpretation von Testergebnissen wird durch bekannte klinische Daten verfälscht
Incorporation Bias	Ergebnisse des Indextests werden zur Bestätigung der Diagnose verwendet
Analyse	
Bewertung von uneindeutigen Testergebnissen	Uneindeutige Testergebnisse, die von der Analyse der Testgenauigkeit ausgeschlossen werden, können zur Verzerrung der Ergebnisse führen

Variation

Spectrum Effekt

Unterschiede des Patientenspektrums können zu abweichenden Ergebnissen der Testgenauigkeit führen

Bedingt durch Charakteristika wie Alter, Geschlecht, Ko-Morbidität oder durch unterschiedliche klinische Settings kann die Testgenauigkeit variieren [16, 38, 46, 54]. Auch durch die Einschlusskriterien einer Studie kann das eingeschlossene Patientenspektrum von dem, für die klinische Praxis relevantem, deutlich abweichen.

Unter Spektrum Effekt wird verstanden, wenn die Testgenauigkeit in verschiedenen Patientengruppen variiert, eine Studie aber nur ein eingeschränktes Patientenspektrum evaluiert [16]. Die häufigste Methode Spektrum-Effekte zu vermeiden, besteht daher darin, möglichst klinisch relevante Patientengruppen einzuschließen. Unklar ist noch ob es sich dabei um ei-

nen Bias, oder lediglich um eine mögliche Ursache von Variation handelt. Manche Autoren argumentieren, dass es sich nicht um einen Bias handelt, da die Testgenauigkeit unweigerlich von Patientencharakteristika abhängt. Wichtig ist nur, dass die zugrundeliegenden Ursachen für Spektrum Effekte in Studien identifiziert und klar beschrieben werden [16].

Generalisierbarkeit der Testergebnisse abhängig vom Spektrum der Studienpopulation

Um die Generalisierbarkeit von Studienergebnissen also zu gewährleisten, sollte die Studienpopulation daher jene PatientInnen umfassen, die mit dem fraglichen Test auch im klinischen Alltag untersucht werden. Besonderes Augenmerk sollte deshalb auf die, in Studien verwendeten Ein- und Ausschlusskriterien, und auf die Patientencharakteristika gelegt werden. Bei Letzteren sind besonders wichtig:

- demographische Charakteristika
- Schwere der Erkrankung/Symptome
- Ko-Morbidität
- klinisches Setting
- Krankheitsprävalenz [38].

Demographische Charakteristika

Spektrum Effekt durch:

- Alter, Geschlecht

Charakteristika wie Geschlecht, Alter, ethnische Zugehörigkeit oder Raucherstatus können die Genauigkeit beeinflussen [46].

Schwere der Erkrankung

- Schwere der Erkrankung,

Je nach Stadium der Erkrankung innerhalb der untersuchten Bevölkerungsgruppe kann die Genauigkeit eines Tests variieren: die Sensitivität nimmt mit zunehmender Schwere der Erkrankung zu, wohingegen die Spezifität vermutlich nicht beeinträchtigt wird [46]. Erklärt werden kann dies zum Beispiel so, dass etwa weiter fortgeschrittene Tumore durch bildgebende Verfahren leichter zu diagnostizieren sind, als solche in früheren Stadien [38, 55].

Prävalenz & klinisches Setting

- Prävalenz und klinisches Setting

Die Prävalenz einer Erkrankung ist abhängig vom klinischen Setting und kann so zu Unterschieden der Testgenauigkeit führen [35] (siehe auch Kapitel 4.2.1). So ist es in der allgemeinmedizinischen Praxis, einem Bereich mit niedriger Prävalenz, eher wahrscheinlich, dass ein Testnegativer gesund ist (NPV) wohingegen ein Testpositiver mit geringer Wahrscheinlichkeit krank ist (PPV). Somit besteht also auch ein höheres Risiko eines falsch positiven Testergebnisses. Mit fortschreitender Selektion des Patientengutes – vom niedergelassenen SpezialistIn, über Landeskrankenhäuser bis hin zu Schwerpunktkrankenhäusern, kehrt sich dieser Trend aber um, da die *a priori* - Wahrscheinlichkeit (Prävalenz) immer weiter ansteigt [19, 26]. Von weiterer Bedeutung ist dieser Zusammenhang, da Tests die in der Gesamtbevölkerung mit niedriger Prävalenz durchgeführt werden, nur zu wenigen entdeckten Fällen führen würden. Wird der Test aber in einer Risikopopulation, in der die Krankheitsprävalenz höher ist, durchgeführt, werden mehr „Fälle“, bezogen auf die Anzahl der durchgeführten Tests, entdeckt.

Einen weiteren Einfluss kann die Prävalenz auch auf die Interpretation eines Testergebnisses haben, da UntersucherInnen in Settings mit einer höheren Prävalenz der Erkrankung eher dazu neigen, ein Testergebnis als positiv

zu befunden, als UntersucherInnen in einem Niedrigprävalenzbereich [46]. Etwa wird ein ChirurgIn anhand eines Ultraschallbefundes unter Umständen häufiger eine Appendizitis diagnostizieren, als ein AllgemeinmedizinerIn.

Ko-Morbidität

- Ko-Morbidität

Zusätzlich zur Zielerkrankung vorhandene Erkrankungen können die Testgenauigkeit beeinflussen. Ein Beispiel wäre, wenn etwa ein H2-Atemtest bei Verdacht auf Laktose-Intoleranz aufgrund einer gleichzeitig bestehenden Lungenfunktionsstörung ein falsch negatives Ergebnis zeigen würde [38].

Testdurchführung

Testdurchführung abhängig von Expertise der UntersucherInnen

Indextest und Referenztest müssen ausführlich beschrieben sein, um etwaige, durch die Testdurchführung bedingte Unterschiede berücksichtigen zu können [46]. Vor allem bei Tests, für deren Durchführung ein gewisses Maß an Expertise benötigt wird, ist dies wichtig (z.B. Endoskopie) [46, 55].

Unterschiedliche Technologien

neuere Varianten eines Tests

Bedingt durch technische Erneuerungen und neuere Versionen eines Tests kann sich mit der Zeit die diagnostische Genauigkeit verändern [26, 46, 55].

Beobachtervariabilität

Interpretation der Ergebnisse abhängig von Einschätzung durch BeobachterInnen

Beobachtervariabilität kommt durch Abweichungen in der Interpretation eines Testergebnisses durch UntersucherInnen zustande.

Unter intraobserver Variabilität wird die unterschiedliche Einschätzung mehrerer Testergebnisse eines einzelnen Tests durch ein und denselben Beobachter verstanden (z.B. bei der Beurteilung von Röntgenbildern). Gemessen wird ihr Ausmaß, indem identische Testergebnisse (z. B. eine Auswahl von Röntgenbildern) der gleichen Person mehrfach zur Befundung vorgelegt werden.

Im Gegensatz dazu interpretieren bei der interobserver Variabilität zwei verschiedene Beobachter ein Testergebnis unterschiedlich [2, 46].

Beobachtervariabilität ist also dann besonders häufig, wenn die Interpretation eines Ergebnisses von der subjektiven Einschätzung eines Befundes abhängt oder durch die Expertise des UntersucherIn beeinflusst wird [46, 55].

Arbiträre Wahl des Grenzwertes

explizite Grenzwerte, wenn unterschiedliche Definition von „gesund" und „krank"

implizite Grenzwerte durch abweichende Einschätzungen der UntersucherInnen

Wie in Kapitel 4.2.1 erwähnt, verändern sich Sensitivität und Spezifität in Abhängigkeit des Grenzwerts anhand dessen zwischen gesund und krank unterschieden wird. Prinzipiell kann zwischen einem expliziten und einem impliziten Grenzwert differenziert werden. Ein expliziter Grenzwert ist dann gegeben, wenn zum Beispiel Studien unterschiedliche Definitionen für „krank" und „gesund" verwenden. Implizite Grenzwerte können dann bestehen, wenn etwa ein Testergebnisses abhängig von der individuellen, beobachterabhängiger Einschätzung ist und sich die, von den einzelnen UntersucherInnen festgelegte Grenzwerte ab dem ein Test als positiv bewertet wird, unterscheiden [25, 35]. So ist es etwa möglich, dass sich Unterschiede zwischen UntersucherInnen bei der Befundung von Mammographien ergeben. Ein RadiologIn befundet bei Vorliegen einer nur minimalen Verschattung eine Mammographie als unauffällig, während ein anderer zumindest den Verdacht auf Brustkrebs äußert.

Kann in diagnostischen Studien dieser Grenzwert nun beliebig gewählt werden, oder wurde der Grenzwert erst nach Bekanntwerden der Studienergebnisse festgelegt [48], besteht die Gefahr, dass der Grenzwert so gelegt wird, dass Sensitivität und Spezifität innerhalb der Studienpopulation maximiert werden. Die erhaltenen Werte wären dann aber nicht mehr auf andere Patientengruppen übertragbar [35, 46, 55].

Tabelle 4.2-6: Überblick über mögliche Quellen für Variation (adaptiert nach: [46])

Quelle der Variation	Art der Variation
Population	
Spektrumeffekt	
- Demographische Charakteristika	Charakteristika wie Geschlecht, Alter, ethnische Zugehörigkeit können die Testperformance beeinflussen
- Schwere der Erkrankung	Die Schwere der Erkrankung kann zu unterschiedlichen Einschätzungen der Testperformance führen
- Prävalenz	Die Prävalenz einer Erkrankung variiert in Abhängigkeit vom klinischen Setting. In Settings mit hoher Prävalenz einer Erkrankung, tendieren UntersucherInnen daher eher dazu ein Testergebnis als positiv zu bewerten, als in Settings mit niedriger Prävalenz.
- Selektion der Studienpopulation	Je nach Selektionsprozess der Studienpopulation wird diese variieren. Führt dieser Prozess nicht zu der Auswahl einer Population, die repräsentativ für die Population ist, in der der Test in der klinischen Praxis angewandt werden soll, ist die Umlegbarkeit der Studienergebnisse reduziert.
Testprotokoll: Material und Methoden	
Testdurchführung	Diagnostische Genauigkeit eines Tests kann durch Unterschiede in der Ausführung sowohl des Index- als auch des Referenztests zustande kommen
Technologie	Testperformance kann sich durch Verbesserungen der Technologie selbst oder durch zunehmende Erfahrung des UntersucherIn verändern
Interpretation	
Beobachtervariabilität	Intraobserver Variabilität entsteht, wenn ein und derselbe Beobachter zu unterschiedlichen Ergebnisse kommt, Interobserver Variabilität wenn zwei unterschiedliche Beobachter zu abweichenden Ergebnissen gelangen
Analyse	
Arbiträrer Grenzwert	Wird der Grenzwert so gelegt, dass Sensitivität und Spezifität optimiert werden, kann die Testperformance überschätzt werden, und nicht die Testperformance für die geplante Zielpopulation widerspiegeln

4.2.6 Bewertung der methodischen Qualität diagnostischer Studien

für Beurteilung der methodischen Qualität von diagnostischen Genauigkeitsstudien stehen eigene Instrumente zur Verfügung

Dadurch, dass die Bewertung einer diagnostischen Technologie in den meisten Fällen auf diagnostischen Genauigkeitsstudien beruhen wird (siehe Kapitel 4.4.2) kommt der Beurteilung der Validität der Ergebnisse eine besondere Bedeutung zu [56].

Wie bereits erwähnt (siehe Kapitel 4.2.5), kann die interne und die externe Validität diagnostischer Genauigkeitsstudien durch etliche Faktoren beeinträchtigt werden. Zahlreiche Instrumente stehen zur Verfügung, um die einzelnen Aspekte zu bewerten, wobei aber keines als allgemein gültiger Standard akzeptiert ist [56].

Im folgenden Abschnitt sollen drei der in den Literaturstellen am häufigsten angeführten Instrumente kurz vorgestellt werden.

QUADAS-Instrument

QUADAS – Instrument enthält insgesamt 14 Fragen, davon sollten für eigene Fragestellung relevante ausgewählt werden

Das QUADAS (= Quality assessment tool for diagnostic accuracy studies) Instrument (siehe Appendix 8.1) wurde zur Bewertung der methodischen Qualität diagnostischer Genauigkeitsstudien für systematische Übersichtsarbeiten entwickelt [57], wobei unter Qualität sowohl die Einschätzung der internen, als auch der externen Validität verstanden wird. Dies bedeutet, dass auf der einen Seite eine Beurteilung der durch Bias verursachten Verzerrung und auf der anderen Seite die Beurteilung der Übertragbarkeit der Ergebnisse auf die klinische Praxis ermöglicht werden soll (siehe Tabelle 4.2-7).

5 der Fragen sind aber immer zu bewerten

Insgesamt enthält das QUADAS - Instrument 14 Fragen, die aber in Abhängigkeit von der Relevanz für die Forschungsfrage der systematischen Review adaptiert werden sollen. Fünf der 14 Fragen sind aber immer zu bewerten. Diese sind:

1. Ist das eingeschlossene Patientenspektrum repräsentativ für die Zielpopulation im klinischen Alltag?
2. Sind die Einschlusskriterien klar definiert?
3. Wird die Zielerkrankung durch den Referenzstandard richtig diagnostiziert?
4. Wurden nicht interpretierbare Testergebnisse angegeben?
5. Wurden Studienabbrüche begründet?

In den meisten Fällen relevant für die Bewertung der Qualität sind ferner:

6. Erhielten alle PatientInnen unabhängig vom Ergebnis des Indextests den Referenzstandard?
7. War die Ausführung des Referenzstandards detailliert genug beschrieben, um ihn wiederholen zu können?

Die meisten Fragen von QUADAS beziehen sich auf potentielle Bias, zwei auf mögliche Variationen und zwei auf die Qualität der Berichterstattung [57].

Cochrane Collaboration

Cochrane Checkliste basiert auf QUADAS, enthält 11 Fragen

Die Cochrane Collaboration empfiehlt eine auf dem QUADAS Instrument basierende Checkliste zur Qualitätsbewertung, wobei aber nur 11 der insgesamt 14 Fragen im „Handbook for Systematic Reviews of Diagnostic Accuracy“ angeführt werden (siehe Tabelle 4.2-7 und Appendix 8.2) [38].

Von den angeführten Kriterien werden als besonders wichtig erachtet:

1. ob eine repräsentative Patientenpopulation untersucht wurde.
2. die Methode, mit der die Indextestergebnisse verifiziert wurden.
3. ob die Testergebnisse unter Verblindung bewertet wurden.
4. wie mit fehlenden Daten umgegangen wurde.

Weitere neun Fragen stehen zur Verfügung, die je nach Fragestellung hinzugefügt werden können.

STARD – Initiative

STARD, um Berichterstattung von diagnostischen Genauigkeitsstudien zu verbessern und Beurteilung der internen und externen Validität zu erleichtern

Die “Standards for Reporting of Diagnostic Accuracy” (STARD) Initiative wurde gegründet, um die Qualität der Berichterstattung von diagnostischen Genauigkeitsstudien zu verbessern, und so die Beurteilung der internen und externen Validität zu erleichtern [58] (siehe Appendix 8.3). Durch die 25 Punkte der STARD Checkliste soll garantiert werden, dass der Studienbericht eine Einschätzung des Potentials für mögliche Bias und Variationen erlaubt und daher die Bewertung der Relevanz der Ergebnisse für die eigene Fragestellung ermöglicht wird (siehe Tabelle 4.2-7).

Tabelle 4.2-7: Übersicht über in QUADAS - Instrument, Cochrane Collaboration und STARD- Checkliste bewertete Bias

Bias	QUADAS	Cochrane	STARD
Limited Challenge Bias		+	+
Disease Progression Bias & Therapieparadoxon	+	+	+
Mangelhafter Referenzstandard	+	+	+
Verification bias (partial & differential)	+	+	+
Review Bias	+	+	+
Klinischer Review Bias	+	+	
Inkorporations Bias	+	+	
Uneindeutige Testergebnisse	+	+	+

Weitere Instrumente

In einer systematischen Übersichtsarbeit wurden insgesamt 90 verschiedene Instrumente identifiziert, die zur Bewertung von diagnostischen Genauigkeitsstudien angewandt werden können [56] (für weitere Beispiele siehe Appendix 8.4).

4.2.7 Systematische Reviews & Meta-Analysen zu diagnostischer Genauigkeit

prinzipielle Abläufe von systematischen Reviews und Meta-Analysen zu diagnostischer Genauigkeit gleich wie bei therapeutischen Interventionen

Wie bereits erwähnt werden auch bei Studien zu diagnostischer Genauigkeiten systematische Übersichtsarbeiten als höchste Evidenzstufe angesehen (siehe Kapitel 4.2.2). Obwohl der reine Nachweis der diagnostischen Genauigkeit nicht ausreichen sollte, um ein diagnostisches Untersuchungsverfahren in die klinische Praxis zu übernehmen, sondern zumindest der Nutzen für PatientInnen etabliert sein sollte, sind diese Studien kaum verfügbar [6], sodass systematische Reviews über Testgenauigkeit häufig sind.

Diese systematischen Reviews unterscheiden sich in den grundlegenden Abläufen nicht von denen zu therapeutischen Interventionen: Definition einer Fragestellung, von Ein- und Ausschlusskriterien, Literatursuche, Auswahl und Bewertung der methodischen Qualität der eingeschlossenen Publikationen, Datenextraktion und – analyse und letztlich Evidenzsynthese [21, 36, 59].

besonders zu beachten: klare Definition des Referenzstandards, der Patientenpopulation und des klinischen Settings

Trotz allem müssen bei systematischen Reviews über diagnostische Genauigkeitsstudien zahlreiche Besonderheiten beachtet werden (siehe auch Kapitel 4.2.5).

- *Einschlusskriterien:* Da die Genauigkeit eines Tests kein absoluter Wert ist, sondern zum Beispiel abhängig vom Patientenspektrum (siehe Kapitel 4.2.5) ist, sollten der Referenzstandard, die Population und das klinische Setting als erste Einschlusskriterien klar in der Fragestellung formuliert werden [36, 59].

Bewertung der methodischen Qualität anhand der erwähnten Instrumente

- *Methodische Qualität:* Zahlreiche mögliche Bias von diagnostischen Genauigkeitsstudien können zu einer Verzerrung der Ergebnisse führen (siehe Kapitel 4.2.5). Instrumente zur Bewertung der methodische Qualität diagnostischer Genauigkeitsstudien sind daher nötig (siehe Kapitel 4.2.6) [6, 60], wobei Studien von minderer Qualität erst gar nicht berücksichtigt werden müssen [21].

Übertragbarkeit der Ergebnisse auf eigenes Setting

- *Übertragbarkeit der Ergebnisse:* Die Testgüte wird durch Charakteristika der Studienpopulation (Schweregrad der Erkrankung, klinisches Setting, Vortestung, etc.) beeinflusst, sodass die Übertragbarkeit diagnostischer Genauigkeitsstudien auf andere klinische Settings (externe Validität) geprüft werden muss. Auch hierfür eignet sich unter anderem das QUADAS-Tool. Wichtig für die Aussagekraft einer systematischen Review ist daher, ob die in den Studien eingeschlossenen PatientInnen, mit der in der Fragestellung definierten Population vergleichbar ist [6].

Meta-Analysen bei homogenen Ergebnissen

- *Datenpooling:* Wie auch bei Interventionsstudien können die Daten aus den Primärstudien im Rahmen von Meta-Analysen zusammengefasst werden, um dadurch eine verlässlichere Einschätzung der Effektgröße zu erhalten [28]. Generell gilt, dass die Durchführung von Meta-Analysen diagnostischer Genauigkeitsstudien abhängig von der Anzahl und der Qualität der Primärstudien, als auch vom Ausmaß der Heterogenität (Unterschiedlichkeit) ist [21], da Meta-Analysen generell nur durchgeführt werden sollten, wenn die Ergebnisse der Primärstudien homogen sind, also ähnlich sind.

Heterogenität auch durch Unterschiede, der in Studien verwendeten Grenzwerte möglich

Mittels statistischer Methoden sollte zunächst bewertet werden, wie ähnlich die Studienergebnisse sind, beziehungsweise ob die Unter-

schiede nur durch Zufallsschwankungen verursacht wurden. Findet man Hinweise, dass die Ergebnisse deutlich voneinander abweichen, so ist es bei Diagnosestudien besonders wichtig, zu bewerten, ob diese Unterschiede möglicherweise durch verschiedene, in den Primärstudien verwendete Grenzwerte bedingt sind (siehe Kapitel 4.2.5).

Für das Poolen von Daten stehen dann, je nachdem wodurch Heterogenität entstanden ist, unterschiedliche statistische Modelle zur Verfügung (für Details zu Meta-Analysen für Diagnosestudien siehe z.B. [25, 36, 61, 62]).

4.3 Level 3 & Level 4 – diagnostischer/therapeutischer Impakt

ob Testergebnis das therapeutische oder diagnostische Prozedere verändert

Die alleinige Bewertung der diagnostischen Genauigkeit eines Tests ist nicht ausreichend, um ein diagnostisches Verfahren zu bewerten, da das weitere Patientenmanagement nicht zwangsläufig durch ein Testergebnis beeinflusst werden muss [63]. Zum Beispiel kann der durch den Indextest erzielte Informationsgewinn nicht ausreichen, um den KlinikerIn zu einer Änderung des diagnostischen oder therapeutischen Vorgehens zu bewegen.

diagnostischer Impakt durch z.B. schnellen oder weniger invasiven Test

Der Einfluss des Indextests auf das diagnostische Prozedere kann etwa ein genaueres, schnelleres oder kostengünstigeres Testergebnis sein, oder, dass der Indextest weniger invasiv als ein anderer Tests ist und daher von den PatientInnen leichter angenommen wird. Auf der anderen Seite können aber auch differentialdiagnostische Überlegungen verändert werden [5]. Studien, die den diagnostischen Impakt untersuchen, vergleichen also unterschiedliche Aspekte der Diagnosefindung - einmal mit und einmal ohne den Indextest [20].

therapeutischer Impakt, wenn Therapieplan durch Testergebnis verändert wird

Hinterfragt werden kann auch, ob das therapeutische Prozedere durch den Indextest beeinflusst wird, z.B. ob eine Therapie früher eingeleitet wird oder ob es zu einer Veränderung des Therapieplans kommt, beziehungsweise ob eine völlig andere Therapie verabreicht wird [20, 63].

4.3.1 Studiendesigns

bevorzugtes Design: RCT, aber selten

Das beste Studiendesigns um den diagnostischen/therapeutischen Impakt zu untersuchen, sind RCTs [20, 63]. Da für RCTs allerdings große Patientenzahlen und eine gewisse Vorlaufzeit benötig werden, bis UntersucherInnen mit dem neuen Untersuchungsverfahren vertraut sind und diagnostische Technologien auch einen relativ kurzen Lebenszyklus haben, sind sie *in realiter* oft nicht durchführbar oder sinnvoll [63] und daher auch nur sehr selten verfügbar [64].

Alternative: Vorher-Nachher Studien

Als Alternative stehen dann Vorher-Nachher-Studien zur Verfügung, bei denen die von KlinikerInnen erstellten diagnostischen und therapeutischen Strategien vor dem Indextestergebnis, mit denen nach Bekanntwerden des Indextestergebnisses verglichen werden [20, 63]. Diesem Studiendesign sind allerdings zahlreiche methodische Schwächen inhärent, von denen einige durch sorgfältige Planung und Durchführung der Studie vermieden werden können, andere aber nicht. So ist dieses Studiendesign zwar geeignet um ei-

nen einzelnen Add-On Test zu bewerten (siehe Kapitel 4.2.4), aber nicht um mehrere Tests miteinander zu vergleichen [35, 63]. Ebenso kann sich das von den ÄrztInnen angegebene diagnostische/therapeutische Vorgehen davon unterscheiden, was sie tatsächlich im klinischen Alltag gemacht hätten. Gerade bei invasiven Therapien, wie etwa Operationen, kann es leichter sein, als geplantes Prozedere ein operatives Vorgehen als Therapie der Wahl anzugeben, da der/die Arzt/Ärztin weiß, dass er diese Entscheidung in Abhängigkeit des Testergebnisses noch revidieren könnte. Wenn sich KlinikerInnen gleich (ohne Test), entscheiden müssten, würde sie die Indikation für eine Operation möglicherweise restriktiver stellen. Außerdem kann ein kausaler Zusammenhang zwischen Testergebnis und einhergehenden Veränderungen des Patientennutzens durch dieses Design nicht etabliert werden [63].

bedingt durch methodische Schwächen ist Stellenwert dieser Studien unklar

Bedingt durch diese methodischen Schwächen sind diese Studienarten nicht unumstritten. Eine Veränderung des therapeutischen Vorgehens impliziert nicht automatisch Verbesserungen für PatientInnen und auch wenn das Testergebnis Verhaltensweisen von ÄrztInnen nicht verändern kann, gibt dies eher Aufschluss über die behandelnden ÄrztInnen, als über den Test selbst und sein Potenzial patientenrelevante Endpunkte positiv beeinflussen zu können [1, 66].

4.4 Level 5 – patientenrelevanter Nutzen

patientenrelevanter Nutzen eines Tests anhand klinischer Endpunkte, die direkt durch den Test selbst, aber auch durch nachfolgende Therapie beeinflusst werden

Die Verbesserung von patientenrelevanten Endpunkten ist letztlich das eigentliche Ziel jedwedes diagnostischen Untersuchungsverfahren [65] wobei in erster Linie klinische Parameter als Endpunkte herangezogen werden (z.B. Lebenserwartung, Funktionsfähigkeit, Schmerz) [4, 7, 65]. Der aus einem Test resultierende Netto-Nutzen wird einerseits durch den Test selbst bestimmt, indem etwa invasivere Untersuchungsmethoden vermieden werden, oder der Test mit schwerwiegenden Komplikationen einhergehen kann. Andererseits wird der klinische Nutzen eines Tests durch die, einem Testergebnis nachfolgende Therapie bestimmt, die neben positiven Effekten aber auch wiederum mit Nebenwirkungen vergesellschaftet ist [45]. Neben den Konsequenzen für richtig positive und richtig negative Befunden, sollten in diesem Zusammenhang immer auch die aus falsch positiven oder falsch negativen Befunden resultierenden Konsequenzen berücksichtigt werden.

wichtig sind auch Konsequenzen von falschen Befunden

weitere Endpunkte: soziale, emotionale, kognitive Endpunkte

Neben klinischen Endpunkten können aber auch emotionale, soziale, verhaltenstechnische und kognitive Endpunkte relevant sein: die Art der Testdurchführung und die Testergebnisse selbst können neben Ängsten und Depressionen auch zu Veränderungen in Beziehungen führen, oder Auswirkungen auf die Compliance und damit auch auf klinische Ergebnisse (z.B. Häufigkeit mit der Nachfolgeuntersuchungen in Anspruch genommen werden) haben. Diese Aspekte sind vor allem bei solchen Tests wichtig, für die nur geringe Auswirkungen auf klinische Endpunkte gefunden wurden, da dann ein minimaler positiver Einfluss auf klinisch relevante Ergebnisse zunichte gemacht werden kann [7].

