

**Deutsche Agentur für Health Technology Assessment
des Deutschen Instituts für Medizinische Dokumentation
und Information**

**Informationssystem Health Technology Assessment (HTA)
in der Bundesrepublik Deutschland**

**Methoden der Metaanalyse von diagnostischen
Genauigkeitsstudien**

Dr. med. Elke Raum, MPH

PD Dr. med. Matthias Perleth, MPH

September 2002

Korrespondenzadresse:
SiTi – Berliner Institut für Gesundheitsmanagement GmbH
c/o Dr. med. Matthias Perleth, MPH
Ahlener Weg 11h
12207 Berlin
Tel.: 030 / 82703587
Mobil: 0174 / 5724959
Email: Matthias.Perleth@gmx.de

Inhaltsverzeichnis

1	Einleitung	7
2	Hintergrund und Zielstellung	18
3	Methodik des Berichts.....	21
3.1	Literaturrecherche	21
3.2	Kapitelübersicht	22
4	Ergebnisse und Diskussion	23
4.1	Diagnostische Technologien und Arten diagnostischer Studien.....	23
4.1.1	Bewertung diagnostischer Technologien	23
4.1.2	Überblick über die Arten diagnostischer Studien	24
4.2	Metaanalysen von Studien zur diagnostischen Genauigkeit.....	29
4.2.1	Festlegung von Ziel und Umfang der Metaanalyse	31
4.2.2	Identifizierung der relevanten Literatur	31
4.2.3	Datenextraktion und Präsentation der Daten.....	34
4.2.4	Abschätzung der Testgüte	36
4.2.5	Einschätzung der Konsequenzen von Variationen der Studienvalidität bei der Bestimmung der Testgüte/ methodische Qualität	37
4.2.6	Einschätzung der Konsequenzen von Variationen von Patientencharakteristika und des Tests auf die Bestimmung der Testgüte (Generalisierbarkeit).....	39
4.2.7	Verdeutlichung anhand ausgewählter Beispiele	41
4.2.8	Zusammenfassung.....	43
4.3	Methoden der Metaanalyse von Studien zur diagnostischen Genauigkeit	46
4.3.1	Fixed-Effects- und Random-Effects-Modelle.....	48
4.3.2	Diagnostische Odds Ratio (DOR).....	50
4.3.3	Mittelwerte der Sensitivität und Spezifität.....	50
4.3.4	Summary Receiver Operating Characteristics (SROC) - Kurve.....	51
4.3.5	Latent-Scale Regression Modelle	58
4.3.6	Modell für ordinale Daten.....	60
4.3.7	Standardisierte Mittelwertdifferenzen.....	61
4.3.8	Weitere statistische Methoden	67
4.3.8.1	Imperfekte Standards.....	67
4.3.8.2	Likelihood Ratios.....	69
4.3.8.3	Neuere statistische Ansätze.....	70
4.3.9	Zusammenfassung.....	71
4.4	Bestimmung der Heterogenität.....	74
4.4.1	Quellen der Heterogenität	74

4.4.1.1 Studiendesign	74
4.4.1.2 Methodische Qualität	75
4.4.1.3 Patientenmerkmale	76
4.4.1.4 Merkmale des Tests.....	76
4.4.2 Messung der Heterogenität	76
4.4.2.1 Ausreißer	77
4.4.2.2 Subgruppen.....	78
4.4.3 Zusammenfassung.....	78
4.5 Methoden zur Abschätzung von Publikationsbias	79
4.5.1 Verzerrungsquellen	80
4.5.2 Funnelplot	81
4.5.3 Tests auf Bias	82
4.5.4 Zusammenfassung.....	85
4.6 Weitere Aspekte von Metaanalysen diagnostischer Studien	85
4.6.1 Sensitivitätsanalysen	85
4.6.2 Qualitative Auswertungen diagnostischer Studien	86
5 Weitere Überlegungen zur Bewertung diagnostischer Tests und Schlussbemerkungen	87
Anlage	92
1 Literaturverzeichnis:	92
2 Literaturrecherche	99
3 Indexwörter oder Textwörter	102

Abkürzungsverzeichnis

BSW	Between-Study-Variation
EBM	Evidenzbasierte Medizin
DOR	Diagnostic Odds Ratio
FEM	Fixed-Effects-Modell
FNR	falsch negativ Rate / false negative rate
FÜR	falsch positiv Rate / false positive rate
HTA	Health Technology Assessment
RCT	randomisierte kontrollierte Studie / randomised controlled trial
REM	Random-Effects-Modell
ROC	Receiver Operating Characteristics
SROC	Summary Receiver Operating Characteristics
TNR	richtig negativ Rate / true negative rate
TPR	richtig positiv Rate / true positive rate
WSV	Within-Study-Variation

Tabellenverzeichnis

Tabelle 1: Hierarchisches Modell der Evaluierung diagnostischer Tests (nach Fryback and Thornbury, 1991 ²⁶).....	24
Tabelle 2: Beispiel für eine Hierarchie der Evidenz (nach Perleth und Antes, 1999 ⁶⁰)	25
Tabelle 3: Kriterien zur Bewertung der internen Validität individueller Studien (nach Harris et al, 2001 ²⁹).....	27
Tabelle 4: Schritte bei der Durchführung einer Metaanalyse diagnostischer Testverfahren (nach Irwig et al, 1994 ³⁴)	30
Tabelle 5: Anforderungen an die Berichtsqualität von Ebene-2-Studien	45
Tabelle 6: Vierfeldertafeln von Studien zur diagnostischen Genauigkeit von Röntgenübersichtsaufnahmen im Vergleich zur Punktion bei akuter Sinusitis	56
Tabelle 7: Merkmale prospektiver und retrospektiver Studien zum diagnostisch-therapeutischen Impact diagnostischer Test.....	86

Abbildungsverzeichnis

Abbildung 1: Vierfeldertafel für binäre Outcomes diagnostischer Studien.....	46
Abbildung 2: SROC-Kurve für die Metaanalyse von Röntgen der Nasennebenhöhlen vs. Punktion.....	57
Abbildung 3: Vierfeldertafel für binäre Outcomes diagnostischer Studien.....	62
Abbildung 4: Mittelwertdifferenzen und Konfidenzintervalle für sechs verschiedene nicht invasive diagnostische Methoden der Karotisstenose	66
Abbildung 5: Erwartete Wahrscheinlichkeitsverteilung der Personen: Testergebnis vs. wahrer Krankheitsstatus	68
Abbildung 6: Erwartete Häufigkeit der Ergebnisse für Studie k in Metaanalysen; Indextest vs. Referenztest	68
Abbildung 7: Vereinfachtes Schema einer umfassenden Evaluation diagnostischer Tests...	91

Abstract

FRAGESTELLUNG: Ziel dieses Berichts ist es, die derzeitigen Ansätze für die Metaanalyse diagnostischer Genauigkeitsstudien zu beschreiben, den Einfluss von Kovariablen zu untersuchen und Empfehlungen für die Anwendung von derartigen Metaanalysen abzuleiten.

METHODIK: Es wurde eine systematische Literaturrecherche nach Studien zur Methodenentwicklung von Metaanalysen zu diagnostischen Genauigkeitsstudien durchgeführt. Aus der Literatur wurden die wichtigsten Modelle beschrieben und hinsichtlich ihrer Tauglichkeit bewertet.

ERGEBNISSE UND BEWERTUNG: Diagnostische Technologien werden bewertet und die Arten diagnostischer Studien dargestellt. Die von einer internationalen Arbeitsgruppe entwickelten Leitlinien für die Evaluation von diagnostischen Testverfahren werden erläutert. Mit welcher Methode die diagnostische Genauigkeit von Primärstudien am besten zusammengefasst werden kann, hängt von Annahmen ab, die getroffen werden, um die beobachteten Unterschiede zu erklären. Wenn sich die Studien nicht in ihrem Grenzwert und in ihrer diagnostischen Genauigkeit unterscheiden, können die Felderbesetzungen der Vierfeldertafel jeder Primärstudie gepoolt und gemeinsame Parameter für FPR und TPR aller eingeschlossenen Primärstudien berechnet werden. Wenn der diagnostische Test mit der gleichen Genauigkeit in allen Primärstudien durchgeführt wird, aber unterschiedliche Grenzwerte definiert werden, würden die Studien am besten mit einer summary receiver operating characteristics - Kurve zusammengefasst werden. Am häufigsten wird die Erstellung einer SROC-Kurve nach der Methode von Moses et al. für Fixed-Effects-Modelle verwendet. Eine weitere Methode für Metaanalysen von diagnostischen Tests mit binären Ergebnissen und Goldstandardinformationen ist eine „latent-scale“ logistische Regression, um eine oder mehrere SROC-Kurven anzupassen. Methoden zur Berücksichtigung von kontinuierlichen oder wenigstens ordinal skalierten Daten sollten für Metaanalysen eingesetzt werden, wenn in den Primärstudien derartige Daten präsentiert werden. Hierzu zählen die Erstellung einer ROC-Kurve für jede Primärstudie und einer SROC-Kurve mittels ordinaler Regressionstechniken, die Kalkulation der standardisierten Differenz der empirischen Mittelwerte und die Erstellung von Ergebnis spezifischen Likelihood Ratios.

Es hat sich gezeigt, dass offenbar eine Reihe von Kovariablen für die Heterogenität zwischen Studien verantwortlich sind. Der Einfluss dieser auf die Ergebnisse von Metaanalysen muss überprüft werden. Publikationsbias ist ein Problem von Studien zu diagnostischen Tests. Das

Vorliegen von Publikationsbias sollte graphisch (Funnelplot) und statistisch überprüft werden. Weitere Aspekte von Metaanalysen diagnostischer Studien, wie Sensitivitätsanalysen oder die qualitative Auswertung werden im Rahmen dieses Kurz-HTA-Berichts umrissen.

SCHLUSSFOLGERUNG: Systematische Reviews und Metaanalysen zu diagnostischen Genauigkeitsstudien leisten für alle Beteiligten im Gesundheitswesen einen wertvollen Beitrag, in dem sie wissenschaftliche Evidenz übersichtlich verfügbar machen.

Abstract

OBJECTIVE: Objective of this report is to describe the current methods for meta-analysis of diagnostic accuracy studies and to give recommendations for their application.

METHODS: A systematic literature research for methodological studies of meta-analysis for diagnostic tests was conducted.

RESULTS: Diagnostic technologies and types of diagnostic studies are described. Guidelines for meta-analysis evaluating diagnostic tests are systematically explained. Whether a meta-analysis can be conducted depends on the number and the methodological quality of the included primary studies and the degree of heterogeneity of the estimates for the diagnostic accuracy. Which method can be used to aggregate the diagnostic accuracy of primary studies depends on the assumptions, which are made, to explain the observed differences. When studies do not differ in regard to the diagnostic threshold and diagnostic accuracy, the two-by-two tables of every primary study can be pooled and one summary parameter for the false positive rate (FPR) and true positive rate (TPR) can be calculated. When the diagnostic test is conducted with the same accuracy in all primary studies but with different thresholds, the primary studies are aggregated appropriately using an summary receiver operating characteristics-curve (SROC). The most common method was developed by Moses et al for fixed-effects models. Another method for meta-analysis of diagnostic tests with binary tests and gold standard information is the “latent-scale” logistic regression (LSLR), to fit one or several SROC-Curves. Methods which take into account continuous or at least ordinal scaled data should be used whenever primary studies report such data. Those methods are: calculating a ROC-curve for every category and an SROC-curve using logistic regression techniques, the calculation of standardised mean differences, or the calculation of result specific likelihood ratios. Heterogeneity between studies can be caused by several co-variates. The influence of these and other potentially relevant variables on the results of meta-analysis has to be considered. Publication bias is a problem of studies of diagnostic tests. The existence of publication bias should be tested graphically (funnel plot) and statistically. More aspects of meta-analysis of diagnostic accuracy studies, like sensitivity analysis or qualitatively evaluation are described in this report.

CONCLUSION: Systematic reviews and meta-analyses of diagnostic accuracy studies provide valuable information for all parts of the health care sector, because they make scientific evidence available.

Executive Summary

FRAGESTELLUNG: Systematische Übersichtsarbeiten und die Synthese veröffentlichter Evidenz der Genauigkeit diagnostischer Tests sind im Laufe der Jahre immer notwendiger geworden. Die Informationen solcher Arbeiten sind Schlüsselemente zur Entscheidungsfindung im Gesundheitswesen und in der Klinik. Unter einem systematischen Review versteht man eine Übersichtsarbeit auf der Basis von Primärstudien zu einer klar formulierten Fragestellung, bei der systematisch und anhand expliziter Kriterien relevante Literatur identifiziert, kritisch bewertet und einer qualitativen und eventuell auch quantitativen Analyse (Metaanalyse) unterzogen wird. Ziel dieses Berichts ist es, die derzeitigen Ansätze für die Metaanalyse diagnostischer Genauigkeitsstudien zu beschreiben, anhand von empirischen Daten die Brauchbarkeit und Vergleichbarkeit darzustellen, den Einfluss von Kovariablen zu untersuchen und Empfehlungen für die Anwendung von derartigen Metaanalysen abzuleiten.

METHODIK: Es wurde eine systematische Literaturrecherche nach Studien zur Methodenentwicklung von Metaanalysen zu diagnostischen Genauigkeitsstudien in den Datenbanken MEDLINE, EMBASE, Current Contents, CINAHL; BIOSIS Previews durchgeführt. Die Jahrgänge 2000 und 2001 folgender Zeitschriften wurden per Handsuche durchgesucht: *Statistics in Medicine*, *Medical Decision Making*, *Journal of Clinical Epidemiology*. Weiterhin wurden Referenzlisten publizierter Original- und Übersichtsarbeiten nach weiterführenden Literaturstellen überprüft, und Berichte, Bücher in der aktuellsten Auflage, sowie Dissertationen und Habilitationsschriften einbezogen.

ERGEBNISSE UND BEWERTUNG: Bevor auf die eigentlichen Methoden der Metaanalysen zu diagnostischen Genauigkeitsstudien eingegangen wird, werden diagnostische Technologien bewertet und die Arten diagnostischer Studien dargestellt. Von einer internationalen Arbeitsgruppe wurden Leitlinien⁵ für die Evaluation von diagnostischen Testverfahren entwickelt. Folgenden Schritte werden systematisch erläutert: Festlegung von Ziel und Umfang der Metaanalyse, Identifizierung der relevanten Literatur, Datenextraktion und Präsentation der Daten, Abschätzung der Testgüte, Einschätzung der Konsequenzen von Variationen der Studienvalidität bei der Bestimmung der Testgüte, Einschätzung der Konsequenzen von Variationen von Patientencharakteristika und des Tests auf die Bestimmung der Testgüte (Generalisierbarkeit).

Ob eine Metaanalyse durchgeführt werden kann oder nicht, hängt von der Anzahl und der methodischen Qualität der eingeschlossenen Primärstudien und dem Grad der Heterogenität der Schätzer der diagnostischen Genauigkeit ab.

Diagnostische Studien mit binären Ergebnissen (Test positiv oder negativ) stellen den häufigsten Fall in der Literatur dar. Die Standardgrößen Sensitivität (TPR, true positive rate), Spezifität ($1 - \text{FPR}$, false positive rate) und die Likelihood Ratios können entsprechend berechnet werden. Eine häufig benutzte Maßzahl ist das diagnostische Odds Ratio. Die DOR ist eine Maßzahl für die diskriminatorische Fähigkeit eines Tests. Diese ist um so höher, je größer das DOR ausfällt.

Mit welcher Methode die diagnostische Genauigkeit von Primärstudien am besten zusammengefasst werden kann, hängt von Annahmen ab, die getroffen werden, um die beobachteten Unterschieden zu erklären.

Das einfachste aber zugleich auch das restriktivste Modell geht von der Annahme aus, dass sich die Studien weder in ihrem Grenzwert noch in ihrer diagnostischen Genauigkeit unterscheiden. Unter diesen Annahmen können die Felderbesetzungen der Vierfeldertafel jeder Primärstudie gepoolt und gemeinsame Parameter für FPR und TPR aller eingeschlossenen Primärstudien berechnet werden. Ein weniger restriktives Modell geht von der Annahme aus, dass der diagnostische Test mit der gleichen Genauigkeit in allen Primärstudien durchgeführt wird, aber unterschiedliche Grenzwerte definiert werden. Unter diesen Annahmen würden die Studien am besten mit einer summary receiver operating characteristics - Kurve zusammengefasst werden. Am häufigsten wird die Erstellung einer SROC-Kurve nach der Methode von Moses et al ⁶ für Fixed-Effects-Modelle verwendet. Diese geht von der Annahme aus, dass ein linearer Zusammenhang zwischen dem $\text{logit}(\text{TPR})$ und dem $\text{logit}(\text{FPR})$ besteht. Die abhängige Variable D stellt den Logarithmus des diagnostischen Odds Ratios dar. Eine weitere Methode für Metaanalysen von diagnostischen Tests mit binären Ergebnissen und Goldstandardinformationen ist eine „latent-scale“ logistische Regression (LSLR) ⁷, um eine oder mehrere SROC-Kurven anzupassen. Das LSLR-Modell geht von zwei latenten logistischen Verteilungen der dichotomen Testergebnisse aus, eine für die erkrankte und eine für die nicht erkrankte Population.

Die Dichotomisierung von Testergebnissen, um Sensitivität und Spezifität zu erhalten, führt zu einem Informationsverlust. Methoden zur Berücksichtigung von kontinuierlichen oder wenigstens ordinal skalierten Daten sollten für Metaanalysen eingesetzt werden, wenn in den

Primärstudien derartige Daten präsentiert werden. Wenn in den Primärstudien für die gleiche Anzahl von Kategorien Daten vorhanden sind, kann für jede Primärstudie eine ROC-Kurve erstellt werden und mittels ordinaler Regressionstechniken eine Gesamt-ROC-Kurve ⁴.

Als weitere Methode für Metaanalysen von Studien, die (annähernd normalverteilte) kontinuierliche Testergebnisse berichten, wird die Kalkulation der standardisierten Differenz der empirischen Mittelwerte vorgestellt ³. Dabei wird d als Effektgröße berechnet. d ist ein Maß für die Diskriminierungsfähigkeit oder Wirksamkeit (test effectiveness score) des untersuchten Tests. Um so größer d , desto größer ist die diskriminatorische Fähigkeit des Tests. Als weitere Methode für kontinuierliche Testergebnisse zeigen Irwig et al ⁴ die Erstellung von Ergebnis spezifischen Likelihood Ratios auf.

Es hat sich gezeigt, dass offenbar eine Reihe von Kovariablen für die Heterogenität zwischen Studien verantwortlich sind. Hierzu gehören u.a. die Qualität der Studien, das Publikationsjahr und die verwendete Untersuchungstechnik. Der Einfluss dieser und anderer potentiell relevanter Kovariablen auf die Ergebnisse von Metaanalysen muss überprüft werden. Die Untersuchung der Heterogenität im Rahmen der diagnostischen Forschung gibt dem Wissenschaftler wertvolle Indikatoren über die wichtige Quellen der Variabilität in der Genauigkeit diagnostischer Studien.

Publikationsbias ist ein Problem von Studien zu diagnostischen Tests. Viele Studien zur diagnostischen Genauigkeit nutzen Daten, die primär im Rahmen der klinischen Routine gesammelt wurden. Studien, die zur Veröffentlichung gelangen, sind häufiger verzerrt und überschätzen wahrscheinlich die Testgenauigkeit. Die Tests von Beggs und Mazumdar sowie Egger ^{1,2} gehören zu den gebräuchlichsten, um das Vorhandensein von Publikationsbias zu überprüfen. Neben der Durchführung eines statistischen Tests auf Publikationsbias ist die Erstellung eines Funnelplots in jeder Metaanalyse sinnvoll. Weitere Aspekte von Metaanalysen diagnostischer Studien, wie Sensitivitätsanalysen oder die qualitative Auswertung werden im Rahmen dieses Kurz-HTA-Berichts umrissen.

SCHLUSSFOLGERUNG: Medizinische Praxis versucht ständig effektiver, effizienter und nebenwirkungsärmer zu werden, in dem sie wissenschaftliche Evidenz standardisiert zusammenfasst, in dem sie diese Evidenz in standardisierter Weise in praktische Leitlinien übersetzt und in dem sie diese wissenschaftliche Evidenz für alle Beteiligten im Gesundheitswesen verfügbar macht. Systematische Reviews und Metaanalysen zu diagnostischen Test können dabei in vieler Hinsicht wertvolle Beiträge leisten. Sie ermöglichen es, schlechte oder nutzlo-

se Tests zu eliminieren bevor sie weitverbreitet Anwendung finden. Sie stellen eine verbesserte Qualität der Information über diagnostische Tests zur Verfügung, in dem sie das Patientenspektrum darstellen oder relevante Subgruppen mit dem Ziel einer verbesserten Patientenversorgung analysieren.

LITERATURVERZEICHNIS

1. *Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088-99.*
2. *Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;629-34.*
3. *Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull 1995;117:167-78.*
4. *Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1995;48:119-30.*
5. *Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers C et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667-76.*
6. *Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12:1293-316.*
7. *Rutter, CM, Gatsonis, C. Regression methods for meta-analysis of diagnostic test data. Acad Radiol 1995;2,: S48-S56.*

Executive Summary (Englisch)

OBJECTIVE: Systematic reviews and the synthesis of published evidence of the accuracy of diagnostic tests have become more and more necessary over the years. The information provided by such publications is one important element for decision making in the health care sector and for clinical settings. Systematic reviews should consist of a clearly defined objective, a systematic identification of relevant literature for explicitly set criteria, a critical assessment and a qualitative or quantitative analysis (meta-analysis). Objective of this report is to describe the current methods for meta-analysis of diagnostic accuracy studies and to give recommendations for their application.

METHODS: A systematic literature research for methodological studies of meta-analysis for diagnostic tests was conducted using the following computerised bibliographic systems: MEDLINE, EMBASE, Current Contents, CINAHL and BIOSIS Previews. The journals "Statistics in Medicine", "Medical Decision Making" and "Journal of Epidemiology" from 2000 and 2001 were hand searched for relevant publications. Reference lists of published original publications and reviews were screened for more references. Reports, the latest edition of books and doctoral theses were included into this report.

RESULTS: Diagnostic technologies and types of diagnostic studies are described. Guidelines⁵ for meta-analysis evaluating diagnostic tests are systematically explained: determination of the objective and the scope of the meta-analysis, retrieval of relevant literature, extraction and display of the data, estimation of diagnostic accuracy, assessment of the effect of variation in study validity on estimates of diagnostic accuracy, assessment of the effect of variation in the characteristics of patients and test on estimates of diagnostic accuracy (generalisability). Whether a meta-analysis can be conducted or not, depends on the number and the methodological quality of the primary studies and on the degree of heterogeneity of the estimates for the diagnostic accuracy. Diagnostic studies mainly report binary results (test positive or negative). The standard measures sensitivity, specificity, likelihood ratios and diagnostic odds ratio can be calculated, respectively. Another measure which, is often used, is the diagnostic odds ratio, which describes the discriminating ability of the test. What method should be used to aggregate the diagnostic accuracy of primary studies depends on the assumptions, which are made, to explain the observed differences.

The simplest but also most restrictive model is based on the assumptions, that the studies do not differ in regard to the diagnostic threshold and diagnostic accuracy. For these assump-

tions, the two-by-two tables of every primary study can be pooled and one summary parameter for the false positive rate (FPR) and true positive rate (TPR) can be calculated. A less restrictive model assumes, that the diagnostic test is conducted with the same accuracy in all primary studies but with different thresholds. Then the primary studies are aggregated appropriately using an summary receiver operating characteristics-curve (SROC). The most common method was developed by Moses et al ⁶ for fixed-effects models. The assumption is made, that there is a linear association between logit (TPR) and logit (FPR). The dependent variable D is the logarithm of the diagnostic odds ratio. Another method for meta-analysis of diagnostic tests with binary tests and gold standard information is the “latent-scale” logistic regression (LSLR) ⁷, to fit one or several SROC-Curves. The LSLR assumes two latent logistic distributions for the binary test results, one for the diseased and non diseased study population.

The dichotomising of test results to obtain sensitivity and specificity reduces the available information. Methods which take into account continuous or at least ordinal scaled data should be used whenever primary studies report such data. When primary studies present the same number of categories, a ROC-curve can be calculated for every primary study and an SROC-curve using logistic regression techniques ⁴. Another methods for meta-analysis of studies with almost normally distributed test results is the calculation of standardised mean differences ³. The estimated effect size d measures the discriminating ability or the effectiveness of the tests (test effectiveness score). The larger d, the better the discriminating ability of the test. For continuous data also the calculation of result specific likelihood ratios can be used ⁴.

Heterogeneity between studies can be caused by several co-variables, like study quality, year of publication or used examination technique. The influence of these and other potentially relevant variables on the results of meta-analysis has to be considered. The investigation of heterogeneity provides valuable information about important sources of variability in the accuracy of diagnostic tests.

Publication bias is a problem of studies of diagnostic tests. Many primary studies use data, which were originally collected during clinical routine. Studies which are published are often biased and overestimate the diagnostic accuracy. The tests of Beggs and Mazumdar and Egger ^{1,2} are the ones, mostly used to test the existence of publication bias. It is also very useful to generate a funnel plot to address this question.

More aspects of meta-analysis of diagnostic accuracy studies, like sensitivity analysis or qualitatively evaluation are described in this report.

CONCLUSION: Medicine tries continuously to become more effective, more efficient and less harmful, by accumulating standardised scientific evidence, by transforming this evidence into guidelines and making this scientific evidence available for all parts of the health care sector. Systematic reviews and meta-analyses of diagnostic accuracy studies provide valuable information in this regard. They make it possible to eliminate poor or useless test before they are widely established and they provide a better quality of information with the objective to improve medical care.

REFERENCE LIST

1. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;**50**:1088-99.
2. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;629-34.
3. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;**117**:167-78.
4. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;**48**:119-30.
5. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers C *et al.* Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;**120**:667-76.
6. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;**12**:1293-316.
7. Rutter CM, Gatsonis, C. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995;**2**:S48-S56

1 Einleitung

Health Technology Assessment (HTA) stellt sich die Aufgabe, eine medizinische Technologie möglichst vollständig, systematisch und transparent zu bewerten. Ihr Ziel ist es, es den Akteuren im Gesundheitswesen zu ermöglichen, mittels evidenzbasierter Information Entscheidungen zu treffen, die die bevölkerungsbezogene Gesundheit in eine günstige Richtung beeinflussen.

Zur Generierung evidenzbasierter Informationen gehört neben der Suche und Auswertung von Primärstudien auch die Nutzung hochwertiger und glaubwürdiger Übersichtsarbeiten. Der Wert dieser systematischen Übersichten für die evidenzbasierte Entscheidungsunterstützung in allen Bereichen des Gesundheitswesens kann dabei nicht hoch genug eingeschätzt werden. Sie beschleunigen die Entscheidungsfindung und führen zur Umsetzung der Evidenzbasierten Medizin (EBM) am Krankenbett. Mit dem Bedarf an systematischen Übersichten wachsen die Anforderungen an die Qualität der Übersichten. Insbesondere Nutzer von systematischen Übersichten, zu denen auch HTA-Einrichtungen gehören, sollten die Methodik von systematischen Übersichten sorgfältig bewerten, bevor sie die Ergebnisse verwenden.

Praktische Bedeutung im Rahmen von HTA haben derzeit vor allem die Ebenen diagnostische Genauigkeit und diagnostischer bzw. therapeutischer Impact. Die technische Qualität wird selten im Rahmen von HTA betrachtet. Für die Nutzenbetrachtung aus Patientenperspektive bzw. aus der Perspektive der Gesellschaft liegen in der Regel keine aussagekräftigen Studien vor. In diesem Bericht soll daher diese Ebene betrachtet werden, in der normalerweise die meisten Studien und Übersichtsarbeiten vorliegen, die Ebene der Studien zur diagnostischen Genauigkeit.

2 Hintergrund und Zielstellung

Systematische Übersichtsarbeiten und die Synthese veröffentlichter Evidenz der Genauigkeit diagnostischer Tests sind im Laufe der Jahre immer notwendiger geworden. Die Informationen solcher Arbeiten sind Schlüsselemente zur Entscheidungsfindung im Gesundheitswesen und in der Klinik.

Von Entscheidungsträgern im Gesundheitswesen wird erwartet, den „Gesamtwert“ eines diagnostischen Tests zu bewerten, ihn mit seinen Alternativen zu vergleichen und zu entscheiden, ob der Test eingeführt werden soll oder nicht.

Von Klinikern wird erwartet, eine Diagnose zu stellen, eine Prognose abzugeben und eine Behandlung einzuleiten. Diagnosen werden aufgrund von Patientencharakteristika, der Anamnese, der klinischen Untersuchung und diagnostischer Tests gestellt und sind der Eckstein für gute klinische Versorgung⁴¹. Deshalb brauchen Kliniker Hilfestellungen für effiziente diagnostische Strategien und deshalb schließt Evidenzbasierte Medizin evidenzbasierte Diagnostik ein.

Die Entscheidung beider sollte auf einer gründlichen Evaluation der jeweiligen diagnostischen Tests basieren. Der entscheidende Schritt hierzu ist, die diagnostische Genauigkeit eines Tests zu bestimmen, d. h. die Fähigkeit eines diagnostischen Verfahrens, richtig das Vorhandensein oder das Nichtvorhandensein einer Krankheit zu bestimmen. Dazu wird das Ergebnis des Tests mit dem eines Referenztests verglichen, dem sogenannten Goldstandard. Die Schätzer der diagnostischen Genauigkeit (z. B. Sensitivität und Spezifität) unterscheiden sich aber häufig zwischen den einzelnen zu diesem Test veröffentlichten Untersuchungen. Dies kann unterschiedliche Gründe haben, z. B. können sich die Studien hinsichtlich der Studienpopulation unterscheiden oder es wurden unterschiedliche Grenzwerte, die ein positives Testergebnis markieren, gewählt.

Unter einem systematischen Review versteht man eine Übersichtsarbeit auf der Basis von Primärstudien zu einer klar formulierten Fragestellung, bei der systematisch und anhand expliziter Kriterien relevante Literatur identifiziert, kritisch bewertet und einer qualitativen und eventuell auch quantitativen Analyse (Metaanalyse) unterzogen wird. Übersichtsarbeiten und Metaanalysen reduzieren eine schier unbegrenzte Menge von Information auf eine überschaubare Quantität, sie liefern einen effizienten Weg, nützliche Informationen zu extrahieren

und klinisch zu implementieren, und sie schätzen die Validität und Reproduzierbarkeit der Ergebnisse ab ⁵⁶.

Daher stellen systematische Übersichten¹ (synonym: systematische Reviews) wie sie nach mehr als einer Dekade methodischer Entwicklung unter anderem von der Cochrane Collaboration erstellt werden, die Basis der auf klinischer evidenzbasierter Entscheidungsfindung dar. Damit ist nicht nur die bloße Verfügbarkeit von mittlerweile mehreren Tausend hochwertigen systematischen Reviews gemeint, sondern gerade der Prozess der Erstellung derartiger Übersichten. Die zunehmende Verfügbarkeit systematischer Reviews hat die Verbreitung der evidenzbasierten Medizin stark gefördert, nicht zuletzt dadurch, dass viele Proponenten der EBM selbst in der Erstellung systematischer Reviews involviert waren und sind. Dieses „Ownership-Prinzip“ hat auch zu einer schnellen und selten harmonischen Akzeptanz der Arbeitsweise der Cochrane Collaboration als Standard beigetragen. Die Durchführung von systematischen Reviews wird mittlerweile als originäre Forschung anerkannt und entsprechend gewertet.

Metaanalysen, die vorliegende Genauigkeitsstudien zusammenfassen, sind ein hilfreiches Instrument, die Genauigkeit diagnostischer Tests zu beurteilen. Metaanalysen verhelfen dazu,

1. eine „Overall Summary“ der diagnostischen Genauigkeit zu erstellen,
2. zu bestimmen, ob die Schätzer der diagnostischen Genauigkeit von den Studiencharakteristika der originären Studien abhängen (Studiengültigkeit),
3. zu bestimmen, ob sich die diagnostische Genauigkeit in einzelnen Untergruppen unterscheidet und
4. weiteren Forschungsbedarf zu identifizieren ³⁴.

Zu den Prinzipien von systematischen Übersichtsarbeiten oder Metaanalysen gehören die Fokussierung der Fragestellung, die Erstellung eines Reviewprotokolls, inklusive Festlegung von Ein- und Ausschlusskriterien für die Primärstudien, eine auf Vollständigkeit zielende Literaturrecherche, die Qualitätsbewertung der eingeschlossenen Studien, die detaillierte Ex-

¹ Das systematische Vorgehen und die Fokussierung auf eine relevante Fragestellung oder mehrere miteinander verbundene Fragestellungen unterscheidet systematische von sogenannten narrativen Reviews. Zu den narrativen Übersichten gehören die praxisorientierten Übersichten in Zeitschriften, die in der Regel ein Krankheitsbild oder eine Intervention (z. B. ein diagnostisches oder therapeutisches Verfahren) behandeln. Es handelt sich dabei um Zusammenstellungen von Experten, die nicht auf systematischen Recherchen und Analysen beruhen. Die Methodik des Vorgehens wird meist nicht beschrieben. In diese Kategorie fallen auch viele Lehrbuchkapitel, die zudem an den jeweiligen Leserkreis adaptiert sein können. Eine grundsätzliche Kritik an narrativen Re-

traktion der Studiendaten in Tabellen, die Durchführung der Metaanalyse, der Test auf Robustheit und Verzerrungsfreiheit der Ergebnisse sowie die Ableitung von Schlussfolgerungen auf der Basis der vorliegenden Evidenz.

Im Gegensatz zu therapeutischen Studien ist das Instrumentarium zur Durchführung von Metaanalysen diagnostischer Genauigkeitsstudien wesentlich weniger weit entwickelt. Die oft zahlreich vorliegenden Genauigkeitsstudien werden immer häufiger, auch oft unkritisch, in Form von Metaanalysen gepoolt, um die statistische Aussagekraft zu erhöhen. Um Fehlschlüsse zu vermeiden, ist es notwendig, methodische Fehler zu minimieren.

Bisher wurde international kein Konsens darüber hergestellt, welche Methode unter welchen Bedingungen die zuverlässigsten Resultate liefert.

Dies bezieht sich auf die folgenden Vorgehensweisen:

- Identifizierung von Studien,
- Bewertung der Qualität,
- Auswertung bzw. Datenextraktion,
- Metaanalyse,
- Tests auf Heterogenität und Publikationsbias,
- Identifizierung relevanter Kovariablen mit Einfluss auf das Ergebnis der Metaanalyse.

Ziel dieses Berichts ist es, die derzeitigen Ansätze für die Metaanalyse diagnostischer Genauigkeitsstudien zu beschreiben, anhand von empirischen Daten die Brauchbarkeit und Vergleichbarkeit darzustellen, den Einfluss von Kovariablen zu untersuchen und Empfehlungen für die Anwendung von derartigen Metaanalysen abzuleiten. Es werden vor allem Methoden zur Zusammenfassung von Studien mit binären Testergebnissen vorgestellt, da diese den häufigsten Fall in der Literatur darstellen. Eine kurzen Überblick über Methoden bei ordinalen und stetigen Testergebnissen geben die Kapitel 4.3.6 - 4.3.8.

views besteht in ihrer Anfälligkeit für Bias durch selektives Zitieren der die eigene These stützenden Literatur. Dies wurde u. a. für Aufsätze des bekannten Chemikers Pauling nachgewiesen^{42,57}.

3 Methodik des Berichts

3.1 Literaturrecherche

Es wurde eine systematische Literaturrecherche in den Datenbanken MEDLINE, EMBASE, Current Contents, CINAHL; BIOSIS Previews durchgeführt. Die Suchstrategie wurde zunächst unspezifisch formuliert, um keine relevanten Publikationen zu übersehen. In einem ersten Schritt wurde nach einer Reihe von Begriffen gesucht: „diagnosis“, „diagnostic“, „false negativ“, „accuracy“, „specificity“, „ROC curve“, „meta-analysis“, „metaanalysis“, „meta-analytic“, „method“, „methods“, „technique“, „techniques“, „technic“, „technics“. Diese wurden nach inhaltlichen Gesichtspunkten verknüpft und die Anzahl relevanter Literatur so reduziert. Es wurden Veröffentlichungen von 1995 bis 2001 berücksichtigt (siehe Anlage 2). Diese Suchergebnisse (3 – 347 Treffer) wurden anhand des Titels und/oder des Abstrakts hinsichtlich der thematischen und methodischen Relevanz manuell selektiert. Die Jahrgänge 2000 und 2001 folgender Zeitschriften wurden zusätzlich per Handsuche durchsucht: *Statistics in Medicine* (Volume 19, 2000; 1 – 24; Volume 20, 2001; 1 – 20 (30.10.01)); *Medical Decision Making* (Volume 20, 2000; 1 – 6, Volume 21, 2001; 1 – 5 (Sept./Oct.)); *Journal of Clinical Epidemiology* (Volume 53, 2000; 1 – 12, Volume 54, 2001; 1- 10 (Oct.)). Auch hier wurden die Titel und Abstrakts hinsichtlich der thematischen und methodischen Relevanz manuell selektiert.

Falls die Titel oder die Abstrakts keine eindeutige inhaltliche Zuordnung ergaben, wurde der gesamte Artikel zur Entscheidungsfindung bzgl. seiner Relevanz herangezogen.

Weiterhin wurden Referenzlisten publizierter Original- und Übersichtsarbeiten nach weiterführenden Literaturstellen überprüft, und Berichte, Bücher in der aktuellsten Auflage, sowie Dissertationen und Habilitationsschriften einbezogen.

Eingeschlossen wurden: Studien zur Methodenentwicklung von Metaanalysen zu diagnostischen Genauigkeitsstudien. Hinsichtlich der Publikationstypen wurden keine Restriktionen vorgenommen. Eingeschlossen wurden Monographien, Dissertationen, Kongressbeiträge, Editorials, Kommentare, Simulationsstudien und Übersichtsarbeiten zu diesem Thema.

Ausgeschlossen wurden: Reviews und Metaanalysen zu einzelnen diagnostischen Tests sowie Publikationen, die jeweils keine neuen, weiterführenden Aspekte zu der Thematik beinhalten (z. B. bloße Wiederholungen bereits publizierter Sachverhalte).

3.2 Kapitelübersicht

Bevor auf die eigentlichen Methoden der Metaanalysen zu diagnostischen Genauigkeitsstudien eingegangen wird, werden in Kapitel 4.1 diagnostische Technologien bewertet und die Arten diagnostischer Studien dargestellt.

Von einer internationalen Arbeitsgruppe wurden Leitlinien für die Evaluation von diagnostischen Testverfahren entwickelt³⁴. Diese werden im Rahmen dieses Berichts (Kapitel 4.2) zunächst dargestellt und systematisch besprochen.

In der Literatur werden verschiedene Ansätze für Metaanalysen diagnostischer Studien diskutiert. Hierzu gehören vor allem das *diagnostic odds ratio* (DOR), die Mittelwertdifferenzmethode³⁰ und die *summary receiver operating characteristics curve* (SROC-Kurve)⁵⁵. Diese Methoden werden ausführlich in Kapitel 4.3 dargestellt.

Es hat sich gezeigt, dass offenbar eine Reihe von Kovariablen für die Heterogenität zwischen Studien verantwortlich sind. Hierzu gehören u.a. die Qualität der Studien, das Publikationsjahr und die verwendete Untersuchungstechnik⁷⁷. Der Einfluss dieser und anderer potentiell relevanter Kovariablen auf die Ergebnisse von Metaanalysen und die Methoden zur Überprüfung werden in Kapitel 4.4 beschrieben.

Schließlich sollen verschiedene in der Literatur diskutierte Tests auf Publikationsbias dargestellt und auf ihre Anwendbarkeit hin überprüft werden. Hierzu werden insbesondere die von Therapiestudien entlehnte Methode der Konstruktion von Funnelplots sowie verschiedene statistische Tests zur Überprüfung der Asymmetrie der Funnelplots dargestellt (Kapitel 4.5).

In Kapitel 4.6 werden weitere Aspekte von Metaanalysen diagnostischer Studien, wie Sensitivitätsanalysen oder die qualitative Auswertung kurz umrissen. In einem abschließenden 5. Kapitel wird eine Zusammenfassung der Übersichtsarbeit gegeben und die wichtigsten Aussagen wiedergegeben und diskutiert.

4 Ergebnisse und Diskussion

4.1 Diagnostische Technologien und Arten diagnostischer Studien

4.1.1 Bewertung diagnostischer Technologien

Von verschiedenen Autoren bzw. Arbeitsgruppen wurden Klassifikationen von diagnostischen Studien hinsichtlich der Qualität, der jeweiligen Evaluationsphase bzw. der Generalisierbarkeit ihrer Ergebnisse vorgeschlagen. Ein Memorandum der Deutschen Gesellschaft für Medizinische Dokumentation, Informatik und Statistik e.V. (jetzt: Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie) aus dem Jahre 1989 fordert, in Analogie zu den vier Phasen der Evaluation von Arzneimitteln, eine phasenweise Evaluation diagnostischer Tests. Die vier Phasen beinhalten technische und methodische Voruntersuchungen (Phase I); Schätzung der Sensitivität (bei Kranken) und Spezifität (bei Gesunden) (Phase II); in Phase III wird eine kontrollierte diagnostische Studie im Vergleich zum etablierten Goldstandard durchgeführt; in der letzten Phase soll die Wirksamkeit hinsichtlich der Auswirkung auf den Krankheitsverlauf der Patienten überprüft werden⁴³.

Die Einteilung findet sich auch in einem Vorschlag von Fryback und Thornbury (1991), der die Diskussion der 1970er und 80er Jahre aufnimmt und erweitert. Das Modell beinhaltet eine sechsstufige Hierarchie (siehe Tabelle 1) von Studiendesigns entsprechend den Charakteristika der jeweiligen Testphase.

Bei der Einteilung diagnostischer Tests in ihre jeweilige Evaluationsphase sind die Anforderungen an die Berichtsqualität entsprechend zu berücksichtigen. Praktische Bedeutung im Rahmen von HTA haben derzeit vor allem die Ebenen diagnostische Genauigkeit und diagnostischer bzw. therapeutischer Impact. Die technische Qualität wird selten im Rahmen von HTA betrachtet. Für die Nutzenbetrachtung aus Patientenperspektive bzw. aus der Perspektive der Gesellschaft liegen in der Regel keine aussagekräftigen Studien vor.

Tabelle 1: Hierarchisches Modell der Evaluierung diagnostischer Tests (nach Fryback and Thornbury, 1991²⁶).

Level 1: Technische Qualität	<ul style="list-style-type: none"> • Demonstration der Korrelation der Diagnose (pathologisch gesichert) mit dem Testergebnis • Untersuchung der Inter- und Intra-Rater-Reliabilität • Eindeutige Auswertungskriterien für den Test müssen vorliegen
Level 2: Diagnostische Genauigkeit	<ul style="list-style-type: none"> • Bestimmung von Sensitivität und Spezifität an ausreichend großen Stichproben bzw. mit Hilfe von Metaanalysen • Repräsentation eines möglichst breiten Spektrums von Patienten / Krankheitsstadien • Etablierung von Referenzwerten
Level 3: Diagnostischer Impact	<ul style="list-style-type: none"> • Vergleich von zwei Tests bei einem Patienten in zeitlich naher Abfolge und zufälliger Reihenfolge • Verblindete (d. h. ohne Kenntnis von Krankheitszustand und Ergebnis des jeweils konkurrierenden Tests) Auswertung der Testergebnisse • Vergleich mit Goldstandard
Level 4: Therapeutischer Impact	<ul style="list-style-type: none"> • Demonstration therapeutischer Konsequenzen im Vergleich mit Hilfe klinischer Studien (vorzugsweise RCTs) • Verwendung expliziter Kriterien zur Demonstration des therapeutischen Impacts
Level 5: Nutzen aus der Perspektive des Patienten	<ul style="list-style-type: none"> • wie therapeutischer Impact, aber Betonung auf patientenrelevante Endpunkte wie funktioneller Status, Schmerzstatus, Lebensqualität • Demonstration mit Hilfe von RCTs, aber auch retrospektiver Studien (ethisch weniger problematisch), Entscheidungsanalyse
Level 6: Nutzen aus gesellschaftlicher Perspektive	<ul style="list-style-type: none"> • Nutzen und Kosten-Nutzen aus gesellschaftlicher Sicht

4.1.2 Überblick über die Arten diagnostischer Studien

Als Maßstab für die „Güte“ der Evidenz haben sich die sogenannten Levels of Evidence durchgesetzt. Dabei handelt es sich um eine Hierarchie von Studiendesigns (siehe Tabelle 2).

Tabelle 2: Beispiel für eine Hierarchie der Evidenz (nach Perleth und Antes, 1999⁶⁰)

Stufe	Evidenz-Typ
I	starke Evidenz: wenigstens ein systematischer Review auf der Basis methodisch hochwertiger kontrollierter, randomisierter Studien (RCTs)
II	wenigstens ein ausreichend großer, methodisch hochwertiger RCT
III	methodisch hochwertige Studien ohne Randomisierung bzw. nicht prospektiv (Kohorten-, Fall-Kontroll-Studien)
IV	mehr als eine methodisch hochwertige nicht experimentelle Studie
V	Meinungen und Überzeugungen von angesehenen Autoritäten (aus klinischer Erfahrung), Expertenkommissionen, beschreibende Studien

Soweit feststellbar, wurde eine Hierarchie der Evidenz zuerst von der Canadian Task Force on the Periodic Health Examination konsequent verwendet ¹¹. Die Task Force war damit beauftragt, Empfehlungen zur Prävention zu erarbeiten. Um die Sicherheit der Empfehlungen einschätzen zu können, wandten die Autoren ein System der Hierarchie der Evidenz an. Aus diesem System wurde dann die Stärke der Empfehlungen abgeleitet. Die Stärke der Empfehlung (class / strength of recommendation) wurde dabei in drei Grade eingeteilt (gut, mittel, schlecht). Diese auch von der US Preventive Services Task Force übernommene Vorgehensweise wurde seither in vielen Varianten weiterentwickelt und dadurch zum Standard einer Bewertung von medizinischen Technologien.²

Die Hierarchie bezieht sich nur auf die interne Validität von Studien und ordnet diese entsprechend. Die interne Validität sagt etwas über die Nähe des beobachteten zum wahren Effekt aus, oder anders ausgedrückt, den Grad der Freiheit von systematischen Fehlern, die verzerrend auf das Studienergebnis wirken. Eine Studie auf dem Evidenz-Level I, also eine leger artis durchgeführte (und publizierte!) randomisierte kontrollierte Studie (systematische Übersichten sollen zunächst unberücksichtigt bleiben), kann demnach als Studie mit hoher Aussagekraft im Sinne der internen Validität gewertet werden. Eine über die Validität hinausgehende Differenzierung von Studien ist alleine anhand einer solchen Skala nicht möglich.

Die dritte U. S. Preventive Services Task Force ²⁹ hat die Evidenz-Level um einige entscheidende Punkte ergänzt:

² Neuere Ansätze der Ableitung von Empfehlungsstärken berücksichtigen neben der Validität noch mindestens zwei weitere Aspekte. Dies ist zum einen die Größe des Effekts bzw. die Schwelle, an der der Nutzen den potentiellen Schaden übersteigt; hierbei können auch die Kosten berücksichtigt werden. Die Number-needed-to-treat bzw. die Number-needed-to-harm können zur Einschätzung des Nutzens hilfreich sein. Zum anderen ist die Präzision des Effektschätzers, ausgedrückt im Konfidenzintervall, eine wichtige Größe, um die Wirksamkeit einer Maßnahme einschätzen zu können (siehe auch die Fassung des Centre for Evidence-based Medicine in Oxford [<http://cebm.jr2.ox.ac.uk/>]).