Letztlich gilt es also festzustellen, ob der Nutzen eines Tests größer ist, als dessen Risiken [20].

4.4.1 Direkte Evidenz

patientenrelevante Endpunkte von Indexteststrategie und Therapie werden mit Referenzteststrategie und Therapie verglichen

Das beste Design, um den für PatientInnen entstehenden Nutzen zu erheben wären – wie auch bei Interventionsstudien - RCTs, die direkt die patientenrelevanten Ergebnisse einer Strategie mit dem herkömmlichen Test, inklusive seiner therapeutischen Konsequenzen mit denen einer Strategie bestehend aus dem Indextest, einschließlich seiner therapeutischen Konsequenzen, vergleicht [20, 64-66]. Die Bewertung der methodischen Qualität dieser Studien unterscheidet sich dann nicht von der von Interventionen.

Andere Studiendesigns von minderer Qualität sind Kohortenstudien, Fall-Kontroll-Studien oder Fallserien [5, 65].

4.4.2 Linked evidence

indirekter Nachweis patientenrelevanten Nutzens mittels „linked Evidence"

Wie erwähnt, wird der Einfluss eines diagnostischen Verfahrens auf patientenrelevante Ergebnisse am besten mittels RCT untersucht. Aus Ressourcengründen werden diese Studien aber selten durchgeführt. Unter gewissen Voraussetzungen lassen sich indirekt Rückschlüsse auf den Patientennutzen eines diagnostischen Test ziehen, wenn die Ergebnisse von diagnostischen Genauigkeitsstudien mit den Ergebnissen von Wirksamkeitsstudien verknüpft werden. Diese Methode wird „linked Evidence" genannt [64, 67].

= Verknüpfung von diagnostischer Genauigkeitsstudie mit Wirksamkeitsstudie

Das Prinzip von „linked Evidence" ist also die Verknüpfung von Wirksamkeitsdaten aus hochwertigen, vergleichenden Therapiestudien mit übertragbarer und hochwertiger Evidenz aus diagnostischen Genauigkeitsstudien, die den Indextest mit der derzeit existierenden Teststrategie vergleichen (siehe Abbildung 4.4-1) [67].

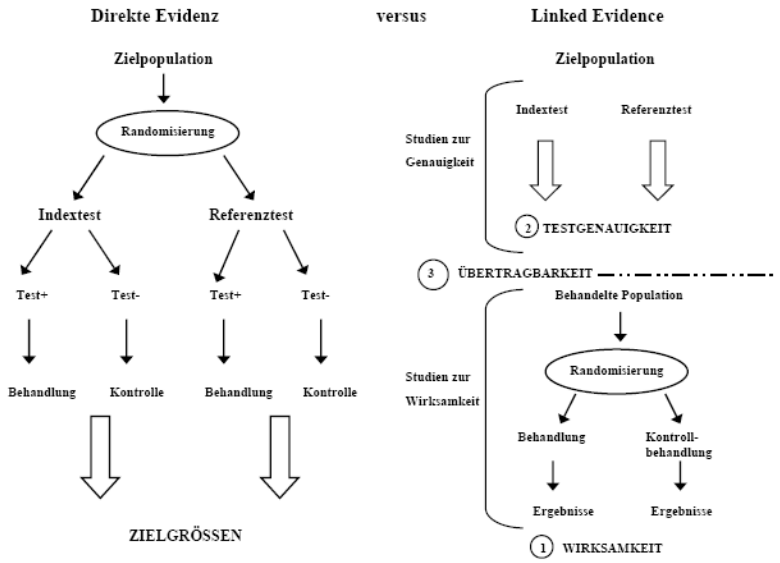


Abbildung 4.4-1: Direkte Evidenz versus „linked Evidenz" (Quelle: [68])

Folgende Voraussetzungen für „linked Evidence“ müssen gegeben sein:

Voraussetzungen:

therapeutische Effektivität in hochwertigen Studien belegt

1. Der Nutzen der Therapie für eine über ein spezifisches Testergebnis definierte Population muss durch hochwertige Studien, idealerweise durch RCTs, nachgewiesen sein. Ist dies nicht der Fall, dann kann dieser Nachweis gegebenenfalls durch systematische Reviews erbracht werden [67]. In manchen Fällen muss zusätzlich erhoben werden, ob ein Testergebnis auch tatsächlich das weitere diagnostische/therapeutische Management beeinflusst. Dies ist aber dann unnötig, wenn der Indextest als Ersatz eines anderen Tests gedacht ist, da dann bereits ein klarer Zusammenhang zwischen Testergebnis und Therapieentscheidung vorausgesetzt werden kann.

diagnostische Genauigkeit in hochwertigen Studien belegt

2. Die diagnostische Genauigkeit des Indextests zur Erkennung der in den Therapiestudien spezifizierten Population muss in hochwertigen Studien im Vergleich zum Referenztest belegt sein.

Populationen dieser beiden Studien sind vergleichbar

3. Die Population der Wirksamkeitsstudie muss hinsichtlich der, die Variation beeinflussenden Kriterien (siehe Kapitel 4.2.5), mit der Population in der diagnostischen Genauigkeitsstudie vergleichbar sein. Dies wäre zum Beispiel nicht der Fall, wenn durch ein positives Indextestergebnis eine Therapie in einem früheren Krankheitsstadium, oder eine gänzlich andere Therapie eingeleitet werden würde.

allgemein anerkannter Referenzstandard existiert

4. Es existiert ein allgemein anerkannter Referenzstandard über den Wirksamkeitsstudie und diagnostische Genauigkeitsstudie verknüpft werden können, wobei idealerweise das Ergebnis des Indextest Einschlusskriterium in der Wirksamkeitsstudie ist [64]. Ansonsten sollte gelten:

    - die Population der Wirksamkeitsstudie wurde mittels Referenztest identifiziert.
    - die diagnostische Genauigkeit des Indextest wurde durch den Vergleich mit dem Referenztest etabliert (Lord, Irwig et al. 2006; Eikermann 2009).

„linked Evidence“:

Wie erwähnt wird der Nachweis für die therapeutische Wirksamkeit am besten mittels RCTs belegt. Diese können – ohne die Studienergebnisse von neuen RCTs abwarten zu müssen - mit diagnostischen Genauigkeitsstudien verknüpft werden, wenn:

wenn Index- und Vergleichstest gleiche Sensitivität besitzen

- der Indextest *gleich sensitiv* ist wie der Vergleichstest (siehe Abbildung 4.4-2), aber andere positive Attribute (sicherer, spezifischer, kostengünstiger) besitzt. Ist er weniger sensitiv oder spezifisch, hat aber andere positive Attribute, dann müssen die Vor- und Nachteile, die mit diesen Unterschieden einhergehen, gegeneinander abgewogen werden, wofür sich entscheidungsanalytische Modelle eignen (siehe Kapitel 4.5).

wenn neuer Test sensitiver ist und Wirksamkeitsstudien für dieses neu entdeckte Patientenkollektiv vorhanden sind

- der neue Test *sensitiver*, aber ähnlich spezifisch wie der Vergleichstest ist, für die zusätzlich als erkrankt diagnostizierten Personen stehen aber ebenfalls Wirksamkeitsstudien zur Verfügung. Umgekehrt bedeutet das, dass sich „linked Evidence“ generell nicht für Triage Tests, die für Ruling Out (siehe Kapitel 4.2.3 und 4.2.1) eingesetzt werden sollen, eignet [69] – außer es stehen eben Wirksam-

keitsstudien für das neue mit dem Indextest diagnostiziertes Patientenkollektiv zur Verfügung [64] (siehe Abbildung 4.4-2).

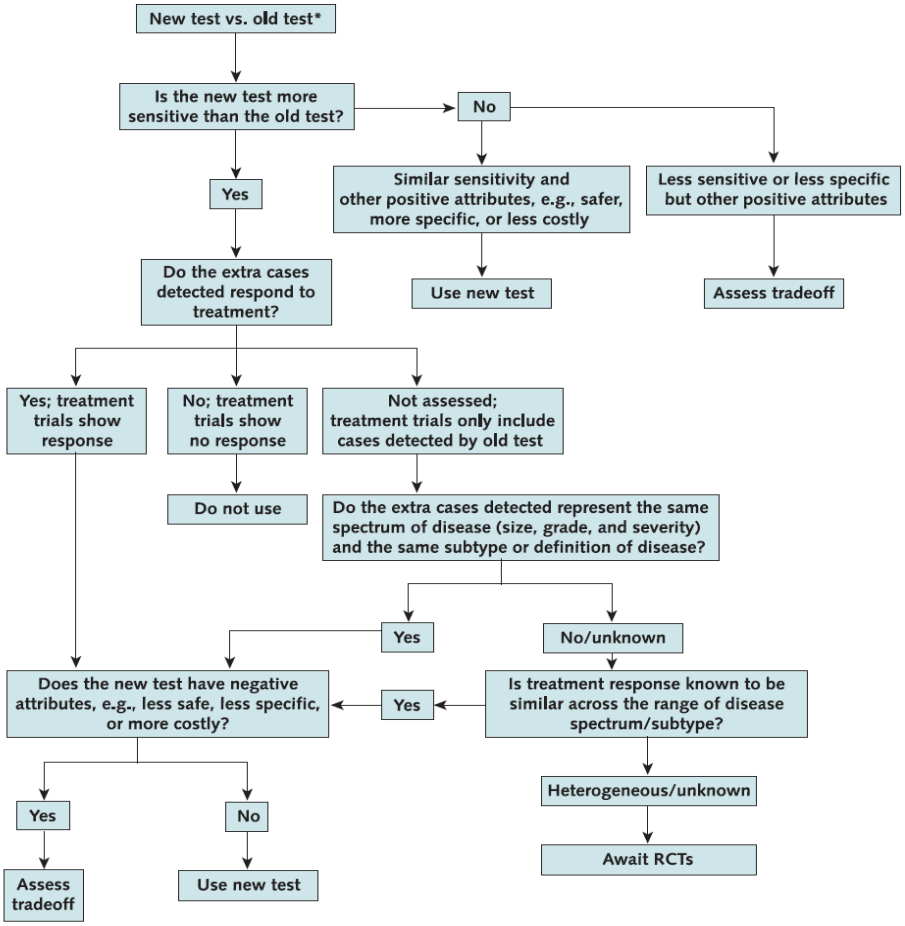


Abbildung 4.4-2: Bewertung neuer Tests im Rahmen von "linked Evidence" und benötigte Evidenz [64]

Eine Erweiterung des Konzepts wird von *Lord et. al* [45] vorgestellt: Je nachdem wo der neue Test in eine bestehende diagnostische Strategie eingebettet werden soll (siehe Kapitel 4.2.4) und in Abhängigkeit davon welcher Abschnitt des Diagnose-Behandlungspfades (diagnostische Genauigkeit, Änderungen im diagnostischen/therapeutischen Managements, Unterschied der therapeutischen Wirksamkeit, Vermeidung anderer Tests) am entscheidendsten durch den Indextest verändert wird, können neben RCTs zur therapeutischen Wirksamkeit und neben diagnostischen Genauigkeitsstudien bei denen alle PatientInnen mittels Referenzstandard verifiziert werden, auch andere Studiendesigns für die Evaluation eines diagnostischen Verfahrens herangezogen werden (siehe dazu Kapitel 4.2.4). Neue Wirksamkeitsdaten aus RCTs werden aber wieder gefordert, wenn durch einen sensitiveren Indextest zusätzlich Fälle entdeckt werden, oder, wenn durch einen Triage-Test mehr PatientInnen als „negativ" klassifiziert werden [45].

„linked Evidence" erlaubt keine Aussagen für testnegative Personen

Berücksichtigt werden muss auch, dass bei der Verwendung von „linked Evidence" keine Aussagen für testnegative PatientInnen getroffen werden können, da weder die Nebenwirkungen des Tests selbst, noch die Schäden oder Vorteile, die mit dem Behandeln, beziehungsweise dem Nicht-Behandeln assoziiert sind, bewertet werden können [70]. Darüber hinaus ist es auch nicht möglich unerwünschte Nebenwirkungen der gesamten Strategie (vom Test bis hin zur Therapie) zu beurteilen, da PatientInnen, die durch den Indextest erhebliche Nebenwirkungen erfahren, möglicherweise gar nicht erst in den therapeutischen Wirksamkeitsstudien eingeschlossen werden [70].

„linked Evidence" nur nach Rechtfertigung, ob alle Kriterien erfüllt sind

Bedingt durch die Voraussetzungen, die erfüllt, und die Annahmen die bei linked Evidence getroffen werden müssen, kann diese Methode in der Praxis aber häufig nicht angewandt werden, wenn aber doch, dann nur nach eingehender Erörterung und Rechtfertigung, ob alle Kriterien erfüllt sind.

4.5 Level 6 – Nutzen aus gesellschaftlicher Sicht

Bewertung des Kosten-Nutzen-Verhältnisses, um gesellschaftlichen Nutzen zu bestimmen

Die letzte Stufe für den Nachweis der Wirksamkeit von diagnostischen Technologien wird laut Fryback et al. [5] durch Studien erhoben, die aus einer gesellschaftlichen Perspektive heraus, den aus einem diagnostischen Verfahren resultierenden Nutzen und das Risiko in Relation zu den damit verbundenen Kosten setzen. Gesellschaftlicher Nutzen einer diagnostischen Technologie ist nach diesem Ansatz dann gegeben, wenn damit eine „effiziente Ressourcenallokation" erreicht wird.

andere Faktoren für gesellschaftlichen Nutzen, z.B. ethische, rechtliche Konsequenzen

Hinzugefügt werden soll, dass der gesellschaftliche Nutzen aber auch über das reine Kosten-Nutzen-Verhältnis hinaus definiert werden kann, indem etwa ethische oder rechtliche Konsequenzen, aber auch Aspekte von Verteilungsgerechtigkeit berücksichtigt werden [8, 71].

4.5.1 Studiendesigns

Netto-Nutzen eines Verfahrens wird durch direkt aus Test resultierende Konsequenzen aber auch durch therapeutische Konsequenzen bestimmt

Wenn nun EntscheidungsträgerInnen bestimmen sollen, ob ein neues Verfahren eingesetzt werden soll, müssen für die Bestimmung des gesellschaftlichen Nutzens, die damit einhergehenden Vorteile gegen die Nachteile abgewogen werden, wobei nicht nur die direkt aus einem Test resultierenden Konsequenzen berücksichtigt werden sollten, sondern auch die der anschließenden Therapie.

Mit diesen Entscheidungen sind aber zahlreiche Unsicherheiten verbunden, etwa über die tatsächliche Wirksamkeit einer Therapie, die Testgenauigkeit, der Trade-Off von Nutzen und Risiken eines diagnostischen Verfahrens im Vergleich zu einem anderen Test oder auch die Konsequenzen von falsch positiven und falsch negativen Befunden.

Trade-Off zwischen Nutzen und Risiken anhand von entscheidungs-analytischen Modellen

Die Vor- und Nachteile, als auch die finanziellen Konsequenzen diagnostischer Verfahren, können entweder direkt durch klinische Daten oder aber mittels Entscheidungsanalysen bestimmt werden [72, 73]. Bedingt durch die zahlreichen Möglichkeiten für den Einsatz eines diagnostischen Verfahrens innerhalb von diagnostischen Strategien (siehe Kapitel 4.2.4), können klinische Studien aber häufig nicht alle möglichen Variationen erfassen [6, 73].

Zusätzlich können Studien aber auch aufgrund ethischer Überlegungen unzulässig sein, oder sind bedingt durch lange Follow-Up-Zeiten oder der benötigten Studiengröße nicht durchführbar [72, 73].

Ziel von Entscheidungsanalysen - ein „systematischer, expliziter und quantitativer Ansatz zur Entscheidungsfindung unter Unsicherheit“ [74] - ist die Nutzenmaximierung, indem der Nutzen, die Risiken und gegebenenfalls auch die Kosten unterschiedlicher Strategien miteinander verglichen werden, wobei dabei aber andere Aspekte wie ethische oder rechtliche nicht berücksichtigt werden. Je nachdem aus welcher Perspektive (z.B. Patientenperspektive, Gesundheitssystem) Entscheidungsanalysen durchgeführt werden, können unterschiedliche Parameter herangezogen werden wie etwa Lebensqualität, Komplikationsraten, Gesamtüberleben oder auch Kosten. Am häufigsten werden Entscheidungsbäume oder Markov-Modelle für Entscheidungsanalysen herangezogen (Siebert 2003).

Bei diagnostischen Verfahren eignen sich Entscheidungsanalysen unter anderem dazu, um

Modelle geeignet um:

- Studien mit Kurzzeitergebnissen mit Studien mit Langzeitergebnissen zu verknüpfen (z.B. RCT mit intermediären Ergebnissen zu klinischer Effektivität werden mit Beobachtungsstudien, die einen längeren Zeitraum untersucht haben verknüpft, um Aussagen zu Langzeitmorbidität oder –mortalität treffen zu können),

  Kurz- und Langzeitergebnisse von Studien zu verknüpfen,

- Patientenpräferenzen hinsichtlich des Wertes der diagnostischen Information gegen die möglichen nachteiligen Effekte (z.B. direkt durch den Test entstehende Nebenwirkungen, Angst, Stress, Depression [7]) abzuwägen,

  Patientenpräferenzen abzuwägen,

- Studienergebnisse an nationale Gegebenheiten anzupassen (z.B. abweichende Prävalenz der Erkrankung),

  Studienergebnisse an nationale Gegebenheiten anzupassen,

- modifizierte, neuere Varianten eines Tests zu bewerten, oder Unterschiede in der Verwendung eines Tests zu erheben (z.B. bei Screening-Programmen Unterschiede der Screening-Intervalle, der Population, der Screeningdauer),

  modifizierte Varianten von Tests zu untersuchen,

- Daten aus RCTs unter *Ideal*bedingungen an *Real*bedingungen anzupassen,

  Anpassung von Ergebnissen unter Idealbedingungen an Realbedingungen,

- unterschiedliche diagnostische Pfade zu zu vergleichen (z.B. Add-On, Ersatz, Triage) (Schwartz, Ledil et al. 2003; Siebert 2003; Lord, Irwig et al. 2006; Sutton, Cooper et al. 2008; Trikalinos, Siebert et al. 2009).

  mehrere diagnostische Pfade zu vergleichen

Allerdings sind mit der Modellierung von diagnostischen Strategien auch etliche Herausforderungen verbunden. So entstehen Probleme, wenn für die wichtigsten zu modellierenden Parameter nur sehr ungenaue Daten zur Verfügung stehen. Besonders zu nennen sind hierbei ungenaue Angaben zu Prävalenz, Angaben zu Testgenauigkeit und zu Wirksamkeit und Nebenwirkungen der Therapie sowohl bei tatsächlich Erkrankten, aber auch bei tatsächlich gesunden PatientInnen. Die Testgenauigkeit kann vor allem bei ungenauen Referenzstandard kompromittiert werden (Trikalinos, Siebert et al. 2009).

Problem: mangelnde Qualität der den Modellen zu Grunde liegenden Daten

Um eine effiziente Ressourcenallokation zu gewährleisten, kommt entscheidungsanalytischen Modellen im Rahmen von ökonomischen Evaluationen eine besondere Bedeutung zu.

entscheidungsanalytische Modelle im Rahmen von ökonomischen Evaluationen

Dabei werden Kosten und Konsequenzen von unterschiedlichen Strategien miteinander verglichen. Die Konsequenzen können in Form unterschiedlicher Endpunkte, wie zum Beispiel Lebensjahre, die Anzahl von diagnostizierten Karzinomen oder von Lebertransplantationen, angegeben werden (Kosten-Effektivitätsanalyse) [4, 72]. Ein Maß, das einen Vergleich über verschiedene Krankheitsgruppen hinweg erlaubt, ist das qualitätskorrigiertes Lebensjahr (QALY) im Rahmen von Kosten-Nutzwertanalysen [5, 72].

zahlreiche mögliche Endpunkte

QALY für Vergleich von verschiedenen Krankheiten

Berücksichtigung falscher Testergebnisse wichtig

Wie bei der reinen Nutzenbewertung auch, sollten bei ökonomischen Evaluationen neben den Konsequenzen von richtigen Diagnosen, unbedingt die Konsequenzen falsch positiver und falsch negativer Testergebnisse berücksichtigt werden: werden falsch positive PatientInnen behandelt, können durch die Therapie unerwünschte Nebenwirkungen entstehen, die mit beträchtlichen finanziellen Konsequenzen einhergehen können. Umgekehrt sind auch die Konsequenzen zu beachten, wenn ein falsch negativer PatientIn nicht behandelt wird, obwohl eine wirksame Therapie zur Verfügung gestanden hätte. Außerdem sollten auch die, mit den Tests selbst einhergehenden, negativen Konsequenzen und deren Kosten nicht vergessen werden, da zum Beispiel bei invasiven Untersuchungsverfahren wie etwa endoskopischen Untersuchungen, Komplikationen wie iatrogen verursachte Perforationen zu erheblichen Folgekosten führen können [6, 72, 75, 76].

Sensitivitätsanalysen um unterschiedliche Prävalenzen zu modellieren

Da Kosteneffektivitätsanalysen meist spezifisch auf den, für EntscheidungsträgerInnen, relevanten Kontext zugeschnitten sind (z.B. Einschlusskriterien der Primärstudien), ist die Umlegbarkeit der Ergebnisse auf andere Settings limitiert [76]. Sensitivitätsanalysen bieten aber eine Möglichkeit, dieses Problem zu umgehen [6] und erlauben auch die Modellierung für Populationen mit unterschiedlichen Prävalenzen [64].

Schwierigkeiten, wenn Patientennutzen lediglich anhand von diagnostischen Genauigkeitsstudien einem Testergebnis zugeordnet werden soll

Modelle können sehr umfangreich werden (z.B. unterschiedliche Grenzwerte, mehrere Verwendungsmöglichkeiten des Tests)

Neben den bereits erwähnten Herausforderungen und neben methodischen Überlegungen, die bei ökonomischen Evaluationen immer angestellt werden müssen [77], können Kosteneffektivitätsanalysen von diagnostischen Verfahren dadurch verkompliziert werden, indem der resultierende Patientennutzen lediglich anhand von diagnostischen Genauigkeitsstudien einem Testergebnis zugeordnet werden soll. Die Einschätzung von finanziellen Konsequenzen für falsche Befunde wird dadurch erheblich erschwert [6]. Zusätzlich können Modelle durch unterschiedliche Grenzwerte oder durch zahlreiche, mögliche Verwendungszwecke innerhalb von diagnostischen Strategien sehr umfangreich werden [72, 76]. Eine andere Herausforderung ist, wenn diagnostische Genauigkeitsstudien nicht in dem Setting durchgeführt wurden, in dem der Test angewandt werden soll, oder wenn der Test an einer anderen Position der diagnostischen Strategie eingesetzt worden war [73].

4.6 Zusammenfassung

Die Bewertung eines Tests kann auf unterschiedlichen Ebenen erfolgen, da die klinische Effektivität und die ökonomischen Konsequenzen diagnostischer Verfahren nicht nur durch die Testgenauigkeit bestimmt wird, sondern auch durch den Einfluss des Testergebnisses auf diagnostische und therapeutische Entscheidungen und letztlich durch die Wirksamkeit der nachfolgenden Therapie.

Für jede Ebene innerhalb der Evidenzhierarchie gibt es bevorzugte Studiendesigns, die jeweils mit eigenen methodischen Herausforderungen vergesellschaftet sein können (siehe Tabelle 4.6-1).

Tabelle 4.6-1: Bevorzugte Studiendesigns in der Evidenzhierarchie zur Evaluation von diagnostischen Verfahren (Quelle: [4])

Evidenzhierarchie	Studiendesign
Level 1 - technische Qualität	Nicht von Bedeutung
Level 2 – diagnostische Genauigkeit	Verblindete Querschnittsstudie
Level 3 diagnostischer Impakt	RCTs, Vorher-Nachher Studie
Level 4 therapeutischer Impakt	RCTs, Vorher-Nachher Studien
Level 5 – patientenrelevanter Nutzen	Direkt: RCTs, Kohortenstudien, Fall-Kontroll Studien, Fallserien Indirekt: RCT zur therapeutischen Wirksamkeit + diagnostische Genauigkeit aus Querschnittsstudie Entscheidungsanalysen
Level 6 – Nutzen aus gesellschaftlicher Sicht	Kosten-Nutzen, Kosten-Nutzwert, Kosten-Effektivitätsanalysen

Neben den mit den einzelnen Studiendesigns vergesellschafteten Problemen, wurden weitere kritische Aspekte identifiziert:

- Level 2:
  - Parameter der diagnostischen Genauigkeit sind keine festen Größen, sondern variieren in Abhängigkeit von Prävalenz, Patientencharakteristika und Grenzwerten.
  - Die diagnostische Genauigkeit eines neuen Tests sollte im Vergleich mit dem Referenzstandard ermittelt werden. Probleme entstehen, wenn es keinen Referenzstandard gibt, oder der Referenzstandard selber ungenau ist.
  - Für ein diagnostisches Untersuchungsverfahren gibt es oft zahlreiche Verwendungsmöglichkeiten, da ein Test als Ersatz oder zusätzlich zu einem anderen Verfahren verwendet wer-

den kann. Studien müssen den genauen Einsatzort des Indextests innerhalb der diagnostischen Strategie abbilden

- Die Ergebnisse von diagnostischen Genauigkeitsstudien können durch eine Reihe von Einflussfaktoren beeinträchtigt werden, sodass eigene Instrumente zur Bewertung der methodischen Qualität benötigt werden.

- Level 3 & Level 4:
  - Häufigstes Studiendesign ist die diagnostische Vorher-Nachher Studie, mit zahlreichen Limitationen. Bedingt durch diese methodischen Schwächen, ist der Stellenwert dieser Studien allerdings unklar.
  - Vorher-Nachher Studien sind für einzelne Add-On Tests geeignet, nicht aber für den Vergleich mehrerer Tests.