- Die interne Validität wird innerhalb jedes Evidenz-Levels in drei Kategorien (gut, mittel, schlecht) eingeteilt. Um diese Einteilung vornehmen zu können, wurden Parameter für das jeweilige Studiendesign entwickelt (siehe Tabelle 3). Eine gute Studie erfüllt alle Kriterien ein, eine mittlere Studie erfüllt nicht alle Kriterien, hat aber keine wesentlichen Fehler, eine schlechte Studie hat nicht akzeptable Fehler. Eine gut geplante Kohortenstudie mag besser sein als eine inadäquat durchgeführte RCT. Aber auch gut geplante und durchgeführte Studien liefern nicht notwendigerweise die benötigte Evidenz, wenn sie eine hochselektierte Population untersuchen, d. h. nicht auf die Gesamtpopulation übertragen werden können (fehlende Generalisierbarkeit, externe Validität).
- Die Qualität der vorhandenen Evidenz wird für die Forschungsfrage in einen analytischen Rahmen gesetzt. Hier werden weitere Kriterien, wie Kohärenz bzw. Konsistenz, ergänzt. Kohärenz bedeutet, dass die vorhandene Evidenz sich in ein verständliches Modell einfügt. Die Konsistenz unterschiedlicher Studien muss dafür nicht notwendigerweise vorhanden sein, doch unterstützt Konsistenz die Kohärenz.

Eine Reihe von Limitationen des Konzepts der Hierarchie der Evidenz sollten beachtet werden ⁶³:

- Evidenz-Stufen berücksichtigen nicht das Verhältnis von Studiendesign und Fragestellung. In den gängigen Hierarchiemodellen sind verschiedene Studiendesigns eingeordnet, die nur für jeweils bestimmte Fragestellungen geeignet sind. Die Studientypen in den Evidenzskalen beinhalten sowohl prospektive, experimentelle Designs wie auch retrospektive und beobachtende Studientypen. Jedes Design bietet Vorteile und Nachteile für verschiedenen Fragestellungen.

Tabelle 3: Kriterien zur Bewertung der internen Validität individueller Studien (nach Harris et al., 2001²⁹)

Studiendesign	Kriterien
Systematische Reviews	<ul style="list-style-type: none"> • Ausführlichkeit der Quellen, der benutzten Suchstrategien • Standardbeurteilung der eingeschlossenen Studien • Validität der Schlussfolgerungen • Aktualität und Relevanz
Fall-Kontroll-Studien	<ul style="list-style-type: none"> • genaue Ermittlung der Fälle • nicht verzerrte Selektion der Fälle und Kontrollen anhand von für beide geltenden Ausschlusskriterien • Responserate • Diagnostische Test bei beiden Gruppen gleich durchgeführt • angemessene Aufmerksamkeit für potentielle Confounder³
RCTs und Kohortenstudien	<ul style="list-style-type: none"> • Initiale Rekrutierung vergleichbarer Gruppen <ul style="list-style-type: none"> - für RCTs: adäquate Randomisierung, einschließlich Verblindung und Gleichverteilung potentieller Confounder zwischen den Gruppen - für Kohortenstudien: Berücksichtigung potentieller Confounder entweder durch Restriktion oder Erhebung für eine Adjustierung im Rahmen der Analyse, Berücksichtigung der Anfangskohorte • Erhalt vergleichbarer Gruppen (Abrieb, Crossover, Kontamination, Adhärenz) • Unterschiede im Loss to follow-up zwischen den Gruppen, oder insgesamt hoher Loss to follow-up • Erhebung: gleich, reliabel, und valide (einschließlich Verblindung der Outcomeerhebung) • klare Definition der Intervention • Berücksichtigung aller wichtigen Outcomes • Analyse: Adjustierung für potentielle Confounder bei Kohortenstudien, Intention-to-treat-Analyse bei RCTs
diagnostische Genauigkeitsstudien	<ul style="list-style-type: none"> • relevanter Screeningtest, verfügbar für die Primärversorgung, adäquat beschrieben • Studie benutzt einen akzeptierten Referenztest, durchgeführt unabhängig vom Ergebnis des Indextests⁴ • Interpretation des Referenztests unabhängig vom Screeningtest • benutzt intermediäre Ergebnisse vernünftig • Spektrum der in die Studie eingeschlossenen Patienten

³ Als Confounder werden Faktoren bezeichnet, die nicht Ziel einer Untersuchung sind, aber zum einen kausal auf die Zielgröße wirken und zum anderen gleichzeitig und unabhängig mit der Exposition assoziiert sind. Sie führen zu einer systematischen Verzerrung der Ergebnisse.

⁴ Als Indextest wird in der Literatur der Test bezeichnet, dessen Testgüte gegen einen etablierten Referenztest (Goldstandard) überprüft werden soll.

	<ul style="list-style-type: none">• Größe der Studienpopulation• Einsatz eines reliablen Screeningtests
--	--

- Evidenz-Stufen lassen keine Aussage über die Adäquanz von Studien zu. Die Adäquanz (klinische Angemessenheit) einer Studie bezeichnet die Verwendbarkeit einer Studie in einer konkreten klinischen Situation. Eine Studie auf methodisch hohem Niveau kann klinisch dennoch unangemessen sein, etwa wenn für den Patienten völlig irrelevante Outcomes untersucht worden sind. Kriterien der Adäquanz sind u. a. die Übertragbarkeit der Studienbedingungen auf die Bedingungen des klinischen Alltags, Auswahl der Endpunkte, Akzeptanz durch Ärzte und Patienten, technische und finanzielle Umsetzbarkeit ⁶⁴.
- Externe Validität, Konsistenz der Studienergebnisse, klinische Relevanz und Qualität der Effekte werden in der Regel nicht in Evidenzskalen abgebildet. Unter externer Validität versteht man die Frage, ob und wie in Studien gewonnene Ergebnisse auch außerhalb der Studienpopulation Anwendung finden können. Diese kann bei einer bevölkerungsbasierenden Beobachtungsstudie (niedriger Evidenzgrad) deutlich höher sein als bei einem RCT mit höchstem Evidenzgrad⁵. Die Konsistenz der Studienergebnisse, die strenggenommen nur dann beurteilt werden kann, wenn alle verfügbaren Studien zu einer Fragestellung vorliegen, macht den potentiell irreführenden Effekt von Evidenzskalen besonders deutlich. Einerseits hilft es nicht weiter, wenn mehrere methodisch gute Studien zur gleichen Fragestellung vorliegen, aber entgegengesetzte Ergebnisse liefern. Noch problematischer ist der Fall, dass eine Entscheidung aufgrund einer Studie mit einem positiven Ergebnis getroffen wurde, gleichzeitig aber eine Studie auf dem gleichen Evidenz-Level mit gegenteiligem Ergebnis nicht berücksichtigt wurde. Klinisch irrelevante Ergebnisse aus Level-I-Studien sind nicht immer höher zu gewichten als dramatische Ergebnisse aus Studien mit niedrigerem Evidenzgrad. Insbesondere bei widersprüchlichen Studien kann eine Klärung nur über den Weg systematischer Übersichten erfolgen, in denen sämtliche vorhandenen Studien gewürdigt werden. In die Hierarchie der Evidenz geht in den meisten Vorschlägen auch nicht die Qualität der zugrundeliegenden Studien ein. Lediglich die Skalen von Jovell und Navarro-Rubio ³⁷ und die Fassung des Centre for Evidence-based Medicine in Oxford berücksichtigen die Studienqualität.

⁵ Die Ergebnisse nicht randomisierter Studien weichen von den Ergebnissen auch nicht konsistent ab ^{4;10;13;44}. Das heißt, wenn potentiell verzerrende Einflüsse in nicht randomisierten Studien adäquat berücksichtigt werden, dann sind die Ergebnisse vergleichbar mit den Ergebnissen aus RCTs für die gleichen Interventionen.

Eine systematische Übersicht ist oft höher zu bewerten als eine einzelne Studie, unabhängig vom Design. Systematische Übersichten stellen letztlich für jeden Evidenz-Level den Goldstandard dar, was zu einer „horizontalen“ Betrachtungsweise von Evidenzhierarchien führt. Am weitesten ist die Evidenzskala des Center for Evidence-based Medicine in Oxford in dieser Hinsicht entwickelt (siehe Anhang).

4.2 Metaanalysen von Studien zur diagnostischen Genauigkeit

Zu den Prinzipien von Metaanalysen gehören, wie bereits oben erwähnt, die Fokussierung der Fragestellung, die Erstellung eines Reviewprotokolls inklusive Festlegung von Ein- und Ausschlusskriterien für die Primärstudien, eine auf Vollständigkeit zielende Literaturrecherche, die Qualitätsbewertung der eingeschlossenen Studien, die detaillierte Extraktion der Studiendaten in Tabellen, die Durchführung der Metaanalyse, der Test auf Robustheit und Verzerrungsfreiheit der Ergebnisse sowie die Ableitung von Schlussfolgerungen auf der Basis der vorliegenden Evidenz.

Es gibt keinen internationalen Konsens darüber, wie viele Studien mindestens vorhanden sein müssen, um eine Metaanalyse durchführen zu können. Es konnte bisher auch kein einheitliches Vorgehen bei Heterogenität innerhalb und zwischen den Studien festgelegt werden. Allerdings wurden von einer internationalen Arbeitsgruppe Leitlinien für die Evaluation von diagnostischen Testverfahren entwickelt³⁴. Diese empfehlen ein schrittweises Vorgehen, legen sich aber nicht auf ein spezielles Verfahren fest (siehe Tabelle 4). Im folgenden sollen diese Schritte zunächst systematisch erläutert werden.

Ein wesentliches Merkmal von systematischen Übersichten ist die eng begrenzte oder zumindest stark fokussierte Fragestellung bzw. Studienhypothese. Da es sich bei systematischen Übersichten um originäre wissenschaftliche Arbeiten handelt, ist die wohlformulierte Hypothese ein wesentliches Element. Eine sorgfältig ausgearbeitete Fragestellung determiniert bereits zu einem Teil die Einschlusskriterien für die Primärstudien.

Tabelle 4: Schritte bei der Durchführung einer Metaanalyse diagnostischer Testverfahren (nach Irwig et al., 1994³⁴)

1	<p><i>Festlegung von Ziel und Umfang der Metaanalyse</i></p> <p>Gibt es eine eindeutige Feststellung zu folgenden Aspekten: dem zur Diskussion stehenden Test, der zu diagnostizierenden Krankheit und dem Referenztest („Goldstandard“), der klinischen Fragestellung und dem klinischem Setting?</p> <p>Ist es Ziel, einen einzelnen Test zu evaluieren oder mehrere Tests zu vergleichen?</p>
2	<p><i>Identifizierung der relevanten Literatur</i></p> <p>Sind die Details der Literaturrecherche inklusive Such- und Verbindungswörtern angegeben? Sind Ein- und Ausschlusskriterien definiert?</p>
3	<p><i>Datenextraktion und Präsentation der Daten</i></p> <p>Wurden die Studien von zwei oder mehr Personen bewertet? Erklären die Autoren, wie Meinungsverschiedenheiten geklärt wurden? Wurde für jede Primärstudie eine komplette Auflistung der Studiencharakteristika und der Testgüte angegeben?</p>
4	<p><i>Abschätzung der Testgüte</i></p> <p>Berücksichtigt die Methode der Datensynthese von Sensitivität und Spezifität die gegenseitige Abhängigkeit dieser Werte? Falls multiple Testkategorien verfügbar sind, wurden diese in der Datensynthese berücksichtigt?</p>
5	<p><i>Einschätzung der Konsequenzen von Variationen der Studienvolidität bei der Bestimmung der Testgüte</i></p> <p>Wurde die Beziehung zwischen der Bestimmung der Testgüte und der Validität der Studien für jedes der folgenden Kriterien untersucht: Angemessener Referenztest, unabhängige Bewertung von Test und Referenztest, Vermeidung von Verifikationsbias</p> <p>Wurden in Vergleichsstudien alle Tests bei jedem einzelnen Patienten angewendet oder die Patienten zufällig einem Test zugeteilt? Wurden analytische Methoden eingesetzt, um abzuschätzen, inwiefern methodische Mängel von Primärstudien die Testgüte beeinflussen?</p>
6	<p><i>Einschätzung der Konsequenzen von Variationen von Patientencharakteristika und des Tests auf die Bestimmung der Testgüte (Generalisierbarkeit)</i></p> <p>Wurde die Beziehung zwischen der Bestimmung der Testgüte und Patienten- bzw. Testcharakteristika untersucht? Wurden analytische Methoden eingesetzt, um zwischen Einflüssen auf die Testgüte und die Grenzwerte zu unterscheiden?</p>

4.2.1 Festlegung von Ziel und Umfang der Metaanalyse

1	<p><i>Festlegung von Ziel und Umfang der Metaanalyse</i></p> <p>Gibt es eine eindeutige Feststellung zu folgenden Aspekten:</p> <ul style="list-style-type: none"> dem zur Diskussion stehenden Test, der zu diagnostizierenden Krankheit und dem Referenztest („Goldstandard“), der klinischen Fragestellung und dem klinischem Setting? <p>Ist es Ziel, einen einzelnen Test zu evaluieren oder mehrere Tests zu vergleichen?</p>
---	--

Eine Fokussierung der Fragestellung ist für folgende Punkte notwendig ³⁴:

- Der diagnostische Tests, der evaluiert wird, muss klar beschrieben sein, ebenso wie der Referenztest (Goldstandard), der zur Bestimmung der diagnostischen Genauigkeit herangezogen wird.
- Die Krankheit, zu deren Erkennung der Test eingesetzt wird, muss definiert sein.
- Der klinische Kontext, in dem die diagnostische Genauigkeit evaluiert wird, muss beschrieben werden, da die Testgenauigkeit mit der Population variieren kann, in der der Test eingesetzt wird (Studienpopulation, Prävalenz der Zielkrankheit, Verteilung der Krankheitsstadien). Das klinische Setting kann als Primär-, Sekundär- oder Tertiärversorgung beschrieben werden oder hinsichtlich der Studienpopulation als Hochrisikokollektiv oder Kollektiv mit nur geringem Risiko in bezug auf die Zielkrankheit ⁵⁸.

Auch wenn mehrere Tests gleichzeitig verglichen werden, gelten die oben beschriebenen Bedingungen und der Kontext, in dem die Tests durchgeführt wurden, muss für alle Tests vergleichbar sein ³⁴.

4.2.2 Identifizierung der relevanten Literatur

2	<p><i>Identifizierung der relevanten Literatur</i></p> <p>Sind die Details der Literaturrecherche inklusive Such- und Verbindungswörtern angegeben?</p> <p>Sind Ein- und Ausschlusskriterien definiert?</p>
---	---

Eine systematische Übersichtsarbeit oder eine Metaanalyse sollte alle verfügbare Evidenz beinhalten. Um diese Evidenz zusammenfassen zu können, müssen zunächst die relevanten Veröffentlichungen identifiziert werden. Viele Autoren betonen die Bedeutung einer akkuraten Suchstrategie, die sowohl eine hohe Sensitivität als auch eine hohe Spezifität beinhaltet ^{35;48}. Wenn jedoch Publikationen für eine systematische Übersichtsarbeit oder eine Metaanalyse gesammelt werden, müssen auch Suchstrategien mit einer niedrigen Spezifität akzeptiert werden ¹⁷.

Die Suche von Primärstudien, die die Einschlusskriterien für einen systematischen Review erfüllen, wurde vielfach beschrieben und Standardprozeduren für die wichtigsten Datenbankanbieter wurden entwickelt. Eine elementare Erkenntnis aus zahlreichen Untersuchungen ist, dass eine Recherche in MEDLINE (der sicher wichtigsten biomedizinischen Datenbank) alleine nicht ausreichend ist, um alle potentiell relevanten Studien zu identifizieren. Die folgenden Quellen sollten recherchiert werden, um eine verzerrte Aussage durch das Übersehen wichtiger Datenquellen zu vermeiden (z.B. English Language Bias⁶): biomedizinische Datenbanken (MEDLINE, EMBASE, SciSearch u.a.m.), Literaturverzeichnisse bereits identifizierter Studien, Handsuche von Kongressbänden, Bücher, nicht in Datenbanken gelistete Zeitschriften und Referenzlisten, Studienregister, Kontakte zu Herstellerfirmen und zu Forschern.

Beispiel: Suche nach RCTs, die Ballonangioplastie mit Stenting bei koronarer Herzkrankheit vergleichen (nach Perleth 1998⁵⁹):

Im März 1998 wurde die erste Literaturrecherche zu der Fragestellung begonnen. Dabei wurden zunächst MEDLINE, EMBASE, HealthStar und die Cochrane Library nach RCTs und systematischen Reviews bzw. Metaanalysen durchsucht (Zeitraum jeweils seit 1990). Anschließend wurden die Kongressbände der Jahre 1995 bis 1997/8 von Hand nach relevanten RCTs recherchiert, Projektlisten von HTA-Einrichtungen sowie Referenzlisten der gefundenen Studien überprüft. Dadurch wurden insgesamt 31 RCTs identifiziert.

Die folgende Übersicht gibt einen Überblick über die Verteilung der Treffer:

	RCTs (zusätzlich zu MEDLINE)
MEDLINE	6
EMBASE	0
HealthStar	0
Cochrane Library	2
Kongresse	20
Projektlisten	0
sonstige Quellen (u. a. Expertenkontakte und Informationsdienste im Internet)	3

Der bei weitem aufwendigste Suchschritt war die Handsuche in den Kongressbänden, die mehrere Tage in Anspruch nahm. Alle 20 Erstautoren von Abstracts zusätzlich identifizierter Studien wurden schriftlich kontaktiert und um weitere Informationen gebeten. Hierdurch konnten zusätzliche Informationen zu 10 dieser Studien, z. B. neu erschienene oder im Druck befindliche Publikationen, unpublizierte Zwischenberichte oder Manuskripte, erhalten werden. Die Recherchephase erstreckte sich bis zum aktuellen Stand über ein halbes Jahr. Am ergiebigsten war dabei die Suche in den Kongressbänden. Hierdurch wurden laufende Studien identifiziert, deren Publikation für zukünftige Updates berücksichtigt werden können.

⁶ Zu den unterschiedlichen Bias-Möglichkeiten in systematischen Übersichten siehe Davey Smith et al, 1997¹⁴. Zumindest für Publikationen aus dem deutschsprachigen internistischen Bereich läßt sich zeigen, daß Studien mit signifikanten positiven Effekten mit höherer Wahrscheinlichkeit in englischsprachigen Zeitschriften publiziert werden, während nicht signifikante Ergebnisse eher in deutschsprachigen Zeitschriften zur Veröffentlichung gelangen²⁴, siehe auch Kapitel 4.5.

Suchstrategien zur Evaluation diagnostischer Tests sind nach Irwig et al.³⁴ weniger untersucht als Suchstrategien für Literatur zu klinischen Studien.

Literaturrecherchen in Datenbanken dürfen sich nicht alleine auf Schlagwörter beschränken, da nicht adäquat verschlagwortete Studien übersehen werden können. Die Freitextsuche kann sehr schwierig sein, wenn für einen bestimmten Begriff zahlreiche Synonyme und Homonyme existieren. Eine neuere Untersuchung aus der Nuklearmedizin zeigte, dass alleine für den Tracer Fluorodeoxyglucose (FDG) 56 verschiedene Schreibweisen in den Datenbanken MEDLINE, EMBASE und Current Contents verwendet werden⁵².

Berichte über diagnostische Forschung, ältere Veröffentlichungen im besonderen, sind in elektronischen Datenbanken häufig schlecht verschlagwortet. Es ist daher hilfreich, Pilotsuchen durchzuführen, in dem man subjektspezifische Strategien einsetzt. Dieser Prozess kann iterativ durchgeführt werden, nachdem man zusätzliche Schlüsselwörter und Freitextwörter identifiziert und eingeschlossen hat. Studien, die nur in Referenzlisten gefunden und mit der Suchstrategie übersehen wurden, sollten in der Datenbank anhand des Titels und des Erstautors gesucht werden. Falls die Studie gefunden wird, können ihre Schlüsselwörter der Suchstrategie hinzugefügt werden. Das Verfolgen von Zitaten kann weitere Studien identifizieren. Der Science Citation Index kann „Forward in time“ durchsucht werden, um Artikel zu identifizieren, die relevante Publikationen zitieren¹⁹. Nützliche Indexwörter oder Textwörter, die hilfreich sind Studien zur diagnostischen Genauigkeit zu finden, wurden von Deeks¹⁵ zusammengestellt und finden sich im Anlage 3.

Es ist auf jeden Fall ratsam, professionelle Hilfe bei der Literaturrecherche in Anspruch zu nehmen. Die Ergebnisse der Literatursuche müssen mit ausreichenden Details dokumentiert werden.

Nach Abschluss der Literatursuche müssen die Überschriften und Abstrakts der identifizierten Literatur anhand festgelegter Ein- und Ausschlusskriterien hinsichtlich ihrer Relevanz bewertet werden. Die Einschlusskriterien können an einer Stichprobe der Veröffentlichungen vorgetestet werden¹⁹. Die Auflistung der Gründe, die zum Ausschluss bestimmter Studien geführt haben, hilft dem Leser der Metaanalyse zu verstehen, wie die Kriterien eingesetzt wurden.

Als mögliche Kriterien hinsichtlich des Ein- bzw. Ausschlusses von identifizierter Literatur sind zu nennen:

- **Referenztest:** Die Genauigkeit eines diagnostischen oder Screeningtests sollte evaluiert werden, in dem seine Ergebnisse mit einem Goldstandard, dem Referenztest, verglichen werden, der als der beste verfügbare Test anerkannt ist. Der Referenztest kann ein einzelner Test, eine Kombination verschiedener Tests oder der klinische Follow-up der Patienten sein. Die Veröffentlichung sollte den Referenztest beschreiben, da er ein *conditio sine qua non* für die Evaluation des diagnostischen Tests darstellt ¹⁹. Eine fehlende Beschreibung des Referenztest kann ein mögliches Ausschlusskriterium darstellen.
- **Studienteilnehmer:** Detaillierte Informationen über die Studienpopulation der Primärstudie fehlen häufig. Teilnehmer sollten ausdrücklich u. a. hinsichtlich folgender Merkmale definiert sein: Alter, Geschlecht, Beschwerden, Symptome und Krankheitsdauer. Eine Beschreibung der Teilnehmer mit und ohne Krankheit (definiert mit dem Goldstandard) sollte gegeben werden. Neben den demographischen Daten muss auch die Selektion der Studienteilnehmer beschrieben werden, um eine potenzielle Verzerrung der Ergebnisse durch Spektrum Bias (s. u.) abschätzen zu können. Ein Fehlen dieser Angaben muss als mögliches Ausschlusskriterium in Betracht gezogen werden. Die statistische Power der Studie bzw. die notwendige Teilnehmerzahl ist abhängig vom Studiendesign, den Schätzern der diagnostischen Genauigkeit und der Präzision, mit der diese Parameter bestimmt werden ¹⁹.
- **Outcomedaten:** In jeder Primärstudie sollte soviel Information zur Verfügung stehen, dass eine Vierfeldertafel mit den Häufigkeiten von richtig positiven, falsch negativen, falsch positiven und richtig negativen Testergebnissen erstellt werden kann.
- **Setting:** Das Spektrum der Patienten unterscheidet sich hinsichtlich der unterschiedlichen Ebenen der Krankenversorgung, daher können diagnostische Reviews sich auf ein spezifisches Setting konzentrieren oder alle Ebenen einschließen. Diese Informationen sind wichtig, wenn Subgruppenanalysen bei Heterogenität durchgeführt werden (s. u.).

Jede verfügbare Evidenz sollte unabhängig von der Publikationssprache verwendet werden. Es ist schwieriger nicht englische Veröffentlichung zu identifizieren, da sie häufig in den elektronischen Datenbanken nicht aufgenommen sind.

4.2.3 Datenextraktion und Präsentation der Daten

3	<p><i>Datenextraktion und Präsentation der Daten</i></p> <p>Wurden die Studien von zwei oder mehr Personen bewertet?</p> <p>Erklären die Autoren, wie Meinungsverschiedenheiten geklärt wurden?</p>
---	---

Wurde für jede Primärstudie eine komplette Auflistung der Studiencharakteristika und der Testgüte angegeben?
--

Die Relevanz der Studie zu beurteilen und die Daten zu extrahieren, erfordert die Bewertung der Studie. Hierfür wurden unterschiedliche Prozeduren vorgeschlagen, um die Gefahr von verzerrten Ergebnissen zu minimieren:

- Unabhängige Bewertung durch mehrere Gutachter und Klärung von Unstimmigkeiten durch einen weiteren Gutachter bzw. Diskussion der beiden und
- Verblindung der Gutachter hinsichtlich Autorenschaft und Studienergebnisse ³⁴.