- Level 5:
  - Obwohl das beste Studiendesign zum Nachweis der klinischen Effektivität eines diagnostischen Untersuchungsverfahrens ein RCT ist, sind diese Studien nur selten verfügbar.
  - Mittels „linked Evidence“ kann indirekt ein Zusammenhang zwischen Test und Patientennutzen etabliert werden, wobei aber neben den bei Level 2 angeführten Problemen, weitere methodische Herausforderungen vergesellschaftet sind.
  - Aufgrund der Bedingungen, die für „linked Evidence“ erfüllt werden müssen, ist dieser Ansatz in der Praxis häufig nicht umzusetzen.
  - Entscheidungsanalysen eignen sich um Nutzen und Risiken von richtigen, aber auch von falschen Befunden gegeneinander abzuwägen. Ihre Aussagekraft wird aber durch die Qualität der zu Grunde liegenden Parameter bestimmt.

- Level 6:
  - Ökonomische Evaluationen eignen sich, um den Nutzen, die Risiken und die Kosten von Tests gegeneinander abzuwägen. Ihre Aussagekraft wird aber durch die Qualität der zu Grunde liegenden Parameter bestimmt.
  - Vorteile von ökonomischen Evaluationen sind unter anderem, dass viele, potenziell mögliche Testkombinationen und Testvarianten modelliert werden können. Allerdings können diese Modelle bedingt durch die Vielzahl der Möglichkeiten sehr komplex werden.

5 Methodensynthese der ausgewählten Institutionen

Die vier ausgewählten Institutionen (MSAC, NICE, IQWiG, EUnetHTA) wurden hinsichtlich ihrer Methoden zur Bewertung von diagnostischen Verfahren untersucht und ihre Vorgehensweisen anhand des Evidenzmodells nach Fryback dargestellt (siehe Appendix 11).

Methodenmanuale von MSAC, NICE, IQWiG und EUnetHTA wurden untersucht

Die Methodenmanualen der untersuchten Institutionen unterscheiden sich in Bezug auf Umfang, der Genauigkeit der Ausführungen, sowie hinsichtlich ihres Fokus. Zwei Manuale, das von MSAC [67] und EUnetHTA [78], konzentrieren sich ausschließlich auf die Bewertung von diagnostischen Verfahren und sind somit die umfangreichsten Beschreibungen. NICE befasst sich in einem Methodenmanual zur Erstellung von klinischen Richtlinien mit diagnostischen Verfahren [79] und pilotiert derzeit einen ersten Methodenentwurf zur Bewertung von diagnostischen Verfahren, dessen Schwerpunkt die Bewertung der Kosteneffektivität von Untersuchungsmethoden ist [80]. Das IQWiG erwähnt im Rahmen seiner themenübergreifenden „Allgemeinen Methoden“ [81], wie diagnostische Maßnahmen zu evaluieren sind.

Unterschiede in Bezug des Umfangs und Fokus

Im Folgenden sollen nun die Vorgehensweisen der ausgewählten Institutionen überblicksmäßig dargestellt werden und Gemeinsamkeiten, sowie Unterschiede der methodischen Ansätze erläutert werden.

5.1 Allgemeine Methodik

Evidenzlevel für Evaluationen

zahlreiche Ähnlichkeiten zwischen Institutionen

Die grundlegenden methodischen Vorgehensweisender Institutionen ähneln sich über weite Teile (siehe Tabelle 5.1-1). So ist ersichtlich, dass keine der Institutionen diagnostische Verfahren lediglich anhand ihrer Fähigkeit zwischen gesund und krank zu unterscheiden, beurteilt, sondern, wie auch bei Interventionen, vor allem der für PatientInnen entstehende Nutzen (Level 5) zentraler Bestandteil der Bewertung ist und somit die Grundlage für Entscheidungen über Ablehnung oder Adoption einer neuen Technologie bildet. Von drei Institutionen (MSAC, IQWiG, EUnetHTA) werden dabei aber nicht nur Studiendesigns berücksichtigt, die direkt den aus einem diagnostischen Verfahren resultierenden Nutzen untersuchen, sondern der Nutzen kann auch indirekt, unter Verwendung von Studien des Levels 2, 3 und 4, etabliert werden.

alle bewerten diagnostische Verfahren mindestens anhand des patientenrelevanten Nutzens

die Meisten auch noch anhand von Kosten-Nutzen

Im Methodenhandbuch von NICE wird zwar auch die reine Bewertung der diagnostischen Genauigkeit (Level 2) erwähnt, allerdings basieren die resultierenden Empfehlungen immer auch auf Kosten- Nutzwertanalysen (vorzugsweise unter Verwendung von QALYs) und entsprechen somit Level 6 der Evidenzhierarchie. Kosteneffektivitätsanalysen werden auch von MSAC und im EUnetHTA Core Modell erwähnt.

Studiendesign

Methode der Wahl zur Evaluation ist systematische Übersichtsarbeit

Die Methode der Wahl zur Evaluation diagnostischer Verfahren ist bei allen Institutionen eine systematische Übersichtsarbeit, die prinzipiell dieselben Qualitätsstandards wie Reviews über Interventionen erfüllen muss (z.B. Literaturauswahl und Datenextraktion durch zwei unabhängige WissenschafterInnen, systematische Literatursuche, etc.). Bei der Formulierung der Fragestellung und des Reviewprotokolls sollten relevante Gruppen wie PatientInnen, ÄrztInnen oder Angehörige der Pharmaindustrie miteinbezogen werden.

Forschungsfrage

Forschungsfrage anhand der PICO Kriterien, eventuell Einschluss von „prior tests" als P(P)ICO

Wie in den vorangegangenen Kapiteln dargestellt, können Tests häufig an unterschiedlichen Positionen der diagnostischen Strategie verwendet werden (siehe Kapitel 4.2.4). Ebenso kann die diagnostische Genauigkeit, bedingt durch Patientencharakteristika oder die Wahl des Vergleichstests (siehe Kapitel 4.2.5), variieren, sodass die genaue Beschreibung der Forschungsfrage ein wesentlicher Schritt bei der Evaluation von Testverfahren ist.

Die Formulierung der Fragestellung folgt bei allen Institutionen dem PICO Schema, wodurch die zu untersuchende *P*opulation, die *I*ntervention (=Indextest), der *K*(C)omparator (= Vergleichstest), sowie relevante *O*utcomes festgelegt werden. Bis auf EUnetHTA erwähnen alle Institutionen auch explizit, dass bereits in der Forschungsfrage *S*tudiendesigns festgelegt werden können, die in der systematischen Review berücksichtigt werden sollen (PICOS). MSAC's Richtlinien erwähnen zusätzlich ein zweites P, folgen also einem PPICO Schema, wobei das zweite P für „*P*rior Tests", also für etwaige bereits erfolgte Untersuchungen, steht.

Hintergrund

wichtige Fragestellungen, um Forschungsfrage und Studienprotokoll zu definieren

Überlegungen, die bei der Entwicklung und der Formulierung des Studienprotokolls und der PICO Frage hilfreich und als Hintergrundteil in einem Bericht enthalten sein können, ähneln sich zwischen den Institutionen und können folgende Fragen beinhalten:

Population: in welcher Population, bei welchen Symptomen soll der Indextest verwendet werden? wie kann die Zielerkrankung klar definiert werden? verändert sich die diagnostische Genauigkeit des Indextests in unterschiedlichen Patientengruppen? Prognose? Krankheitsverlauf? was ist die Inzidenz/Prävalenz der Zielerkrankung, was sind deren Konsequenzen (z. B. Mortalität, Krankenständen, Frühpensionierungen)?

Intervention: was ist der geplante Einsatz des Indextests innerhalb des diagnostischen Pfades (Ersatz, Add-on)? ist der Indextest genauer als der Referenzstandard? Was sind personelle/technische Voraussetzungen für Einsatz des Indextests? was sind die technischen Angaben (Schwellenwerte für negative/positive Ergebnisse)? Inter-/Intraobserver Variabilität, technische Unterschiede (z.B. Art und Menge des verwendeten Kontrastmittels, bei Röntgenbilder Expositionszeit, etc.)?

Komparator: was ist der Referenzstandard? Wie geeignet ist der Referenzstandard zur Diagnose der Zielerkrankung? welche alternative diagnostische Verfahren(-strategien) gibt es? mit welchem Verfahren soll der Indextest

verglichen werden? entspricht der gewählte Vergleichstest dem Referenztest?

Outcomes: je nachdem welcher Evidenzlevel bewertet wird, können diagnostische Genauigkeit, Sicherheit, Wirksamkeit oder Kosten relevante Endpunkte sein

Studiendesign: je nach Evidenzlevel werden unterschiedliche Studiendesigns bevorzugt (siehe Tabelle 5.1-1.)

Weitere Faktoren, die berücksichtigt werden können, sind der derzeitige Zulassungsstatus (international, national), der Lebenszyklus in dem sich die Technologie befindet (neu/etabliert/obsolet), soziale Gerechtigkeit und ob etwaige Ausnahmeregelungen für spezielle Patientengruppen zu treffen sind.

Tabelle 5.1-1.: Übersicht über Evidenzlevel, die von den ausgewählten Institutionen bei der Evaluierung von diagnostischen Verfahren Verwendung finden, sowie relevante Studiendesigns

	MSAC	IQWiG	NICE	EUnetHTA
Methodik	Systematische Review	Systematische Review	Systematische Review	Systematische Review
Evaluation basierend auf	Sicherheit, Wirksamkeit, Kosteneffektivität	Sicherheit, Wirksamkeit	Sicherheit, Wirksamkeit, Kosteneffektivität	Sicherheit, Wirksamkeit, Kosteneffektivität
Relevante Evidenzlevel der Evaluation	5, 6	5	2, 5, 6	5, 6
Verwendete Evidenzlevel	2, 3, 4, 5, 6	2, (3, 4) 5, 6	2, 5, 6	2, 3, 4, 5, 6
Bevorzugte Studiendesigns nach Evidenzlevel				
Level 2	Verblindete Querschnittstudien in konsekutiv eingeschlossenen PatientInnen Verblindete Querschnittsstudie in nichtkonsekutiv eingeschlossenen PatientInnen Fallkontrollstudie	Verblindete Querschnittsstudie RCT mit zufälliger Zuordnung der PatientInnen zu Index- oder Referenztest	Querschnittsstudie Fallkontrollstudien	Querschnittsstudie RCT mit zufälliger Zuordnung der PatientInnen zu Index- oder Referenztest Studiendesigns mit limitierter Verifizierung
Level 3	Diagnostische Vorher-Nachher Studien	Unklarer Stellenwert zur Nutzenbewertung	-	Vorher-Nachher Studien, Zeitserien
Level 4	Diagnostische Vorher-Nachher Studien	Unklarer Stellenwert zur Nutzenbewertung	-	Vorher-Nachher Studien, Zeitserien

Level 5	Wirksamkeit			
	RCTs nicht-randomisiert kontrollierte Studien Kohortenstudien Fallkontrollstudien	RCTs nicht-randomisiert kontrollierte Studien	RCTs	RCTs Kohortenstudien Fallkontrollstudien
	Sicherheit			
	Alle Studiendesigns inklusive Fallserien, Registern	Kontrollierte Interventionsstudien bevorzugt	RCTs bevorzugt	RCTs, aber auch andere Studiendesign, inklusive Fallserien, Register
	Wirksamkeit + Sicherheit			
	Entscheidungsanalysen	-	Entscheidungsanalysen	Entscheidungsanalysen
Level 6	Einfache Kostenaufstellung Kosten-Minimierungsanalyse Kosten-Nutzen/Nutzwert/Effektivitätsanalyse	-	Kosten-Nutzwertanalyse Kosten-Effektivitätsanalyse	Kostenminimierungsanalyse Kosten-Nutzenanalyse/Nutzwert/Effektivitätsanalyse

5.2 Nutzenbewertung diagnostischer Verfahren

5.2.1 Level 2 - Bewertung der diagnostischen Genauigkeit

Bewertung der diagnostischen Genauigkeit zentraler Bestandteil bei allen Institutionen

Obwohl der aus einem diagnostischen Verfahren resultierende Nutzen in den meisten Fällen nicht ausschließlich über die diagnostische Genauigkeit etabliert werden kann, ist deren Bewertung aber zentraler Bestandteil der methodischen Richtlinien aller dargestellten Institute (siehe Tabelle 5.2-1.).

diagnostische Genauigkeit ausreichend, wenn gleich sensitiver, kostengünstiger Indextest einen anderen Test ersetzen soll

Die Bewertung eines diagnostischen Verfahrens lediglich anhand der diagnostischen Genauigkeit wird von zwei Institutionen (MSAC, EUnetHTA) dann als ausreichend angesehen, wenn der Indextest kostengünstiger und ein nicht-invasiver *Ersatz* für einen anderen Test ist und eine ähnliche Sensitivität aufweist, wie der zu ersetzende Test. Grund dafür ist, dass bei einer als Ersatz geplanten Technologie davon ausgegangen werden kann, dass das weitere therapeutische Management und damit die resultierenden patientenrelevanten Konsequenzen dieselben bleiben, wie bei der Verwendung des alten Verfahrens. Das IQWiG erwähnt, dass Level 2 Studien genügen, wenn es sich bei dem Indextest lediglich um eine modifizierte Variante eines bereits etablierten Tests handelt (siehe Kapitel 5.2.3).

Methode

Wie erwähnt, ist die Methode der Wahl zur Erhebung der diagnostischen Genauigkeit eine systematische Übersichtsarbeit zu einer klar definierten PICO - Frage, wobei die im vorherigen Kapitel (siehe Kapitel 5.1) dargestellten Überlegungen und Fragestellungen im Bericht enthalten sein sollten.

Studiendesign

bevorzugtes Studiendesign: diagnostische Querschnittstudie

Bevorzugtes Studiendesign ist bei allen Institutionen eine diagnostische Querschnittsstudie (siehe Tabelle 5.1-1.), wobei aber auch andere Designs erwähnt werden, wie etwa eine diagnostische Genauigkeitsstudie, mit randomisierter Zuteilung der PatienInnen entweder zu Indext- oder zu Referenztest.

Endpunkte

relevante Endpunkte Sensitivität, Spezifität, LR, ROC, AUC

Relevante Endpunkte, die von allen Institutionen erwähnt werden, sind Sensitivität/Spezifität, LR und ROC-Kurven mit den dazugehörigen AUCs. NICE gibt als möglichen Parameter der diagnostischen Genauigkeit auch PVs an, während MSAC und EUnetHTA sie zwar erwähnen, aufgrund ihrer Abhängigkeit von der Prävalenz der Erkrankung aber als nicht relevant für systematische Reviews betrachten. Das IQWiG hingegen beschreibt nur das „allgemein anerkannte Testgütekriterien“ verwendet werden können (siehe Tabelle 5.2-1).

PV aufgrund der Prävalenzabhängigkeit nur bedingt geeignet

Tabelle 5.2-1: Methodenübersicht zur Evaluation der diagnostischen Genauigkeit

	MSAC	IQWiG	NICE	EUnetHTA
Methode	Systematische Review	Systematische Review	Systematische Review	Systematische Review
Hintergrund	✪ Indextestbeschreibung ✪ Geplante Verwendung, Indikation(en) ✪ Klinischer Bedarf: Krankheit, Symptome, Stadien, Inzidenz, Risikofaktoren, Mortalität ✪ Population: Anzahl der zu testenden Population ✪ Therapie ✪ Referenztest: Beschreibung alternativer diagnostischer Strategien, Auswahl der für den eigenen Kontext relevantesten Komparatoren ✪ - derzeitige Kostenerstattung, Marketingstatus des Indextests	✪ Zielerkrankung ✪ Diagnostische Verfahren ✪ Therapeutische Optionen ✪ Indextest	✪ Indextestbeschreibung: Verwendung, Grenzwerte, technische Beschreibung ✪ Verfügbare Vergleichstests ✪ Patientenpopulation ✪ Therapien ✪ - Outcomes	✪ Zielerkrankung/Zielpopulation: Inzidenz, Prävalenz, Mortalität, Krankenständen, Frühpensionierungen, Subgruppen ✪ -Einsatz der Technologie: regionale/internationale Unterschiede, bereits im klinischen Alltag in Verwendung ✪ Krankheitsmanagement: existierende Diagnosepfade, geplante Verwendung des Indextest, alternative diagnostische Strategien ✪ - Technologie: Beschreibung der Funktionsweise, der technischen Charakteristika, Unterschiede verschiedener Versionen der Technologie, personelle/strukturelle/finanzielle Voraussetzungen
Fragestellung	PPICO(S)	PICO(S)	PICO(S)	PICO
Outcome	Sensitivität, Spezifität, LR, DOR, ROC	Sensitivität, Spezifität, LR, PV, DOR, ROC	Sensitivität, Spezifität, PV, LR, ROC-Kurven	Sensitivität, Spezifität, LR, DOR, ROC und AUC
Datenpräsentation	Narrativ Variable mit Konfidenzintervallen	Narrativ Variable mit Konfidenzintervallen	Narrativ numerische Zusammenfassung	Narrativ Variable mit Konfidenzintervallen
Instrumente zur Bewertung der Studienqualität	QUADAS, STEP, Cochrane Handbook, Checkliste nach Jaeschke Bewertung einzelner Studien nach: Adäquater Vergleich: in-/direkter Vergleich von Indextest und Vergleichstest Adäquate Studienpopulation: wurde die Studie in einer für die Zielerkrankung repräsentativen Population durchgeführt Studienqualität: Studiendesigns klassifiziert nach NHMCR + methodische Qualität	Diagnose: QUADAS Prognose: Checkliste nach Hayden, Laupacis	Diagnose: QUADAS, STARD, Cochrane Prognose: Checkliste nach Hayden	Cochrane Handbuch + 10 weitere mögliche Fragen

Datenextraktion	1. Studiencharakteristika - Autor, Publikationsjahr - Land, klinisches Setting, Studiendauer - Studiendesign 2. Studienpopulation und Testverfahren - Patientenzahl - Krankheitsprävalenz in der Studien- population - Methoden und Kriterien der Patientenselektion - Details zu Indextest und Vergleichstest sowie zu Referenzstandard - Schwellenwert für ein positives Testergebnis 3. Studienbewertung - Studienqualität, methodische Schwäche und deren mögliche Auswirkungen auf Bias (siehe Studienbewertung) - Bewertung der Umlegbarkeit der Studienergebnisse und mögliche Limitationen für den eigenen Kontext 4. Studienergebnisse - TP, FN, TN, FN, Sens/Spez, NPV, Resultate mit 95% Konfidenzintervalle, Ergebnisse für Patientensubgruppen	Nicht näher spezifiziert	1. Studientyp 2. Studienqualität 3. Patientenzahl 4. Prävalenz: Anzahl der PatientInnen mit der Zielerkankung in der gesamten beobachteten Population 5. Patientencharakteristika: neben Alter, Geschlecht, auch etwa klinisches Setting 6. Indextest: inklusive Grenzwert 7. Referenztest 8. Sens/Spez 9. PPV/NPV 10. Funding	11. Patientenpopulation, Prävalenz der Zielerkrankung 12. vorangegangene Tests 13. Indextest, Grenzwerte 14. Referenztest 15. Testergebnisse (4-Felder Tafel) 16. Sensitivität/spezifität mit 95% Konfidenzintervallen 17. Andere Kenngrößen der diagnostischen Genauigkeit 18. Studienqualität
Datensynthese	Meta-Analyse bei homogenen Studienergebnissen oder erklärbarer Heterogenität (z.B. Grenzwerteffekt)	Meta-Analyse bei homogenen Studienergebnissen oder erklärbarer Heterogenität (z.B. Grenzwerteffekt)	Meta-Analyse bei homogenen Studienergebnissen oder erklärbarer Heterogenität (z.B. Grenzwerteffekt)	Meta-Analyse bei homogenen Studienergebnissen oder erklärbarer Heterogenität (z.B. Grenzwerteffekt)
Besonderheiten	Diagnostische Genauigkeit reicht aus, wenn der Indextest kostengünstiger und ein nicht-invasiver Ersatz für einen anderen Test ist.	Diagnostische Genauigkeit reicht aus, wenn lediglich Varianten eines bereits etablierten Tests untersucht werden.	-	Diagnostische Genauigkeit reicht aus, wenn der Indextest Ersatz für einen anderen Test bei etablierter Standardtherapie ist.

Bewertung der methodischen Qualität

am häufigsten genannte Instrumente sind QUADAS und Cochrane Checkliste

Die kritische Bewertung von Diagnosestudien umfasst die Beurteilung der internen, als auch der externen Validität, da die Qualität und die Generalisierbarkeit von Ergebnissen aus Diagnosestudien durch Fehler im Studiendesign, in der Studiendurchführung oder durch die Einschlusskriterien reduziert sein kann (siehe Kapitel 4.2.5). Eigene Instrumente stehen dafür zur Verfügung, wobei es aber noch kein allgemein anerkanntes Instrument gibt. Die von den Institutionen am häufigsten erwähnten Tools sind das QUADAS - Instrument und der im Cochrane Handbuch für diagnostische Studien angeführte Fragenkatalog (siehe Appendix 8).

davon, *a priori* Auswahl von für eigene Fragestellung am relevantesten Fragen

Alle Institutionen betonen, dass aus den in den Instrumenten angeführten Fragen nicht alle für die eigene Forschungsfrage relevant sein werden, sondern *a priori* jene ausgewählt werden sollten, die am ehesten die Qualität der Ergebnisse beeinträchtigen könnten. Wenn die methodische Qualität einer Studie als sehr niedrig eingestuft wird, kann dies auch als Ausschlusskriterium herangezogen werden. MSAC betont, dass die wichtigste Komponente bei Diagnosestudien, die Auswahl eines geeigneten und zuverlässigen Referenzstandards ist.

MSAC stellt zusätzlich noch ein Bewertungsschema vor, anhand dessen die Studienqualität, die Relevanz und die Übertragbarkeit der Studienergebnisse auf die Fragestellung zusammengefasst wird (siehe Appendix 11.1). Eine Gesamtdarstellung der Evidenzstärke wird von NICE zwar erwähnt, allerdings unter dem Zusatz, dass ein valides Instrument erst entwickelt wird.

Tools für Prognosestudien werden von IQWiG und NICE genannt

Als mögliches Tool zur Bewertung der Qualität von Prognosestudien erwähnen IQWiG und NICE eine Checkliste nach Hayden (siehe Appendix 8.4.2)

Datenextraktion

Übereinstimmung in Bezug auf zu extrahierende Daten

Drei der vier Institutionen (MSAC, NICE, EUnetHTA) präsentieren standardmäßig aus den Primärstudien zu extrahierende Punkte, die weitgehend deckungsgleich sind:

- Patientenanzahl, Patientencharakteristika
- Prävalenz der Zielerkrankung
- klinisches Setting, vorangegangene Tests
- Indextest, Grenzwerte
- Referenztest
- Sensitivität/Spezifität (+95% Konfidenzintervall)
- Andere Kenngrößen der diagnostischen Genauigkeit
- Studienqualität

Daten-Analyse

narrative Beschreibung der Ergebnisse unter Berücksichtigung der Studienqualität

Neben der tabellarischen Darstellung der Ergebnisse (inklusive der 95% Konfidenzintervalle), ist auch die narrative Beschreibung der Resultate unter Berücksichtigung der methodischen Qualität der zugrundeliegenden Studien bei allen Institutionen angezeigt.

Meta-Analysen gewünscht

Ebenso einheitlich werden Meta-Analysen der diagnostischen Genauigkeit gewünscht, sofern die Studien homogen sind, oder heterogene Studiener-

gebnisse durch Zufallsschwankungen oder durch einen Grenzwerteffekt erklärt werden können.

Je nachdem wodurch Heterogenität letztlich verursacht wird, stehen unterschiedliche statistische Modelle für das Poolen von Daten zur Verfügung, wobei die Vorgehensweisen der Institute sowohl in Bezug auf die Testung der Heterogenität, als auch beim Datenpooling selbst, relativ deckungsgleich sind.

5.2.2 Level 3 & Level 4 – diagnostischer/therapeutischer Impakt

Unterschiede in der Bedeutung von Studien zu Level 3 & 4

Drei der Institutionen erwähnen Studien der Evidenzlevel 3 und 4, wobei diese bei MSAC und EUnetHTA im Rahmen von „linked Evidence" Verwendung finden, das IQWiG hingegen den Stellenwert dieser Studien, aufgrund der mit diesen Designs assoziierten methodischen Schwächen, als unklar ansieht.

EUnetHTA's Methodenbeschreibung erwähnt darüber hinaus, dass diese Studien besonders wichtig sind:

- bei Add-On Tests, nicht aber bei Ersatztests, da bei Ersteren erst der Nachweis erbracht werden muss, dass das Testergebnis auch tatsächlich zu Veränderungen von Managemententscheidungen führt.
- wenn andere Einflüsse, wie zum Beispiel Patientenpräferenzen nachfolgende therapeutische Entscheidungen bestimmen können.
- wenn sich die Population der Diagnosestudie in Bezug auf Prävalenz oder Schwere der Erkrankung von der in der Wirksamkeitsstudie unterscheidet.
- der Wert, der durch den Test gewonnenen Information, unsicher ist.

5.2.3 Level 5 - patientenrelevanter Nutzen

patientenrelevanter Nutzen entscheidend bei der Evaluation von Test

Alle untersuchten Institutionen stimmen darin überein, dass bei der Bewertung von diagnostischen Untersuchungsverfahren letztlich entscheidend ist, ob durch den Test der patientenrelevante Nutzen verbessert werden kann. Da der Nutzen eines Tests in den meisten Fällen erst durch die Verabreichung einer wirksamen Therapie entsteht, müssen Testergebnis und therapeutische Konsequenzen zusammen betrachtet werden. Dazu gibt es zwei Möglichkeiten (siehe Tabelle 5.2-2).