Unabhängig von der Wahl der Methode muss dargestellt werden, wie die Bewertung durchgeführt wurde. Diese kann an einer Stichprobe der identifizierten Studien getestet werden. Bei unterschiedlicher Bewertung von Artikeln und bei ungenügenden Informationen kann ein weiterer Gutachter eingeschaltet oder die ganze Veröffentlichung konsultiert werden ¹⁹.

Die für Zufall adjustierte Übereinstimmung (Kappa-Statistik) der Gutachter sollte dargestellt werden.

Die Maßzahlen für die diagnostische Genauigkeit und die Studiencharakteristika sollten für jede Primärstudie aufgelistet werden, um dem Leser die Möglichkeit zu geben, die Bewertung zu beurteilen, Reanalysen mit anderen statistischen Verfahren durchzuführen oder Studien hinzuzufügen, um die Metaanalyse zu aktualisieren ³⁴.

Die diagnostische Genauigkeit kann auf unterschiedliche Weise dargestellt werden. Für die Metaanalyse von dichotomen Tests ist es notwendig Vierfeldertafeln zu konstruieren und die absolute Zahl in den vier Feldern ist notwendig. Die Anzahl der erkrankten und nicht erkrankten Teilnehmer ist notwendig, um die Prätest-Wahrscheinlichkeit des Tests zu berechnen und um die Vierfeldertafel anhand der Angaben zu Sensitivität, Spezifität, Likelihood Ratios⁷, prädiktiver Werte oder ROC-Kurven zu rekonstruieren. Wenn möglich sollten Vierfeldertafeln für alle relevanten Subgruppen gebildet werden können¹⁹.

Folgende Studiencharakteristika sollten daher aufgeführt sein: Studiendesign (prospektiv oder retrospektiv), Anzahl der Patienten, Alter, Selektionskriterien, Art der Erkrankung und deren Prävalenz, das Vorhandensein von Symptomen, Grenzwerte, Art des Referenztests. Diese Studiencharakteristika ermöglichen es auch die Validität der Studien zu beurteilen und erklä-

⁷ Die Likelihood Ratio ist eine Methode, die diagnostische Genauigkeit eines Tests auszudrücken. Es ist die Ratio zwischen der Wahrscheinlichkeit, ein bestimmtes Testergebnis bei erkrankten Personen zu finden, und der Wahrscheinlichkeit, dieses bei nicht erkrankten Personen zu erhalten.

ren z. T. die Variabilität der Ergebnisse. Diese Angaben können auch als Kovariablen zur Analyse der Heterogenität in Metaanalysen dienen.

Weitere Informationen, die extrahiert werden sollten sind: Publikationsjahr, Sprache, Land oder Region, in der die Studie durchgeführt wurde.

4.2.4 Abschätzung der Testgüte

4	<p><i>Abschätzung der Testgüte</i> Berücksichtigt die Methode der Datensynthese von Sensitivität und Spezifität die gegenseitige Abhängigkeit dieser Werte? Falls multiple Testkategorien verfügbar sind, wurden diese in der Datensynthese berücksichtigt?</p>
---	---

Idealerweise diskriminiert ein diagnostischer Test zwischen erkrankten und nicht erkrankten Personen fehlerfrei. Testfehler können auf unterschiedliche Weise beschrieben werden: Hierbei werden die klassischen Testgütekriterien wie Sensitivität und Spezifität ermittelt. Diese beiden Kriterien basieren auf einem Schwellenwert, ein Testergebnis als positiv zu werten. Eine Änderung des Schwellenwerts bedingt eine Änderung von Sensitivität und Spezifität, so dass beide nicht unabhängig von einander betrachtet werden können.

Eine andere Möglichkeit besteht darin, nicht nur eine Sensitivität/Spezifität zu dokumentieren, sondern in Abhängigkeit vom Schwellenwert unterschiedliche Paare von Sensitivität / Spezifität darzustellen. Dies geschieht häufig durch sogenannte *receiver operating characteristics* (ROC-) Kurven, bei denen die Sensitivität gegen $(1 - \text{Spezifität})$ aufgetragen wird.

4.2.5 Einschätzung der Konsequenzen von Variationen der Studienvalidität bei der Bestimmung der Testgüte/ methodische Qualität

5	<p><i>Einschätzung der Konsequenzen von Variationen der Studienvalidität bei der Bestimmung der Testgüte</i></p> <p>Wurde die Beziehung zwischen der Bestimmung der Testgüte und der Validität der Studien für jedes der folgenden Kriterien untersucht:</p> <ul style="list-style-type: none"> angemessener Referenztest, unabhängige Bewertung von Test und Referenztest, Vermeidung von Verifikationsbias? <p>Wurden in Vergleichsstudien alle Tests bei jedem einzelnen Patienten angewendet oder die Patienten zufällig einem Test zugeteilt?</p> <p>Wurden analytische Methoden eingesetzt, um abzuschätzen, inwiefern methodische Mängel von Primärstudien die Testgüte beeinflussen?</p>
---	---

Die Änderung der Wahrscheinlichkeit für das Vorliegen einer Krankheit nach dem Test (ausgedrückt durch die Likelihood Ratio) wird auch als diagnostischer Impact bezeichnet. Die Likelihood Ratio ermöglicht auch die Berücksichtigung von multiplen Testkategorien. Dies vermeidet einen Informationsverlust durch die Dichotomisierung von Ergebnissen. Diese Parameter werden meist im Rahmen konsekutiver / selektiver Fallserien oder prospektiver Studien erhoben.

Validitätskriterien für diagnostische Forschung sind von der Cochrane Methods Group on Screening and Diagnostic Tests (www.cochrane.org/cochrane/sadt.htm) und anderen veröffentlicht worden. Kriterien, die die interne (und auch externe) Validität überprüfen, sollten kodiert und ausdrücklich im Review beschrieben werden. Kriterien für die interne Validität beziehen sich auf Studiencharakteristika, die systematische Fehler oder Bias verhindern. Kriterien der externen Validität geben Hinweise auf die Generalisierbarkeit (siehe Kapitel 4.6) der Studie und überprüfen, ob der Test gemäß anerkannter Standards durchgeführt wurde.

Kriterien der internen Validität sind: valider Referenzstandard, Definition des Grenzwertes für den Referenzstandard, verblindete Messung des Index- und Referenztests, Vermeidung von Verifikationsbias, Indextest wird unabhängig von der klinischen Information bewertet, Studiendesign¹⁹.

Die Studienvalidität kann durch eine Reihe von systematischen Fehlern beeinträchtigt werden, denen in den Primärstudien Rechnung getragen werden muss:

Spektrum Bias: Hiermit wird der Bias aufgrund des Einflusses des Patientenspektrums und der Krankheitsschwere (case mix) bezeichnet. Dieser Bias kann sich auf die Generalisierbarkeit

keit der Ergebnisse auswirken. Zur Minimierung dieses Bias können Überweisungen von Patienten zum Indextest aus unterschiedlichen Quellen vorgenommen werden. Sensitivität und Spezifität sind in sofern nicht von der Prävalenz der Erkrankung unabhängig, da die Sensitivität eines Test mit der Prävalenz der Erkrankung ansteigt. Unterschiede in der Sensitivität können also auf ein unterschiedliches Patientenspektrum (mit unterschiedlicher Prävalenz) zurückzuführen sein ⁷.

Unter Verifikationsbias (auch Workup-Bias genannt) versteht man eine Verzerrung der Ergebnisse dadurch, dass gesunde Patienten in verschiedenen Studien mit unterschiedlich hoher Wahrscheinlichkeit einem (invasiven) Referenztest unterzogen werden. Wenn Patienten mit einem negativen Testergebnis nicht mit dem Referenztest untersucht werden, können bei unterschiedlicher Prävalenz (und damit einer unterschiedlichen Relation Falsch-Negativer zu allen Test-Negativen) falsch hohe Sensitivitäten resultieren. Diese Möglichkeit ist insbesondere dann gegeben, wenn die Auswahl der Patienten, deren Krankheitsstatus mit dem Referenztest verifiziert wurde, nicht randomisiert erfolgt. Man unterscheidet den partiellen Verifikationsbias, wenn nicht alle Teilnehmer den Referenztest erhalten, und den differentiellen Verifikationsbias, wenn unterschiedliche Referenztests in Abhängigkeit des Ergebnisses der Indextests eingesetzt werden. Partieller Verifikationsbias führt zu einer niedrigeren Rate von richtig negativen und falsch negativen Ergebnissen und verzerrt damit die Ergebnisse in Richtung einer höheren Sensitivität und niedrigen Spezifität. Differentieller Verifikationsbias dagegen verzerrt die Ergebnisse in Richtung einer höheren Sensitivität und Spezifität. Dieser Bias kann u. a. vermieden werden, wenn alle Patienten mit dem Referenztest untersucht werden.

Diagnostic Review Bias: Dieser Bias kommt dadurch zustande, dass die Enddiagnose bzw. das Ergebnis des Goldstandards durch das Ergebnis des Indextests beeinflusst wird. Er kann durch eine verblindete Testauswertung vermieden werden.

Inkorporationsbias: Die Durchführung des Referenztests (Goldstandard) wird durch das Ergebnis des Indextests beeinflusst. Dies ist vor allem bei histopathologischer Probenentnahme von Bedeutung.

Studien mit unterschiedlichen Referenztests sollten getrennt analysiert werden.

Eine Möglichkeit mit der Variation der Studienvalidität im Rahmen von Metaanalysen umzugehen, ist, die Studien auszuschließen, die die Kriterien der wissenschaftlichen Validität nicht erfüllen. Diese Methode vermeidet systematische Verzerrungen führt aber zu breiteren Konfi-

denzintervallen. Es können auch getrennte Analysen für Studien mit und ohne Schwächen durchgeführt werden (Sensitivitätsanalyse).

Weitere Qualitätsaspekte ¹⁵: Beide Tests sollten vor dem Behandlungsbeginn durchgeführt werden, sonst kann es zum sogenannten Behandlungsparadox kommen, wenn nämlich Patienten, bei denen die Diagnose einer Krankheit gestellt wird, behandelt und geheilt werden, und dann erst den zweiten Test erhalten, werden sie in Abhängigkeit der zeitlichen Abfolge der Tests fälschlicherweise als falsch positiv oder negativ eingeordnet.

Eine weiterer Aspekt ist, wie mit nicht interpretierbaren Ergebnissen umgegangen wird, ob diese aus der Analyse ausgeschlossen oder einer gesonderten Kategorie zugeordnet wurden ⁶.

Eine Metaanalyse kann die vorhandene Literatur akkumulieren, deren Qualität bewerten, Summaryschätzer bilden und untersuchen, inwieweit diese Schätzer der Testgenauigkeit von der Qualität der Studien beeinflusst werden. Um weitere Forschung anzuleiten, ist es wichtig, dass Metaanalysen die Qualität der Primärstudien begutachten und im Detail darstellen.

4.2.6 Einschätzung der Konsequenzen von Variationen von Patientencharakteristika und des Tests auf die Bestimmung der Testgüte (Generalisierbarkeit)

6	<p><i>Einschätzung der Konsequenzen von Variationen von Patientencharakteristika und des Tests auf die Bestimmung der Testgüte (Generalisierbarkeit)</i></p> <p>Wurde die Beziehung zwischen der Bestimmung der Testgüte und Patienten- bzw. Testcharakteristika untersucht?</p> <p>Wurden analytische Methoden eingesetzt, um zwischen Einflüssen auf die Testgüte und die Grenzwerte zu unterscheiden?</p>
---	--

Valide Schätzer der diagnostischen Genauigkeit müssen nicht notwendigerweise übertragbar auf andere klinische Settings sein. Die Übertragbarkeit muss anhand von bestimmten Charakteristika überprüft werden: Übereinstimmung der Patientenmerkmale, Merkmale, die nicht mit der diagnostischen Genauigkeit assoziiert sind.

Kriterien der externen Validität sind u. a.: Krankheitsspektrum, Setting, vorangegangene Test, Überweisungsfilter, Dauer der Krankheit vor Diagnose, Komorbidität, demographische Informationen, Durchführung des Indextests, Anteil fehlender Werte, Reproduzierbarkeit des Indextests ¹⁹.

Für die Bewertung der Qualität diagnostischer Studien eignet sich der Ansatz von Fryback und Thornbury (1991) und Kent und Larson (1992) ³⁹, der explizit die methodische Qualität diagnostischer Studien in Verbindung mit ihrer Generalisierbarkeit adressiert. Die Qualität

der eingeschlossenen Studien (zur diagnostischen Genauigkeit) kann entsprechend einem international gebräuchlichen Schema bewertet werden²⁵. Die Studien werden danach in 4 Qualitätsstufen klassifiziert, die wie folgt definiert sind:

- A Studien, die auf ein breites Spektrum von Patienten angewandt werden können und die keine gravierenden methodischen Fehler enthalten:
- prospektives Design
 - ≥ 35 Patienten, jeweils mit und ohne Krankheit
 - Patienten stammen aus einer klinisch relevanten Grundgesamtheit, deren klinischen Symptome komplett beschrieben werden
 - Diagnose durch angemessenen Referenz- oder Goldstandard gesichert
 - technisch hohe Qualität und vom Referenzstandard unabhängige Auswertung der Aufnahmen
- B Nur eingeschränkt generalisierbare Studien, die zwar methodische Mängel aufweisen können; diese sind jedoch beschrieben und können hinsichtlich ihrer Belastung auf die Schlussfolgerungen abgeschätzt werden:
- prospektives Design
 - ≥ 35 Patienten jeweils mit und ohne Krankheit
 - eingeschränktes Patientenspektrum, z. B. Universitätskliniken
 - keine weiteren methodischen Fehler, die eine Interaktionen zwischen dem Testergebnis und der Diagnosestellung fördern
- C Studien mit gravierenden methodischen Mängeln
- eine geringe Zahl von Teilnehmern
 - mangelhafte Berichtsqualität
 - retrospektives Design
- D Studien mit unzureichender methodischer Qualität
- fehlender adäquater Referenzstandard
 - Testergebnis und Stellung der endgültigen Diagnose nicht unabhängig
 - Patientenherkunft nicht beschrieben, oder war offensichtlich von den Testergebnissen beeinflusst (Workup-Bias)
 - nicht durch Daten belegte Aussagen

Der Hauptunterschied zwischen A und B ist das bei B eingeschränkte und - wenn man so will - verzerrte Sample.

Übergreifende Qualitätskriterien sind dabei die Reportqualität, die eine Replikation der Studie erlauben, die Angabe von Konfidenzintervallen um TPR- und FPR-Werte, Anzahl und Spektrum der untersuchten Patienten sowie Anwendung von Methoden zur Reduktion von Bias (z. B. Auswertung von Testergebnissen ohne Kenntnis des Krankheitsstatus der Patienten).

4.2.7 Verdeutlichung anhand ausgewählter Beispiele

Beispiel 1: Einhaltung definierter Qualitätskriterien in publizierten Metaanalysen und ihr Einfluss auf die diagnostische Genauigkeit⁴⁶

Lijmer et al. formulierten in einer Untersuchung folgende Qualitätskriterien: Patientenspektrum (klinische Population/Fall-Kontroll), Verifikation (komplett / differentieller Verifikationsbias / partieller Verifikationsbias), Interpretation der Ergebnisse (verblindet / nicht verblindet), Patientenselektion (konsekutiv / nicht konsekutiv), Datensammlung (prospektiv / unbekannt / retrospektiv), Testdetails (wann negativ / wann positiv – ausreichend / nicht ausreichend beschrieben), Referenztestdetails (wann negativ / wann positiv – ausreichend / nicht ausreichend beschrieben), Beschreibung der Patientenpopulation (Alter / Geschlecht / Symptome – ausreichend / nicht ausreichend beschrieben). Nur 6,8% der 218 von ihnen untersuchten Evaluationen erfüllten alle 8 Kriterien, 30% erfüllten 6 oder mehr. Die diagnostische Genauigkeit eines Tests wurde dabei unter folgenden Bedingungen überschätzt:

- bei Studien mit Fall-Kontroll-Design im Vergleich zu Kohortenstudien (DOR = 3, 95%-Konfidenzintervall: 2,0 – 4,5; Spektrum Bias),
- wenn unterschiedliche Referenztests für positive bzw. negative Ergebnisse des zu evaluierenden Tests eingesetzt wurden (DOR = 2,2, 95%-Konfidenzintervall: 1,5 – 3,3),
- wenn das Ergebnis des Indextests bei der Durchführung des Referenztest bekannt war (DOR = 1,3, 95%-Konfidenzintervall: 1,0 – 1,9),
- wenn die diagnostischen Kriterien des Indextests nicht beschrieben wurden (DOR = 1,7, 95%-Konfidenzintervall: 1,1 – 2,5),
- wenn die Teilnehmer nicht ausreichend beschrieben wurden (DOR = 1,4, 95%-Konfidenzintervall: 1,1 – 1,7).

Die beiden letzten Faktoren stehen in keinem direkten Zusammenhang mit dem Studiendesign, die Erklärung für den beobachtenden Zusammenhang könnte darin liegen, dass sie als Prädiktoren für methodische Schwächen der Studien gewertet werden können.

Beispiel 2: Einhaltung definierter Qualitätskriterien in publizierten Metaanalysen⁶⁵

Reid et al. haben Standardkriterien zur Qualitätsbeurteilung definiert und an 112 publizierten Metaanalysen überprüft. Ihre Kriterien waren:

1. Spektrumkomposition: Beschreibung von Alters- und Geschlechtsverteilung, Zusammenfassung der klinischen Symptome oder des Krankheitsstadiums, Beschreibung der Ein- und Ausschlusskriterien der Patienten. Sensitivität, Spezifität, Likelihood Ratio sind davon abhängig.
2. Subgruppenanalysen: Diagnostische Genauigkeit kann sich in Subgruppen unterscheiden. Analysen für relevante Subgruppen.
3. Vermeidung von Workup-Bias (Verifikationsbias): Vermeidung auch durch Follow-up, wenn Goldstandard bei testnegativen Personen nicht durchgeführt werden soll (nicht bei Krankheiten mit langer Latenz). In Kohorten: alle Teilnehmer erhalten beide Tests (entweder Tests oder Follow-up); in Fall-Kontrollstudien: Indextest vor Goldstandard, Goldstandard wird durchgeführt unabhängig vom Testergebnis; Indextest nach Goldstandard: Testergebnisse wurden nach klinischen Faktoren stratifiziert, die zur Durchführung des Goldstandards führten.
4. Vermeidung von Review Bias: Kann entstehen, wenn entweder der Indextest oder der Goldstandard bewertet werden, ohne dass Vorsichtsmaßnahmen getroffen werden, Objektivität in ihrer sequentiellen Interpretation zu schaffen. Wissen um Indextestergebnis führt zu größerer Übereinstimmung, wenn der Goldstandard danach interpretiert wird. Führt auf jeden Fall zu falsch hoher diagnostischer Genauigkeit. Unabhängige Begutachtung.

5. Präzision der Ergebnisse der diagnostischen Genauigkeit: Sensitivität, Spezifität oder Likelihood Ratio können numerisch instabil sein, wenn zu wenige Patienten untersucht wurden. Die quantitative Instabilität spiegelt sich in der Breite der Konfidenzintervalle wieder (je größer die Stichprobe, um so enger die Konfidenzintervalle). Konfidenzintervalle werden angegeben.

6. Präsentation von unklaren Ergebnissen: Bei der Durchführung von diagnostischen Tests kommt es nicht immer zu eindeutigen Ergebnissen. Manche Ergebnisse sind nicht eindeutig oder nicht interpretierbar und brauchen weitere Abklärung. Die Häufigkeit solcher Ergebnisse ist bei neuen Tests wichtig, da solche Test eine niedrige klinische Effektivität aufweisen, wenn Testergebnisse nicht interpretiert werden können. Der Umgang mit derartigen Ergebnissen kann zu einer Verzerrung der Parameter der diagnostischen Genauigkeit führen, wenn diese Ergebnisse entweder ausgeschlossen werden oder als positiv oder negativ gewertet werden. Aufführung der Häufigkeit intermediärer Ergebnisse und der Umgang mit ihnen.

7. Testreproduzierbarkeit: Aufgrund von Variationen in den Laborprozeduren oder der Gutachter kann ein Test nicht immer zu den gleichen Ergebnissen kommen, wenn er wiederholt wird. Manche Testergebnisse sind nicht reproduzierbar, vor allem wenn der Test die Begutachtung durch Personen benötigt. Reproduzierbarkeit wird überprüft und dargestellt in % Übereinstimmung oder Kappa-Statistik.

Nur 27% der untersuchten Studien erfüllten das 1. Kriterium, nur 8% Kriterium 2, 46% erfüllten Kriterium 3, keine der retrospektiven Kohortenstudien oder Fall-Kontroll-Studien, dafür aber 72% der prospektiven Kohortenstudien. Review Bias wurde von 38 % der Studien vermieden. Kriterium 5 wurde von 9 % der Studien erfüllt, 23% erfüllten Kriterium 6 und Kriterium 7. Es fand sich eine ansteigende Qualität über die Zeit. Aber nur 2 jüngere Veröffentlichungen erfüllten 6 Kriterien oder mehr.

Beispiel 3: Implementierung der Leitlinien von Irwig et al. 1994⁸¹

Walter et al. fanden, in einer Untersuchung zu Metaanalysen (1996 – 1997) von Screeningtests, dass nur 23% der gefundenen Metaanalysen Indikatoren aufführten, die auf der Arbeit von Irwig et al.³⁴ für die methodische Qualität oder Studienvolidität in die Primärstudien basierten. In nur 13% der Metaanalysen wurde Auskunft darüber gegeben, welche Methoden zur Beurteilung der Qualität eingesetzt wurden (Standardisierung, wie viele Gutachter, u.s.w.).

Diese Beispiele machen deutlich, dass hinsichtlich der Qualität der veröffentlichten Primärstudien diagnostischer Tests ein deutliches Verbesserungspotential besteht. Ähnlich wie für RCTs (CONSORT-Satement¹), werden zur Zeit für Studien zur diagnostischen Genauigkeit „Standards for Reporting of Diagnostic Accuracy“ (STARD Statement) erarbeitet⁵⁰.

4.2.8 Zusammenfassung

Interne und externe Validität greifen häufig ineinander und beschreiben beide Aspekte zur Beurteilung der Qualität der identifizierten Primärstudien. Interne und externe Validität, die Beschreibung der Teilnehmer, des diagnostischen Test, der Zielkrankheit, der eingesetzten Methoden können in Metaanalysen benutzt werden, den Evidence-Level zu überprüfen. Weiterhin können diese Angaben für Subgruppen- und Sensitivitätsanalysen eingesetzt werden.

Eine strikte Anwendung von Qualitätskriterien einerseits hätte zur Folge, dass nur ein geringer Anteil der verfügbaren Daten genutzt werden kann. Andererseits würde der Einschluss

von nicht perfekten Studien eine Gewichtung der Evidenz hinsichtlich der relativen Bedeutung der Kriterien notwendig machen, die von den einzelnen Studien nicht erfüllt werden.

Der Umgang mit der anhand dieser Kriterien ermittelten Qualität der Primärstudien im Rahmen von Metaanalysen wird nicht einheitlich diskutiert.

Deville et al.²⁰ schlagen vor, mittels der Kriterien zur internen und externen Validität einen Validitätsscore zu bilden, der als zusätzliches Ein- bzw. Ausschlusskriterium dienen kann. Die Entscheidung, Ausreißerstudien auszuschließen, kann auch auf der Grundlage dieses Scores getroffen werden. Ebenso kann der Score zur Stratifikation oder für Sensitivitätsanalysen herangezogen werden. Interne und externe Validität kann getrennt untersucht werden, da sie unterschiedliche Dimensionen abbilden, oder in einem gemeinsamen Score zusammengefasst werden. Die Bedeutung der internen und externen Validität wird deutlich durch die Tatsache, dass bei Deville et al. vor allem die Ausreißer im Rahmen einer Metaanalyse, d. h. die Primärstudien mit extrem abweichenden Ergebnissen, die Studien mit der schlechtesten Qualität waren²⁰.

Irwig et al.³³ sind der Ansicht, dass Studien von schlechter Qualität nicht von vornherein ausgeschlossen werden müssen, sondern dass der Einfluss der dargestellten Schwächen im Studiendesign auf die diagnostische Genauigkeit untersucht werden sollte. Metaanalysen von Subgruppen, stratifiziert nach der Qualität der Primärstudien, oder die Berücksichtigung einzelner Komponenten in den Analysemodellen ermöglichen es, diesen Einfluss abzuschätzen. Da unterschiedliche Schwächen in der Qualität zu einer unterschiedlich gerichteten Verzerrung der Metaanalyseergebnisse führen, betrachten es Irwig et al.³³ als sinnvoller, den Effekt einzelner Schwächen getrennt zu untersuchen, als sie in einem Qualitätsscore zusammenzufassen.

Vamvakas⁷⁸ schlägt 3 unterschiedliche Wege vor, mit der unterschiedlichen Studienqualität umzugehen: Ausschluss von der Analyse, Stratifizierung der Analyse anhand eines Qualitätsscore und getrenntes Aufführen der Ergebnisse anhand der Strata, Inkorporieren des Qualitätsscores jeder Primärstudie in die Gewichtung der Studie.

Insgesamt besteht ein Konsensus, dass, wenn man Studien mit schlechter Qualität in die Analyse aufnimmt, die Qualität in der Analyse berücksichtigt werden muss. Wenn es genügend Studien mit guter Qualität gibt, können solche mit schlechterer Qualität ausgeschlossen werden.

Die Anforderungen an die Berichtsqualität werden noch einmal in Tabelle 5 zusammengefasst:

Tabelle 5: Anforderungen an die Berichtsqualität von Ebene-2-Studien

Abschnitt	Unterabschnitt	Beschreibung
Titel		<ul style="list-style-type: none"> Identifizierung als Studie zur Ermittlung der diagnostischen Genauigkeit inklusive Angabe der zu vergleichenden Tests
Abstrakt		<ul style="list-style-type: none"> Strukturiertes Format
Einführung		<ul style="list-style-type: none"> Präzise Angabe der Fragestellung und des klinischen Problems
Methoden	Design / Protokoll	<ul style="list-style-type: none"> Setting Untersuchungsabfolge, z. B. zufällige Zuordnung zum Test, zeitliche Abfolge Patientenzuteilung zu den Testverfahren Datenerhebung prospektiv oder retrospektiv Ein- / Ausschlusskriterien bzw. Indikation
	Beschreibung des Testverfahrens	<ul style="list-style-type: none"> Technische Charakteristika, z. B. Gerätetyp, Hilfsmittel, Reagenzien, Test-Kits, vorbereitende Maßnahmen, technische Qualitätssicherung Auswertungsalgorithmus bei computergestützten Verfahren
	Referenztest	<ul style="list-style-type: none"> Benennung des Referenztests oder des Goldstandards Methode der Verifizierung, z. B. Pathologie
	Patientenselektion	<ul style="list-style-type: none"> Beschreibung der Studienpopulation, z. B. Stadienverteilung, Komorbidität, Alter, Geschlecht Methode der Rekrutierung der Patienten, z. B. konsekutiv, geschichtet, getrennte Erhebung bei Kranken und Gesunden („Fall-Kontroll-Ansatz“)
	Fallzahlplanung	<ul style="list-style-type: none"> Angabe der Fallzahl für die erwünschte Schätzgenauigkeit (Varianz, Konfidenzintervall) Fallzahlplanung für Subgruppenanalysen
	Auswertung / Interpretation der Tests	<ul style="list-style-type: none"> Kenntnisstand des / der Auswerter(s) in Bezug auf Vor-Testergebnisse und Krankheitsstatus Definition / Klassifikation der Testergebnisse
	Datenanalyse / statistische Auswertung	<ul style="list-style-type: none"> Datenaufbereitung, Beschreibung und Begründung von Klassenbildung, z. B. Dichotomisierung kontinuierlicher Variablen Berechnung von Effektschätzern, Angabe von statistischen Testverfahren Umgang mit unklaren oder nicht interpretierbaren Befunden
Ergebnisse	Patientenfluss	<ul style="list-style-type: none"> Anzahl der untersuchten Patienten bzw. Organe Vollständigkeit der Testdurchführung
	Datenpräsentationen	<ul style="list-style-type: none"> Anzahl der korrekt und nicht korrekt durch den Test identifizierten Entitäten (Vierfeldertafel) Angabe von abgeleiteten Effektschätzern, z. B. Sensitivität, Spezifität, prädiktive Werte sowie die Konfidenzintervalle Angabe von Kurvenfläche und Konfidenzintervallen bei ROC-Kurven für quantitative Tests
Diskussion	Diskussion designtypischer Biasformen	<ul style="list-style-type: none"> Spektrum-Bias (s.o.) Verifikationsbias (Workup-Bias, s. o.) Diagnostic / Review Bias (s. o.) Inkorporationsbias (s.o.)
	Generalisierbarkeit (externe Validität)	<ul style="list-style-type: none"> Reproduzierbarkeit der Testergebnisse in anderen Settings bzw. Abhängigkeit von der Interpretation Abhängigkeit bzw. Änderung der Richtung der Ergebnisse, z. B. von Krankheitsstadium, Komorbidität, Alter, Geschlecht Repräsentativität der untersuchten Patientenpopulation Zufallsfehler

4.3 Methoden der Metaanalyse von Studien zur diagnostischen Genauigkeit

Ob eine Metaanalyse durchgeführt werden kann oder nicht, hängt von der Anzahl und der methodischen Qualität der eingeschlossenen Primärstudien und dem Grad der Heterogenität der Schätzer der diagnostischen Genauigkeit ab.

Jede statistische Methode zur Durchführung einer Metaanalyse sollte

- es ermöglichen, den Einfluss der Unterschiede zwischen den Primärstudien (Studienqualität, Grenzwerte) auf die diagnostische Genauigkeit abzuschätzen,
- Daten zur Generalisierbarkeit der Ergebnisse für Subgruppen liefern und
- adäquat die verbleibende Variabilität zwischen den Studien darstellen.

Diagnostische Studien mit binären Ergebnissen (Test positiv oder negativ) stellen den häufigsten Fall in der Literatur dar. Die Daten von Studien mit binären Outcomes zur diagnostischen Genauigkeit können in Vierfeldertafeln entsprechend dem Schema in Abbildung 1 extrahiert werden; auf jeden Fall müssen die Raten für Personen mit einem richtig positiven (TP), einem falsch negativen (FN), einem falsch positiven (FP) und einem richtig negativen (TN) Ergebnis aus den Studien extrahiert werden. Die Standardgrößen Sensitivität, Spezifität und die Likelihood Ratios werden entsprechend daraus berechnet.

Testergebnis	Krankheitsstatus	
	positiv	Negativ
positiv	TP	FP
negativ	FN	TN
	$N_1 = TP + FN$	$N_2 = FP + TN$

TP = richtig positiv
 FN = falsch negativ
 FP = falsch positiv
 TN = richtig negativ

Abbildung 1: Vierfeldertafel für binäre Outcomes diagnostischer Studien

$$\text{Sensitivität} = \frac{TP}{TP + FN} = \text{TPR} = \text{true positive rate}$$

$$\text{Spezifität} = \frac{TN}{FP + TN} = \text{TNR} = \text{true negative rate}$$

$$\text{FPR} = \text{false positive rate} = 1 - \text{Spezifität}$$

$$\text{Positive Likelihood Ratio} = \frac{\text{TPR}}{\text{FPR}} = \frac{\text{Sensitivität}}{1 - \text{Spezifität}}$$

$$\text{Negative Likelihood Ratio} = \frac{FNR}{TNR} = \frac{1 - \text{Sensitivität}}{\text{Spezifität}}$$

Bei Metaanalysen diagnostischer Tests ist es wichtig, zwei Aspekte zu berücksichtigen:

- die Fähigkeit des diagnostischen Tests zwischen kranken und gesunden Personen zu diskriminieren
- die Wahl des Grenzwerts, der ein positives Ergebnis definiert.

Unterschiedliche Stringenz in der Wahl des Grenzwertes führt zu einem Trade-off zwischen Sensitivität und Spezifität, aber hat keine Einfluss auf die diskriminatorische Fähigkeit des Tests.

Mit welcher Methode die diagnostische Genauigkeit von Primärstudien am besten zusammengefasst werden kann, hängt von Annahmen ab, die getroffen werden, um die beobachteten Unterschiede zu erklären.

Das einfachste aber zugleich auch das restriktivste Modell geht von der Annahme aus, dass sich die Studien weder in ihrem Grenzwert noch in ihrer diagnostischen Genauigkeit unterscheiden, d. h. alle Primärstudien implementieren den diagnostischen Test identisch, benutzen einen identischen Grenzwert und setzen den Test bei identischen Populationen ein⁶⁸. Midgette et al.⁵¹ schlagen vor, in diesem Fall zunächst die Heterogenität für TPR und FPR mittels eines χ^2 -Tests (oder Fisher-Exakt-Test) getrennt zu überprüfen. Wenn beide Parameter keine signifikante Heterogenität aufweisen, können die Felderbesetzungen der Vierfeldertafel jeder Primärstudie gepoolt werden und ein gemeinsamer Parameter für FPR und TPR aller eingeschlossenen Primärstudien berechnet werden¹⁵. Standardfehler und Konfidenzintervalle können dann mit den üblichen Methoden für Binomialverteilungen erstellt werden.

Ein weniger restriktives Modell geht von der Annahme aus, dass der diagnostische Test mit der gleichen Genauigkeit in allen Primärstudien durchgeführt wird, aber unterschiedliche Grenzwerte definiert werden. Mit der Berechnung der Korrelationskoeffizienten zwischen Sensitivität und Spezifität lässt sich untersuchen, ob es Unterschiede in der Wahl der Grenzwerte zwischen den einzelnen Primärstudien gibt¹⁵. In diesem Falle würde man eine negative Korrelation zwischen Sensitivität und Spezifität über die Primärstudien erwarten. Wenn Grenzwerte variieren (implizit oder explizit)⁸, würden einfach gepoolte mittlere Sensitivitäten

⁸ Der Grenzwert kann implizit variieren, wenn z. B. die Einordnung eines Testergebnisses als negativ oder positiv von der Beurteilung von Gutachtern abhängig ist und deren Einordnung von der Prävalenz der Krankheit in

und Spezifitäten zu einer Unterschätzung der diagnostischen Genauigkeit führen, weil die Mittelwerte einer linearen Funktion folgen, während der Trade-off zwischen Sensitivität und Spezifität durch Veränderung des Grenzwertes einer curvilinearen Funktion folgt³³. Unter diesen Annahmen würden die Studien am besten mit einer *summary receiver operating characteristics* - Kurve⁹ zusammengefasst werden.

Das am wenigsten restriktivste Modell geht von den Annahmen aus, dass die beobachteten Unterschiede nicht nur unterschiedliche Grenzwerte reflektieren, sondern auch unterschiedliche Diskriminationsfähigkeiten.

Da die Daten der Primärstudien aus TPR und FPR Paaren bestehen, können diese im Sinne einer ROC-Kurve (FPR; TPR) aufgetragen werden. Eine derartige Graphik gibt zum einen erste Hinweise hinsichtlich der Genauigkeit des diagnostischen Tests. Je näher die Punkte an der linken oberen Ecke liegen, desto besser ist die diskriminatorische Fähigkeit des Tests. Zum andern auch Hinweise darauf, welche quantitative Methode zur Zusammenfassung am geeignetsten ist⁶⁸.

In der Literatur werden entsprechend der gewählten Modelle Ansätze für Metaanalysen diagnostischer Studien diskutiert, die im Folgenden dargestellt werden sollen.

4.3.1 Fixed-Effects- und Random-Effects-Modelle

Generell wird bei allen Methoden zwischen Fixed-Effects- und Random-Effects-Modellen unterschieden:

Bei *Fixed-Effects-Modellen* wird angenommen, dass allen Studien die gleiche (wahre, aber unbekannte) diagnostische Genauigkeit zugrunde liegt. Variationen in den einzelnen Studienergebnissen sind dann als Streuung um einen gemeinsamen wahren Wert der diagnostischen Genauigkeit zu betrachten und lassen sich durch Unterschiede in der Stichprobenziehung erklären oder auf unterschiedlich gewählte Grenzwerte zurückführen (Within-Study-Variation, WSV). Je größer die Stichprobengröße der Primärstudie, um so kleiner wird die Within-

der Studienpopulation beeinflusst wird und so bei einem offiziell identischen Grenzwert, doch unterschiedliche Sensitivitäten bzw. Spezifitäten geschätzt werden.

⁹ Receiver Operating Characteristic (ROC) Kurven werden im Rahmen der Evaluation diagnostische Tests eingesetzt, um die gegenseitige Abhängigkeit von Sensitivität und Spezifität für ‚positive‘ oder ‚abnormale‘ Testergebnisse zu veranschaulichen. Die TPR wird dazu gegen die FPR aufgetragen. Für verschiedene Grenzwerte lässt sich die jeweilige Sensitivität und Spezifität ermitteln. Eine Erhöhung des Grenzwertes erhöht die Sensitivität und verringert die Spezifität und umgekehrt. Die Fläche unter der Kurve ist ein Maß für die Genauigkeit eines Tests, mit 1 als Idealwert. Werte unter 0,5 entsprechen einer zufälligen Zuordnung der Testergebnisse. Die Methode wurde in den 1950er Jahren entwickelt, um elektromagnetische Signale optimal orten zu können. Seit den 70er Jahren werden ROC-Analysen auch in der Medizin eingesetzt.

Study-Variation. Allen Studien liegt eine gemeinsame, wahre (fixed) diagnostische Genauigkeit (ausgedrückt durch Sensitivität und Spezifität) zugrunde.

Random-Effects-Modelle beruhen auf der Annahme, dass der einer Einzelstudie zugrundeliegende Effekt zufällig von einer mittleren diagnostischen Genauigkeit abweichen kann⁶⁶. Sie müssen angenommen werden, wenn die Streuung der Studienergebnisse nicht allein durch die zufällige Auswahl (oder Unterschiede in den Grenzwerten) erklärt werden kann, sondern wenn diese durch Unterschiede in Patientencharakteristika, technische Details bei der Ausführung des Tests, der Qualität des Studiendesigns oder der Studienanalyse zustande kommen (Between-Study-Variation, BSV)⁷⁸. Daher muss angenommen werden, dass die wahre diagnostische Genauigkeit des Indextest in Abhängigkeit der gewählten Umstände zwischen den einzelnen Studien variiert. Random-Effects-Modelle für Metaanalysen sollen die WSV und BSV trennen und es ermöglichen, sowohl systematische als auch zufällige Komponenten zu untersuchen⁶⁶. Die primäre Motivation, Random-Effects-Modelle einzusetzen, ist die Notwendigkeit, so vollständig wie möglich, die Variabilität zwischen den Studien zu berücksichtigen. Dies gelingt durch Fixed-Effects-Modelle nur bedingt.

Der Hauptunterschied zwischen Fixed-Effects- und Random-Effects-Modellen betrifft die Größe der Varianz, die die wahre Position der SROC-Kurve umgibt. In Fixed-Effects-Modell wird eine gemeinsame wahre diagnostische Genauigkeit allen Studien zugrunde gelegt und nur die Variation innerhalb der Studien beeinflusst die „Unsicherheit“ im Ergebnis der Metaanalysen. In Random-Effects-Modellen wird die „Unsicherheit“ um die Ergebnisse nicht nur von diesen Variationen innerhalb der Studien sondern auch zwischen Studien (unterschiedliche Studienpopulation, technische Durchführung (Between-Study-Variation)⁷⁸ bestimmt. In Random-Effects-Modellen werden die Schätzer stärker von kleineren Studien beeinflusst und die Gefahr der Verzerrung der Ergebnisse durch Publikationsbias (s. u.) ist damit größer.

Eine ungewichtete Analyse stellt die extremste Random-Effects-Voraussetzung dar, in dem die Between-Study-Variation die Within-Study-Variation dominiert, denn alle Studien tragen unabhängig ihrer Stichprobengröße in gleichem Maße zur Positionierung der SROC-Kurve bei⁷⁸. Daher liefern Random-Effects-Modelle Ergebnisse, die zwischen denen von Fixed-Effects-Modellen, die nach der Stichprobengröße gewichtete Werte einsetzen, und völlig ungewichteten Analysen liegen. Üblicherweise resultieren Random-Effects-Modelle in weiteren Konfidenzintervallen als Fixed-Effects-Modelle.

4.3.2 Diagnostische Odds Ratio (DOR)

Eine häufig benutzte Maßzahl ist das *diagnostische Odds Ratio*. Es stellt eine Funktion der TPR und FPR dar und beschreibt das Verhältnis der Chance eines positiven Testergebnisses bei erkrankten Personen zur Chance eines positiven Testergebnisses bei nicht erkrankten Personen:

$$DOR = [TPR/(1 - TPR)] / [FPR/(1-FPR)]$$

Die DOR ist eine Maßzahl für die diskriminatorische Fähigkeit eines Tests. Diese ist um so höher, je größer das DOR ausfällt. Je nachdem wie der Grenzwert gewählt wird, kann diese diskriminatorische Fähigkeit genutzt werden, um den Test sensitiver zu machen (dafür aber weniger spezifisch) oder spezifischer (dafür aber weniger sensitiv)³³.

Wenn ausreichende Homogenität (s. u.) besteht und TPR und FPR nicht positiv miteinander korrelieren (d. h. keine unterschiedlichen Grenzwerte benutzt wurden), kann die DOR direkt gepoolt werden.

Die Methoden, die entwickelt wurden, um diagnostische Daten mit unterschiedlichen Grenzwerten zusammenzufassen, beruhen auf Regressionsmodellen mit der DOR als abhängiger Variable nach logarithmischer Transformation. Da die Beobachtungseinheit aber keine individuellen Patienten sondern Primärstudien sind, wird diese Methode auch als Meta-Regression bezeichnet. Eine Schwierigkeit, diese Vorgehensweise anzuwenden, liegt darin, dass DORs sich mit dem gewählten Grenzwert verändern können. Um dies zu berücksichtigen schlagen Moses et al.⁵⁵ vor, die Ergebnisse jeder Primärstudie mittels einer *receiver operating characteristic* Kurve darzustellen, bei der für jede Primärstudie die TPR gegen die FPR aufgetragen wird. Diese Modelle ermöglichen dann Analysen zur Abhängigkeit der Variabilität der DOR von den Grenzwerten und somit von spezifischen Charakteristika der Primärstudien, die auch mögliche Confounder darstellen. Die Kurven sind symmetrisch zur „Sensitivität = Spezifität“-Linie, wenn das Odds Ratio nicht mit dem Grenzwert variiert und asymmetrisch, wenn sich das Odds Ratio bei unterschiedlichen Grenzwerten ändert.

4.3.3 Mittelwerte der Sensitivität und Spezifität

Die einfachste Methode, Primärstudien zur diagnostischen Genauigkeit zusammenzufassen, ist es, die Felderbesetzungen der Vierfeldertafel jeder Primärstudie zu poolen und gemeinsame Parameter für FPR und TPR¹⁵ aller eingeschlossenen Primärstudien zu berechnen. Diese Methode kann nur eingesetzt werden, wenn es keine Variabilität des Grenzwertes zwischen den Studien gibt, d. h. falls es einen einheitlichen Grenzwert gibt, und eine ausreichende Ho-

mogenität zwischen den Studien besteht. Dies sollte sowohl graphisch als auch statistisch überprüft werden, bevor diese Methode eingesetzt wird (s.o.)¹⁵.

Die Mittelwerte errechnen sich folgendermaßen:

Wenn man die Sensitivität oder die Spezifität in jeder Studie i als Proportion betrachtet:

$$p_i = \frac{y_i}{n_i};$$

dann ist

Sensitivität_i = Anzahl der richtig positiv Personen_i / Anzahl der erkrankte Personen_i

Spezifität_i = Anzahl der richtig negativen Personen_i / Anzahl der gesunden Personen_i

Die Gesamtproportion errechnet sich aus

$$p = \Sigma y_i / \Sigma n_i$$

wobei Σy_i die Summe aller richtig positiven Personen (für die Sensitivität) oder aller richtig negativen Personen (für die Spezifität) und Σn_i die Summe aller erkrankten (für die Sensitivität) oder aller gesunden Personen (für die Spezifität) ist.

Für große Stichproben kann die Standardabweichung mit folgender Formel berechnet werden:

$$Se(p) = \sqrt{p(1-p) / \Sigma n_i}$$

Diese Methode ist identisch mit der, einen mit dem Kehrwert der Varianz gewichteten Mittelwert für FPR und TPR zu berechnen⁵¹.

Diese Methode hat jedoch mehrere Limitationen: Zum einen impliziert ein großer p-Wert bei der Überprüfung der statistischen Heterogenität nicht notwendigerweise einen starken Hinweis für Homogenität, denn dieser kann auch auf einer fehlenden statistischen Power beruhen (s. u.). Zum anderen wird keine Korrektur für multiples Testen integriert. Wenn ein Parameter heterogen ist, würden diese gepoolten Maße zur Unterschätzung der wahren diagnostischen Genauigkeit führen.

4.3.4 Summary Receiver Operating Characteristics (SROC) - Kurve

Für Metaanalysen mehrerer binärer Tests werden *Summary Receiver Operating Characteristics* - Kurven (SROC) erstellt. Am häufigsten wird die Erstellung einer SROC-Kurve nach der Methode von Moses et al.⁵⁵ für Fixed-Effects-Modelle verwendet. Ein ähnliches Modell wurde bereits von Kardaun und Kardaun³⁸ entwickelt, das ebenfalls mit einer logistischen

Transformation der Raten arbeitet und das bivariate Modell mit der Maximum-Likelihood Methode abschätzt. Beide kommen zu ähnlichen Ergebnissen und stehen konzeptuell in Beziehung, so dass die Ergebnisse des einen Modells in die des anderen übertragen werden können⁶⁶. Im folgenden soll das Modell von Moses et al.⁵⁵ dargestellt werden.

Die Interpretation einer Proportion oder Rate (P) ist oft leichter, wenn die Rate logistisch transformiert wird. Dann werden Werte kleiner als 0,5 als negative Werte dargestellt, Werte größer als 0,5 als positive Werte. Die Graphik ist symmetrisch um $P = 0,5$, so dass Werte wie 0,7 und 0,3 den gleichen Abstand von der 0 haben. TPR und die FPR werden somit als

$$\text{logit}(TPR) = \ln[TPR/(1-TPR)] \text{ und}$$

$$\text{logit}(FPR) = \ln[FPR/(1-FPR)] \text{ definiert.}$$

Die Methode nach Moses et al.⁵⁵ geht von der Annahme aus, dass ein linearer Zusammenhang zwischen dem $\text{logit}(TPR)$ und dem $\text{logit}(FPR)$ besteht. Im speziellen Falle, dass die Varianzen von $\text{logit}(TPR)$ und $\text{logit}(FPR)$ gleich sind, würde sich eine Gerade mit einer Steigung von 45° ergeben. Eine horizontale Gerade entsteht in diesem Falle, wenn $[\text{logit}(TPR) - \text{logit}(FPR)]$ gegen $[\text{logit}(TPR) + \text{logit}(FPR)]$ ¹⁰ aufgetragen wird.

Die Regressionsgleichung hat die Form

$$D = \alpha + \beta S$$

mit

$$D = \text{logit}(TPR) - \text{logit}(FPR) = \ln[\text{odds}(TPR)/\text{odds}(FPR)] = \ln(\text{DOR})$$

$$S = \text{logit}(TPR) + \text{logit}(FPR)$$

α = Schnittpunkt mit der y-Achse, Intercept

β = Regressionskoeffizient für S.