Tabelle 5.2-2: Methodenübersicht zur Evaluation des klinischen Nutzens

	MSAC	IQWiG	NICE	EUnetHTA
Direkte Evidenz				
Definition	Studien mit patientenrelevanten Endpunkten, die Indextest und nachfolgende therapeutische Konsequenzen mit denen aus dem Vergleichstest und nachfolgenden Therapien vergleichen Evidenzlevel 5			
Methode	Systematische Review mit standardisiertem Vorgehen wie bei der Evaluation von Interventionen			
Indirekt				
Definition	= linked evidence = Evidenzlevel 2 + Evidenzlevel 3, 4 + Wirksamkeitsstudie die Wirksamkeit einer Therapie, belegt durch hochwertige Studien, wird mit Ergebnissen aus hochwertigen Studien zur Testgenauigkeit des Indextests im Vergleich zum Referenzstandard verknüpft	= diagnostische Kette = Evidenzlevel 2 + Wirksamkeitsstudie Ergebnisse aus hochwertigen Studiendesigns zu Testgenauigkeit, im besten Fall diagnostischen Querschnittsstudien, werden mit Ergebnissen aus (in den meisten Fällen) randomisierten Interventionsstudien verknüpft	-	= linked evidence = Evidenzlevel 2 + Evidenzlevel 3, 4 + Wirksamkeitsstudie Ergebnisse zu diagnostischer Genauigkeit werden mit randomisiert kontrollierten Interventionsstudien verknüpft
Bedingungen	✪ der Nutzen der Intervention muss durch hochwertige Studien (RCT, oder systematischen Übersichtsarbeit) nachgewiesen sein. ✪ der Indextest diagnostiziert exakt die Zielerkrankung die in der Wirksamkeitsstudie untersucht worden war und das Ergebnis des Tests beeinflusst die Therapie. ✪ die Population der Wirksamkeitsstudie muss mit der in der diagnostischen Genauigkeitsstudie vergleichbar sein.	✪ Populationen von diagnostischen Genauigkeitsstudien und Therapiestudien sind vergleichbar ✪ der Nutzen der Therapie ist in hochwertigen Studien belegt ✪ die diagnostische Genauigkeit des Indextests wurde in hochwertigen Studien etabliert ✪ Ergebnis des Indextests ist Einschlusskriterium in der Wirksamkeitsstudie	-	✪ Diagnose- und Wirksamkeitsstudie sind vergleichbar in Bezug auf Patientenspektrum, der Erkrankung, dem Test und anderen Charakteristika

Besonderheiten	Mittels Vorher-Nachher Studien muss belegt werden, dass die durch ein Testergebnis gewonnene Information ausreicht, um therapeutische/diagnostische Entscheidungen beeinflussen zu können. Diese Studien sind überflüssig, wenn der Indextest als Ersatztest geplant ist und eine Standardtherapie für die Zielerkrankung klar definiert ist	Diagnostische Kette unnötig, wenn lediglich Varianten eines bereits etablierten Tests untersucht werden		Mittels Vorher-Nachher Studien muss belegt werden, dass die durch ein Testergebnis gewonnene Information ausreicht, um therapeutische/diagnostische Entscheidungen beeinflussen zu können. Diese Studien sind überflüssig, wenn der Indextest als Ersatztest geplant ist und eine Standardtherapie für die Zielerkrankung klar definiert ist

Direkte Evidenz

Patientennutzen kann direkt etabliert werden

dann unterscheidet sich die Bewertung dieser Studien nicht von der von Interventionen

In den einfachsten Fällen sind vergleichende Studien vorhanden, die den Indextest und anschließende Therapien mit einem existierenden diagnostischen Verfahren und einer daran anschließende Therapien hinsichtlich ihrer Unterschiede auf patientenrelevanten Endpunkten untersuchen. Als bestes Studiendesign werden von allen Instituten dabei RCTs gesehen (siehe Tabelle 5.1-1). Das Prozedere bei der Bewertung des Nutzens und der Risiken unterscheidet sich dann nicht von dem zur Bewertung von Interventionen.

NICE erwähnt aber Besonderheiten

NICE diskutiert allerdings Besonderheiten von Endpunkten diagnostischer Verfahren (siehe Appendix 11.3):

- selbst wenn direkte Evidenz die Verbesserung von patientenrelevanten Ergebnissen durch eine an ein Testergebnis anschließende Therapie belegt, sind in der Studienpopulation auch falsch negative PatientInnen enthalten, die keine Therapie erhalten. Ebenso können auch nachteilige Effekte entstehen, wenn PatientInnen mit falsch positivem Testergebnis behandelt werden, obwohl sie tatsächlich gesund sind.
- nachteilige Effekte auch durch dem Indextest nachfolgende Tests entstehen können.
- Therapien selbst mit unerwünschten Nebenwirkungen vergesellschaftet sind.
- der Wert prognostischer Informationen für den individuellen PatientIn schwer zu quantifizieren ist.
- der Zeitpunkt an dem ein Test durchgeführt wird, beeinflusst einerseits die diagnostische Genauigkeit und andererseits aber auch die Wirksamkeit einer Therapie.

EUnetHTA: eigenes Kapitel zur Bewertung der Sicherheit

EUnetHTA's Methodenbuch behandelt darüber hinaus in einem eigenen Kapitel auch Fragen zur Sicherheit von diagnostischen Verfahren (siehe Appendix 11.4). Sicherheitsaspekte sind besonders wichtig, wenn

- mit der Technologie besondere Risiken vergesellschaftet sind.
- das Nutzen-Risiko Profil nicht eindeutig ist.
- mehrere Tests mit ähnlicher Genauigkeit, aber mit unterschiedlichen Sicherheitsprofilen zur Diagnose ein- und derselben Erkrankung verwendet werden können.
- die Zahl der falsch positiven Ergebnisse hoch ist.
- Nebenwirkungen die Akzeptanz und den Einsatz des Tests untergraben könnten.

Als Methode, mittels derer Nutzen und Schaden gegeneinander aufgewogen werden können, erwähnen MSAC, NICE und EUnetHTA entscheidungsanalytische Modelle (siehe Kapitel 5.2.4).

Da aber Studien, die direkt Auswirkungen des Indextests auf patientenrelevante Endpunkte erheben und mit denen des Referenztests vergleichen selten verfügbar sind, wird in den meisten Fällen keine direkte Evidenz vorhanden sein.

Linked Evidence

MSAC, IQWiG und EUnetHTA erwähnen „linked Evidence"

Von den drei Institutionen, die „linked Evidence" in ihren Methoden erwähnen, das sind MSAC, IQWiG und EUnetHTA, beziehen sich sowohl das IQWiG, als auch das „Core Model" von EUnetHTA auf das MSAC Handbuch. NICE erwähnt in seinen Richtlinien zwar, dass sich der Nutzen eines diagnostischen Verfahrens erst durch eine Beeinflussung von patientenrelevanten Endpunkten entfaltet, erwähnt „linked Evidence" aber nicht. Auch in dem vorläufigen Methoden Bericht zur Bewertung von Diagnostika findet sich kein Hinweis darauf.

Definition

Definition

Unter „linked Evidence" wird die Verknüpfung von Wirksamkeitsdaten aus hochwertigen, vergleichenden Studien mit übertragbarer und hochwertiger Evidenz aus diagnostischen Genauigkeitsstudien, die den Indextest mit dem Referenzstandard vergleichen, verstanden (siehe Kapitel 4.4.2).

Folgende Fragen sind daher vor „linked Evidenz" zu beantworten:

für „linked Evidence" zu beantwortende Fragen

1. Ist die Therapie wirksam?

2. Entspricht die durch den Indextest entdeckte Erkrankung, der Erkrankung, die in den Wirksamkeitsstudien untersucht wurde?

3. Sind die Ergebnisse der diagnostischen Genauigkeitsstudie auf die Wirksamkeitsstudien übertragbar?

Diese drei Fragen müssen vor der Verwendung von „linked Evidence" erörtert werden, um deren Anwendung zu rechtfertigen.

„Linked Evidence" ist also gerechtfertigt ist, wenn

- der neue Test kostengünstiger, nicht-invasiv und ein Ersatz für einen bereits bestehenden Test oder Teststrategie ist. Dann reicht die alleinige Bewertung der diagnostischen Genauigkeit aus.
- Wenn eine wirksame Therapie für die durch den Indextest erkannte Erkrankung existiert, der Indextest die Zielerkrankung in genau demselben Krankheitsstadium wie der bereits existierende Test entdeckt, das Krankheitsspektrum der Population von Diagnose- und Wirksamkeitsstudie vergleichbar ist.

Nachweis, dass Test Behandlungsentscheidungen verändern kann wird von Institutionen unterschiedlich diskutiert

Normalerweise muss bei „linked Evidence" auch nachgewiesen werden, dass ein neuer Test Behandlungsentscheidungen überhaupt beeinflussen kann – dies würde Studien der Level 3 und 4 entsprechen. Am besten eignen sich dafür Vorher-Nachher Studien (siehe auch Kapitel 4.3.1). Der Nachweis einer Änderung des Patientenmanagements ist aber unnötig, wenn der Indextest als Ersatz für einen anderen Test dienen soll und eine Standardtherapie für die Erkrankung klar definiert ist. Wie erwähnt bestehen aber Unterschiede zwischen den Institutionen in der Bedeutung, die diesen Studien beigemessen wird (sieh Kapitel 5.2.2).

Ungelöste Fragen sind, wenn

offene Fragen bei „linked Evidence"

- als Vergleichstest ein ungenauer Referenzstandard herangezogen wird, dann ist die Einschätzung der diagnostischen Genauigkeit unzuverlässig. In diesen Fällen wird direkte Evidenz benötigt.

Fehlen eines etablierten Referenzstandards,

- Indextest und Referenztest in zwei unterschiedlichen Studien untersucht wurden. Die Übertragbarkeit der Ergebnisse ist dann nur bei hochwertigen Studiendesigns die vergleichbare Populationen untersuchen möglich.

Index- und Referenztest wurden nicht direkt miteinander verglichen,

- Der in Studien untersuchte Einsatz des Indextests nicht mit dem übereinstimmt, der in der Fragestellung formuliert worden war.

Studienpopulation von Wirksamkeits- und diagnostischer Genauigkeitsstudie unterscheiden sich

- wenn sich die Studienpopulation der diagnostischen Genauigkeitsstudien von der in Wirksamkeitsstudien unterscheidet.

5.2.4 Entscheidungsanalyse

Entscheidungsanalysen um Nutzen und Risiken richtiger, aber auch falscher Befunde gegeneinander abzuwägen von drei Institutionen erwähnt

Drei Institutionen (MSAC, NICE, EUNnetHTA) erwähnen (siehe Tabelle 5.1-1), dass entscheidungsanalytische Modelle bei der Abwägung von Nutzen gegen die mit einem Verfahren vergesellschafteten Risiken hilfreich sein können, wobei dabei auch die Konsequenzen falsch positiver und falsch negativer Befunde berücksichtigt werden können, oder auch Unterschiede der zugrundeliegenden Krankheitsprävalenz modelliert werden können.

5.2.5 Level 6 – Nutzen aus gesellschaftlicher Sicht

Kosten-Nutzen-Berechnung von 3 Institutionen

Bis auf das IQWiG berücksichtigen alle Organisationen auch die mit diagnostischen Verfahren vergesellschafteten Kosten.

Laut NICE sind einfache Kostenabschätzungen dann ausreichend, wenn die neue Technologie einen höheren Nutzen und geringere Nebenwirkungen als die Vergleichstechnologie hat; laut MSAC dann, wenn der Indextest kostengünstiger und „wirksamer" ist. Andernfalls sind volle ökonomische Evaluationen gewünscht, wobei Kosten-Nutzwertanalysen bevorzugt werden (siehe Tabelle 5.1-1).

generell keine Unterschiede zu Interventionen

Generell sind dieselben Vorgehensweisen wie bei Interventionen einzuhalten, wobei aber das MSAC spezifisch erwähnt, dass sich etwa Sensitivitätsanalysen eignen können, um Unterschiede in Krankheitsprävalenzen, Sensitivität und Spezifität bei unterschiedlichen Grenzwerten oder Unsicherheiten in Bezug auf die therapeutische Konsequenzen zu untersuchen.

NICE erwähnt, dass Modelle von ökonomischen Evaluationen von Tests sehr komplex sein können

Der vorläufige Methodenbericht von NICE zur Bewertung diagnostischer Verfahren, befasst sich am ausführlichsten mit den bei der Modellierung von Kosten und Nutzen diagnostischer Verfahren verbundenen Schwierigkeiten (siehe Appendix 11.3). Die generelle Vorgehensweise ist zwar wiederum dieselbe wie bei Interventionen auch, allerdings zeichnen sich ökonomische Evaluationen diagnostischer Verfahren durch eine wesentliche höher Komplexität aus, da

1. der gesamte Behandlungspfad modelliert werden muss.
2. Variationen in Patientensubgruppen in Betracht gezogen werden sollten.
3. alle relevanten Teststrategien und –alternativen, die sehr zahlreich sein können, in dem Modell zu berücksichtigen sind.
4. je nach Verwendungszweck des Tests (z.B. Screening, Diagnose) unterschiedliche Faktoren, wie etwa der Zeitpunkt wann ein Screening durchgeführt wird oder die Konsequenzen falsch positiver/falsch negativer Befunde in die Kosteneffektivitätsanalysen mit einfließen sollten.

Methoden um Modelle zu vereinfachen: wenn Langzeitdaten verfügbar sind, bereits Modelle vorhanden sind, direkte Evidenz existiert

Modelle können allerdings vereinfacht werden, wenn

- direkte Evidenz aus randomisierten, kontrollierten Studien vorhanden ist.
- bereits Modelle für einzelne Schritte vorhanden sind.
- Studien mit Langzeitfollow-up vorhanden sind, da dadurch die Modellierung von Zwischenstufen unnötig wird.

oder durch

- reverse Modellierung, wobei anhand eines Modells die minimal erforderliche Testgenauigkeit berechnet wird, die notwendig ist, um Kosteneffektivität zu erreichen.

5.2.6 Empfehlungen

Empfehlungen der Institutionen basieren auf Abwägung von Nutzen, Risiken und Kosten

Schlussfolgerungen bei der Bewertungen diagnostischer Verfahren basieren bei allen Institutionen auf der Abwägung von Nutzen und Risiken einer Technologie und, in den meisten Fällen, auch auf der Berücksichtigung der finanziellen Konsequenzen, die mittels entscheidungsanalytischer und ökonomischer Modelle untersucht werden können.

Bei der Formulierung von Empfehlungen sollte die Stärke der Evidenz, die Größe des Effekts und die Konsistenz der Ergebnisse berücksichtigt werden. Andere Überlegungen wie soziale, ethische, rechtliche oder institutionelle Aspekte können ebenfalls eine Rolle spielen (siehe Appendix 11).

5.3 Zusammenfassung

- Die Methode der Wahl zur Evaluierung diagnostischer Verfahren ist eine systematische Übersichtsarbeit.
- Eine genaue Formulierung der PICO Frage ist gerade bei diagnostischen Maßnahmen besonders wichtig, da Patientencharakteristika, Krankheitsprävalenz, der geplante Einsatz des Indextest in bestehenden diagnostischen Strategien, aber auch die unterschiedlichen Versionen eines Tests die Testgenauigkeit beeinträchtigen können.
- Die Bewertung des mit einem Test verbundenen Nutzens sollte auf patientenrelevanten Endpunkten (Level 5) beruhen.

- Patientenrelevante Endpunkte können direkt durch vergleichende Studien (Level 5), die PatientenInnen zufällig zu Index- und Vergleichstest zuordnen, erhoben werden. Die Unterschiede in den patientenrelevanten Endpunkten in Abhängigkeit nachfolgender therapeutischer Konsequenzen werden dann zwischen den beiden Tests verglichen.
- Ist keine direkte Evidenz vorhanden, dann bietet „linked Evidence" eine Möglichkeit diagnostische Genauigkeitsstudien (Level 2) mit Wirksamkeitsstudien zu verknüpfen.
- Studien, die den Einfluss des Indextestergebnisses auf das diagnostische/therapeutische Management untersuchen (Level 3 & Level 4) werden im Rahmen von „linked Evidence" nicht benötigt, wenn es sich bei dem Indextest um einen Ersatztest handelt und die Standardtherapie klar definiert ist.
- Diagnostische Genauigkeitsstudien alleine sind ausreichend, wenn es sich bei dem Indextest um einen kostengünstigeren, nichtinvasiven Ersatztest handelt, der eine ähnliche Sensitivität wie der zu ersetzende Test besitzt.
- Diagnosestudien sollten direkt den Indextest mit dem Referenzstandard vergleichen.
- Die Bewertung von Diagnosestudien sollte mittels des QUADAS – Instruments oder durch die von der Cochrane Collaboration vorgeschlagenen Checkliste erfolgen.
- Die finanziellen Konsequenzen, die mit diagnostischen Verfahren einhergehen, sollten entweder durch einfache Kostenaufstellungen oder durch ökonomische Evaluationen (Level 6) erhoben werden, um letztlich entscheiden zu können, welche gesellschaftlichen Auswirkungen mit der Adaption eines neuen Verfahrens zu erwarten sind.
- Es bestehen noch zahlreiche ungeklärte methodische Fragestellungen bei der Evaluation von diagnostischen Verfahren, die Gegenstand wissenschaftlicher Forschung sind (z.B. ungenauer Referenzstandard, Studien nicht genau den geplanten Verwendungszweck des Indextests abbilden).

6 Fragenkatalog

Besonderheiten bei der Evaluation von diagnostischen Verfahren

Bedingt durch die bisherigen Ausführungen, wird ersichtlich, dass bei der Bewertung von diagnostischen Verfahren zahlreiche Eigenheiten zu berücksichtigen sind, da etwa

- der Nachweis eines patientenrelevanten Nutzens meist nicht direkt, sondern indirekt erfolgt.
- Diagnosestudien zahlreiche spezielle methodische Mängel aufweisen können, die die Qualität der Ergebnisse kompromittieren können.
- die Übertragbarkeit von Studienergebnissen auf andere Populationen sorgfältig geprüft werden muss.

anhand der kausalen Kette Spezifizierung von Möglichkeiten wo Nutzen und Schaden entstehen kann und an welchen Stellen Verbindung nachgewiesen werden muss

Um die einzelnen Schritte, die für den Nutzennachweis eines diagnostischen Verfahrens erfüllt werden müssen, als auch mögliche Stellen, an denen Nutzen und Risiken eines Tests entstehen können, zu verdeutlichen, können diese mittels einer kausalen Kette dargestellt werden (siehe Abbildung 6-1). Fehlt direkte Evidenz (entspricht in Abbildung 6-1 dem abgewinkelten Pfeil), dann müssen im Rahmen von „linked Evidence“ als Schlüsselfragen zumindest die diagnostische Genauigkeit und die Wirksamkeit der Therapie beantwortet werden. Außerdem sollten sowohl die Nebenwirkungen des Tests selbst (und auch etwaiger nötiger Nachfolgeuntersuchungen), als auch die Nebenwirkungen der Therapie in Betracht gezogen werden. Neben den Konsequenzen für richtige Befunde, sollten dabei auch immer die Konsequenzen von falschen Testergebnissen berücksichtigt werden.

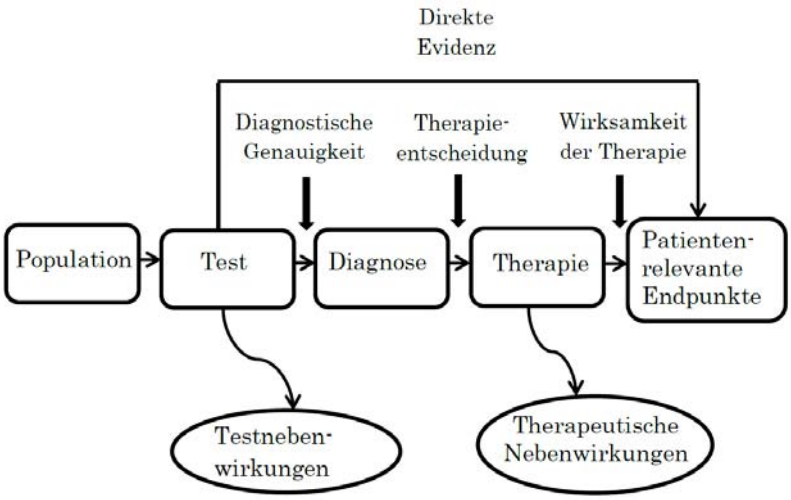


Abbildung 6-1: Kausale Kette und Determinanten der klinischen Effektivität von diagnostischen Verfahren (adaptiert nach: [8, 67])

aus Gemeinsamkeiten der Institutionen und der identifizierten Herausforderungen wurde ein Fragenkatalog abgeleitet, der EntscheidungsträgerInnen erlauben soll, Überblick über vorhandene Evidenzlage zu bekommen

Stehen KostenträgerInnen nun vor Kostenerstattungsentscheidungen über diagnostische Verfahren, sollten Überlegungen zu der Evidenzlage zu Nutzen, Schaden und Kosten angestellt werden. Basierend auf den Gemeinsamkeiten der Institutionen und der dargestellten spezifischen Herausforderungen, die mit der Evaluation von diagnostischen Verfahren einhergehen, wurde nun folgende Fragen abgeleitet, die für österreichische EntscheidungsträgerInnen bei der Strukturierung dieser Überlegungen dienlich sein können, indem sie erlauben, einen Überblick über die vorhandene Evidenzlage und deren kontextspezifische Relevanz zu gewinnen (siehe Tabelle 6-1).

Tabelle 6-1: Fragenkatalog zur Beurteilung der Evidenzlage von diagnostischen Verfahren

Testcharakteristika	
Testname, Version, Funktion, technische Angaben (z.B. Schwellenwerte, Inter-/Intraobserver Variabilität, Art und Menge des verwendeten Kontrastmittels, benötigte Expositionszeit, etc)?	Siehe Kapitel 4.2.5
Beschreibung der Testdurchführung, benötigte Expertise?	
Geplante Verwendung (Einsatzort des Tests innerhalb existierender diagnostischer Strategien, Add-On/Ersatz/ Triage)?	Siehe Kapitel 4.2.4
Zulassungsstatus in Österreich und anderen EU Ländern?	
Welche alternativen diagnostischen Verfahren existieren?	
Beschreibung des Referenzstandards, wie valide ist der Referenzstandard?	Siehe Kapitel 4.2.5
Patientencharakteristika	
Definition der Zielerkrankung, Stadium, Prävalenz, Inzidenz, Prognose, Krankheitsverlauf?	Siehe Kapitel 4.2.1
Definition des Patientenspektrums in dem der Test durchgeführt werden soll (Alter, Geschlecht, Ko-Morbidität, vorangegangen Tests)?	Siehe Kapitel 4.2.5
Beschreibung besonders zu berücksichtigender Subgruppen?	
Therapeutische Optionen, Standardtherapie?	
Nutzen	
Wie wurde der Nutzen etabliert (direkte vs indirekte Evidenz)	

Direkte Evidenz	Siehe Kapitel 4.4.1
Studiendesign?	
methodische Qualität?	
Ist die eingeschlossene Studienpopulation repräsentativ für die, in der der Test angewandt werden soll?	
Welcher Vergleichstest wurde gewählt? ist der gewählte Vergleichstest relevant für die klinische Praxis?	
Welche Endpunkte wurden erhoben (Wirksamkeit und Sicherheit)?	
95% Konfidenzintervalle, p- Werte, klinisch relevante Effekte, konsistente Studienergebnisse?	
Ergebnisse von Meta-Analysen?	
Indirekte Evidenz	
Evidenzlage der therapeutischen Effektivität	Siehe Kapitel 4.4.2
Studiendesign(s)?	
methodische Qualität?	
Berücksichtigte Endpunkte (Wirksamkeit und Sicherheit)?	
Studienergebnisse (inkl. 95% Konfidenzintervalle, p-Wert), klinisch relevante Effekte, konsistente Ergebnisse, Größe des Effekts?	
Evidenzlage Diagnosestudien	
Studiendesign?	Siehe Kapitel 4.2.2
methodische Qualität (anhand des QUADAS- Instruments, wobei a priori die für die Frage am wichtigsten Punkte festgelegt werden sollen)?	Siehe Kapitel 4.2.6
Mit welchem Test wurde der Indextest verglichen, entspricht der Vergleichstest dem Referenzstandard? Wenn nicht: ist der Vergleichstest relevant für die klinische Praxis, wurden die Tests mittels Referenzstandard verifiziert?	Siehe Kapitel 4.2.5
Wurden die Tests in den Studien direkt miteinander verglichen?	Siehe Kapitel 4.2.2
Wurden Subgruppen, in denen die diagnostische Genauigkeit variieren kann, berücksichtigt?	Siehe Kapitel 4.2.5
Endpunkte (inkl. 95% Konfidenzintervalle, p-Wert)? Nebenwirkungen des Tests	Siehe Kapitel 4.2.1
Sind die Ergebnisse konsistent?	

Übertragbarkeit	
Sind die Charakteristika der Studienpopulationen von Diagnose- und Wirksamkeitsstudie vergleichbar?	Siehe Kapitel 4.4.2
Entspricht die, durch den Indextest diagnostizierte Erkrankung, der Erkrankung für die die Wirksamkeit der Therapie nachgewiesen wurde?	
Entsprechen die untersuchten Studienpopulationen klinischen relevanten Populationen?	Siehe Kapitel 4.2.5
Entspricht die Verwendung des Indextest in den Studien, der geplanten Verwendung?	Siehe Kapitel 4.2.4
Kosten	
Listenpreis der Technologie Kosten von zusätzlich benötigtem Material Personalaufwand Bei Geräten: jährliche Wartungskosten Durchschnittliche Kosten/ Testdurchführung Entstehen die Kosten zusätzlich, oder ersetzt der Test ein bereits etabliertes Untersuchungsverfahren (ggf. Kosten des zu ersetzenden Tests)	Siehe Kapitel 4.5
Evidenz	
Gibt es Evidenz zu Kosteneffektivität?	Siehe Kapitel 4.5
Wurden klinisch relevante Behandlungs- und Diagnosepfade modellierten? Wurden relevante Endpunkte berücksichtigt? Wurden Unterschiede der Prävalenz, der Sensitivität und Spezifität in Abhängigkeit unterschiedlicher Grenzwerte, der diagnostischen Genauigkeit in Subgruppen berechnet? Wurden Konsequenzen falsch positiver/falsch negativer Befunde berücksichtigt?	Die hier angeführten Fragen sind immer in Kombination mit Standardchecklisten (siehe dafür z.B. Appendix 10) für ökonomische Evaluationen zu sehen

7 Appendix: Suchstrategie

7.1 Cochrane Database

Suche am 05. November 2009

#1 MeSH descriptor Diagnostic Techniques and Procedures, this term only

#2 MeSH descriptor Reproducibility of Results explode all trees

#3 MeSH descriptor Sensitivity and Specificity explode all trees

#4 MeSH descriptor Process Assessment (Health Care) explode all trees

#5 (#2 OR #3 OR #4)

#6 (#1 AND #5)

#7 MeSH descriptor Evidence-Based Medicine explode all trees

#8 (#5 AND #7)

#9 MeSH descriptor Diagnostic Techniques and Procedures explode all trees

#10 (#8 AND #9)

#11 MeSH descriptor Methods explode all trees with qualifiers: DU,DI,ST

#12 (#5 AND #11)

#13 (#9 AND #12)

#14 (#6 OR #10 OR #13)

7.2 CRD Datenbank

Suche am 5. November 2009

MeSH Diagnostic Techniques and Procedures

MeSH Reproducibility of Results EXPLODE 1 2 3 4

MeSH Sensitivity and Specificity EXPLODE 1 2 3 4 5 6 7

MeSH Process Assessment (Health Care) EXPLODE 1 2

#2 OR #3 OR #4

#1 AND #5

MeSH Evidence-Based Medicine EXPLODE 1 2

#5 AND #7

MeSH Diagnostic Techniques and Procedures EXPLODE 1

#8 AND #9

MeSH Methods EXPLODE 1

#5 AND #11

#9 AND #12

#6 OR #10 OR #13

7.3 Embase

Suche am 4. November 2009

#17. #7 OR #11 OR #16

#16. #1 AND #15

#15. #4 AND #14

#14. 'methods'/de OR 'methods'

#11. #1 AND #10

#10. #8 AND #9

#9. #4 OR #6

#8. 'evidence-based medicine'/exp

#7. #1 AND #4 AND #6

#6. 'process assessment (health care)'/syn

#4. #2 OR #3

#3. 'sensitivity and specificity'/exp

#2. 'reproducibility of results'/exp

#1. 'diagnostic techniques and procedures'/mj

7.4 Medline

Suche am 4. November 2009

1 *"Diagnostic Techniques and Procedures"/ (948)

2 exp "Reproducibility of Results"/ (192178)

3 exp "Sensitivity and Specificity"/ (303308)

4 exp "Process Assessment (Health Care)"/ (2337)

5 3 or 2 or 4 (431832)

6 1 and 5 (280)

7 exp Evidence-Based Medicine/ (36023)

8 7 and 5 (1939)

9 8 and 1 (11)

10 exp Methods/ (500186)

11 10 and 5 (15850)

12 1 and 11 (60)

13 6 or 9 or 12 (280)

8 Appendix: Instrumente zur Qualitätsbewertung diagnostischer Genauigkeitsstudien

8.1 Das QUADAS Tool

The QUADAS tool

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

Whiting *et al. BMC Medical Research Methodology* 2003 **3**:25 doi:10.1186/1471-2288-3-25

Abbildung 8.1-1: Das QUADAS – Instrument zur Bewertung von diagnostischen Genauigkeitsstudien in systematischen Reviews (Quelle: [57])

8.2 Checkliste nach Cochrance Collaboration

Table 9.1 Recommended quality items derived from QUADAS tool (Whiting 2003)

1.	Was the spectrum of patients representative of the patients who will receive the test in practice? (representative spectrum)
2.	Is the reference standard likely to classify the target condition correctly? (acceptable reference standard)
3.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? (acceptable delay between tests)
4.	Did the whole sample or a random selection of the sample, receive verification using the intended reference standard? (partial verification avoided)
5.	Did patients receive the same reference standard irrespective of the index test result? (differential verification avoided)
6.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (incorporation avoided)
7.	Were the reference standard results interpreted without knowledge of the results of the index test? (index test results blinded)
8.	Were the index test results interpreted without knowledge of the results of the reference standard? (reference standard results blinded)
9.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (relevant clinical information)
10.	Were uninterpretable/ intermediate test results reported? (uninterpretable results reported)
11.	Were withdrawals from the study explained? (withdrawals explained)

Abbildung 8.2-1: Checkliste der Cochrane Collaboration zur Bewertung der methodischen Qualität von diagnostischen Genauigkeitsstudien (Quelle: [38])

8.3 STARD Checkliste

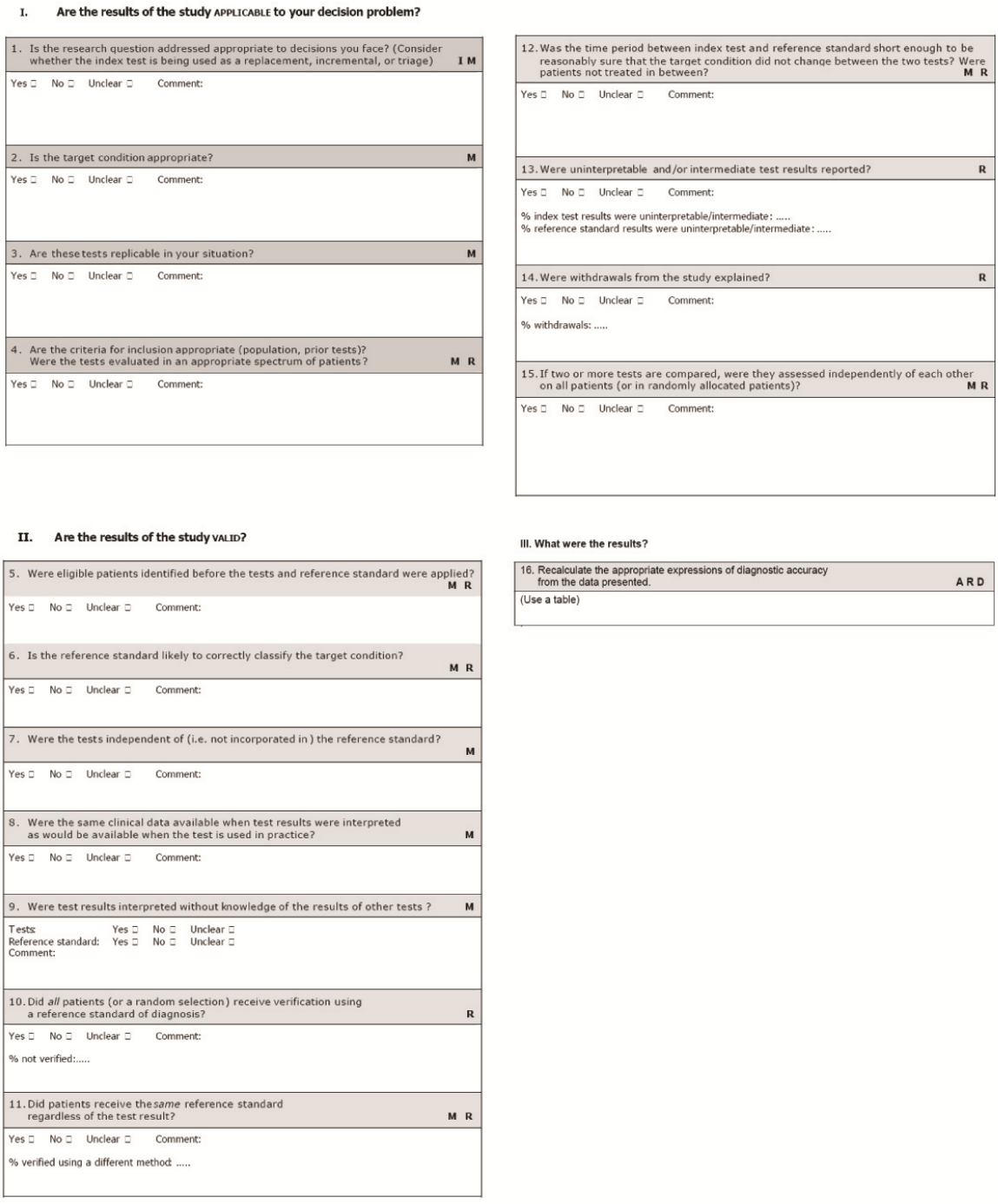
STARD checklist for the reporting of studies of diagnostic accuracy.
First official version, January 2003.

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS			
Participants	3	Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	
	4	Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Describe participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.	
	6	Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	Describe the reference standard and its rationale.	
	8	Describe technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Describe definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.	
	10	Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.	
	11	Describe whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	
	13	Describe methods for calculating test reproducibility, if done.	
RESULTS			
Participants	14	Report when study was done, including beginning and ending dates of recruitment.	
	15	Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	
	16	Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	
Test results	17	Report time interval from the index tests to the reference standard, and any treatment administered between.	
	18	Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Report any adverse events from performing the index tests or the reference standard.	
Estimates	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	
	22	Report how indeterminate results, missing responses and outliers of the index tests were handled.	
	23	Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Report estimates of test reproducibility, if done.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

Abbildung 8.3-1: Checkliste zur Berichterstattung diagnostischer Genauigkeitsstudien nach STARD (Quelle: [47])

8.4 Weitere Instrumente zur Qualitätsbewertung

8.4.1 Diagnostic Test Appraisal Form of the Screening and Test Evaluation Programme (STEP)



I. Are the results of the study APPLICABLE to your decision problem?

1. Is the research question addressed appropriate to decisions you face? (Consider whether the index test is being used as a replacement, incremental, or triage) **I M**
Yes ☐ No ☐ Unclear ☐ Comment:

2. Is the target condition appropriate? **M**
Yes ☐ No ☐ Unclear ☐ Comment:

3. Are these tests replicable in your situation? **M**
Yes ☐ No ☐ Unclear ☐ Comment:

4. Are the criteria for inclusion appropriate (population, prior tests)? Were the tests evaluated in an appropriate spectrum of patients? **M R**
Yes ☐ No ☐ Unclear ☐ Comment:

12. Was the time period between index test and reference standard short enough to be reasonably sure that the target condition did not change between the two tests? Were patients not treated in between? **M R**
Yes ☐ No ☐ Unclear ☐ Comment:

13. Were uninterpretable and/or intermediate test results reported? **R**
Yes ☐ No ☐ Unclear ☐ Comment:
% index test results were uninterpretable/intermediate: .....
% reference standard results were uninterpretable/intermediate: .....

14. Were withdrawals from the study explained? **R**
Yes ☐ No ☐ Unclear ☐ Comment:
% withdrawals: .....

15. If two or more tests are compared, were they assessed independently of each other on all patients (or in randomly allocated patients)? **M R**
Yes ☐ No ☐ Unclear ☐ Comment:

II. Are the results of the study VALID?

5. Were eligible patients identified before the tests and reference standard were applied? **M R**
Yes ☐ No ☐ Unclear ☐ Comment:

6. Is the reference standard likely to correctly classify the target condition? **M R**
Yes ☐ No ☐ Unclear ☐ Comment:

7. Were the tests independent of (i.e. not incorporated in) the reference standard? **M**
Yes ☐ No ☐ Unclear ☐ Comment:

8. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? **M**
Yes ☐ No ☐ Unclear ☐ Comment:

9. Were test results interpreted without knowledge of the results of other tests ? **M**
Tests: Yes ☐ No ☐ Unclear ☐
Reference standard: Yes ☐ No ☐ Unclear ☐
Comment:

10. Did *all* patients (or a random selection) receive verification using a reference standard of diagnosis? **R**
Yes ☐ No ☐ Unclear ☐ Comment:
% not verified:.....

11. Did patients receive the *same* reference standard regardless of the test result? **M R**
Yes ☐ No ☐ Unclear ☐ Comment:
% verified using a different method: .....

III. What were the results?

16. Recalculate the appropriate expressions of diagnostic accuracy from the data presented. **A R D**
(Use a table)

Abbildung 8.4-1:Bewertung diagnostischer Genauigkeitsstudien nach STEP (Quelle: [47])

8.4.2 Checkliste nach Hayden zur Bewertung der methodischen Qualität von Prognosestudien

Study identification *Include author, title, reference, year of publication*				
Guideline topic:		**Review question no:**		
Checklist completed by:				
		Circle one option for each question		
1.1	The study sample represents the population of interest with regard to key characteristics, sufficient to limit potential bias to the results	Yes	No	Unclear
1.2	Loss to follow-up is unrelated to key characteristics (that is, the study data adequately represent the sample), sufficient to limit potential bias	Yes	No	Unclear
1.3	The prognostic factor of interest is adequately measured in study participants, sufficient to limit potential bias	Yes	No	Unclear
1.4	The outcome of interest is adequately measured in study participants, sufficient to limit bias	Yes	No	Unclear
1.5	Important potential confounders are appropriately accounted for, limiting potential bias with respect to the prognostic factor of interest	Yes	No	Unclear
1.6	The statistical analysis is appropriate for the design of the study, limiting potential for the presentation of invalid results	Yes	No	Unclear

Abbildung 8.4-2 Checkliste zur Bewertung der methodischen Qualität von Prognosestudien (Quelle: [79])

9 Appendix: Evidenzhierarchien von Studien zur Bewertung diagnostischer Verfahren

9.1 National Health and Medical Research Council - Evidence Hierarchy

Level	Intervention [1]	Diagnostic accuracy [2]
I [4]	A systematic review of level II studies	A systematic review of level II studies
II	A randomised controlled trial	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among consecutive persons with a defined clinical presentation[6]
III-1	A pseudorandomised controlled trial (i.e. alternate allocation or some other method)	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among non-consecutive persons with a defined clinical presentation[6]
III-2	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial[9] ▪ Cohort study ▪ Case-control study ▪ Interrupted time series with a control group	A comparison with reference standard that does not meet the criteria required for Level II and III-1 evidence
III-3	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study[10] ▪ Interrupted time series without a parallel control group	Diagnostic case-control study[6]
IV	Case series with either post-test or pre-test/post-test outcomes	Study of diagnostic yield (no reference standard)[11]

Abbildung 9.1-1: Evidenzhierarchie nach National Health and Medical Research Council (NHMRC) (Quelle: [41])

9.2 Centre for Evidence Based Medicine – Levels of Evidence

Level	Prognosis	Diagnosis
1a	SR (with homogeneity*) of inception cohort studies; CDR" validated in different populations	SR (with homogeneity*) of Level 1 diagnostic studies; CDR" with 1b studies from different clinical centres
1b	Individual inception cohort study with > 80% follow-up; CDR" validated in a single population	Validating** cohort study with good" " " reference standards; or CDR" tested within one clinical centre
1c	All or none case-series	Absolute SpPins and SnNouts" "
2a	SR (with homogeneity*) of either retrospective cohort studies or untreated control groups in RCTs	SR (with homogeneity*) of Level >2 diagnostic studies
2b	Retrospective cohort study or follow-up of untreated control patients in an RCT; Derivation of CDR" or validated on split-sample§§§ only	Exploratory** cohort study with good" " " reference standards; CDR" after derivation, or validated only on split-sample§§§ or databases
2c	"Outcomes" Research	
3a		SR (with homogeneity*) of 3b and better studies
3b		Non-consecutive study; or without consistently applied reference standards
4	Case-series (and poor quality prognostic cohort studies***)	Case-control study, poor or non-independent reference standard
5	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"	Expert opinion without explicit critical appraisal, or based on physiology, bench research or "first principles"

Abbildung 9.2-1: Centre for Evidence Based Medicine Evidenzhierarchie für diagnostische und prognostische Verfahren [42]

* By homogeneity we mean a systematic review that is free of worrisome variations (heterogeneity) in the directions and degrees of results between individual studies. Not all systematic reviews with statistically significant heterogeneity need be worrisome, and not all worrisome heterogeneity need be statistically significant. As noted above, studies displaying worrisome heterogeneity should be tagged with a "-" at the end of their designated level.

" Clinical Decision Rule. (These are algorithms or scoring systems that lead to a prognostic estimation or a diagnostic category.)

§§§ Split-sample validation is achieved by collecting all the information in a single tranche, then artificially dividing this into "derivation" and "validation" samples.

" " An "Absolute SpPin" is a diagnostic finding whose Specificity is so high that a Positive result rules-in the diagnosis. An "Absolute SnNout" is a diagnostic finding whose Sensitivity is so high that a Negative result rules-out the diagnosis.

" " " Good reference standards are independent of the test, and applied blindly or objectively to applied to all patients. Poor reference standards are haphazardly applied, but still independent of the test. Use of a non-independent reference standard (where the 'test' is included in the 'reference', or where the 'testing' affects the 'reference') implies a level 4 study.

** Validating studies test the quality of a specific diagnostic test, based on prior evidence. An exploratory study collects information and trawls the data (e.g. using a regression analysis) to find which factors are 'significant'.

*** By poor quality prognostic cohort study we mean one in which sampling was biased in favour of patients who already had the target outcome, or the measurement of outcomes was accomplished in <80% of study patients, or outcomes were determined in an unblinded, non-objective way, or there was no correction for confounding factors.

10 Appendix: Checklisten für ökonomische Evaluationen - Beispiele

10.1 British Medical Journal Checklist

Referees' checklist (also to be used, implicitly, by authors)

Item	Yes	No	Not clear	Not appropriate
Study design				
(1) The research question is stated	☐	☐	☐	
(2) The economic importance of the research question is stated	☐	☐	☐	
(3) The viewpoint(s) of the analysis are clearly stated and justified	☐	☐	☐	
(4) The rationale for choosing the alternative programmes or interventions compared is stated	☐	☐	☐	
(5) The alternatives being compared are clearly described	☐	☐	☐	
(6) The form of economic evaluation used is stated	☐	☐	☐	
(7) The choice of form of economic evaluation is justified in relation to the questions addressed	☐	☐	☐	
Data collection				
(8) The source(s) of effectiveness estimates used are stated	☐	☐	☐	
(9) Details of the design and results of effectiveness study are given (if based on a single study)	☐	☐	☐	☐
(10) Details of the method of synthesis or meta-analysis of estimates are given (if based on an overview of a number of effectiveness studies)	☐	☐	☐	☐
(11) The primary outcome measure(s) for the economic evaluation are clearly stated	☐	☐	☐	
(12) Methods to value health states and other benefits are stated	☐	☐	☐	☐
(13) Details of the subjects from whom valuations were obtained are given	☐	☐	☐	☐
(14) Productivity changes (if included) are reported separately	☐	☐	☐	☐
(15) The relevance of productivity changes to the study question is discussed	☐	☐	☐	☐
(6) Quantities of resources are reported separately from their unit costs	☐	☐	☐	
(17) Methods for the estimation of quantities and unit costs are described	☐	☐	☐	
(18) Currency and price data are recorded	☐	☐	☐	
(19) Details of currency of price adjustments for inflation or currency conversion are given	☐	☐	☐	
(20) Details of any model used are given	☐	☐	☐	☐
(21) The choice of model used and the key parameters on which it is based are justified	☐	☐	☐	☐
Analysis and interpretation of results				
(22) Time horizon of costs and benefits is stated	☐	☐	☐	
(23) The discount rate(s) is stated	☐	☐	☐	☐
(24) The choice of rate(s) is justified	☐	☐	☐	☐
(25) An explanation is given if costs or benefits are not discounted	☐	☐	☐	☐
(26) Details of statistical tests and confidence intervals are given for stochastic data	☐	☐	☐	☐
(27) The approach to sensitivity analysis is given	☐	☐	☐	☐
(28) The choice of variables for sensitivity analysis is justified	☐	☐	☐	☐
(29) The ranges over which the variables are varied are stated	☐	☐	☐	☐
(30) Relevant alternatives are compared	☐	☐	☐	
(31) Incremental analysis is reported	☐	☐	☐	☐
(32) Major outcomes are presented in a disaggregated as well as aggregated form	☐	☐	☐	
(33) The answer to the study question is given	☐	☐	☐	
(34) Conclusions follow from the data reported	☐	☐	☐	
(35) Conclusions are accompanied by the appropriate caveats	☐	☐	☐	

Abbildung 10.1-1: Checkliste für ökonomische Evaluationen nach British Medical Journal (Quelle: [82])

10.2 Checkliste der gesundheitsökonomischen Projektgruppen München, Hannover, Ulm

Checkliste methodische Qualität		
Autoren, Titel und Publikationsorgan:	1 = Kriterium erfüllt 1/2 = Kriterium teilweise erfüllt 0 = Kriterium nicht erfüllt nr = nicht relevant	1, ½, 0, nr
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 univariate Sensitivitätsanalysen für die relevanten Parameter durchgeführt? 43. Wurden multivariate Sensitivitätsanalysen für die relevanten Parameter durchgeführt? 44. Wurde Sensitivitätsanalysen für die relevanten strukturellen Elemente durchgeführt? 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?	
Diskussion 48. Wurde die Datenqualität kritisch beurteilt? 49. Wurden Richtung und Größe des Einflusses unsicherer oder verzerrter Parameterschätzung auf das Ergebnis konsistent diskutiert? 50. Wurde Richtung und Größe des Einflusses struktureller Modellannahmen auf das Ergebnis konsistent diskutiert? 51. Wurden die wesentlichen Einschränkungen und Schwächen der Studie diskutiert? 52. Wurden plausible Angaben zur Generalisierbarkeit der Ergebnisse gemacht? 53. Wurden wichtige ethische und Verteilungsfragen diskutiert? 54. Wurde das Ergebnis sinnvoll im Kontext mit unabhängigen Gesundheitsprogrammen diskutiert?	
Schlussfolgerungen 55. Wurden in konsistenter Weise Schlussfolgerungen aus den berichteten Daten/Ergebnissen abgeleitet? 56. Wurde eine auf Wissensstand und Studienergebnissen basierende Antwort auf die Fragestellung gegeben?	

Abbildung 10.2-1: Checkliste zur Beurteilung der methodischen Qualität gesundheitsökonomischer Studien entwickelt im Konsensusverfahren von den gesundheitsökonomischen Projektgruppen München, Hannover, Ulm) (Quelle: [83])

11 Appendix: Methoden ausgewählter Institutionen für die Nutzenbewertung diagnostischer Verfahren

11.1 MSAC

Das australische "Medical Services Advisory Committee" (MSAC) wurde 1998 als beratendes Organ für den Gesundheitsminister gegründet, um Kostenerstattungsentscheidungen von neuen oder bereits existierenden Technologien zu erleichtern.

Die Bewertung diagnostischer Verfahren beruht auf einer systematischen Review der besten verfügbaren Evidenz über die relative Sicherheit, Wirksamkeit und Kosteneffektivität des Indextests im Vergleich zu einem bereits existierenden diagnostischen Verfahren für die Zielerkrankung [84].

Die Vorgehensweisen zur Bewertung diagnostischer Verfahren ist in einer im August 2005 publizierten Richtlinie beschrieben [67].

11.1.1 Allgemeine Methodik

Die Methode zur Evaluation von diagnostischen Verfahren ist eine systematische Übersichtsarbeit, bei der prinzipiell dieselben Methoden angewandt werden, die auch bei systematischen Reviews über Interventionen angezeigt sind (z.B. Auswahl relevanter Literaturstellen und Datenextraktion durch zwei unabhängige WissenschafterInnen, Bewertung der methodischen Qualität der eingeschlossenen Studien, wenn möglich, Präsentation der Ergebnisse durch Meta-Analysen).

Die „klinische Effektivität" eines diagnostischen Verfahrens hängt davon ab, ob durch seinen Einsatz (Add-On oder Ersatz) die Gesamtgenauigkeit verbessert wird und ob ein Einfluss auf therapeutische Entscheidungen gegeben ist und wird letztlich von der Wirksamkeit der anschließenden Therapie bestimmt (siehe Abbildung 11.1-1). Ausnahmen stellen Verfahren dar, die für rein prognostische Aussagen durchgeführt werden, da deren Wirksamkeit von nachfolgenden Therapien unabhängig ist. Bei der Bewertung diagnostischer Verfahren stellen sich also spezielle methodische Herausforderungen, die in weiterer Folge dargestellt werden (Medical Services Advisory Committee 2005).

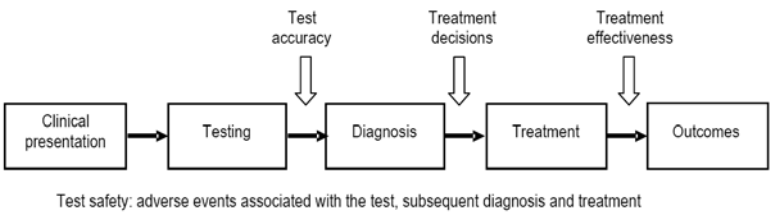


Abbildung 11.1-1 Kausaler Zusammenhang und Determinanten, die die klinische Effektivität eines diagnostischen Tests bedingen (Quelle: [67])

Forschungsfrage

Ein Studienprotokoll und damit die Formulierung der Fragestellung wird in Zusammenarbeit mit KlinikerInnen, WissenschafterInnen, und PatientInnen erstellt, wobei die Forschungsfrage dem **PPICO-Schema** folgen soll. Wie bei Interventionen auch, werden dadurch die **P**opulation, die **I**ntervention, der K**(C)**omparator und die zu untersuchenden **O**utcomes (Endpunkte) festgelegt. Das zweite **P** (prior tests) dient der genauen Definition des Kontexts, also der zu testenden Population, beziehungsweise des Krankheitsspektrums, das untersucht werden soll. Neben der Bestimmung des Vergleichstests (z.B. der in der klinischen Praxis relevanteste Test) muss auch festgelegt werden, was der derzeitige Referenzstandard ist, mit dem die diagnostische Genauigkeit des Indextest erhoben werden soll.

Auch sollte bereits im Studienprotokoll festgehalten werden, ob „linked Evidence" - bei Abwesenheit direkter Evidenz – angewandt werden wird.

Hintergrundinformationen

Im Hintergrundteil des Berichts sollten also folgende Punkte erörtert werden:

1. Intervention: Was sind personelle/technische Voraussetzungen für Einsatz des Indextests, was sind die technischen Angaben (Schwellenwerte für negative/positive Ergebnisse)? in welcher Population, bei welchen Symptomen soll der Indextest verwendet werden? ist der Indextest genauer als der Referenzstandard? verändert sich die diagnostische Genauigkeit des Indextests in unterschiedlichen Patientengruppen? wie hängen Indextest und patientenrelevanter Nutzen zusammen?
2. Verwendungszweck: Darstellung des geplanten Einsatzes und möglicher Outcomes mittels klinischer Behandlungspfade, ist der Test ein Add-on oder ein Ersatz-Test? welche Tests wurden bereits im Vorfeld durchgeführt?
3. Referenzstandard: Definition des Referenzstandards gegen den der Indextest verglichen werden soll? Wie geeignet ist der Referenzstandard zur Diagnose der Zielerkrankung?
4. bereits existierende Tests: gegen welchen Test/welche Teststrategien soll verglichen werden?
5. Zielerkrankung: Definition und Klassifizierung der Zielerkrankung, Inzidenz, Prävalenz, Morbidität, Mortalität, Einfluss auf Lebensqualität, welche Informationen kann der Indextest in Bezug auf die Zielerkrankung liefern?
6. Therapieoptionen: welche Therapieoptionen existieren? was ist die Standardtherapie für die Zielerkrankung? ist diese evidenzbasiert?
7. Potentieller Einfluss des Indextest: ist der Indextest genauer? werden dadurch neue PatientInnengruppen identifiziert, wodurch zusätzliche therapeutische Optionen entstehen?
8. Marketing Status und derzeitige Kostenerstattung

Ein klinisches Flow-Chart kann ein hilfreiches Mittel darstellen, um den geplanten Verwendungszweck des Indextests innerhalb des diagnostischen

Pfades, aber auch die Zusammenhänge zwischen Indextest und den gesundheitlichen Konsequenzen graphisch darzustellen.

Weitere Überlegungen können zu klinischem Bedarf, den Konsequenzen falsch positiver/falsch negativer Befunde und sozialer Ausgewogenheit angestellt werden.