Um zu vermeiden, dass aufgrund der Nichtbesetzung eines Feldes der Vierfeldertafel die Transformation nicht durchgeführt werden kann, kann zu jeder Felderbesetzung der Wert 0,5 addiert werden. Für jede Studie muss dann $\text{logit}(TPR)$ und $\text{logit}(FPR)$ plus deren Differenz

¹⁰ Aufgrund von Effekten eines Measurement Errors der unabhängigen Variablen würde man nach der kleinsten Quadrat-Methode unterschiedliche Regressionsgeraden (die Verteilung der Fehler unterscheidet sich von Studie zu Studie) erhalten, je nachdem ob man $\text{logit}(TPR)$ oder $\text{logit}(FPR)$ als abhängige Variable wählen würde, um dieses Problem zu überwinden, schlagen Moses et al vor, dass $[\text{logit}(TPR) - \text{logit}(FPR)]$ als lineare Funktion von $[\text{logit}(TPR) + \text{logit}(FPR)]$ modelliert wird.

und Summe berechnet werden. Aus diesen wird dann die Regressionsgleichung mit der Methode der kleinsten Quadrate berechnet.

Die abhängige Variable D stellt den Logarithmus des diagnostischen Odds Ratios dar.

Die Variable S ist ein Maß für den Grenzwert, um ein Testergebnis als positiv zu bezeichnen, sie nimmt den Wert 0 an, wenn Sensitivität und Spezifität identisch sind. Sie ist positiv, wenn ein Grenzwert benutzt wird, der die Sensitivität erhöht (und die Spezifität senkt) und ist negativ, wenn ein Grenzwert festgelegt wird, der zu einer niedrigeren Sensitivität (und höheren Spezifität) führt³³.

Der Schnittpunkt mit der Y-Achse α ist das geschätzte Log Odds Ratio, wenn $S = 0$ ist. Je größer α ist, desto näher verläuft die ROC-Kurve an der linken oberen Ecke, d. h. desto größer ist die diagnostische Genauigkeit des Test⁷⁸.

Der Regressionskoeffizient β ist ein Maß dafür, inwieweit das Log Odds Ratio von dem benutzten Grenzwert abhängig ist. Wenn der Regressionskoeffizient sich der Null annähert oder nicht signifikant davon abweicht, kann die Testgenauigkeit jeder Primärstudie mit einem gemeinsamen Odds Ratio zusammengefasst werden ($= \alpha$) und die DOR hängt nicht vom Grenzwert ab (Konstantes-Odds-Ratio Modell). Ein konstantes Odds Ratio beinhaltet, dass die Ergebnisse aller Studien auf einer symmetrischen logistischen ROC-Kurve liegen. In diesem Fall können im Rahmen von Metaanalysen Methoden zur Zusammenfassung von DORs eingesetzt werden, wie z. B. die Mantel-Haenszel-Methode oder vorzugsweise Random-Effects-Methoden, wie von DerSimonian und Laird¹⁶ vorgeschlagen, die die Heterogenität (s. u.) berücksichtigen, die häufig in Studien zur diagnostischen Testgenauigkeit gefunden wird³³. Die Methode von DerSimonian und Laird benutzt eine 1-Schritt-Approximation, um die Between-Study-Varianz zu schätzen, und kalkuliert dann den gepoolten Schätzer als korrekt gewichteten Durchschnitt der studienspezifischen Parameter.

Wenn der Regressionskoeffizient β sich signifikant von Null unterscheidet, schätzen diese Modelle ein Odds Ratio, das nicht konstant über die einzelnen Studien ist. Die ROC-Kurve verläuft asymmetrisch und die DOR ist alleine nicht ausreichend die Testgenauigkeit zu beschreiben, da die Genauigkeit vom Grenzwert abhängig ist⁶⁶.

Auch andere Variablen können in die Gleichung aufgenommen werden, um den Einfluss z. B. der Studienqualität oder von Patientencharakteristika auf die Testgenauigkeit oder den Grenzwert zu untersuchen und um für Confounder zu adjustieren. Es resultieren dann unter-

schiedliche SROC-Kurven, wobei jede die Anpassung für eine unterschiedliche Kombination an Kovariaten repräsentiert. In diesem Falle kann die BSV-Komponente überprüft werden, indem man eine SROC-Kurve für die kombinierten Gruppen anpasst und die Residuen der beiden Gruppen mittels t-Test vergleicht.

Eine essentielle Komponente für dieses Modell ist, dass eine relevante Spannbreite a priori für TPR und FPR definiert wird, denn in den Grenzbereichen mit hoher TPR und FPR oder niedriger TPR und FPR (beide Kombinationen sind nicht klinisch relevant), beeinflussen diese Punkte disproportional die Steigung der Geraden. Dies bedeutet, dass man eine Obergrenze für die FPR-Werte und eine Untergrenze für die TPR-Werte spezifizieren muss (bei Moses et al. 0,5). Sind diese Grenzen festgelegt, hat dies folgende Konsequenzen: nur Studien mit Ergebnissen innerhalb dieser Grenzen werden in die Analyse eingeschlossen und die SROC-Kurve wird nur für diese Spannbreite berechnet. Dies kann zu einer Verzerrung der geschätzten Linie nach oben führen (Truncation Bias). Aber dieser Bias und der Bias, der durch die Addition von 0,5 zur Besetzung der Vierfeldertafel eingeführt wird, haben gegensätzliche Wirkung und gleichen sich nahezu aus^{11 55}. Nach Irwig et al. birgt diese Vorgehensweise einige Nachteile³³:

- Häufig ist es schwierig zu entscheiden, welche Bandbreite klinisch relevant ist.
- Die Methode schließt Punkte ein oder aus, die sich innerhalb oder außerhalb dieser Bandbreite aufgrund von „random sampling variability“ befinden.
- Punkte können in oder aus dem Gebiet verschoben werden, in dem der offensichtliche Grenzwert durch die zufällige Auswahl von testnegativen Personen zur Verifikation verändert wird.

Neben der Methode der kleinsten Quadrate präsentieren Moses et al. ein robustes Modell: Dies erhält man auch dadurch, dass man die (S,D)-Paare der Primärstudien nach der Größe von S ordnet und in drei gleich große Gruppen teilt. Die Mediane von S und D des oberen und unteren Drittels ergeben zwei Punkte aus denen sich die Steigung der Geraden ergibt. Das Intercept erhält man, indem man die Gerade so parallel verschiebt, dass die Hälfte der Punkte über und unter der Geraden liegen.

Empirisch zeigt sich aber, dass ungewichtete, gewichtete (s. u.) oder robuste Methoden im allgemeinen ähnliche Schätzer für α und β ergeben.

Das Modell mittels der Methode der kleinsten Quadrate kann ungewichtet, d. h. jede Studie geht zu gleichen Anteilen in die Analyse ein, oder gewichtet, z. B. durch den Kehrwert der Varianz von D und damit für die Stichprobengröße der Primärstudien, angepasst werden.

Beim Einsatz einer gewichteten Methode der kleinsten Quadrate kann eine einzige sehr große Studie die Schätzung dominieren. Nach Moses et al. ist es nicht ratsam, Between-Study-Variation zu ignorieren⁵⁵. Daher ist es wichtig, keine Schätzer einzusetzen, die diese Interstudienkomponente nicht berücksichtigen. Deshalb empfehlen sie die ungewichtete Methode der kleinsten Quadrate, als die geeignetste Methode, wenn die Variation innerhalb der Studien gegenüber der zwischen den Studien vernachlässigt werden kann, umgekehrt ist die gewichtete Variante einzusetzen, wenn es keine Variationen zwischen den Studien gibt, aber Within-Study-Variation.

Durch eine Rücktransformation der errechneten Regressionslinien, lässt sich eine konventionelle ROC-Kurve als SROC-Kurve erstellen. Die Werte der Regressionsfunktion werden dann mit Hilfe einer Exponentialfunktion graphisch in einer SROC-Kurve dargestellt. Die von Moses et al.⁵⁵ vorgeschlagene Formel lautet:

$$Q = \left[1 + e^{-\alpha/(1-\beta)} \left(\frac{1-P}{P} \right)^{(1+\beta)/(1-\beta)} \right]^{-1}$$

wobei gilt $P = \text{FPR}$ und $Q = \text{TPR}$.

Gesamtwerte für TPR und FPR sind nicht direkt verfügbar, sondern für einen ausgewählten FPR-Wert lässt sich über die SROC-Kurve der korrespondierende TPR-Wert ermitteln und umgekehrt.

Als Schätzer für die Gesamtgüte eines Tests wird Q^* berechnet (bzw. graphisch ermittelt) als Schnittpunkt der geschätzten ROC-Kurve mit der Geraden, an der Sensitivität und Spezifität gleich sind ($\text{TPR} + \text{FPR} = 1$). Q^* stellt nur eine Funktion von α dar.

$$Q^* = [1 + e^{-\alpha/2}]^{-1}$$

Ein Wert von Q^* nahe der „1“ impliziert einen Kurvenverlauf der ROC-Kurve nahe der oberen linken Ecke. Dieser Wert wird anstelle der sonst bei ROC-Kurven üblichen Berechnung der Fläche unter der Kurve verwendet. Dieser Wert eignet sich, um verschiedene Tests zu vergleichen, in dem man die Q^* -Werte und ihre Standardfehler vergleicht.

¹¹ Nur wenn die Stichproben sehr klein sind oder der diagnostische Test eine sehr hohe Genauigkeit besitzt, dominiert der Bias durch Addition von 0,5⁶⁸.

Der Wert für

$$\frac{(Q_1^* - Q_2^*)}{\sqrt{SE^2(Q_1^*) + SE^2(Q_2^*)}}$$

kann mit der Normalverteilungstabelle verglichen werden, wenn für jeden Test wenigstens 10 Studien vorliegen und die Studien statistisch unabhängig sind, d. h. es keine Überschneidungen in der Studienpopulation gibt.

Beispiel: Diagnostische Genauigkeit von Röntgenübersichtsaufnahmen im Vergleich zu Punktion bei akuter Sinusitis bei Erwachsenen (nach Perleth et al., 1999 ⁶²)

Im Rahmen einer systematischen Übersicht im deutschen HTA-Projekt wurden sämtliche recherchierbaren Studien zur diagnostischen Genauigkeit verschiedener Verfahren bei akuter Sinusitis bei Erwachsenen ausgewertet. Aus dieser Studie sind im folgenden die Ergebnisse für den Vergleich der Röntgenübersichtsaufnahme mit der Punktion als Goldstandard dargestellt.

Tabelle 6: Vierfeldertafeln von Studien zur diagnostischen Genauigkeit von Röntgenübersichtsaufnahmen im Vergleich zur Punktion bei akuter Sinusitis

Studie	N (Sinus)	TP	FN	FP	TN	TPR	FPR
Ballantyne 1946	93	59	1	22	11	0,98	0,67
Vuorinen et al. 1962	272	154	4	46	68	0,97	0,40
McNeill 1963	242	113	35	23	71	0,76	0,24
Axelsson et al. 1970	301	140	59	45	57	0,70	0,44
Evans et al. 1975	68	34	0	1	33	1,00	0,03
Revonta 1980 (Serie I)	170	56	23	7	84	0,71	0,08
Revonta 1980 (Serie III)	60	28	7	5	20	0,80	0,20
Gwaltney et al. 1983	58	34	2	14	8	0,94	0,64
Kuusela et al. 1983	156	68	14	21	53	0,83	0,28
van Buchem et al. 1995 (Serie 1)	62	10	6	4	42	0,63	0,09
van Buchem et al. 1995 (Serie 2)	125	43	8	35	39	0,84	0,47
Savolainen et al. 1997	234	174	13	18	29	0,93	0,38
Laine et al. 1998	72	14	9	1	48	0,61	0,02

In der Metaanalyse wurden die Daten von 1908 Sinus (1108 mit Krankheit, 800 ohne Krankheit) einbezogen. Das gepoolte Ergebnis der Metaanalyse für die Sensitivität betrug 0,85 (95%-CI 0,77 – 0,90), für die Spezifität 0,72 (95%-CI 0,61 – 0,81). Q^* für diesen Vergleich betrug 0,82. Aus der Metaanalyse wurde eine SROC-Kurve konstruiert, die in Abbildung 2 dargestellt ist.

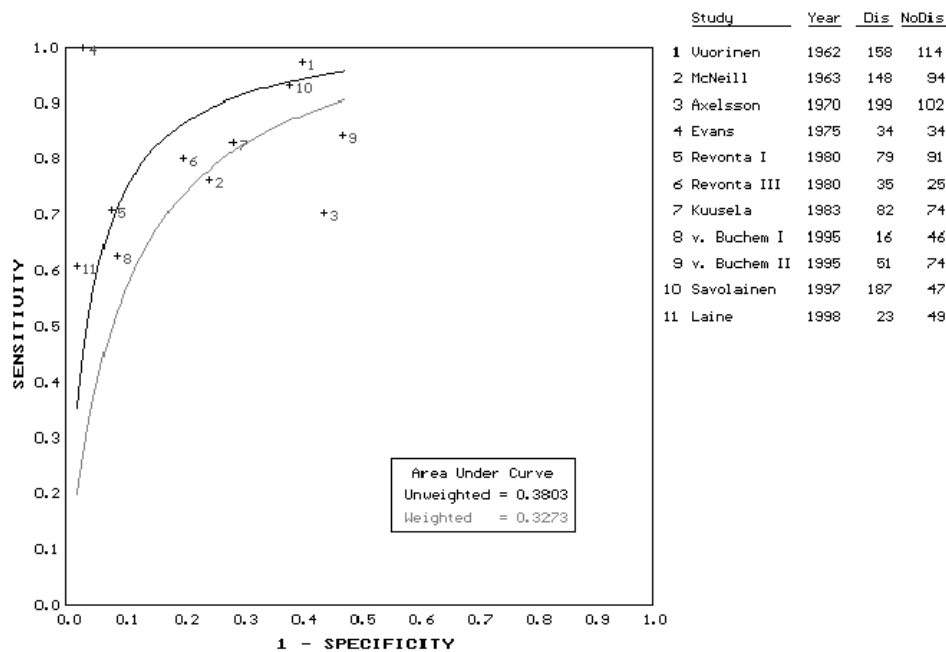


Abbildung 2: SROC-Kurve für die Metaanalyse von Röntgen der Nasennebenhöhlen vs. Punktion. Die hellgraue Linie zeigt die SROC-Kurve nach Gewichtung der Studien mit dem jeweiligen Kehrwert der Varianz, was erwartungsgemäß einen etwas flacheren Verlauf ($Q^* = 0,79$) ergibt.

Zusammenfassend lässt sich feststellen, dass sich Erstellung und Interpretation von SROC-Kurven beträchtlich vereinfacht, wenn diese symmetrisch um die Gerade $TPR = 1 - FPR$ liegen⁶⁸. In diesem Fall, ist die Diskriminationsfähigkeit des diagnostischen Test bei allen Grenzwerten gleich und die ROC-Kurve wird durch einen Parameter (α) bestimmt, der ein Maß für diese ist.

Diese Methode ist für Studien geeignet, deren Ergebnisse in Form von Vierfeldertafeln dichotomisiert präsentiert werden und denen ein Vergleich mit einem Goldstandard zugrunde liegt. Das Verfahren besteht durch seine intuitive Erfassbarkeit bedingt durch die graphische Darstellung. Letztlich resultiert eine Kurve, die sofort die Güte des Testverfahrens erkennen lässt. Das Moses-Modell scheint weniger anfällig für „Ausreißer“ zu sein, wie ein Vergleich von drei verschiedenen Fixed-Effects-Modellen zeigte⁶⁶. Ein weiterer Vorteil dieser Methode ist, dass es nicht die Annahme erfordert, dass die Varianzen der zugrundeliegenden kontinuierlichen Verteilungen der richtig positiven und negativen Ergebnisse gleich sind³⁰.

Die Kalkulation kann u. a. mit Hilfe des Programms Metatest® durchgeführt werden, das von Dr. Joseph Lau (New England Medical Center, Boston, Massachusetts) entwickelt wurde und unentgeltlich zur Verfügung gestellt wird.

4.3.5 Latent-Scale Regression Modelle

Rutter et al. ⁶⁶ schlagen für Metaanalysen von diagnostischen Tests mit binären Ergebnissen und Goldstandardinformationen eine „latent-scale“ logistische Regression (LSLR) vor, um eine oder mehrere SROC-Kurven anzupassen. Das LSLR-Modell geht von zwei latenten logistischen Verteilungen der dichotomen Testergebnisse aus, eine für die erkrankte und eine für die nicht erkrankte Population. Die Modelle schließen Studienindikatorvariablen als Kovariaten ein, um der Within-Study-Variation in Fixed-Effects-Modellen Rechnung zu tragen. Random-Effects-Modelle liefern zusätzlich den Rahmen, um explizit sowohl die Within-Study-Variation als auch die Between-Study-Variation zu berücksichtigen.

Rutter et al. gehen in ihrem Modell für binäre Testergebnisse von der Annahme aus, dass die Anzahl der negativen Testergebnisse binomial verteilt ist und die Wahrscheinlichkeit eines negativen Testergebnisses in folgender Weise vom wahren Krankheitsstatus abhängt:

$$\text{logit}[P(Y = 1 | D)] = (\theta - \alpha D) e^{-\beta D}$$

Y = Testergebnis; 1 = negativ, 2 = positiv

D = wahrer Krankheitsstatus

θ = Grenzwert, modelliert den „Trade-off“ zwischen TP und FP

α = Genauigkeitsparameter

β = Regressionskoeffizient

Diese LSLR hat folgende intuitive Interpretationsmöglichkeit: Das diagnostische Ergebnis eines bestimmten Patienten kann als die dichotomisierte Version der zugrundeliegenden zufälligen logistischen Variable Y^* beschrieben werden, mit einem Mittelwert von αD und einer Standardabweichung von $e^{\beta D}$. Wenn man von den zwei latenten Verteilungen für die TP-Fälle und die TN-Fälle ausgeht, nimmt das Testergebnis Y den Wert 1 (negativ) an, wenn Y^* unterhalb des Grenzwertes θ liegt und den Wert 2 (positiv), wenn Y^* oberhalb des Grenzwertes liegt. Der Parameter α liefert dabei ein Maß, wie weit die latenten Verteilungen der wirklich kranken und gesunden Populationen von einander getrennt liegen. Je weiter die beiden latenten Verteilungen von einander getrennt liegen, desto besser diskriminiert der Test zwischen wirklich Kranken und Gesunden. Der Parameter β reflektiert Unterschiede in der Variabilität von Y^* für TP- und TN- Patienten. Die gesamte Separation der latenten Verteilungen ist eine Funktion von α und β .

Die für die Erstellung üblicher ROC-Kurven notwendigen Parameter, Schnittpunkt mit der y-Achse und Steigung, können berechnet werden, in dem man die Schätzer für α und β benutzt.

Wenn der Parameter β gleich Null ist bzw. sich nicht signifikant davon unterscheidet, korrespondiert das Odds Ratio für ein positives Testergebnis mit α und ist konstant über die Studien.

Weitere Kovariaten können folgendermaßen in dieses LSLR Modell aufgenommen werden:

$$\text{logit}[P(Y = 1 | D, X)] = (\theta - \alpha D - \gamma X) e^{-\beta D}$$

mit X als einem Vektor für Kovariaten auf Studienebene.

In Fixed-Effects-Modellen können die Kovariaten entweder Studienindikatoren sein oder Charakteristika reflektieren, die von mehreren Studien geteilt werden. Rutter et al. beschreiben zwei Modelle, die Studienindikatoren benutzen:

- „Cut-point“-Modell: Unterschiede zwischen den Studien resultieren aus dem Einsatz von unterschiedlichen Grenzwerten. Die diagnostische Genauigkeit ist über die Studien konstant. Dieses Modell liefert eine SROC-Kurve basierend auf einem einzigen Schätzer der Gesamtgenauigkeit.
- „Accuracy“-Modell: Unterschiede zwischen den Studien resultieren aus den Unterschieden der Testgenauigkeit zwischen den Studien, mit einheitlichem Grenzwert zwischen den Studien. Das Accuracy Modell führt zu separaten ROC-Kurven für jede Studie. Eine SROC-Kurve kann gebildet werden basierend auf dem Gesamtgenauigkeitsschätzer α .

SROC-Kurven, die für Fixed-Effects-Modelle, generiert werden, sind für das Modell von Moses et al.⁵⁵ und dem „Cut-point“ Modell nahezu identisch.

Die Formulierung von „random-coefficient latent-scale logistic regression“ (RCLSLR) unter der Random-Effects Voraussetzung geht von folgender hierarchischen Modellstruktur aus:

Um die Within-Study-Variation zu berücksichtigen, wird ein separates LSLR Modell für jede Primärstudie aufgestellt. Für jede *i*te Studie, ist die Wahrscheinlichkeit eines negativen Testergebnisses für den Patienten *j* folgendermaßen modelliert:

$$\text{logit}[P(Y_{ij} = 1 | D_{ij})] = (\theta_i - \alpha_i D_{ij}) e^{-\beta D_{ij}}$$

Um die Between-Study-Variabilität zu berücksichtigen, erlaubt das Modell jeder Studie einen eigenen Grenzwert θ_i und einen eigenen Genauigkeitsparameter α_i . Der Regressionsparameter β (?) wird als konstant zwischen den Studien angenommen.

RCLSLR Modelle erlauben daher, dass sich sowohl der Grenzwert als auch die Genauigkeitsparameter zwischen den Studien unterscheiden. Die zugrundeliegende Verteilung dieser Parameter beschreibt, wie diese zwischen den Studien variieren.

Eine SROC-Kurve für alle Studien kann folgendermaßen abgeleitet werden:

$$\text{logit}[P(Y_{ij} = 1 | D)] = (\hat{\Theta} - \hat{\Lambda} D_{ij}) e^{-\beta D_{ij}}$$

mit Θ als dem erwarteten (oder mittleren) Grenzwertparameter über die Studien und Λ als dem erwarteten (oder mittleren) Genauigkeitsparameter (siehe Rutter et al. 1995⁶⁶). Ein Bayes'scher Ansatz wird eingesetzt, um das Modell anzupassen.

Ein Vorteil der LSLR ist, dass sie keine Kontinuitätskorrektur (+0,5) braucht, und somit die Ergebnisse dadurch nicht verzerrt werden. Aber das LSLR Modell fordert, dass entweder die diagnostische Genauigkeit oder der Grenzwert über die Studien fixiert ist und es macht bestimmte Annahmen hinsichtlich der Verteilung der zugrundeliegenden latenten Variablen.

4.3.6 Modell für ordinale Daten

Die Dichotomisierung von Testergebnissen, um Sensitivität und Spezifität zu erhalten, führt zu einem Informationsverlust und Methoden zur Berücksichtigung von kontinuierlichen oder wenigstens ordinal skalierten Daten sollten für Metaanalysen eingesetzt werden, wenn in den Primärstudien derartige Daten präsentiert werden. Wenn in den Primärstudien für die gleiche Anzahl von Kategorien Daten vorhanden sind, kann für jede Primärstudie eine ROC-Kurve erstellt werden und eine Gesamt-ROC-Kurve mittels ordinaler Regressionstechniken erstellt werden. Mit Hilfe dieser Methoden kann auch für Kovariaten adjustiert werden³³.

Hierfür wird angenommen, dass ein Testergebnis in eine von J Kategorien fallen kann, die in aufsteigender Reihenfolge der Testpositivität angeordnet sind und durch $J - 1$ Grenzwerte getrennt werden. Die Wahrscheinlichkeit, dass ein Testergebnis (Y) in eine gegebene Kategorie (j) oder darunter fällt, kann als lineare Funktion von k erklärenden Variablen (x_1, \dots, x_k) mit Hilfe des proportionalen Odds Modells berechnet werden:

$$\text{logit}[P(Y \leq j | x_1, \dots, x_k)] = \theta_j - (\alpha_1 x_1 + \dots + \alpha_k x_k)$$

Ein separater Term (θ_j) wird für jeden Grenzwert eingeschlossen, welcher die unterschiedlichen Punkte entlang der angenommenen zugrundeliegenden (latenten) kontinuierlichen Skala für jedes Testergebnis reflektiert. Der Koeffizient α ist nicht vom Grenzwert abhängig. Somit ist das Odds Ratio, das als Maß für die Assoziation zwischen der erklärenden Variable und dem Testergebnis benutzt wird, konstant über alle Grenzwerte. Für binäre Tests reduziert sich dieses Modell auf ein logistisches Regressionsmodell.