11.1.2 Nutzenbewertung diagnostischer Verfahren

Die Bewertung diagnostischer Verfahren beruht auf Evidenz der Level 5 und 6. Da Studien, die direkt patientenrelevante Endpunkte untersuchen aber selten sind, können mittels „linked Evidence“ auch Studien der Level 2, 3 oder 4 als Grundlage für die Nutzenbewertung herangezogen werden müssen.

Level 2 – Testgenauigkeit

Studiendesign

Als beste Evidenz für die Testgenauigkeit gelten Querschnittsstudien, die den Indextest mit einem validen Referenzstandard in konsekutiv ausgewählten PatientInnen mit einer klar definierten, klinischen Symptomatik unter Verblindung vergleichen (siehe Tabelle 11.1-1).

Bewertung der Studienqualität

Die Beurteilung der Studienqualität aller durch die Literatursuche identifizierten Studien, sowie die Beurteilung der Generalisierbarkeit der Ergebnisse aus diagnostischen Studien können mittels unterschiedlicher Checklisten wie etwa QUADAS, STEP oder anhand einer Checkliste von Jaeschke erfolgen (siehe Appendix 8). Von diesen Checklisten werden nicht alle Komponenten für die Fragestellung relevant sein, sondern die wichtigsten Kriterien zur kritischen Bewertung sollten *a priori* festgelegt werden.

Als wichtigste Quellen, die Bias oder Variationen verursachen können, werden

1. Methoden und Kriterien, die zur Auswahl der Studienpopulation herangezogen wurden
2. die Auswahl des Referenzstandards
3. die Durchführung des Indextest(strategie), des Komparators und des Referenzstandards
4. die Interpretation von Indextest(strategie), des Komparators und des Referenzstandards
5. die Analyse der Ergebnisse

genannt.

MSAC erachtet von diesen fünf Kriterien Punkt 2, als maßgeblich, da die diagnostische Genauigkeit von Indextest gegen den Referenzstandard verglichen wird. Studien die einen unzureichenden Referenzstandard verwenden, sind daher entweder ganz auszuschließen oder als sehr niedrige Evidenz einzustufen. Methodische Herausforderungen sind daher, wenn

1. der Vergleichstest gleichzeitig der Referenzstandard ist, der Referenzstandard aber ungenauer als der Indextest ist, wodurch die dia-

gnostische Genauigkeit des Indextests kompromittiert wird. Dann wird der direkte Nachweis von Indextest auf patientenrelevante Ergebnisse benötigt.

2. es keine Studien gibt, die Indextest und Vergleichstest direkt miteinander vergleichen. Ein indirekter Vergleich von Studien, die den Indextest bewerten, mit Studien die den Vergleichstest bewerten ist nur dann zulässig, wenn hochwertige Studien beide Tests in ähnlichen Populationen und klinischen Settings bewerten.
3. Studien z.B. nur Abschnitte einer gesamten Teststrategien berücksichtigen, oder die Verwendung des Indextests (z.B. Add-on und nicht Ersatz) eine andere ist, als in der Fragestellung formuliert.

Abschließend werden alle Studien hinsichtlich ihrer Studienqualität, ihrer Relevanz für, und der Übertragbarkeit der Studienergebnisse auf die Fragestellung bewertet (siehe Abbildung 11.1-2).

Validity criteria	Description	Grading System
Appropriate comparison	Did the study evaluate a direct comparison of the index test strategy versus the comparator test strategy?	C1 direct comparison CX other comparison
Applicable population	Did the study evaluate the index test in a population that is representative of the subject characteristics (age and sex) and clinical setting (disease prevalence, disease severity, referral filter and sequence of tests) for the clinical indication of interest?	P1 applicable P2 limited P3 different population
Quality of study	Was the study designed to avoid bias? High quality = no potential for bias based on pre-defined key quality criteria Fair quality = some potential for bias in areas other than those pre-specified as key criteria Poor quality = poor reference standard and/or potential for bias based on key pre-specified criteria	Study design: NHMRC level of evidence Study quality: Q1 high quality Q2 fair quality Q3 poor reference standard poor quality

Abbildung 11.1-2: Bewertungsschema zur Evaluierung diagnostischer Verfahren

Datenextraktion

Wie üblich soll die Datenextraktion auch bei der Bewertung von diagnostischen Verfahren von zwei unabhängigen ReviewerInnen durchgeführt werden. Zu extrahierende Daten sind:

1. Studiencharakteristika
   - Autor, Publikationsjahr
   - Land, klinisches Setting, Studiendauer
   - Studiendesign
2. Studienpopulation und Testverfahren
   - Patientenzahl
   - Krankheitsprävalenz in der Studienpopulation
   - Methoden und Kriterien der Patientenselektion
   - Details zu Indextest, Vergleichstest und Referenzstandard

- Schwellenwert für ein positives Testergebnis

3. Studienbewertung
    - Studienqualität, methodische Schwäche und deren mögliche Auswirkungen auf Bias
    - Bewertung der Umlegbarkeit der Studienergebnisse und mögliche Limitationen für den eigenen Kontext
4. Studienergebnisse
    - Resultate mit 95% Konfidenzintervalle
    - Ergebnisse für Patientensubgruppen

Datenanalyse

Als wichtige Kenngrößen der diagnostischen Genauigkeit für systematische Reviews werden

- Sensitivität, Spezifität (inklusive des 95% Konfidenzintervalls)
- LR

und als zusammengesetzte Kenngrößen

- DOR
- ROC-Kurven

genannt. Durch die Abhängigkeit von der Prävalenz eignen sich PV nicht für systematische Reviews. Idealerweise sollten in allen eingeschlossenen Studien, ausreichend Informationen vorhanden sein, um diese Variablen mittels der Vierfelder-Tafel berechnen zu können.

Für alle Ergebnisse sollten die 95% Konfidenzintervalle angegeben werden und die p-Werte unter Berücksichtigung der Studiengröße, der Validität der verwendeten statistischen Methoden für die Datenanalyse, sowie der Anzahl der statistischen Vergleiche (je mehr Vergleiche, desto wahrscheinlicher ist es einen zufallsbedingten Effekt zu finden) bewertet werden. Neben der Präzision der Ergebnisse, sollten auch die Größe des Effekts kommentiert werden.

Sind die Studienergebnisse heterogen, dann ist die rein deskriptive Beschreibung der Resultate angezeigt; bei homogenen Ergebnissen jedoch erlauben Meta-Analysen eine bessere Einschätzung des Gesamteffekts.

Ob Meta-Analysen durchgeführt werden können, hängt davon ab, ob die Studienergebnisse sehr heterogen sind. Ursachen für Heterogenität wurden in Kapitel 4.2.5 und Instrumente, um Studien hinsichtlich ihres Potentials für Bias und Variationen zu bewerten, in Kapitel 4.2.6 beschrieben. MSAC gibt als mögliche Quellen für Heterogenität Unterschiede der Studien in Bezug auf

1. Definition von Zielerkrankung und Referenzstandard
2. die verwendeten Tests

3. die verwendeten Grenzwerte
4. das Patientenspektrum der erkrankten Population
5. das Patientenspektren der gesunden Population

an.

Kann Heterogenität nicht durch einen dieser Faktoren erklärt werden, sollen Studienergebnisse nicht gepoolt werden.

Etwaige Subgruppenanalysen, die ebenfalls ein adäquates Mittel zur Exploration von Heterogenität darstellen, sollten dann durchgeführt werden, wenn die diagnostische Genauigkeit eines Tests in unterschiedlichen Patientenpopulationen variiert. Relevante Subgruppen sollten bereits im Studienprotokoll festgelegt werden.

Level 3 & Level 4 - diagnostischer/therapeutischer Impakt

Von MSAC erwähnte Studiendesigns, die Änderungen im klinischen Management untersuchen können, sind diagnostische Vorher-Nachher Studien (siehe Tabelle 11.1-1). Diese Studien finden allerdings nur im Rahmen von „linked Evidence" Anwendung, um indirekt einen patientenrelevanten Nutzen etablieren zu können. Studien die Änderungen im Patientenmanagement untersuchen, werden nur dann *nicht* benötigt, wenn der Indextest als Ersatz für einen bereits bestehenden Test gedacht ist und eine Standardtherapie für die Zielerkrankung klar definiert ist.

Level 5 – patientenrelevanter Nutzen

Die Auswirkungen eines Testergebnisses auf patientenrelevante Endpunkte kann durch direkte Evidenz oder durch „linked Evidence" etabliert werden.

Direkte Evidenz

Hochwertige vergleichende Studien liegen vor, die patientenrelevante Endpunkte von aus dem Indextest resultierenden therapeutischen Konsequenzen, mit denen aus dem Standardtest resultierenden vergleichen.

Bevorzugtes Studiendesigns sind RCTs und dann in absteigender Reihenfolge nicht-randomisierte kontrollierte Studien, Kohortenstudien und Fallkontrollstudien. Für die Bewertung von Sicherheitsaspekten können nahezu alle Studiendesigns herangezogen werden (sieheTabelle 11.1-1).

Die Bewertung dieser Studiendesigns unterscheidet sich nicht von der von Wirksamkeitsstudien.

Indirekte Evidenz

Wenn keine Studien vorhanden sind, die die Auswirkungen direkt untersuchen kann der Nutzen von diagnostischen Verfahren indirekt unter Verwendung von „linked Evidence" etabliert werden (siehe Kapitel 4.4.2). Dabei wird der therapeutische Wirksamkeitsnachweis aus hochwertigen Studien mit hochwertigen Diagnosestudien, die Referenzstandard(-strategien) und Indextest(-strategien) miteinander vergleichen, verknüpft.

„Linked Evidence" beruht also auf Studien des Evidenzlevel 2, wobei das Vorgehen zur Bewertung der diagnostischen Genauigkeit in Kapitel 11.1.2 beschrieben wurde.

Nach MSAC sind daher folgende Voraussetzungen sind für „linked Evidence" nötig:

- Die Wirksamkeit der Therapie für die Zielerkrankung wurde anhand hochwertiger Evidenz (RCTs oder systematische Reviews über RCTs) nachgewiesen.
- Der Indextest diagnostiziert exakt die Zielerkrankung, die in der Wirksamkeitsstudie untersucht worden war und das Ergebnis des Tests beeinflusst therapeutische Entscheidungen.
    - Die durch den Indextest diagnostizierte Erkrankung muss exakt mit der Erkrankung für die eine wirksame Therapie besteht übereinstimmen. Würde ein Testergebnis dazu führen, dass eine Therapie früher, eine modifizierte oder eine neue Therapie verabreicht wird, wäre die Verknüpfung zwischen Wirksamkeit und Testgenauigkeit unzulässig.
    - Der genaue Verwendungszweck des Indextests muss festgelegt werden, da nur wenn der Indextest als Ersatz eines bereits bestehenden Tests geplant ist, davon ausgegangen werden kann, dass das Testergebnis auch tatsächlich zu der Verabreichung einer wirksamen Therapie führt. Ist dies nicht der Fall, dann werden Studien benötigt, die beweisen, dass der Indextest therapeutische Entscheidungen auch tatsächlich beeinflussen kann.
- Die Population in der Diagnosestudie ist repräsentativ für die Population der Wirksamkeitsstudie.

    Die Charakteristika der beiden Populationen sollten sorgfältig verglichen werden, um zu gewährleisten, dass beide Studien dasselbe Patientenspektrum untersuchen.

Für MSAC ist „linked Evidence“ also gerechtfertigt ist, wenn

- der neue Test kostengünstiger, nicht-invasiv und ein Ersatz für einen bereits bestehenden Test oder eine Teststrategie ist.
- die Therapie der durch den Indextest erkannten Erkrankung patientenrelevante Ergebnis verbessert, der Indextest die Zielerkrankung in genau demselben Krankheitsstadium wie der bereits existierende Test entdeckt, das Krankheitsspektrum der Population der Diagnosestudie vergleichbar mit dem in der Wirksamkeitsstudie ist.

Tabelle 11.1-1:Relevante Studiendesigns und Endpunkte zur Bewertung von diagnostischen Verfahren

Table 9: Types and sources of evidence

Type of evidence	Source of evidence	Outcomes
Benefits	**Systematic review of evidence**	
Clinical effectiveness of the test and subsequent interventions on patient outcomes relative to the comparator strategy	Controlled studies of test effectiveness Randomised controlled trials Non-randomised controlled trials Cohort studies Case control studies	Patient morbidity Patient mortality Quality of life scores
Therapeutic effectiveness of the index test strategy on changing patient management	Studies of test impact on clinical management Diagnostic pre and post test studies	Changes in patient management
Accuracy of the test relative to comparator strategy	Studies of test accuracy Cross-sectional Test-based enrolment Case referent	Diagnostic accuracy Sensitivity/specificity, likelihood ratio Diagnostic odds ratio Area under the summary ROC curve

Cost effectiveness of the index test relative to comparator strategy	Economic evaluation
	Costing of index test and comparator
	Cost-minimisation
	Cost-benefit, Cost-effectiveness
	Cost-utility
		Simple costing of tests
		$ costs
		$ costs and benefits
		$ cost per event avoided
		$ cost per quality of life year saved

Harms		
Safety of index test and associated procedures versus comparator	Clinical studies including: Case series Registers of adverse events	Rates of adverse events
Impact of misdiagnosis (false positive, false negative)	Cohort studies, Cross-sectional studies Decision analysis	
Clinical need	Epidemiological studies and Australian utilisation data	Incidence, prevalence, morbidity and mortality of target condition. Frequency of use of existing diagnostic procedures for the same indication.

Derzeit noch ungelöste methodische Fragestellungen sind:

- Wenn als Vergleichstest ein ungenauer Referenzstandard herangezogen wird, ist die Einschätzung der diagnostischen Genauigkeit unzuverlässig. In diesen Fällen wird direkte Evidenz benötigt.
- Indextest und Vergleichstest werden in zwei unterschiedlichen Studien bewertet. Die Übertragbarkeit von Ergebnissen ist dann nur bei hochwertigen Studiendesigns mit vergleichbaren Populationen möglich.
- Die in Studien untersuchten diagnostischen Strategien sind nicht deckungsgleich mit dem in der Fragestellung formulierten Verwendungszweck.
- Die Anwendung von „linked Evidence“, wenn sich die Studienpopulation der Diagnosestudien von der in Wirksamkeitsstudien unterscheidet.

Entscheidungsanalyse

MSAC empfiehlt entscheidungsanalytische Methoden unter Verwendung von Entscheidungsbäumen, um Nutzen und Nebenwirkungen gegeneinander abzuwägen, um so eine Aussage zum Nettonutzen machen zu können. Auch können so Trade-offs zwischen Sensitivität und Spezifität, beziehungsweise die unterschiedlichen Konsequenzen in Bezug auf falsch positive und falsch negative Befunde modelliert werden. Der Arm des Entscheidungsbaumes, der den höchsten Nutzenwert bringt, entspricht der besten diagnostischen Strategie. Die Robustheit des Modells soll mittels Sensitivitätsanalysen und, wenn möglich, durch Vergleich der Ergebnisse mit anderen Modellen, geprüft werden.

Level 6 – Nutzen aus gesellschaftlicher Sicht

Wenn die systematische Review den Nachweis erbracht hat, dass der Indextest zumindest genauso sicher und wirksam wie der Referenztest ist, sollte laut MSAC eine ökonomischen Evaluation aus einer breiten gesundheitssystemischen Perspektive heraus, durchgeführt werden.

Festzulegen gilt:

- Zielpopulation, diagnostische Optionen, vorangegangene Tests
- Patientenrelevanten Endpunkte, inklusive validierter Surrogatparameter
- Zeithorizont
- Sensitivitätsanalyse von inkrementellen Kosten, Nutzen und Kosteneffektivität für Prävalenz, Sensitivität/Spezifität mit unterschiedlichen Schwellenwerten, Unsicherheiten in Bezug auf Behandlungspfade, Behandlungseffekte und die damit verbundenen Änderungen des Ressourcenverbrauchs

Der Fokus sollte sowohl auf der Fähigkeit des Indextests patientenrelevante Ergebnisse (z.B. gewonnene Lebensjahre, QALYs) verbessern zu können, als auch auf den damit verbundenen Ressourcenimplikationen liegen und nicht auf der diagnostischen Genauigkeit. Wenn sich in einer einfachen Kostenaufstellung bereits zeigt, dass der Indextest das Vergleichsverfahren dominiert (also kostengünstiger und wirksamer), ist eine volle ökonomische Evaluation hinfällig. Ökonomische Evaluationen eignen sich auch, um Unterschiede von Sensitivität und Spezifität zwischen Indextest und Referenztest und die damit verbundenen Änderungen der mit falsch positiven und falsch negativen Ergebnissen verbundenen Konsequenzen zu berücksichtigen.

11.1.3 Empfehlungen

Die durch die systematische Suche gefundene Evidenz wird von MSAC in drei Stufen beurteilt:

- die Qualität der einzelnen Studien und deren Relevanz.
- die Genauigkeit, die Größe des Effekts und die klinische Relevanz der primären Endpunkte.
- der zu erwartende Nettonutzen des neuen diagnostischen Verfahrens mit Bezug auf die klinische Praxis in Australien.

Die Empfehlungen sollten dann Überlegungen zu mit dem Indextest assoziierten patientenrelevanten Nutzen, dem Schaden und auch dem Trade-offs im Vergleich zu anderen Strategien beinhalten.

Kostenerstattungsentscheidungen werden dann basierend auf der gefundenen Evidenz getroffen und können neben Kostenübernahme, Ablehnung der Technologie auch eine zeitlich limitierte Kostenübernahme sein.

11.2 Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Das Deutsche „Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen“ (IQWiG) wurde 2004 mit dem Ziel gegründet, Arzneimittel, nicht-medikamentöse Behandlungen, sowie Diagnose- und Screeningverfahren im Rahmen von evidenzbasierten Gutachten zu prüfen. Diese Gutachten dienen den Gemeinsamen Bundesausschuss als Entscheidungshilfe für Kostenerstattungsentscheidungen, die für alle gesetzlich Krankenversichten bindend sind (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2010).

Im Rahmen des Manuals „Allgemeine Methoden“ werden neben der Bewertung von Interventionen auch Vorgehensweisen zur Evaluierung von diagnostischen Verfahren behandelt (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2008).

Grundlage für Berichte zur Nutzenbewertung stellen systematische Übersichtsarbeiten dar, die

> „eine medizinische Maßnahme im Vergleich zu einer anderen klar definierten aktiven Maßnahme oder Scheinmaßnahme oder keiner Maßnahme bezüglich ihrer Auswirkungen auf definierte patientenrelevante Endpunkte in ihrem (Zusatz-)Nutzen und Schaden zusammenfassend beschreiben (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 2008)“.

11.2.1 Allgemeine Methodik

Ein vorläufiger Berichtsplan wird in Zusammenarbeit mit externen Sachverständigen entworfen. Im Anschluss erhält die Öffentlichkeit, unter anderem ÄrztInnen, PharmazeutInnen, Arzneimittelhersteller aber auch PatientInnen, die Möglichkeit eine Stellungnahme zu dem vorläufigen Berichtsplan abzugeben.

Die prinzipielle Vorgehensweise zur systematischen Bewertung der Evidenz beinhaltet die Formulierung einer Fragestellung, die Literatursuche und -auswahl, die Qualitätsbewertung der Einzelstudien und die Datenextraktion. Letztlich soll anhand der Evidenzlage eindeutig festgestellt werden, ob mit einem Verfahren ein höherer Nutzen, kein Nutzen verbunden ist, oder ob der Nutzen nicht eindeutig belegt ist.

Die Grundlage für diese Feststellungen bildet die Zusammenschau von

- qualitativer (Studiendesign, Bias, etc) und quantitativer (abhängig vom Stichprobenumfang wie z.B. Größe der Studienpopulation, Anzahl der für die systematische Review gefundenen Primärstudien, Variabilität) Ergebnissicherheit,
- Größe des beobachteten Effektes,
- Konsistenz der Effekte.

Forschungsfrage

Das endgültige Studienprotokoll muss eine klar definierte Fragestellung, Ein- und Ausschlusskriterien für primäre Studien, (patientenrelevanten) Endpunkten, die Methodik zur Beschaffung der Informationen, sowie deren Bewertung beinhalten.

Mit der genauen Beschreibung der Population, der Intervention, der Vergleichsbehandlung und der zu bewertenden Endpunkten, folgt die Fragestellung dem PICO Schema. Weitere Ein- und Ausschlusskriterien können die Art des Studiendesigns, die Studiendauer oder andere, *a priori*, definierte Charakteristika sein.

Die eingeschlossenen diagnostischen Studien können dann, je nach Studiendesign, in unterschiedliche Evidenzgrade eingeteilt werden (sieheAbbildung 11.2-1).

<table>
<tr><th>Klassifikations-schema
mögliche Endpunkte</th><th>Köbberling et al. [1]</th><th>Fryback & Thornbury [28]</th><th>Evidenzklassifi-zierung G-BA [36]</th></tr>
<tr><td rowspan="3">Patientenrelevante Zielgrößen</td><td rowspan="3">Phase 4:
Wirksamkeit</td><td>6. Stufe:
Auswirkung auf Systemebene</td><td rowspan="3">Evidenzstufe I</td></tr>
<tr><td>5. Stufe:
Auswirkung im Hinblick auf patientenrelevante Endpunkte</td></tr>
<tr><td>4. Stufe:
Auswirkung im Hinblick auf das therapeutische Denken des Behandelnden</td></tr>
<tr><td>Likelihood Ratio (LR), prädiktive Werte</td><td></td><td>3. Stufe:
Auswirkung im Hinblick auf das (differenzial)diagnostische Denken</td><td></td></tr>
<tr><td>Sensitivität (SN), Spezifität (SP), Likelihood Ratio (LR), prädiktive Werte</td><td>Phase 3:
Diagnostische Genauigkeit bei nicht bekanntem Krankheitsstatus</td><td rowspan="2">2. Stufe:
Auswirkung im Hinblick auf dieDiskriminationsfähigkeit</td><td>Evidenzstufe II</td></tr>
<tr><td>Sensitivität (SN), Spezifität (SP), Likelihood Ratio (LR)</td><td>Phase 2:
Diagnostische Genauigkeit bei bekanntem Krankheitsstatus</td><td>Evidenzstufe III</td></tr>
<tr><td>Analytische Sensitivität, Spezifität, Reproduzierbarkeit</td><td>Phase 1:
Technische Voruntersuchungen</td><td>1. Stufe:
Technische Auswirkungen</td><td></td></tr>
</table>

Abbildung 11.2-1: Klassifizierung von Studien bei der Evaluation diagnostischer Verfahren (Quelle: [85])

Hintergrundinformationen

In IQWiG's "Allgemeinen Methoden" wird zwar nicht eindeutig erläutert, welche Hintergrundinformationen ein Bericht enthalten soll, anhand mehrerer publizierter Berichte zu diagnostischen Verfahren [68, 85] wird aber ersichtlich, dass Angaben zu

1. Zielerkrankung: Definition, Epidemiologie und Krankheitslast, Ursachen, Klassifizierung, Krankheitsverlauf
2. Diagnostische Verfahren: Optionen, Referenz-, Goldstandard
3. Therapieoptionen: Standardtherapie in Deutschland
4. Indextest

gemacht werden sollen.

11.2.2 Nutzenbewertung diagnostischer Verfahren

Der Nutzen diagnostischer Verfahren wird laut IQWiG durch die Wirksamkeit und die Sicherheit bestimmt. Nach Fryback und Thornbury also anhand von Evidenz des Levels 5. Studien der Stufe 2 können im Rahmen der „diagnostischen Kette“ (siehe Kapitel „Level 5 – patientenrelevanter Nutzen“) ebenfalls als Grundlage für die Nutzenbewertung herangezogen, der Stellenwert von Studien der Level 3 und 4 ist aufgrund der reduzierten methodischen Qualität aber nicht eindeutig geklärt.

Level 2 – diagnostische Genauigkeit

Studiendesign

Als beste Evidenz, um diagnostische Verfahren zu vergleichen, sieht das IQWiG verblindete, diagnostische Querschnittsstudien mit einer zufälligen Zuordnung der Reihenfolge von Index- und Referenztest in denselben PatientInnen, oder solche, die Tests zufällig auf unterschiedliche PatientInnen aufteilen.

Bewertung der Studienqualität

Die Auswahl relevanter Publikationen wird durch zwei unabhängige WissenschafterInnen durchgeführt, wobei das STARD - Instrument als Entscheidungshilfe über Ein- oder Ausschluss von nicht als Volltext publizierten Literaturstellen herangezogen werden kann. Die Bewertung der methodischen Qualität von Studien zu diagnostischer Genauigkeit, erfolgt anhand der QUADAS – Kriterien, die projektspezifisch angepasst und ergänzt werden sollen.

Für die Bewertung von Prognosestudien wird die Checkliste nach Hayden (siehe Appendix 8.4.2) oder nach Laupacis erwähnt.

Datenextraktion

Genaue Angaben welche Daten aus den identifizierten Studien extrahiert werden sollen, werden im Methodenhandbuch nicht gemacht.

Datenanalyse

Als Kenngrößen der diagnostischen Genauigkeit werden im Methodenhandbuch nur „allgemein verwendete Testgütekriterien“ genannt, worunter die in Kapitel 4.2.1 beschriebenen Parameter fallen.

Das Vorgehen bei der Datenanalyse wird nicht speziell für diagnostische Verfahren erläutert, aber generell sollten alle Variablen, deren Konfidenzintervalle und ihr Standardfehler angegeben werden und darüber hinaus ihre statistische Signifikanz und klinische Relevanz beurteilt werden.

In der Regel fordert das IQWiG als Beleg für den Nutzen und Schaden eines Verfahrens Meta-Analysen, die ergebnissichere und statistisch signifikante Effekte für patientenrelevante Endpunkte zeigen. Ist die Heterogenität der Studien aber zu groß, um eine Meta-Analyse durchführen zu können, dann sollten mindesten zwei Studien mit großer Ergebnissicherheit und entsprechendem statistisch signifikantem Ergebnis vorliegen, deren Ergebnisse nicht durch andere Studien in Frage gestellt werden.

Level 3 & Level 4 - diagnostischer/therapeutischer Impakt

Studiendesigns, die den Einfluss von Testergebnissen auf das weitere Patientenmanagement untersuchen, werden vom IQWiG zwar erwähnt, bedingt durch die hohe Verzerrungsanfälligkeit, vor allem bei retrospektiven Studien, wird diesen Designs aber keine wesentliche Bedeutung eingeräumt.

Level 5 – patientenrelevanter Nutzen

Eine Unterscheidung zwischen direkter und indirekter Evidenz wird zwar nicht explizit gemacht, implizit wird den damit verbundenen methodischen Ansätzen aber Rechnung getragen, da diagnostische Verfahren und die daraus resultierenden patientenrelevanten Endpunkte vorzugsweise direkt bewertet werden (Stufe 5 nach Fryback und Thornbury) sollten, oder, wenn es dazu keine Daten gibt, mittels der diagnostischen Kette, die sich auf Studien des Levels 2 stützt.

Direkte Evidenz

Als höchsten Evidenzgrad bei der Bewertung von Diagnostika, aber auch von Screening- oder präventiven Maßnahmen, und damit auch als bevorzugte Grundlage für die Evaluierung, beschreibt das IQWiG prospektiv geplante, kontrollierte Studien, die patientenrelevante Endpunkte von zwei Patientengruppen, einmal mit, einmal ohne die zu untersuchende diagnostische Technologie, erheben. Unter patientenrelevante Endpunkte fallen primär Mortalität, Morbidität und gesundheitsbezogene Lebensqualität, die einerseits durch Vermeidung von mit Interventionen assoziierten Risken und andererseits durch Einleitung gezielter Interventionen beeinflusst werden können.

Bei der Bewertung von Studien, die direkt den Einfluss eines diagnostischen Verfahrens auf patientenrelevante Endpunkte erheben, werden zur Abschätzung der Ergebnissicherheit drei Komponenten bewertet:

- das Studiendesign
- die interne Validität
- die Größe des beobachteten Effektes.

Darüber hinaus wird auch die Konsistenz der Ergebnisse von mehreren Studien bewertet. Eine zusammenfassende Bewertung der Evidenzlage wird anhand von standardisierten Vorgehensweisen gemacht, wobei GRADE explizit erwähnt wird.

Indirekte Evidenz

Sind „direkte“ Studien nicht verfügbar, können mittels der „diagnostischen Kette“ Ergebnisse aus hochwertigen Studiendesigns, im besten Fall Querschnittsstudien, zur Testgüte herangezogen werden und mit Ergebnissen aus (in den meisten Fällen) randomisierten Interventionsstudien verknüpft

werden. Voraussetzung dafür ist, dass das Ergebnis des Indextests ein Einschlusskriterium der Wirksamkeitsstudie ist. Das IQWiG bezieht sich dabei auf die MSAC Richtlinien, sodass das Prinzip der diagnostischen Kette mit dem von „Linked evidence“ gleichgesetzt werden kann.

Bei Screeningtests sollten Studien zur diagnostischen Genauigkeit mit randomisierten Interventionsstudien, die den Nutzen einer frühzeitigeren Intervention zu dem aus einer späteren Intervention resultierenden Nutzen erheben, verknüpft werden.

Wenn das fragliche diagnostische Verfahren nicht als singulärer Test eingesetzt wird, sondern in eine bereits bestehende diagnostische Strategie inkorporiert werden soll (z.B. Triage-Tests oder Add-on- Tests (siehe Kapitel 4.2.4), kann durch die Kombination unterschiedlicher Verfahren das Patientenspektrum von dem, aus den Einzeltests resultierenden Populationen abweichen und so zu Unterschieden in der Wirksamkeit der nachfolgenden Therapie führen. Wenn dies nicht sicher ausgeschlossen werden kann, dann werden Studien benötigt, die die diagnostischen Strategien einmal mit, einmal ohne Indextest vergleichen.

Diagnosestudien, die zwei (oder mehrere Tests) nur hinsichtlich bestimmter Testcharakteristika vergleichen, reichen dann aus, wenn lediglich modifizierte Varianten eines Tests untersucht werden sollen, deren Nutzen bereits eindeutig belegt ist. Als beste Evidenz gelten dann entweder diagnostische Querschnittsstudien, oder solche, die Tests zufällig auf unterschiedliche PatientInnen aufteilen.

Level 6 – Nutzen aus gesellschaftlicher Sicht

Obwohl das IQWiG seit 2007 beauftragt werden kann, neben dem Nutzen einer medizinischen Technologie auch die damit verbundenen Kosten zu bewerten, werden diagnostische Verfahren in den „Allgemeine Methoden zur Bewertung von Verhältnissen zwischen Nutzen und Kosten“ [86] nicht erwähnt.

11.2.3 Empfehlungen

Basierend auf der Ergebnissicherheit der Resultate, der Größe des Effektes und der Konsistenz der Ergebnisse kann eine der folgenden Empfehlungen formuliert werden:

1. der Beleg für einen (Zusatz-)Nutzen bzw Schaden liegt vor.
2. Hinweise liegen vor, das ein (Zusatz-)Nutzen bzw Schaden vorhanden ist.
3. Der Beleg für das Fehlen eines (Zusatz-) Nutzen bzw Schadens liegt vor.
4. Hinweise liegen vor, das kein (Zusatz-)Nutzen bzw Schaden vorhanden ist.
5. Kein Beleg für und kein Hinweis auf einen (Zusatz-)Nutzen bzw Schaden liegen vor.

11.3 National Institute for Health and Clinical Excellence

Das englische „National Institute for Health and Clinical Excellence“ (NICE) wurde 1999 gegründet, um Empfehlungen zu neuen oder bereits existierenden Gesundheitstechnologien für das National Health Service (NHS), dem englischen Krankenversicherungsträger, zu entwickeln. Produkte von NICE umfassen neben der Bewertung von einzelnen Technologien auch die Entwicklung klinischer Richtlinien über die Behandlung und Pflege von PatientInnen mit spezifischen Erkrankungen. Die zu untersuchenden Themen werden vom Gesundheitsministerium ausgewählt, wobei die von NICE abgegebenen Empfehlungen die Grundlage für die Kostenübernahme durch das NHS bilden. Positive Empfehlungen und damit die Finanzierung von Technologien, sind von allen NHS Organisationen innerhalb von drei Monaten verpflichtend umzusetzen.

Die in den Richtlinien formulierten Empfehlungen basieren auf einer Review der besten verfügbaren Evidenz von dem mit einer Technologie verbundenem Nutzen und deren Kosteneffektivität.

Da in das Methodenmanual für die Erstellung von klinischen Richtlinien auch die Ergebnisse von Bewertungen einzelner Technologien einfließen, werden im Weiteren die in diesem Manual beschriebenen Vorgehensweisen angeführt [79]. Zusätzlich arbeitet NICE an der Entwicklung einer eigenen Methodik für die Bewertung von diagnostischen Verfahren, wobei ein erster Entwurf derzeit im Rahmen eines Pilotprojektes getestet wird [80]. Endgültige Richtlinien können 2011 erwartet werden.

11.3.1 Allgemeine Methodik

Die Bewertung des Nutzens von medizinischen Technologien beruht auf einer systematischen Review, wobei die Erstellung des Studienprotokolls und die Formulierung der PICO(S) - Fragestellung durch ein multidisziplinäres Team aus WissenschafterInnen, KlinikerInnen und PatientInnen erfolgt. Die grundlegenden Abläufe sind wiederum dieselben wie bei Interventionen (Auswahl relevanter Literaturstellen und Datenextraktion durch zwei unabhängige WissenschafterInnen, Bewertung der methodischen Qualität der eingeschlossenen Studien, Präsentation der Ergebnisse durch Meta-Analysen).

Forschungsfrage

Mit der PICO(S)- Frage werden neben der zu untersuchenden Population, dem Index- und Vergleichstest und der Definition relevanter Endpunkte auch die einzuschließenden Studiendesigns beschrieben. Daneben sollten laut NICE auch die Ziele, die Methoden zur Literatursuche und die Reviewstrategie im Protokoll enthalten sein.

Hintergrundinformationen

Bei der Formulierung der PICO-Frage können folgende Überlegungen angestellt werden, und als Hintergrundinformation im Bericht enthalten sein:

- Population: in welcher Patientenpopulation soll der Test durchgeführt werden, in welchem klinischen Setting? wie wird diese Gruppe am besten beschrieben? sind bestimmte Subgruppen besonders zu berücksichtigen?
- Intervention: genaue Beschreibung des zu evaluierenden Indextests unter Berücksichtigung unterschiedlicher Grenzwerte, Inter-/Intraobserver Variabilität, technischer Unterschiede (z.B. Art und Menge des verwendeten Kontrastmittels, bei Röntgenbildern Expositionszeit, etc).
    - Berücksichtigung der Testgenauigkeit von unterschiedlichen Kombinationen oder Abfolgen von mehreren Tests entweder durch Primärdaten oder durch Berechnung der Gesamtgenauigkeit.
    - Zeitlich bedingte Unterschiede: Veränderung der Testgenauigkeit bei Fortschreiten einer Erkrankung, aber auch Unterschiede in der Wirksamkeit einer Therapie in Abhängigkeit vom Krankheitsstadium (bei Screening zusätzlich Lead-Time und Length-Time Bias (siehe Kapitel „Nutzen aus gesellschaftlicher Sicht“).
- Komparator: welche diagnostischen Tests stehen derzeit zur Verfügung, ersetzt der neue Test einen anderen? mit welchem Test soll der Indextest verglichen werden (üblicherweise der Referenzstandard, der aber nicht mit dem in der klinischen Praxis am häufigsten verwendeten Test übereinstimmen muss)
- Zielerkrankung: genaue Beschreibung der Erkrankung, des Stadiums oder des Subtyps der Erkrankung, der mittels Indextest und Referenzstandard erkannt werden soll
- Outcome: Als mögliche relevante Endpunkte diagnostischer Verfahren werden in erster Linie
    - direkt durch ein Testergebnis resultierende patientenrelevante Endpunkte angesehen.
    - Nebenwirkungen des Indextests.
    - Nebenwirkungen von auf den Indextest anschließenden Untersuchungen.
    - therapiebedingte Nebenwirkungen.
    - der Wert der prognostischen Information.
    - Einfluss auf das zeitliche Management, was wiederum Auswirkungen auf die Testgenauigkeit, die therapeutische Wirksamkeit und auf den PatientIn selbst haben kann, wenn dadurch etwa Konsequenzen wie die stationäre Aufnahme entstehen.
    - Kosten, wobei nicht nur die Kosten für den Indextest, sondern auch für Folgeuntersuchungen, Therapie und Follow-Up berücksichtigt werden sollten.

Weitere Überlegungen können soziale Ungleichheiten, etwaige Ausnahmeregelungen oder Unterschiede in der klinischen Praxis betreffen.

11.3.2 Nutzenbewertung diagnostischer Verfahren

NICE bewertet diagnostische Verfahren sowohl anhand des Nutzens, als auch anhand der Kosteneffektivität. Der Nutzen wird bevorzugt direkt, anhand von patientenrelevanten Endpunkten erhoben (=Level 5), wobei aber auch Studien des Levels 2 ausreichend sein können. Immer wird aber eine Kosteneffektivitätsanalyse durchgeführt (Level 6).

Level 2 – diagnostische Genauigkeit

Studiendesign

Als beste Evidenz für Reviews zu Testgenauigkeit werden von NICE Querschnittsstudien in denen Indextest und Referenzstandard in ein- und derselben Population miteinander verglichen werden, angesehen. Fallkontrollstudien, die anfälliger für Bias sind, können aber auch zur Bewertung der diagnostischen Genauigkeit verwendet werden.

Bewertung der Studienqualität

Erwähnte Checklisten zur Bewertung von diagnostischen Studien sind STARD, QUADAS und Methodenmanual der Cochrane Collaboration. Die vorläufige Version des NICE Handbuchs empfiehlt eine modifizierte Version von QUADAS zur Bewertung der Studienqualität. Nichtsdestotrotz wird darauf hingewiesen, dass diese Checklisten nur eine Hilfe für eine erste Qualitätsbewertung darstellen, vor Verwendung individueller Studien für Modellierung (siehe Kapitel „Nutzen aus gesellschaftlicher Sicht“) sollte aber deren Relevanz für die spezielle Fragestellung geprüft werden.

Datenextraktion

Zu extrahierende Daten bei diagnostischen Genauigkeitsstudien sind:

1. Studiencharakteristika: Studiendesign, Studienqualität, Finanzierung
2. Studienpopulation: Patientenzahl, Charakteristika, Prävalenz
3. Indextest
4. Referenzstandard
5. Sensitivität, Spezifität, PV

Bei Prognosestudien:

1. Studiencharakteristika: Studientyp, Studienqualität
2. Studienpopulation: Patientenzahl, Patientencharakteristika
3. Prognosefaktor
4. Dauer des Follow-Up
5. Endpunkte
6. Ergebnisse
7. Finanzierung

Datenanalyse

Als Kenngrößen für die diagnostische Genauigkeit werden

- Sens/Spez
- PV
- LR
- ROC-Kurven genannt.

Erwähnt wird, dass die GRADE- Gruppe zwar an einer Methodik zur Bewertung der Gesamtstärke der Evidenz von Diagnosestudien arbeitet, diese aber noch nicht fertig ist, sodass validierte Methoden zur zusammenfassenden Darstellung der Evidenzlage nach wie vor ausständig sind. Die Studienergebnisse von diagnostischen als auch von prognostischen Studien, sollten deshalb unter Berücksichtigung der methodischen Qualität in narrativer und in numerischer Form dargestellt werden.

Meta-Analysen von Sensitivität und Spezifität sind in jedem Fall wünschenswert, sind allerdings nur dann angezeigt, wenn Studien mit gleichem Goldstandard und ähnlichen Patientenpopulationen weitgehend homogene Ergebnisse zeigen. Sensitivitätsanalysen stellen vor allem bei sehr heterogenen Studien eine wertvolle Alternative zu Meta-Analysen dar.

Level 3 & Level 4 – diagnostischer/therapeutischer Impakt

Studien, die die Auswirkungen eines Testergebnisses auf das diagnostische/therapeutische Vorgehen untersuchen, werden von NICE nicht erwähnt.

Level 5 – patientenrelevanter Nutzen

Der Nachweis eines direkten, patientenrelevanten Nutzens durch einen Test wird laut NICE mittels „Test and Treat" Studien, also idealerweise RCTs, erbracht. Bei der Bewertung des direkten klinischen Nutzens werden dieselben standardisierten Methoden empfohlen, wie sie auch zur Evaluation von Interventionen angewandt werden sollen. Fragen zur Sicherheit eines Verfahrens sind genauso zu behandeln, wie Fragen zur klinischen Relevanz eines Tests und sollten ebenfalls am besten mittels RCTs beantwortet werden.

Obwohl festgehalten wird, dass sich der Nutzen eines Tests erst durch Beeinflussung patientenrelevanter Endpunkte entfaltet, wird das Prinzip von „linked Evidence" weder im Methodenmanual, noch im vorläufigen Bericht zur Bewertung von diagnostischen Verfahren erwähnt.

Prognose: Ein eigenes Kapitel befasst sich mit Tests, die für Prognosen verwendet werden. Die Testinformationen dienen etwa dazu, um:

- PatientInnen mit prognostischen Informationen zu versorgen.
- PatienInnen unterschiedlichen Risikogruppen und damit unterschiedlichen Therapien zuzuordnen.
- Subgruppen mit unterschiedlichem Ansprechen auf Therapien zu definieren.

Als bestes Studiendesign werden prospektive Kohortenstudien genannt, in denen die Häufigkeit eines Ereignisses zwischen PatientInnen *mit* und PatientInnen *ohne* prognostischen Faktor verglichen werden. Zur Bewertung der Qualität von prognostischen Studien wird die Checkliste nach Hayden empfohlen (siehe Appendix 8.4.2).

Im vorläufigen Bericht zur Evaluierung von Diagnostika geht NICE spezifisch auf Probleme von Endpunkten bei der Bewertung von Tests ein. Diese sind:

- selbst wenn direkte Evidenz die Verbesserung von patientenrelevanten Ergebnissen durch eine an ein Testergebnis anschließende Therapie belegt, sind in der Studienpopulation auch falsch negative PatientInnen enthalten, die keine Therapie erhalten. Ebenso können auch nachteilige Effekte entstehen, wenn PatientInnen mit falsch positivem Testergebnis behandelt werden, obwohl sie tatsächlich gesund sind.
- nachteilige Effekte auch durch dem Indextestergebnis nachfolgende Tests entstehen können.
- Therapien selbst mit unerwünschten Nebenwirkungen vergesellschaftet sind.
- der Wert prognostischer Informationen für den individuellen PatientIn schwer zu quantifizieren ist.
- der Zeitpunkt wann ein Test durchgeführt wird, beeinflusst einerseits die diagnostische Genauigkeit, andererseits aber auch die Wirksamkeit einer Therapie.

Entscheidungsanalytische Modelle

Zur Abwägung von Nutzen, Nebenwirkungen und Kosten, eignen sich entscheidungsanalytische Modelle.

Level 6 – Nutzen aus gesellschaftlicher Sicht

Die von NICE abgegebenen Empfehlungen basieren immer auf den Ergebnissen von ökonomischen Evaluationen, wobei die bevorzugte Methode die Kosten-Nutzwert-Analyse ist, bei der die Kosten pro QALY berechnet werden.

NICE bevorzugt zur Messung der Lebensqualität das EuroQuol Instrument EQ-5D, wobei dieses aber möglicherweise nicht sensitiv genug ist, um kurzfristige und leichtere Nebenwirkungen, die mit Tests vergesellschaftet sein können, zu erfassen. Unklar ist ebenfalls noch, wie die psychologischen Auswirkungen von Testergebnissen, sei es in Form von Angst, oder Erleichterung, erfasst werden können.

Ökonomischen Evaluierungen sollen aus Sicht des NHS durchgeführt werden und nur in Ausnahmefällen können auch die Kostenkonsequenzen für andere, nicht- NHS Organisationen berücksichtigt werden.

Die Methoden zur Berechnung der Kosten und des Nutzens eines diagnostischen Verfahrens unterscheiden sich prinzipiell nicht von denen zur Evaluation von Interventionen, wobei Modellierung von Kosten und Konsequenzen durch einfache Entscheidungsbäume oder durch aufwendigere Markov-Modelle erfolgen können. Einfache Kostenabschätzungen sind dann ausreichend, wenn eine Technologie einen höheren Nutzen und geringere Nebenwirkungen als die Vergleichstechnologie aufweist.

In der Praxis zeichnen sich Modelle zu Diagnoseverfahren aber durch eine wesentlich größere Komplexität aus, da bei der Modellierung folgende Faktoren berücksichtigt werden müssen:

1. Modellierung des Nutzens: der Nutzen einer diagnostischen Untersuchung entsteht in den seltensten Fällen direkt durch den Test selbst (z.B. direkte Testnebenwirkungen, Patientennutzen durch die gewonnene Information), sondern meist indirekt, sodass etwaige nachfolgende Tests und die einem Testergebnis folgende Therapie ebenfalls im Modell berücksichtigt werden müssen. Wenn ein Test in Bezug auf Testcharakteristika, Kosten und unmittelbaren Outcomes allen anderen Verfahren überlegen, muss aber nicht mehr der gesamte Behandlungspfad modelliert werden.
2. Modellierung aller relevanten, alternativen Test(strategien): alle relevanten Alternativen sollten im Modell enthalten sein. Dies wird allerdings dadurch erschwert, dass Alternativen nicht nur durch unterschiedliche diagnostische Verfahren gegeben sind, sondern, auch durch

   - unterschiedliche Grenzwerte/Erfahrung des Testdurchführenden,
   - die Reihenfolge mehrerer Tests, der Zeitpunkt der Testung,
   - die fehlende Unabhängigkeit von nacheinander durchgeführten Tests (falsche Diagnose eines ersten Tests beeinflusst Ergebnisse von nachfolgenden Tests)

   entstehen können.
3. Modellierung für unterschiedliche Patientengruppen: Patientencharakteristika, Unterschiede in der *a- priori* Wahrscheinlichkeit (Prävalenz), und ein unterschiedliches Therapieansprechen in Patientensubgruppen sollten bei der Modellierung berücksichtigt werden.
4. Modellierung in Abhängigkeit des Verwendungszwecks:

- Screening: Berücksichtigung von „Length – Bias“ (= aggressivere Formen von Krankheiten werden weniger häufig entdeckt, da sie häufig im Intervall zwischen zwei Screenings symptomatisch werden; die gescreente Population erscheint dadurch einen größeren Nutzen zu haben, als die ungescreente) und „Lead-time-Bias“ (=die gescreente Population scheint länger zu überleben, wobei dies aber nicht einem tatsächlichen Zuwachs an Gesamtüberleben entspricht, sondern die Erkrankung nur in einem früheren Entwicklungsstadium entdeckt worden war) mittels Sensitivitätsanalysen. Berücksichtigung von wiederholtem Testen und Compliance bei mehrmaligem Testen; bei einem positiven Testergebnis werden weitere Tests nachfolgen, um die Diagnose zu bestätigen, Berücksichtigung von falsch negativen Befunden.
  - Diagnostische Tests: Modellierung von falsch negativen Befunden und den damit vergesellschafteten Konsequenzen.
  - Monitoring: zu berücksichtigende Faktoren sind wieder, wie bei Screening, Compliance, der zeitliche Abstand zwischen den Tests und Lead-time-Bias.

Laut NICE können diese Modelle aber deutlich vereinfacht werden, wenn

- ✪ direkte Evidenz aus randomisierten, kontrollierten Studien vorhanden ist.

- bereits Modelle für einzelne Schritte vorhanden sind.
- Studien mit Daten zu Langzeitergebnissen vorhanden sind, da dadurch nicht alle Zwischenschritte modelliert werden müssen.
- anhand des Modells die minimal erforderliche Testgenauigkeit berechnet wird, die notwendig ist, um Kosteneffektivität zu erreichen (reverse Modellierung).

11.3.3 Empfehlungen

Neben der Abwägung von Nutzen und Risiken, entweder qualitativ oder quantitativ mittels entscheidungsanalytischer Modelle, sowie von Nutzen und Kosten, wird auch die Qualität der zugrundeliegenden Evidenz bei der Formulierung von Empfehlungen berücksichtigt. Optionen sind positive Empfehlungen, negative Empfehlungen oder die Empfehlung eine Technologie nur im Rahmen von „Only-in-research“ zuzulassen.

11.4 European network for Health Technology Assessment

Das europäische Projekt „European network for Health Technology Assessment“ (EUnetHTA) wurde 2006 gegründet, um die Zusammenarbeit und den Informationsaustausch von europäischen HTA Organisationen zu erleichtern. 2008 wurde das HTA Core Model für die Bewertung von diagnostischen Technologien publiziert [78], welches die wichtigsten Komponenten zur Bewertung von diagnostischen Verfahren beschreibt. Neben Methoden zur Bewertung der Genauigkeit, Sicherheit, Wirksamkeit und Kosten-Effektivität umfasst es auch ethische, rechtliche, soziale und institutionelle Aspekte.

11.4.1 Allgemeine Methodik

Die prinzipielle Methode zur Bewertung von Diagnostika ist eine systematische Review, wobei Sicherheitsaspekte in Zusammenschau mit Wirksamkeit die Basis für die Bewertung von diagnostischen Verfahren formen. Darauf aufbauend können dann Kosten und auch andere mögliche Konsequenzen, wie etwa rechtliche, soziale oder organisatorische miteinbezogen werden. Explizit erwähnt wird, dass je nachdem welcher Aspekt bearbeitet werden soll, unterschiedliche Kriterien für die Auswahl relevanter Publikationen und auch unterschiedliche Studiendesigns und Tools zur Bewertung der Studienqualität verwendet werden sollen.

Die grundlegenden Abläufe sind wieder Formulierung einer Forschungsfrage, Definition von Ein- und Ausschlusskriterien, Literatursuche und Bewertung der Studienqualität.

Forschungfrage

Die Forschungsfrage folgt dem PICO Schema.

Hintergrundinformationen

Neben der genauen Definition der PICO Frage, können auch andere Überlegungen miteinbezogen werden und in dem Hintergrundteil des Berichts enthalten sein:

1. Zielerkrankung/Zielpopulation: genaue Definition des Krankheitsbildes, Prognose, Krankheitsverlauf, Definition spezieller Subgruppen, Inzidenz/Prävalenz der Zielerkrankung als auch deren Konsequenzen wie zum Beispiel Mortalität, Krankenständen, Frühpensionierungen ?
2. Einsatz der Technologie: wie weit verbreitet wird der Test bereits eingesetzt, gibt es geographische Unterschiede (national, international)?
3. Diagnosepfade: existierende Diagnosepfade/klinische Richtlinien, Indextest als Add-on/Ersatz Test, paralleles Testen? Vergleichstest
4. Technologie: Beschreibung der Funktionsweise, der technischen Charakteristika, Unterschiede in verschiedenen Versionen der Technologie, personelle/strukturelle/finanzielle Voraussetzungen und Änderungen, die mit der Verwendung des neuen Tests einhergehen?
5. Überblick über bestehende Behandlungsrichtlinien
6. Lebenszyklus der Technologie: neue Technologie, etablierte, obsolete Technologie
7. Zulassungsstatus: wurde das diagnostische Verfahren bereits in anderen Ländern zugelassen? Kostenübernahmen in anderen Ländern?

11.4.2 Nutzenbewertung diagnostischer Verfahren

Die Nutzenbewertung eines diagnostischen Verfahrens erfolgt anhand von Studien des Evidenzlevels 5. Sind keine direkten Studien vorhanden, dann können wiederum Studien des Levels 2 mit Wirksamkeitsstudien verknüpft werden. Die Bewertung der Kosteneffektivität (Level 6) wird ebenfalls beschrieben.

Level 2 – diagnostische Genauigkeit

Studiendesign

Als bevorzugtes Studiendesign für diagnostische Genauigkeit werden diagnostische Querschnittstudien genannt, bei der eine Patientengruppe sowohl mit Indextest als auch mit Referenztest untersucht wird. In Abwesenheit eines geeigneten Referenztests kann eine Referenzdiagnose durch eine vordefinierte Abfolge mehrerer Untersuchung, durch Expertenkonsens oder aber auch durch statistische Modelle erstellt werden. Bei ungenauem Referenzstandard werden entweder Sensitivitätsanalysen, oder Daten über das Ausmaß und den Zusammenhang der Ungenauigkeit herangezogen, um

trotz dieser Ungenauigkeit Aussagen über die diagnostische Genauigkeit des Indextest treffen zu können.