Um die Testgenauigkeit zu bestimmen, muss das Ergebnis des Referenztests in das Modell integriert werden. Wenn diese Variable durch x_I dargestellt wird (0 = nicht erkrankt, 1 = erkrankt), dann erhält man durch $\exp(\alpha_I)$ eine Schätzung für das diagnostische Odds Ratio. Dieses Odds Ratio gibt ein nützliches Maß für die Testgenauigkeit und kann dazu benutzt werden eine ROC-Kurve zu erstellen, in dem man für ausgewählte Werte von FPR die korrespondierenden TPR-Werte berechnet ³³.

Für weitere Informationen zu Modellen mit variablem Odds Ratio und weiteren Modellen wird auf die Veröffentlichung von Irwig et al., 1995 ³³ verwiesen.

4.3.7 Standardisierte Mittelwertdifferenzen

Als weitere Methode kann auch die Berechnung der standardisierten Mittelwertdifferenzen angewandt werden. Für Metaanalysen von Studien, die (annähernd normalverteilte) kontinuierliche Testergebnisse berichten, wurde die Kalkulation der standardisierten Differenz der empirischen Mittelwerte vorgeschlagen ^{5;30;73}.

Die allgemeine Formel hierfür lautet:

$$d = (M_1 - M_2) / s$$

wobei M_1 der Mittelwert der nicht erkrankten Population,

M_2 der Mittelwert der erkrankten Population und

s die Standardabweichung der gepoolten Stichprobe ist.

d wird auch als Effektgröße bezeichnet und ist ein Maß für die Diskrimination zwischen zwei normalverteilten Subpopulationen mit gleicher Varianz und unterschiedlichen Mittelwerten. d ist ein Wert für die Diskriminierungsfähigkeit oder Wirksamkeit (*test effectiveness score*) des untersuchten Tests ⁷³.

Wie schon beschrieben, werden häufig kontinuierliche Ergebnisse eines Test als dichotome Outcomes zusammengefasst, in dem man einen Grenzwert (c) wählt und Testergebnisse, die

über diesem Wert liegen als positiv klassifiziert, alle anderen als negativ. Sensitivität und Spezifität lassen sich dementsprechend als Terme dieser Verteilungen beschreiben: Sensitivität als der Anteil der Erkrankten, die oberhalb von c liegen und Spezifität als der Anteil der Gesunden, die unterhalb von c liegen. Diagnostische Testergebnisse, die innerhalb der Subpopulationen (erkrankt / gesund) normalverteilt sind, ergeben für eine relative große Bandbreite von c einen nahezu identischen Wert für die Summe der logits von Sensitivität und Spezifität, $\log[S_n/(1 - S_n)] + \log[S_p/(1 - S_p)]$, d. h. diese Summe ist nahezu unabhängig vom Grenzwert unter der Bedingung, dass die Testergebnisse normalverteilt sind³⁰. Diese Annahme gilt auch, wenn die Testergebnisse eine logistische Verteilung haben, die häufig als Annäherung an die Normalverteilung benutzt wird. Wenn zwei Verteilungen logistisch sind und gleiche Varianzen haben, ist die Summe der Logarithmen der Odds für Sensitivität und Spezifität unabhängig von c und entspricht dem Logarithmus des Odds Ratio der Vierfeldertafel (AD/BC, siehe unten), um die Testperformance zusammenzufassen. Hasselblad et al.³⁰ leiten her, dass dieses log Odds Ratio (bzw. die Summe der logits von Sensitivität und Spezifität) eine Konstante multipliziert mit der standardisierten Mittelwertdifferenz ist.

		Krankheitsstatus	
Testergebnis		Positiv	Negativ
	positiv	A	B
	negativ	C	D

Abbildung 3: Vierfeldertafel für binäre Outcomes diagnostischer Studien

Daher ergibt sich:

$$d = \frac{\sqrt{3}}{\pi} \left(\log \frac{\text{Sensitivität}}{1 - \text{Sensitivität}} + \log \frac{\text{Spezifität}}{1 - \text{Spezifität}} \right)$$

oder

$$d = \sqrt{3} [\log(A) + \log(D) - \log(B) - \log(C)] / \pi$$

Um zu vermeiden, dass d nicht berechnet werden kann, wenn ein Feld der Vierfeldertafel nicht besetzt ist, schlagen Hasselblad et al.³⁰ vor, zu jeder Felderbesetzung 0,5 zu addieren.

Die beiden Annahmen, dass die zugrundeliegenden kontinuierlichen Verteilungen logistisch (d. h. nahezu normal) sind und gleiche Varianzen haben, müssen erfüllt sein, da es sonst zur Verzerrung der Ergebnisse kommt.

Auch Hasselblad et al. unterscheiden zwischen Fixed-Effects- und Random-Effects-Kombinationen³⁰.

Das Fixed-Effects-Modell geht davon aus, dass jede Studie eine gemeinsame wahre standardisierte Mittelwertdifferenz abschätzt. In diesem Fall werden die Maßzahlen häufig mittels der Inversen-Varianz-Gewichtungsformel zusammengefasst.

So ergeben angenommene m Studien d_1, d_2, \dots, d_m Schätzer eines Parameters mit einem Standardfehler von $v_1^{0.5}, \dots, v_m^{0.5}$. Diese Schätzer werden in der Regel gewichtet zusammengefasst, in dem Studien mit kleinerem Standardfehler mehr Gewicht gegeben wird.

Der gewichtete Mittelwert errechnet sich dann folgendermaßen:

$$\hat{d} = \frac{\sum_{j=1}^m w_j d_j}{\sum_{j=1}^m w_j}$$

wobei $w_j = \frac{1}{v_j}$ ist.

Die Varianz für diesen kombinierten Schätzer ist

$$Var(\hat{d}) = \frac{1}{\sum_{j=1}^m w_j}$$

Mit der Formel

$$\chi^2 = \sum_{j=1}^m w_j (d_j - \hat{d})^2$$

kann getestet werden, ob die Effektgrößenparameter d_j konstant über die Studien sind. Diese Formel folgt einer χ^2 -Verteilung mit $m-1$ Freiheitsgraden. Bei großen Werten muss die Homogenität der Studien abgelehnt werden.

Wenn Heterogenität zwischen den Studien angenommen werden kann, d. h. eine Between-Study-Variation-Komponente vorliegt, ist es angebrachter, Random-Effects-Modelle einzusetzen. Random-Effects-Modelle unterscheiden sich von Fixed-Effects-Modellen, in dem ein weiterer Parameter v^* als Maß für die Variation zwischen den Studien in Berechnung aufgenommen wird, um hierfür zu gewichten.

Der gewichtete Mittelwert errechnet sich dann folgendermaßen:

$$d^* = \frac{\sum_{j=1}^m w_j^* d_j}{\sum_{j=1}^m w_j^*}$$

wobei $w_j^* = \frac{1}{v_j + v^*}$ ist.

Die Varianz für diesen gemeinsamen Schätzer ist

$$Var(d^*) = \frac{1}{\sum_{j=1}^m w_j^*}$$

v^* berechnet sich folgendermaßen:

$$v^* = \left[\chi^2 - (m-1) \right] / \left[\sum_{j=1}^m w_j - \left(\sum_{j=1}^m w_j^2 / \sum_{j=1}^m w_j \right) \right]$$

Der Wert von v^* ist normalerweise größer als 0, und w_j^* somit kleiner als w_j und die Varianz von d^* (Random-Effects gewichtete Mittelwertdifferenz) ist größer als die der Fixed-Effects gewichteten Mittelwertdifferenz. Daraus resultieren breitere Konfidenzintervalle.

In Metaanalysen von Screening- oder diagnostischen Tests kann d als standardisierte Differenz zwischen den Mittelwerten zweier Populationen interpretiert werden. Je größer d , desto größer ist die diskriminatorische Fähigkeit des Tests. Ein Score von 1 impliziert einen wenig effektiven Test, die Sensitivität und Spezifität betragen dann beide nur 71%, oder eine Sensitivität von 90% (95%) korrespondiert mit einer Spezifität von 40% (24%). Beträgt der Score 3, die Sensitivität und Spezifität betragen dann beide 94%, oder eine Sensitivität von 90% (95%) korrespondiert mit einer Spezifität von 96% (92%).

Auch kann d in eine ROC-Kurve transformiert werden. Für gegebene Spezifitäten und d er rechnen sich die korrespondierenden Sensitivitäten folgendermaßen:

$$S_n = \left[1 + e^{-d} \left(\frac{S_p}{1 - S_p} \right) \right]^{-1}$$

Diese Kurve ist symmetrisch aufgrund der getroffenen Annahmen zur Diagonalen, d. h. die Symmetrie kann nicht mehr interpretiert werden.

Eine der wichtigsten Anwendungen des Effectiveness-Maßes ist der Vergleich unterschiedlicher diagnostischer Tests. Die Methoden hierfür unterscheiden sich je nach gewähltem Modell:

Für Fixed-Effects-Kombinationen gilt: Wenn \hat{d}_1 und Varianz (\hat{d}_1) die standardisierte Mittelwertdifferenz des Tests 1 darstellen und \hat{d}_2 und Varianz (\hat{d}_2) die standardisierte Mittelwertdifferenz des Tests 2, dann kann die Hypothese, dass beide sich in ihrer standardisierten Mittelwertdifferenz nicht unterscheiden mittels der Formel:

$$z = \frac{\hat{d}_1 - \hat{d}_2}{\sqrt{\text{Var}(\hat{d}_1) + \text{Var}(\hat{d}_2)}}$$

berechnet werden. Die Hypothese wird zurückgewiesen, wenn z kleiner als $-1,96$ oder größer als $1,96$ ist (zweiseitiger Test, $\alpha < 0,05$).

Für Random-Effects-Kombinationen ist die Berechnung mit der bei Fixed-Effects-Kombination identisch, nur dass die größere Varianz als Komponente mit aufgenommen wird:

$$z^* = \frac{d_1^* - d_2^*}{\sqrt{\text{Var}(d_1^*) + \text{Var}(d_2^*)}}$$

Die Hypothese wird zurückgewiesen, wenn z^* kleiner als $-1,96$ oder größer als $1,96$ ist (zweiseitiger Test, $\alpha < 0,05$).

Zusammenfassend muss festgestellt werden, dass das Maß d die folgenden Vorteile hat:

- Es gibt eine einzige Maßzahl.
- Es kann sehr leicht berechnet werden.
- Es ist unabhängig vom Grenzwert des Tests.
- Es ist unabhängig von der Prävalenz.
- Es ist näherungsweise normalverteilt, so dass Konfidenzintervalle sehr leicht berechnet werden können.

- Sensitivität und Spezifität können einfach daraus abgeleitet und eine ROC-Kurve erstellt werden.

Es müssen aber die beiden Voraussetzungen Normalverteilung und gleiche Varianzen erfüllt sein. Wenn die Varianz nicht gleich ist, ist die Maßzahl nicht unabhängig vom Grenzwert.

Beispiel für die Berechnung der Mittelwertdifferenz d:

Blakeley et al.⁵ führten eine systematische Übersicht zu nicht invasiven Testverfahren für die Arteria carotis-Stenosen durch. Sechs Testverfahren wurden dabei untersucht (Karotis-Duplex- und Doppler-Sonographie, Magnetresonanztomographie, supraorbitale Doppler-Sonographie, B-Mode-Sonographie und Okuloplethysmographie). Die Literatursuche erstreckte sich auf englischsprachige Artikel in MEDLINE und aus Referenzlisten für die Jahre 1977 bis 1993. Artikel wurden eingeschlossen, wenn a) ein Vergleich mit konventioneller intraarterieller oder digitaler Subtraktionsangiographie durchgeführt wurde; b) die Angiographieergebnisse für Verschlüsse separat dargestellt wurden; c) die Erstellung einer Vierfeldertafel möglich war. Von insgesamt 568 identifizierten Artikeln wurden 70 in die Analyse eingeschlossen. Die statistische Auswertung erfolgte nach einer Stratifizierung der Testergebnisse in totalen Verschluss (100% Stenosen), 70%- und 50%-Stenosen durch Pooling der Sensitivitäten und Spezifitäten, der Konstruktion von SROC-Kurven (für den Schwellenwert von 70%) und der Kalkulation von *d*.

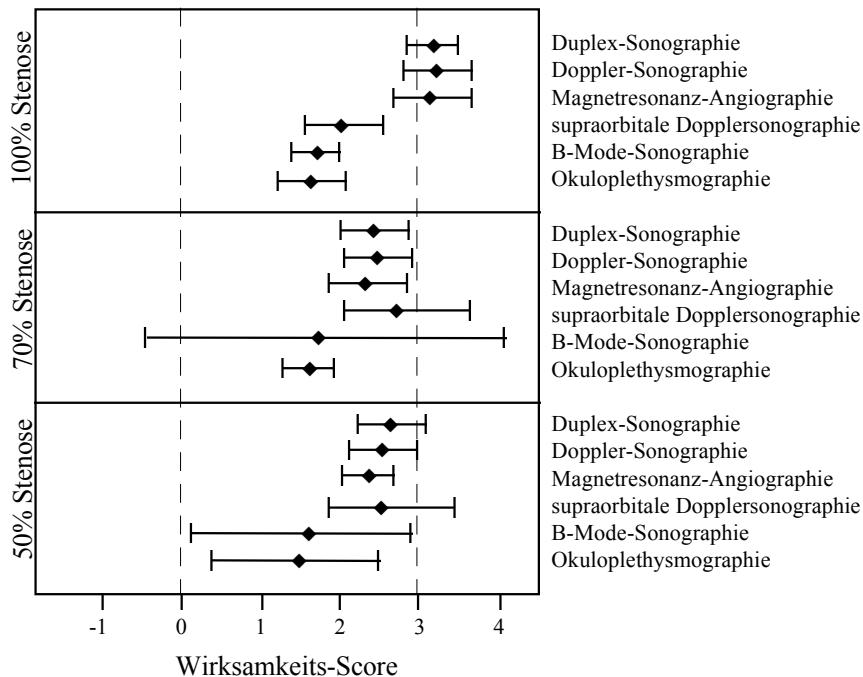


Abbildung 4: Mittelwertdifferenzen und Konfidenzintervalle für sechs verschiedene nicht invasive diagnostische Methoden der Karotisstenose

Insgesamt lagen die Daten für 6.404 Patienten mit 12.265 Arterien-spezifischen Ergebnissen vor. Das Durchschnittsalter betrug 62 Jahre, schätzungsweise 65% der Patienten waren männlich. Die wenigsten Studien erhielten Angaben zu Geschlecht oder Ethnizität. Die Prävalenz für komplette Verschlüsse betrug 10%, für 70%-Stenosen 28% und für 50%-Stenosen 41%. Der Test-Score für totale Verschlüsse war bei drei Verfahren über 3 (Duplex- und Doppler-Sonographie sowie Magnetresonanztomographie), bei 70%- und 50%-Stenose lagen die Werte zwischen 2 und 3, allerdings auch für die supraorbitale Doppler-Sonographie. Auch die anderen Metaanalyse-Ansätze zeigten eine

Überlegenheit der drei genannte Testverfahren, insbesondere für hochgradige Stenosen. Die graphische Darstellung in Abbildung 4 macht die Unterschiede der verschiedenen Diagnoseverfahren deutlich.

4.3.8 Weitere statistische Methoden

4.3.8.1 Imperfekte Standards

In Metaanalysen diagnostischer Tests und in den meisten Primärstudien wird davon ausgegangen, dass der gewählte Goldstandard, gegen den der Indextest getestet wird, fehlerfrei den Krankheitsstatus bestimmt. Die perfekte Bestimmung des Krankheitsstatus ist aber nicht immer möglich, ethisch nicht vertretbar oder zu teuer, so dass nicht perfekte Referenztests eingesetzt werden^{79;31}. Dieser Referenzfehler kann weitreichende Folgen für die Schätzung der diagnostischen Genauigkeit des Indextests haben. Im einfachsten Falle, falls die Referenz- und Indextestfehler unabhängig von einander sind, wird die Testgenauigkeit dabei unterschätzt. Diese Unterschätzung ist eine nicht lineare Funktion der Prävalenz.

Zur Korrektur schlagen Walter et al.⁷⁹ eine dreistufige Vorgehensweise vor:

Zuerst wird ein „latent class“-Modell¹² für die diagnostischen Daten entwickelt, das sich aus den Studien für die Metaanalyse ergibt. Ein „latent class“-Modell geht davon aus, dass der wahre Krankheitsstatus jeder Person unbekannt oder latent ist. Es geht von zwei Annahmen aus: der wahre Krankheitsstatus ist binär (krank oder nicht krank) und die Testfehler sind in Bezug auf den wahren Krankheitsstatus unabhängig. Beobachtungen werden gemacht, in dem man den Indextest und den Referenztest durchführt und jede Person anhand dieser Ergebnisse einem Feld in der Vierfeldertafel zuordnet. Für die Vierfeldertafel ergibt sich dadurch eine Wahrscheinlichkeitsverteilung für die einzelnen Felder, da nicht mehr davon ausgegangen werden kann, dass der wahre Krankheitsstatus ohne Fehler bestimmt werden kann (siehe Abbildung 5).

¹² Der Ausdruck „latent class“ umschreibt die Tatsache, dass der wahre Status der Variable immer verborgen oder unbekannt bleibt, obwohl Wahrscheinlichkeitsschätzungen dafür getroffen werden können⁸⁰.

Testergebnis	Krankheitsstatus	
	positiv	negativ
positiv	$\theta_k (1 - \beta)$	$(1 - \theta_k) \alpha$
negativ	$\theta_k \beta$	$(1 - \theta_k) (1 - \alpha)$

θ_k = Prävalenz der Krankheit in Studie k
 β = FNR des Tests; $(1 - \beta)$ = Sensitivität des Tests
 α = FPR des Tests; $(1 - \alpha)$ = Spezifität des Tests

Abbildung 5: Erwartete Wahrscheinlichkeitsverteilung der Personen: Testergebnis vs. wahrer Krankheitsstatus

Da angenommen wird, dass auch der Referenztest eine FNR und FPR besitzt, müssen α_1 und β_1 für den Indextest und α_2 und β_2 für den Referenztest angenommen werden.

Die erwartete Zellproportion p_{ijk} für die Felder der Vierfeldertafel in Studie k ($i = 1, 2; j = 1, 2$) jeder Studie stellt sich folgendermaßen dar:

Indextest	Referenztest	
	positiv	Negativ
positiv	$N_k [\theta_k (1 - \beta_1) (1 - \beta_2) + (1 - \theta_k) \alpha_1 \alpha_2]$	$N_k [\theta_k (1 - \beta_1) \beta_2 + (1 - \theta_k) \alpha_1 (1 - \alpha_2)]$
negativ	$N_k [\theta_k \beta_1 (1 - \beta_2) + (1 - \theta_k) (1 - \alpha_1) \alpha_2]$	$N_k [\theta_k \beta_1 \beta_2 + (1 - \theta_k) (1 - \alpha_1) (1 - \alpha_2)]$

$N_k = \sum \sum n_{ijk}$ = Gesamtzahl der Personen in Studie k
 θ_k = Prävalenz der Krankheit in Studie k
 $(1 - \beta_1)$ = Sensitivität des Indextests; $(1 - \beta_2)$ = Sensitivität des Referenztests
 $(1 - \alpha_1)$ = Spezifität des Indextests; $(1 - \alpha_2)$ = Spezifität des Referenztests

Abbildung 6: Erwartete Häufigkeit der Ergebnisse für Studie k in Metaanalysen; Indextest vs. Referenztest

Die beobachteten Daten bestehen somit aus beobachteten Häufigkeiten n_{ijk} , die eine multinomiale Stichprobenverteilung mit der Wahrscheinlichkeit p_{ijk} korrespondierend zur n_{ijk} Zelle haben. Zusammengefasst haben die Daten aller beitragenden Studien eine Wahrscheinlichkeit, die durch das Produkt der multinominalen Wahrscheinlichkeiten der Studien gegeben ist.

Für die mit dem natürlichen Logarithmus transformierte Wahrscheinlichkeit ergibt sich dann:

$$\ln L = \sum_{k=1}^K \sum_{i=1}^2 \sum_{j=1}^2 n_{ijk} \ln(p_{ijk})$$

Mit Hilfe eines Algorithmus werden dann die Maximum Likelihood Schätzer der Parameter iterativ aus der log Likelihood abgeleitet, angepasst werden α_1 , α_2 , β_1 , β_2 und $\{\theta_k, k = 1, 2, \dots, K\}$.

Wenn die Parameterschätzer vorhanden sind, können im 2. Schritt die angepassten Häufigkeiten für wahre erkrankte und nicht erkrankte Personen mit positiven oder negativen Indextestergebnis jeder Studie berechnet werden. Mit diesen angepassten Häufigkeiten lassen sich dann adjustierte Schätzer der Sensitivität und Spezifität für jede Primärstudie berechnen. Bei diesem Schritt wird die Annahme getroffen, dass der Fehler des Indextests und des Referenztests konstant zwischen den Studien ist.

Drittens wird mit den korrigierten Sensitivitäten und Spezifitäten ein Regressionsmodell nach Moses et al.⁵⁵ gebildet und eine adjustierte SROC-Kurve angepasst.

Ein derartig angepasstes Modell führt zu einer substantiellen Verbesserung der diagnostischen Genauigkeit.

Hawkins et al.³¹ führen aber kritisch an, dass die Annahme, die Fehler des Index- und Referenztests sind unabhängig, um ein „Latent-class“ Modell zu konstruieren, häufig nicht eingehalten werden kann. Rutter et al.⁶⁷ merken zu dieser Methode an, dass für die Adjustierung die Sensitivität und Spezifität und die Fehlerrate des Referenztest konstant über die Studien ist. Diese Annahme kann nicht immer getroffen werden.

4.3.8.2 Likelihood Ratios

Als weitere Methode für kontinuierliche Testergebnisse zeigen Irwig et al.³³ die Erstellung von ergebnisspezifischen Likelihood Ratios auf. Wie bereits oben erwähnt ist die Likelihood Ratio als Ratio aus der Wahrscheinlichkeit, dass eine gegebenen Höhe eines Testergebnisses bei erkrankten Personen eintritt, zur Wahrscheinlichkeit, dass dieses Testergebnis bei nicht erkrankten Personen eintritt, definiert und ist vor allem für den Kliniker relevant.

Um ergebnisspezifische Likelihood Ratios zu erhalten, kann der natürliche Logarithmus der posterior Odds der Erkrankung als Funktion der Testergebnisse modelliert und in ein log Likelihood Ratio konvertiert werden, in dem man eine Konstante hinzufügt, die für die Ratio der Anzahl der richtig gesunden Personen zur Anzahl der richtig kranken Personen adjustiert:

$$\log(LR) = \log[N_D/N_{D+}] + \alpha + \beta x$$

wobei

LR = Likelihood Ratio

$\log[N_D/N_{D+}]$ = der Korrekturfaktor, um die log posterior Odds in $\log(LR)$ zu konvertieren

= $\log(\text{Anzahl gesunder Personen}/\text{Anzahl kranker Personen})$

α = Schnittpunkt mit der Y-Achse im logistischen Regressionsmodell mit der posterior Odds als abhängiger Variable

β = Regressionskoeffizient für die Testmessung im logistischen Regressionsmodell mit der posterior Odds als abhängiger Variable

x = Testmessung

Für kontinuierliche Testergebnisse muss angenommen werden, dass die Assoziation linear ist und dass die einzelnen Studien sich nicht in ihrer Kalibrierung unterscheiden.

Bislang gibt es wenige Beispiele für diese Vorgehensweise in der Literatur. Das könnte darauf zurückzuführen sein, dass in der Praxis meist doch Testergebnisse dichotomisiert werden, etwa um die Entscheidungsfindung in der Klinik zu simulieren.

4.3.8.3 Neuere statistische Ansätze

Die folgenden neueren Modelle und Ansätze sollen der Vollständigkeit halber aufgeführt werden, ihre Praktikabilität muss im Einzelfall aber noch unter Beweis gestellt werden:

Bayes'sche Ansätze für die Metaanalyse von ROC-Kurven³²: Während die klassischen Fixed- und Random-Effects-Modelle für Metaanalysen binominaler ROC-Kurven Parameter als fest und Wahrscheinlichkeiten als Häufigkeiten annehmen, benutzt der Bayes'sche Ansatz Wahrscheinlichkeitsverteilungen, um Unsicherheiten von unbekanntem Größen zu quantifizieren. Es werden empirische Bayes'sche Modelle und komplett Bayes'sche hierarchische Modelle zur Analyse von ROC-Kurven für binormale Modelle diskutiert.