Zusätzlich unterscheidet EUnetHTA aber je nach geplantem Einsatz des Indextests noch andere mögliche Designs:

1. Ersatz eines anderen Tests: RCTs, oder indirekter Vergleich, wobei Indextest und Referenzstandard in einer Studie und das zu ersetzende diagnostische Verfahren und Referenzstandard in einer anderen Studie verglichen werden.
2. Triage: Studiendesigns mit limitierte Verifizierung: nur PatientInnen mit einem negativem Indextestergebnis werden mittels Referenztest verifiziert (siehe Kapitel 4.2.4).
3. Add-on: Studiendesigns mit limitierte Verifizierung: nur PatientInnen mit einem negativem Indextestergebnis werden mittels Referenztest verifiziert (siehe Kapitel 4.2.4)

Bewertung der Studienqualität

EUnetHTA empfiehlt die Bewertung der Studienqualität anhand der im Cochrane Handbook erwähnten Fragen vorzunehmen, wobei davon die für die Fragestellung am wichtigsten ausgewählt werden sollten.

Zusätzlich werden weitere mögliche Fragen formuliert:

1. Wenn ein Grenzwert verwendet wurde, wurde dieser *vor* Studienbeginn festgelegt?
2. Ist es wahrscheinlich, dass sich der Indextest seit Studienbeginn in seinen technischen Eigenschaften bereits verändert hat?
3. Wurde in der Studie klar definiert, wie ein positives Testergebnis definiert ist?
4. Wurde die Behandlung *nac*h dem Indextest aber *vor* dem Referenztest begonnen?
5. Wurde die Behandlung *nach* dem Referenztest aber *vor* dem Indextest begonnen?
6. Wurde Angaben zu Beobachtervariabilität gemacht?
7. Wurde Variabilität der Instrumente beschrieben?
8. Wurde relevante Subgruppenergebnisse angegeben?
9. Ist die Größe der Studienpopulation ausreichend?
10. Wurden die Fragestellungen der Studien klar definiert?

Datenextraktion

Nachdem alle relevanten Studien durch zwei WissenschafterInnen ausgewählt wurden, sollten laut EUnetHTA folgende Daten aus den diagnostischen Genauigkeitsstudien extrahiert werden:

- Studiendesign
- Patientenpopulation, Prävalenz der Zielerkrankung
- vorangegangene Tests
- Indextest, Grenzwerte

- Referenztest
- Testergebnisse (Vierfelder -Tafel)
- Sensitivität, Spezifität (+ 95% Konfidenzintervall)
- Andere Kenngrößen der diagnostischen Genauigkeit
- Studienqualität

Zu Sicherheit:

- Art des unerwünschten Effektes
- Schweregrad (1 =mild, 2= mdoerat, 3= schwer, 4= lebensbedrohlich)
- Anzahl der unerwünschten Effekte
- Qualität der Informationen
- Generalisierbarkeit der Daten

Datenanalyse

Die Vierfelder-Tafel dient auch im „Core Model" wieder zur Berechnung der wichtigsten Kenngrößen der diagnostischen Genauigkeit (siehe Kapitel 4.2.1). Näher eingegangen wird auf

- Sensitivität, Spezifität (+ 95% Konfidenzintervall)
- LR
- DOR
- ROC -Kurven
- AUC

Bevor Daten im Rahmen von Meta-Analysen gepoolt werden können, muss bewertet werden, ob Studienergebnisse heterogen sind und wodurch diese Heterogenität verursacht wurde. Mögliche Gründe für Heterogenität sind:

- Zufall
- unterschiedliche Grenzwerte
- abweichende Studiendesigns, -methoden, Referenzstandards
- Variationen in der Studienpopulation
- nicht erklärbare Heterogenität.

Je nachdem wodurch Unterschiede in den Studienergebnissen bedingt wurden, stehen mehrere statistische Modelle für Meta-Analysen zur Verfügung. Kann die Ursache für Heterogenität aber nicht erklärt werden, sollten keine Meta-Analysen durchgeführt werden.

Level 3 & Level 4 – diagnostischer/therapeutischer Impakt

Als mögliche Studiendesigns die im Rahmen von linked Evidence berücksichtigt werden sollten, erwähnt EUnetHTA Vorher-Nachher Studien oder Zeitserien, die Änderungen im Patientenmanagement untersuchen. Diese Studien sind besonders wichtig:

- bei Add-On Tests, da bei Ersatztests davon ausgegangen wird, dass sich Managemententscheidungen nicht verändern.
- wenn andere Einflüsse, wie zum Beispiel Patientenpräferenzen, nachfolgende therapeutische Entscheidungen bestimmen können.
- wenn sich die Population der Diagnosestudie in Bezug auf Prävalenz oder Schwere der Erkrankung von der in der Wirksamkeitsstudie unterscheidet.
- der Wert der durch den Test gewonnenen Information unsicher ist.

Level 5 – patientenrelevanter Nutzen

Der aus einem Test resultierende Nutzen kann laut EUnetHTA Core Model wieder direkt, oder indirekt etabliert werden.

Direkte Evidenz

Als beste verfügbare Evidenz werden RCTs genannt, die patientenrelevante Konsequenzen von Indextest und bestehender Teststrategie miteinander vergleichen. Diagnostische Kohortenstudien oder Fallkontrollstudien werden zwar auch erwähnt, sind aber aufgrund des Studiendesigns anfälliger für systematische Verzerrungen. Zur Bewertung der Studienqualität gelten dieselben Kriterien wie auch bei der Bewertung von Interventionsstudien.

Ein eigenes Kapitel befasst sich zudem noch mit der Bewertung der Sicherheit:

Sicherheit ist besonders dann wichtig, wenn

- mit der Technologie besondere Risiken vergesellschaftet sind.
- das Nutzen-Risiko Profil nicht eindeutig ist.
- mehrere Tests mit ähnlicher Genauigkeit aber mit unterschiedlichen Sicherheitsprofilen zur Diagnose ein und derselben Erkrankung verwendet werden können.
- die Zahl der falsch positiven Ergebnissen hoch ist.
- Nebenwirkungen die Akzeptanz und den Einsatz des Tests untergraben könnten.

Nebenwirkungen können anhand unterschiedlicher Klassifizierungen eingeteilt werden:

- direkte Nebenwirkungen (wie Mortalität und Morbidität, die z. B. durch Strahlenexposition, toxische Kontrastmittel oder durch invasive Untersuchungsmethoden entstehen können) im Gegensatz zu indirekten Nebenwirkungen, die mit einer falschen Diagnose oder einer suboptimalen Patientenselektion vergesellschaftet sein können.
- Nebenwirkungen in Abhängigkeit von UntersucherIn/klinischem Setting (Erfahrungen und Expertise in der Durchführung des Tests) oder bedingt durch Patientencharakteristika, die mit einer erhöhten Wahrscheinlichkeit von Nebenwirkungen einhergehen können.
- Einteilung nach der Schwere der Nebenwirkungen.

Bedingt durch die zahlreichen Nebenwirkungen, die mit Diagnostika vergesellschaftet sein können, wird empfohlen, entweder nur die Häufigsten, die Schwerwiegendsten oder jene, die am öftesten zum Abbruch einer Intervention führen, zu erheben.

Auch für die Sicherheitsaspekte von diagnostischen Verfahren, nennt EUnetHTA RCTs als höchste Evidenz. Da aber seltene, oder erst nach einem längeren Zeitraum auftretende Nebenwirkungen in diesen Studien nur selten identifiziert werden, sollten zusätzlich auch Beobachtungsstudien oder Primärdaten aus Registern als Informationsquellen dienen. Wenn möglich, sollten alle gefundenen Angaben quantitativ zusammengefasst werden und die Häufigkeit, das relative Risiko oder die Number-Needed to Harm berechnet werden. In allen Fällen ist aber die kritische Bewertung der Studienqualität angezeigt, wobei zur Bewertung von Beobachtungsstudien die Newcastle Ottawa Scale [87] oder das STROBE Statement [88] angeführt werden.

Indirekte Evidenz

Als Voraussetzung, dass Testgenauigkeitsstudien zur Abschätzung der klinischen Effektivität herangezogen werden können, gibt EUnetHTA an, dass sich Diagnose- und Wirksamkeitsstudie in Bezug auf Patientenspektrum, Erkrankung, dem Test und anderen Charakteristika ähneln müssen. Diese Voraussetzungen müssen vor der Anwendung von „linked Evidence“ ausreichend gerechtfertigt werden.

Generell muss auch mittels Studien der Level 3 und 4 nachgewiesen werden, dass ein Testergebnis zu einer Änderung des klinischen Managements führt. Dieser Schritt ist allerdings unnötig, wenn der Indextest als Ersatztest geplant ist und für die Zielerkrankung eine Standardtherapie definiert ist. Zahlreiche Studiendesign können also bei „linked Evidence" verwendet werden (siehe Tabelle 11.4-1).

Tabelle 11.4-1: Mögliche Studiendesigns zur Bewertung unterschiedlicher Endpunkte

Studienart	Optimales Studiendesign
Sicherheit	RCTs, aber auch jedes andere Studiendesign, auch Fallserien, Register,…
Diagnostische Genauigkeit	Querschnittsstudien, randomisierte Diagnosestudien, limitierte Verifizierung
Änderung von Patienten-Management	Vorher-Nachher Studien, Zeitserien
Therapeutische Effektivität	RCT

Diagnostische Genauigkeitsstudien alleine sind dann ausreichend, wenn es sich bei dem Indextest um ein sicheres, kostengünstigeres oder spezifischeres Verfahren mit gleicher Sensitivität handelt. Wäre der Indextest dagegen sensitiver würde eine andere Patientengruppe durch den Test diagnostiziert werden.

Entscheidungsanalysen

Wenn es gilt zwischen Sicherheit und Wirksamkeit abzuwägen, wie zum Beispiel, wenn es sich um einen weniger invasiven aber auch weniger spezifischen Test handelt, eignen sich entscheidungsanalytische Modelle, um die mit falsch positiven Befunden einhergehenden Konsequenzen zu berechnen. Mit dieser Methode kann auch die Testeffektivität für Populationen mit unterschiedlicher Prävalenz berechnet werden.

Level 6 – gesellschaftliche Konsequenzen

Der Wichtigkeit von ökonomischen Evaluationen für Prioritätensetzung von EntscheidungsträgerInnen wird in einem eigenen Kapitel des EUnetHTA Methodenbuchs Rechnung getragen. Allerdings sind darin generell bei ökonomischen Evaluierungen zu berücksichtigenden Methoden, Fragestellungen und Vorgehensweisen, die auch bei Interventionen zu beachten sind, genannt und keine Diagnostika-spezifischen Angaben enthalten. Für ökonomische Evaluationen kommen daher je nach Fragestellung Kostenminimierungs-, Kosten-Nutzen-, Kosten-Nutzwert- und Kosten-Effektivitätsanalysen in Frage, wobei der Stellenwert von Kosten-Nutzwertanalysen für EntscheidungsträgerInnen besonders betont wird. Um zu entscheiden, welche Technologie kosteneffektiv ist, wird eine Entscheidungsmatrix präsentiert (siehe Tabelle 11.4-2).

Tabelle 11.4-2: Kosteneffektivitätsmatrix (Quelle: [78])

Eine neue Technologie im Vergleich mit einer alten Technologie	Weniger wirksam	Gleich wirksam	Wirksamer
Kostengünstiger	1. Keine eindeutige Entscheidung, da keine Dominanz → inkrementelle Analyse wird benötigt	4. Adoption der neuen Technologie, schwache Dominanz der neuen Technologie	7. Adoption der neuen Technologie, starke Dominanz der neuen Technologie
Gleiche Kosten	2. Entscheidung zugunsten der alten Technologie, da alte Technologie neue schwach dominiert	5. beide Technologien sind gleichwertig	8. Adoption der neuen Technologie, schwache Dominanz der neuen Technologie
Teurer	3. 2. Entscheidung zugunsten der alten Technologie, da alte Technologie neue eindeutig dominiert	6. Entscheidung zugunsten der alten Technologie, da alte Technologie neue schwach dominiert	91. Keine eindeutige Entscheidung, da keine Dominanz → inkrementelle Analyse wird benötigt

11.4.3 Empfehlungen

Grundlage für die Bewertung von diagnostischen Verfahren ist, wie erwähnt die Abwägung von Sicherheit und Nutzen. Darauf aufbauend können dann andere Aspekte, wie Kosteneffektivität, rechtliche oder organisatorische Konsequenzen berücksichtigt werden. Im Kapitel zu Kosteneffektivität wird aber erwähnt, dass diese letztlich von entscheidender Bedeutung für EntscheidungsträgerInnen ist.

12 Referenzen

1. Lijmer, J.G., M. Leeflang, and P.M. Bossuyt, *Proposals for a phased evaluation of medical tests.* Med Decis Making, 2009. **29**(5): p. E13-21.
2. Knottnerus, J.A., C. Van Weel, and J.W.M. Muris, *Evidence base of clinical diagnosis: Evaluation of diagnostic procedures.* British Medical Journal, 2002. **324**(7335): p. 477-480.
3. Hunink, M., et al., *Decision making in health and medicine - Integrating evidence and values.* 2001, Cambridge: Cambridge University Press.
4. Fryback, D.G., *A conceptual model for output measures in cost-effectiveness evaluation of diagnostic imaging.* J Neuroradiol, 1983. **10**(2): p. 94-6.
5. Fryback, D.G. and J.R. Thornbury, *The efficacy of diagnostic imaging.* Med Decis Making, 1991. **11**(2): p. 88-94.
6. Tatsioni, A., et al., *Challenges in systematic reviews of diagnostic technologies.* Ann Intern Med, 2005. **142**(12 Pt 2): p. 1048-55.
7. Bossuyt, P.M. and K. McCaffery, *Additional patient outcomes and pathways in evaluations of testing.* Med Decis Making, 2009. **29**(5): p. E30-8.
8. Harris, R.P., et al., *Current methods of the US Preventive Services Task Force: a review of the process.* Am J Prev Med, 2001. **20**(3 Suppl): p. 21-35.
9. Gordis, L., *Epidemiology.* 2nd ed. 2000, Philadelphia: W.B. Saunders Company.
10. Weiß, C., *Basiswissen Medizinische Statistik.* 4th ed. 2007, Heidelberg: Springer Medizin Verlag. 336.
11. Bland, M., *An introdcution to medical statistics* 3rd ed. 2000, Oxford: Oxford University Press.
12. Fletcher, R.H., *Interpretation of diagnostic tests.* Indian Journal of Pediatrics, 2000. **67**(1): p. 49-53.
13. Linden, A., *Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis.[see comment].* Journal of Evaluation in Clinical Practice, 2006. **12**(2): p. 132-9.
14. Lee, W.C. and W.C. Lee, *Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance.* International Journal of Epidemiology, 1999. **28**(3): p. 521-5.
15. Centre for Evidence Based Medicine. *Critical Appraisal.* 2009 [cited 2010 28.01.]; Available from: http://www.cebm.net/index.aspx?o=1157.
16. Mulherin, S.A., et al., *Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation.[see comment].* Annals of Internal Medicine, 2002. **137**(7): p. 598-602.
17. Bossuyt, P.M., *Interpreting diagnostic test accuracy studies.* Seminars in Hematology, 2008. **45**(3): p. 189-95.
18. Bossuyt, P.M., et al., *Comparative accuracy: assessing new tests against existing diagnostic pathways.[erratum appears in BMJ. 2006 Jun 10;332(7554):1368].* BMJ, 2006. **332**(7549): p. 1089-92.
19. Schneider, A., G.J. Dinant, and J. Szecsenyi, *Stepwise diagnostic workup in general practice as a consequence of the Bayesian reasoning.* Zeitschrift fur Arztliche Fortbildung und Qualitatssicherung, 2006. **100**(2): p. 121-127.

20. Deeks, J.J., *Using evaluations of diagnostic tests: understanding their limitations and making the most of available evidence.* Annals of Oncology, 1999. **10**(7): p. 761-8.
21. Deeks, J.J., *Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests.[see comment].* BMJ, 2001. **323**(7305): p. 157-62.
22. Bender, R., *[Interpretation of efficacy measures derived from 2 X 2 tables for the evaluation of diagnostic tests and treatment].[erratum appears in Med Klin 2001 Mar 15;96(3):181].* Medizinische Klinik, 2001. **96**(2): p. 116-21.
23. Sheps, S.B. and M.T. Schechter, *The assessment of diagnostic tests. A survey of current medical research.* Jama, 1984. **252**(17): p. 2418-22.
24. Halkin, A., et al., *Likelihood ratios: getting diagnostic testing into perspective.* Qjm, 1998. **91**(4): p. 247-58.
25. Irwig, L., et al., *Guidelines for meta-analyses evaluating diagnostic tests.* Ann Intern Med, 1994. **120**(8): p. 667-76.
26. Irwig, L., et al., *Designing studies to ensure that estimates of test accuracy are transferable.* Bmj, 2002. **324**(7338): p. 669-71.
27. Weinstein, S., N.A. Obuchowski, and M.L. Lieber, *Clinical Evaluation of Diagnostic Tests.* Am J Roentgenol, 2005. **184**: p. 14-19.
28. Deutsches Cochrane Zentrum. *Glossar.* [cited 2010 03.November]; Available from: http://www.cochrane.de/de/glossary.htm#v.
29. Antonelli, P., D. Chiumello, and B.M. Cesana, *Statistical methods for evidence-based medicine: the diagnostic test. Part II.* Minerva Anestesiol, 2008. **74**(9): p. 481-8.
30. Trop, I., et al., *Estimates of diagnostic accuracy efficacy: how well can this test perform the classification task?* Canadian Association of Radiologists Journal, 2003. **54**(2): p. 80-6.
31. Deeks, J.J. and D.G. Altman, *Diagnostic tests 4: likelihood ratios.* Bmj, 2004. **329**(7458): p. 168-9.
32. Akobeng, A.K., *Understanding diagnostic tests 3: Receiver operating characteristic curves.* Acta Paediatrica, International Journal of Paediatrics, 2007. **96**(5): p. 644-647.
33. Soreide, K. and K. Soreide, *Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research.* Journal of Clinical Pathology, 2009. **62**(1): p. 1-5.
34. Lijmer, J.G., et al., *Empirical evidence of design-related bias in studies of diagnostic tests.* Jama, 1999. **282**(11): p. 1061-6.
35. Knottnerus, J.A., *The Evidence Base of Clinical Diagnosis.* 2002, London: BMJ Books.
36. Deville, W.L., et al., *Conducting systematic reviews of diagnostic studies: didactic guidelines.* BMC medical research methodology, 2002. **2**: p. 9.
37. National Health Service Centre for Reviews and Dissemination, *Undertaking Systematic Reviews of Research on Effectiveness.* 2001, NHS CRD: York.
38. Reitsma, J.B., et al., *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* 2009, The Cochrance Collaboration.
39. Oosterhuis, W.P., et al., *Evidence-based guidelines in laboratory medicine: principles and methods.* Clin Chem, 2004. **50**(5): p. 806-18.
40. Sackett, D.L. and R.B. Haynes, *The architecture of diagnostic research.* Bmj, 2002. **324**(7336): p. 539-41.
41. National Health and Medical Research Council (NHMRC). *NHMRC additional levels of evidence and grades for recommendations for de-*

velopers of guidelines. 2009 [cited 2010 03. 02]; Available from: http://www.nhmrc.gov.au/_files_nhmrc/file/guidelines/Stage%202%20Consultation%20Levels%20and%20Grades.pdf.

42. Phillips, B., et al. *Levels of Evidence*. 2009 [cited 2010 3.02.]; Available from: http://www.cebm.net/index.aspx?o=1025.
43. Kleijnen, J., *New methods in the assessment of diagnostic procedures.* Zeitschrift fur Arztliche Fortbildung und Qualitatssicherung, 2006. **100**(7): p. 519-25.
44. Katz, D.L., L. Greci, and H. Nawaz, *Clinical Epidemiology & Evidence-Based Medicine*
Fundamental Principles of Clinical Reasoning and Research. 2001.
45. Lord, S.J., L. Irwig, and P.M. Bossuyt, *Using the principles of randomized controlled trial design to guide test evaluation.* Med Decis Making, 2009. **29**(5): p. E1-E12.
46. Whiting, P., et al., *Sources of variation and bias in studies of diagnostic accuracy: a systematic review.* Ann Intern Med, 2004. **140**(3): p. 189-202.
47. STAndards for the Reporting of Diagnostic accuracy studies - STARD Statement. 2008 [cited 2010 04.02]; Available from: http://www.stard-statement.org/.
48. Bossuyt, P.M., et al., *The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration.* Annals of Internal Medicine, 2003. **138**(1): p. W1-12.
49. Alonzo, T.A., et al., *Using a combination of reference tests to assess the accuracy of a new diagnostic test.[see comment].* Statistics in Medicine, 1999. **18**(22): p. 2987-3003.
50. Hui, S.L., et al., *Evaluation of diagnostic tests without gold standards.* Statistical Methods in Medical Research, 1998. **7**(4): p. 354-70.
51. Rutjes, A.W., et al., *Evaluation of diagnostic tests when there is no gold standard. A review of methods.* Health technology assessment (Winchester, England), 2007. **11**(50): p. iii, ix-51.
52. Glasziou, P., L. Irwig, and J.J. Deeks, *When should a new test become the current reference standard?* Ann Intern Med, 2008. **149**(11): p. 816-22.
53. Ransohoff, D.F. and A.R. Feinstein, *Problems of spectrum and bias in evaluating the efficacy of diagnostic tests.* N Engl J Med, 1978. **299**(17): p. 926-30.
54. Goehring, C., et al., *Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance.[erratum appears in Stat Med. 2005 Jun 15;24(11):1782].* Statistics in Medicine, 2004. **23**(1): p. 125-35.
55. Begg, C.B., *Biases in the assessment of diagnostic tests.* Statistics in Medicine, 1987. **6**: p. 411-423.
56. Whiting, P., et al., *A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools.* Journal of Clinical Epidemiology, 2005. **58**(1): p. 1-12.
57. Whiting, P., et al., *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**: p. 25.
58. Bossuyt, P.M., et al., *Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative.[see comment].* Clinical Chemistry & Laboratory Medicine, 2003. **41**(1): p. 68-73.
59. Leeflang, M.M., et al., *Systematic reviews of diagnostic test accuracy.* Ann Intern Med, 2008. **149**(12): p. 889-97.

60. Mallett, S., et al., *Systematic reviews of diagnostic tests in cancer: review of methods and reporting.[see comment].* BMJ, 2006. **333**(7565): p. 413.
61. Raum, E. and M. Perleth, *Methoden der Metaanalysen von diagnostischen Genauigkeitsstudien.* 2003, Deutsche Agentur für Health Technology Assessment des Deutschen Instituts für Medizinische Dokumentation und Information: Köln.
62. Lijmer, J.G., P.M. Bossuyt, and S.H. Heisterkamp, *Exploring sources of heterogeneity in systematic reviews of diagnostic tests.* Stat Med, 2002. **21**(11): p. 1525-37.
63. Guyatt, G.H., et al., *The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies.* J Chronic Dis, 1986. **39**(4): p. 295-304.
64. Lord, S.J., et al., *When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials?[see comment].* Annals of Internal Medicine, 2006. **144**(11): p. 850-5.
65. Guyatt, G.H., et al., *A framework for clinical evaluation of diagnostic technologies.* Cmaj, 1986. **134**(6): p. 587-94.
66. Schunemann, H.J., et al., *Grading quality of evidence and strength of recommendations for diagnostic tests and strategies.* Bmj, 2008. **336**(7653): p. 1106-10.
67. Medical Services Advisory Committee, *Guidelines for the assessment of diagnostic technologies.* 2005, MSAC: Canberra.
68. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, *Osteodensitometrie bei primärer und sekundärer Osteoporose (Vorbericht).* 2009, IQWiG: Köln.
69. Merlin, T. and S. Lehman, *The Benefits and Flaws of the Linked Evidence Approach to Assess Diagnostic and Screening Tests*, in *HTAi 2010.* 2010: Dublin, Ireland.
70. Eikermann, M. *"Linked Evidence" - Wann lassen sich aus der Verknüpfung der Ergebnisse von Therapie- und Diagnostikstudien Hinweise auf den Patientennutzen diagnostischer Verfahren gewinnen?* in *10. Jahrestagung des Deutschen Netzwerks für Evidenzbasierte Medizin 2009.* 2009. Berlin.
71. Teutsch, S.M., et al., *The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group.* Genet Med, 2009. **11**(1): p. 3-14.
72. Siebert, U., *When should decision-analytic modeling be used in the economic evaluation of health care?* Eur J Health Econom, 2003. **4**: p. 143-150.
73. Trikalinos, T.A., U. Siebert, and J. Lau, *Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations.* Med Decis Making, 2009. **29**(5): p. E22-9.
74. Schwartz, F.W., et al., *Das Public Health Buch.* 2nd ed. 2003, München: Schwartz, F.W.
75. Schweitzer, S.O., *Cost effectiveness of early detection of disease.* Health Serv Res, 1974. **9**(1): p. 22-32.
76. Sutton, A.J., et al., *Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests.* Medical Decision Making, 2008. **28**(5): p. 650-667.
77. Drummond, M.F., et al., *Methods for the Economic Evaluation of Health Care Programmes.* 3rd ed. 2005, Oxford: Oxford University Press.

78. EUnetHTA, *HTA Core Model for Diagnostic Technologies 1.0R*. 2008.
79. National Institute for Health and Clinical Excellence, *The guidelines manual*. 2009, NICE: London.
80. National Institute for Health and Clinical Excellence, *Diagnostics Assessment Programme - Interim methods statement (pilot)*. 2010, NICE: London.
81. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, *Allgemeine Methoden*. 2008, IQWiG.
82. Drummond, M. and T.O. Jefferson, *Guidelines for authors and peer reviewers of economic submissions to the BMJ.* British Medical Journal, 1996. **313**: p. 275-283.
83. Siebert, U., et al., *Entwicklung eines Kriterienkatalogs zur Beschreibung und Bewertung ökonomischer Evaluationsstudien in Deutschland*, in *Ansätze und Methoden der ökonomischen Evaluation. Eine internationale Perspektive*, R. Leidl, J.M. Graf von der Schulenburg, and J. Wasem, Editors. 1999, Nomos Verlag: Baden-Baden.
84. Medical Services Advisory Committee. *Welcome to the Medical Services Advisory Committee*. 2010 [cited 2010 10. January]; Available from: http://www.msac.gov.au/.
85. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, *Wissenschaftliche Bewertung verschiedener Untersuchungsmethoden zur Diagnosestellung eines Asthma bronchiale bei Kindern im ALter von 2 bis < 5 Jahren*. 2009, IQWiG: Köln.
86. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, *Allgemeine Methoden zur Bewertung von Verhältnissen zwischen Nutzen und Kosten - Version 1.0*. 2009.
87. Ottawa Health Research Institute. *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. [cited 2010 12.May]; Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm.
88. STROBE Statement - Strengthening the reporting of observational studies in epidemiology. 2009 [cited 2010 12.May.]; Available from: http://www.strobe-statement.org/index.php?id=strobe-home.