Hierarchisches Regressionsmodell: Rutter et al.⁶⁷ erweitern ihr latent-scale logistic regression Modell durch ein hierarchisches Regressionsmodell. Dieses Modell ermöglicht es, Informationen über Studien hinweg zu poolen und geglättete Schätzer der Effekte der Kovariaten, der Komponenten der Varianz (between und within, systematisch und zufällig) und der individuellen Studienquantitäten zu erhalten. Durch einfache Erweiterung der hierarchischen Struktur können auch Informationen auf Patientenebene eingeschlossen werden, so weit diese verfügbar sind. Die Vorteile dieses hierarchischen SROC-Modells erfordern aber Markov-Ketten Monte Carlo Simulationen, was die Berechnung sehr aufwändig macht.

Kester et al.⁴⁰ beschreiben eine parametrische Methode, um ROC-Kurven zu poolen. Als Input für diese Analyse wird nur die veröffentlichte Kurve, die Anzahl der erkrankten und nicht erkrankten Personen benötigt. Aufbauend auf den Modellen von Kardaun et al.³⁸ und

Moses et al.⁵⁵ entwickeln sie ein Modell, um Studien zu poolen, die eine ROC-Kurve präsentieren und nicht nur ein einziges Sensitivität / Spezifität-Paar.

Einen weiteren Ansatz präsentiert Lloyd in einer Veröffentlichung aus dem Jahr 2000⁴⁷, in der er eine Familie von Regressionsmodellen beschreibt, die die ROC-Kurven mit einem Qualitätsparameter Δ und einem Formparameter μ beschreiben.

4.3.9 Zusammenfassung

Diagnostische Studien mit binären Ergebnissen (Test positiv oder negativ) stellen den häufigsten Fall in der Literatur dar. Die Daten von Studien mit binären Outcomes zur diagnostischen Genauigkeit können in Vierfeldertafeln extrahiert und die Standardgrößen Sensitivität, Spezifität und die Likelihood Ratios entsprechend daraus berechnet werden.

Mit welcher Methode die diagnostische Genauigkeit von Primärstudien am besten zusammengefasst werden kann, hängt von Annahmen ab, die getroffen werden, um die beobachteten Unterschiede zu erklären.

Das einfachste aber zugleich auch das restriktivste Modell geht von der Annahme aus, dass sich die Studien weder in ihrem Grenzwert noch in ihrer diagnostischen Genauigkeit unterscheiden, d. h. alle Primärstudien implementieren den diagnostischen Test identisch, benutzen einen identischen Grenzwert und setzen den Test bei identischen Populationen ein⁶⁸. Unter diesen Annahmen können die Felderbesetzungen der Vierfeldertafel jeder Primärstudie gepoolt und gemeinsame Parameter für FPR und TPR aller eingeschlossenen Primärstudien berechnet werden. Diese Methode hat jedoch mehrere Limitationen: Zum einen impliziert ein großer p-Wert bei der Überprüfung der statistischen Heterogenität nicht notwendigerweise einen starken Hinweis für Homogenität, denn dieser kann auch auf einer fehlenden statistischen Power beruhen. Zum anderen wird keine Korrektur für multiples Testen integriert. Wenn ein Parameter heterogen ist, würden diese gepoolten Maße zur Unterschätzung der wahren diagnostischen Genauigkeit führen.

Ein weniger restriktives Modell geht von der Annahme aus, dass der diagnostische Test mit der gleichen Genauigkeit in allen Primärstudien durchgeführt wird, aber unterschiedliche Grenzwerte definiert werden. Unter diesen Annahmen würden die Studien am besten mit einer *summary receiver operating characteristics* - Kurve zusammengefasst werden. Am häufigsten wird die Erstellung einer SROC-Kurve nach der Methode von Moses et al.⁵⁵ für Fixed-Effects-Modelle verwendet. Diese geht von der Annahme aus, dass eine linearer Zusammenhang zwischen dem $\text{logit}(\text{TPR})$ und dem $\text{logit}(\text{FPR})$ besteht. Die abhängige Variable D

stellt den Logarithmus des diagnostischen Odds Ratios dar. Durch eine Rücktransformation der errechneten Regressionslinie, lässt sich eine konventionelle ROC-Kurve als SROC-Kurve erstellen. Damit sind aber keine Gesamtwerte für TPR und FPR direkt verfügbar, sondern für einen ausgewählten FPR-Wert lässt sich über die SROC-Kurve der korrespondierende TPR-Wert ermitteln und umgekehrt. Diese Methode ist für Studien geeignet, deren Ergebnisse in Form von Vierfeldertafeln dichotomisiert präsentiert werden und denen ein Vergleich mit einem Goldstandard zugrunde liegt. Das Verfahren besticht durch seine intuitive Erfassbarkeit bedingt durch die graphische Darstellung. Letztlich resultiert eine Kurve, die sofort die Güte des Testverfahrens erkennen lässt. Das Moses-Modell scheint weniger anfällig für „Ausreißer“ zu sein, wie ein Vergleich von drei verschiedenen Fixed-Effects-Modellen zeigte⁶⁶. Ein weiterer Vorteil dieser Methode ist, dass es nicht die Annahme erfordert, dass die Varianzen der zugrundeliegenden kontinuierlichen Verteilungen der richtig positiven und negativen Ergebnisse gleich sind³⁰.

Ein Nachteil dieses Modells ist aber, dass eine relevante Spannbreite für TPR und FPR a priori definiert wird. Sind diese Grenzen festgelegt, hat dies folgende Konsequenzen: nur Studien mit Ergebnissen innerhalb dieser Grenzen werden in die Analyse eingeschlossen und die SROC-Kurve wird nur für diese Spannbreite berechnet.

Eine weitere Methode für Metaanalysen von diagnostischen Tests mit binären Ergebnissen und Goldstandardinformationen ist eine „latent-scale“ logistische Regression (LSLR)⁶⁶, um eine oder mehrere SROC-Kurven anzupassen. Das LSLR-Modell geht von zwei latenten logistischen Verteilungen der dichotomen Testergebnisse aus, eine für die erkrankte und eine für die nicht erkrankte Population. Die Modelle schließen Studienindikatorvariablen als Kovariaten ein, um der Within-Study-Variation in Fixed-Effects-Modellen Rechnung zu tragen. Random-Effects-Modelle liefern zusätzlich den Rahmen, um explizit sowohl die Within-Study-Variation und als auch die Between-Study-Variation zu berücksichtigen.

Ein Vorteil der LSLR ist, dass sie keine Kontinuitätskorrektur, wie zum Beispiel die Methode nach Moses et al., braucht, und somit die Ergebnisse dadurch nicht verzerrt werden. Aber das LSLR Modell kann nur eingesetzt werden, wenn entweder die diagnostische Genauigkeit oder der Grenzwert über die Studien fixiert ist und wenn bestimmte Annahmen hinsichtlich der Verteilung der zugrundeliegenden latenten Variablen erfüllt sind⁶⁶.

Die Dichotomisierung von Testergebnissen, um Sensitivität und Spezifität zu erhalten, führt zu einem Informationsverlust. Methoden zur Berücksichtigung von kontinuierlichen oder

wenigstens ordinal skalierten Daten sollten für Metaanalysen eingesetzt werden, wenn in den Primärstudien derartige Daten präsentiert werden. Wenn in den Primärstudien für die gleiche Anzahl von Kategorien Daten vorhanden sind, kann für jede Primärstudie eine ROC-Kurve erstellt werden und eine Gesamt-ROC-Kurve mittels ordinaler Regressionstechniken erstellt werden³³.

Als weitere Methode bei kontinuierlichen Daten kann auch die Berechnung der standardisierten Mittelwertdifferenzen angewandt werden. Für Metaanalysen von Studien, die (annähernd normalverteilte) kontinuierliche Testergebnisse berichten, wurde die Kalkulation der standardisierten Differenz der empirischen Mittelwerte vorgestellt^{5;30;73}. Diese kann eingesetzt werden, wenn die beiden Annahmen, dass die zugrundeliegenden kontinuierlichen Verteilungen nahezu normal (logistisch) sind und gleiche Varianzen haben, erfüllt sind. Sonst kommt es zur Verzerrung der Ergebnisse.

Dabei wird d als Effektgröße berechnet. d ist ein Maß für die Diskriminierungsfähigkeit oder Wirksamkeit (*test effectiveness score*) des untersuchten Tests⁷³. Je größer d , desto größer ist die diskriminatorische Fähigkeit des Tests. Ein Score von 1 impliziert einen wenig effektiven Test. Auch kann d in eine SROC-Kurve transformiert werden. Für gegebene Spezifitäten und d lässt sich die korrespondierende Sensitivität errechnen.

Der *test effectiveness score* d hat zusammenfassend folgende Vorteile: Er ist eine einzige Maßzahl. Er kann sehr leicht berechnet werden. Er ist unabhängig vom Grenzwert des Tests und von der Prävalenz. Er ist näherungsweise normalverteilt, so dass Konfidenzintervalle sehr leicht berechnet werden können. Sensitivität und Spezifität können einfach daraus abgeleitet und eine ROC-Kurve erstellt werden. Wie bereits oben erwähnt, müssen aber die beiden Voraussetzungen Normalverteilung und gleiche Varianzen erfüllt sein. Wenn die Varianz nicht gleich ist, ist die Maßzahl nicht unabhängig vom Grenzwert.

Als weitere Methode für kontinuierliche Testergebnisse zeigen Irwig et al.³³ die Erstellung von ergebnisspezifischen Likelihood Ratios auf. Wie bereits oben erwähnt ist die Likelihood Ratio als Ratio aus der Wahrscheinlichkeit, dass eine gegebenen Höhe eines Testergebnisses bei erkrankten Personen eintritt, zur Wahrscheinlichkeit, dass dieses Testergebnis bei nicht erkrankten Personen eintritt, definiert und ist vor allem für den Kliniker relevant. Für kontinuierliche Testergebnisse muss angenommen werden, dass die Assoziation linear ist und dass die einzelnen Studien sich nicht in ihrer Kalibrierung unterscheiden. Bislang gibt es wenige Beispiele für diese Vorgehensweise in der Literatur. Das könnte darauf zurückzuführen sein, dass

in der Praxis meist doch Testergebnisse dichotomisiert werden, etwa um die Entscheidungsfindung in der Klinik zu simulieren.

In Kapitel 4.3.8 sind noch weitere statistische Methoden aufgeführt werden, ihre Praktikabilität muss aber noch unter Beweis gestellt werden.

4.4 Bestimmung der Heterogenität

Die Ergebnisse von Einzelstudien zeigen oft Heterogenität sowohl innerhalb der Studien, wie auch zwischen den Studien. Studienergebnisse können sich aufgrund von Stichprobenfehlern unterscheiden. Auch wenn der wahre Effekt in jeder Studie gleich ist, werden die Ergebnisse unterschiedlicher Studien um den wahren gemeinsamen Effekt streuen (Varianz innerhalb der Studien, Fixed-Effects-Modell). Jede Studie kann aber auch von einer unterschiedlichen Population hinsichtlich der Patientencharakteristika und der Studienbedingungen gezogen werden. So kann es, auch wenn eine große Anzahl von Patienten in die Studien aufgenommen werden, zu unterschiedlichen Ergebnissen kommen (Varianz zwischen den Studien, Random-Effects-Modell)⁴⁵. Gründe für Heterogenität zwischen den Studien können u. a. unterschiedliches Studiendesign, verschiedene Grenzwerte für positive Testresultate, Unterschiede in der technischen Ausrüstung und verschiedene Patientenpopulationen sein⁶⁶.

Thompson unterscheidet zwischen einer klinischen und einer statistischen Heterogenität^{74;75}. Unter klinischer Heterogenität werden klinisch bedeutsame Unterschiede der Patientencharakteristika und des Patientenmanagements verstanden. Sowohl klinische Heterogenität als auch methodische Unterschiede zwischen den Studien können zu einer Inkompatibilität der quantitativen Ergebnisse führen und damit zu statistischer Heterogenität.

4.4.1 Quellen der Heterogenität

Folgende Quellen der Heterogenität sollen ausführlich dargestellt werden:

4.4.1.1 Studiendesign

Als optimales Design, um die Genauigkeit von diagnostischen Tests zu untersuchen, gilt der prospektive, verblindete Vergleich eines Tests mit einem Referenztest an einer konsekutiven Serie von Patienten aus einer relevanten klinischen Population. Als relevante klinische Population wird eine Gruppe von Patienten verstanden, die das Krankheitsspektrum abdeckt, das dem Test gegenwärtig oder zukünftig ausgesetzt wird⁴⁶. Retrospektive Designs benutzen hauptsächlich bereits vorhandene Routinedaten mit den damit verbundenen Verzerrungsmöglichkeiten und Limitationen.

Das Studiendesign wird dann verkompliziert, wenn unterschiedliche diagnostische Tests verglichen werden: Die Evaluation jedes Tests wird entweder prospektiv oder retrospektiv sein, bei den Patienten können zufällig oder nicht zufällig unterschiedliche Tests durchgeführt werden, oder alle Tests gleichzeitig. Durch die Stratifikation der Studien gemäß des Studiendesigns kann der Einfluss dieser auf die Testgenauigkeit überprüft werden ⁴⁶.

4.4.1.2 Methodische Qualität

Aufgrund von Unterschieden im Referenztest, seinem Grenzwert, der Unabhängigkeit der Beobachtungen oder Verifikationsbias kann die interne Validität einzelner Studien variieren ^{2;34;46}.

- Referenztest und sein Grenzwert: Autoren von Übersichtsarbeiten von Studien zur diagnostischen Genauigkeiten sollten entscheiden, welcher Referenztest der am besten verfügbare ist. Häufig besteht kein Konsens über den zu verwendenden Referenztest und eine Reihe von unterschiedlichen Referenztest wird von den Primärstudien eingesetzt. Wenn kein eindeutiger Referenzstandard verfügbar ist oder nicht bei allen Patienten eingesetzt werden kann, kann der tatsächliche Krankheitsstatus gelegentlich durch weitere Beobachtung des Patienten festgestellt werden. Unvollständigkeit führt aber zu einer Misklassifikation der Patienten ¹⁸. Falls es einen eindeutigen Referenztest (Goldstandard) gibt, können trotzdem unterschiedlich Techniken oder Grenzwerte verwendet werden.
- Unabhängigkeit der Beobachtungen: Um eine Verzerrung zu vermeiden, sollte das Testergebnis unabhängig vom Ergebnis des Referenztests ermittelt werden und umgekehrt. In der Praxis wird der Referenztest, der häufig teurer oder invasiver ist, erst nach dem zu untersuchenden Test durchgeführt. Wie und ob, die Tests „blind“ durchgeführt wurden, muss daher dargestellt werden. Auch das Wissen um die Krankengeschichte oder der klinischen Untersuchung können das Test- und/oder Referenztestergebnis beeinflussen und somit zur Heterogenität beitragen ¹⁸.
- Verifikationsbias: Idealerweise, sollten alle Patienten, bei denen der Indextest durchgeführt werden, dem Referenztest unterzogen werden. Falls dies nicht der Fall ist, sollten die Patienten, bei denen der Referenztest durchgeführt wird, unabhängig vom Ergebnis des Indextests ausgewählt werden. Leider ist dies nicht immer der Fall. Solange die Stichproben zufällig ausgewählt werden und die Größe der einzelnen Stichproben von den Patienten mit verschiedenen Testergebnissen bekannt ist, kann für diesen Bias adjustiert werden ².

4.4.1.3 Patientenmerkmale

Um den Einfluss der Unterschiede zwischen den Patientenpopulationen der Primärstudien abschätzen zu können, sollten Informationen zu demographischen Variablen, Ein- und Ausschlusskriterien, Zeitpunkt des Auftretens von Symptomen, Komorbidität und anderen klinischen und diagnostischen Informationen, und zu spezifischen Details der Erkrankung vorliegen, die die Performance des Indextests beeinflussen könnten¹⁸.

4.4.1.4 Merkmale des Tests

Hier können unterschiedliche Messtechniken, Messeinheiten, die Personen, die den Test durchführen oder interpretieren, Quellen der Heterogenität sein. Auch der Grenzwert kann entscheidend zwischen den Primärstudien variieren. In einigen Studien kann der Grenzwert explizit unterschiedlich gewählt worden sein, in anderen kann es durch natürliche Variationen aufgrund unterschiedlicher Gutachter oder untersuchender Labore zu Unterschieden im Grenzwert kommen. Der Wahl des Grenzwertes kann auch in Abhängigkeit der Prävalenz der Erkrankung getroffen worden sein. Wenn es sich um eine seltene Erkrankung handelt, kann der Schwellenwert niedrig gewählt worden sein, um eine große Anzahl falsch positiver Ergebnisse zu vermeiden.

4.4.2 Messung der Heterogenität

Die Messung und Interpretation der Interstudien-Heterogenität ist allerdings ein noch wenig systematisch bearbeitetes Problem. Fixed-Effects-Modelle (wie auch das hier angewandte von Moses et al. 1993) können Unterschiede zwischen den Studien als Kovariablen berücksichtigen, was aber dazu führt, dass (möglicherweise irrtümlich) alle Unterschiede zwischen den Studien den selektierten Kovariablen zugeschrieben werden. Random-Effects-Modelle berücksichtigen die Heterogenität zwischen Studien zwar stärker, machen aber auch mehr Annahmen hinsichtlich der Verteilung der Variablen. Zudem ist der Nutzen der komplexeren Berechnung eines Random-Effects-Modells unklar, da zumeist vergleichbare Effektschätzer bei lediglich größerem Konfidenzintervall resultieren³⁰.

Eine Möglichkeit, die Interstudien-Heterogenität in Adaptation des Cochran's Q-Test^{13 12} zu bestimmen, ist bei Hasselbad und Hedges (1995) angegeben. Für ein Fixed-Effects-Modell lautet die entsprechende Formel unter der Annahme von m Studien:

¹³ $Q = \sum w_i (y_i - \hat{\mu})^2$, mit $w_i = 1/\text{Varianz der Einzelstudien}$, $y_i = \text{individuelle Effektgröße jeder Primärstudie}$, $\hat{\mu} = \text{Mittelwert der Effektgröße}$.

$$\chi^2 = \sum_{j=1}^m w_j (d_j - \hat{d})^2$$

wobei w_j für die 1/Varianz der Einzelstudien steht,

d_j für die Effektschätzer der Einzelstudien.

\hat{d} der mittlere, gewichtete Effektschätzer, wobei Studien mit kleinerem Standardfehler größeres Gewicht gegeben wird.

Das Ergebnis folgt näherungsweise einer χ^2 -Verteilung mit $m-1$ Freiheitsgraden. Die Nullhypothese, dass alle d_j den gleichen Parameter schätzen (Homogenität), muss dann zurückgewiesen, wenn χ^2 einen hohen Wert annimmt.

Weitere Möglichkeiten sind die Bestimmung der Heterogenität nach Wald, Breslow-Day^{8;9} und der χ^2 -Test.

Wenn die einzelnen Studien eine kleine Fallzahl haben, kann es aufgrund der daraus folgenden niedrigen Power dazu kommen, dass eine Heterogenität nicht statistisch nicht entdeckt wird, obwohl sie vorhanden ist.

Wenn sich statistische Heterogenität bestätigt, sollte nach der Ursache geforscht werden, d.h. Ausreißer identifiziert werden und der Einfluss klinischer oder methodischer Unterschiede zwischen den Studien betrachtet werden¹⁸.

4.4.2.1 Ausreißer

Die Identifikation und der Umgang mit Ausreißern im Rahmen von Analysen ist generell nicht einfach. Wenn eine Ausreißerstudie gefunden wird, sollte man immer nach plausiblen Gründen suchen, diese auszuschließen¹⁸. Die Suche nach Ausreißern kann wertvolle Informationen über den Effekt von Unterschieden im Studiendesign oder von anderen Ursachen der Heterogenität auf diagnostische Genauigkeit liefern.

Es gibt viele unterschiedliche Möglichkeiten, Ausreißer zu entdecken, vor allem graphische Darstellungen sind dabei hilfreich. Eine Möglichkeit ist die $\log(\text{odds}(\text{Sensitivität}))$ gegen die $\log(\text{odds}(1 - \text{Spezifität}))$ aufzutragen oder der sogenannte Galbraithplot, bei dem der $\ln(\text{DOR})/\text{SE}(\ln(\text{DOR}))$ auf der y-Achse gegen den reziproken Wert des Standardfehlers auf der x-Achse aufgetragen wird. Studien, die außerhalb des 95%-Konfidenzintervalls liegen, werden als Ausreißer bezeichnet²⁷.

4.4.2.2 Subgruppen

Subgruppen können gebildet werden, in dem Studien zusammengefasst werden, oder in dem, falls genügend Informationen aus den einzelnen Primärstudien entnommen werden können, bestimmte Patientengruppen (z. B. nach Geschlecht) gebildet werden. Falls unterschiedliche Referenztests benutzt wurden, besteht die Möglichkeit, nur die Studien zu berücksichtigen, die den besten verfügbaren Test eingesetzt haben, oder den, der am häufigsten verwendet wurde. Es können aber auch alle eingeschlossen werden und die Analyse wird stratifiziert durchgeführt.

Falls unterschiedliche Grenzwerte des Indextests die beobachteten Unterschiede nicht erklären, kann ein multiples Regressionsmodell oder eine stratifizierte Analyse benutzt werden, um mögliche Effektmodifizier zu identifizieren oder um homogene Subgruppen zu entdecken¹⁸. Unterschiede zwischen Subgruppen sollten analysiert werden und falls notwendig als Kovariaten in das Regressionsmodell eingeschlossen werden, um für potentielle Confounder zu adjustieren.

Je nach Art der Subgruppenanalysen sind unterschiedliche Punkte (z. B. selektives Berichten) bei der Auswertung zu berücksichtigen. Selektives Berichten stellt ein potentielles Problem in Metaanalysen dar: „Within study selective reporting of subgroups“ beschreibt die Beobachtung, dass Forscher dazu neigen, signifikante Ergebnisse von Subgruppenanalysen eher zu berichten als nicht signifikante²⁸. Es ist schwierig Evidenz für eine derartige Verzerrung zu finden. Hahn et al.²⁸ unterscheiden hierzu zwei Formen:

- (1) Wenn nicht alle Ergebnisse von Subgruppen veröffentlicht werden, sondern nur von einigen und die Auswahl der veröffentlichten von deren Ergebnissen abhängig ist.
- (2) Wenn Subgruppenanalysen durchgeführt werden, aber in Abhängigkeit des Ergebnisses nicht veröffentlicht werden.

Die Richtung, in die die Ergebnisse verzerrt werden, wenn nicht für alle Subgruppen, Ergebnisse aus den Primärstudien zur Verfügung stehen und folglich nur Subgruppenanalysen für einen Teil der Primärstudien durchgeführt werden, kann nicht definiert werden. Wie beim Publikationsbias (s. u.) kann die Erstellung eines Funnelplots Hinweise darauf geben, ob die Subgruppenergebnisse aufgrund eines selektiven Berichtens verzerrt wurden.

4.4.3 Zusammenfassung

Die Analyse der Heterogenität bietet viele Stolpersteine. Der Ausschluss von Ausreißern muss gut begründet sein und kann graphisch untersucht werden. Subgruppenanalysen machen

eine sorgfältige und vorsichtig Interpretation notwendig, besonders dann, wenn diese nicht a priori definiert wurden ⁶⁸. Post hoc Subgruppenanalysen sollten nur zur Generierung weiterer Forschungshypothesen verwendet werden.

Die schlechte Qualität vieler Studien zur diagnostischen Genauigkeit und die begrenzte Information über Studiencharakteristika, die aus den relevanten Veröffentlichungen hervorgeht, macht eine vorsichtige Herangehensweise und Interpretation der Ergebnisse notwendig. Die häufig geringe Anzahl von verfügbaren Primärstudien und die geringe Variation der Daten limitiert die Möglichkeiten der Subgruppenanalyse bzw. der multiplen Regressionsmodelle ¹⁸.

Es bleibt eine Herausforderung die Heterogenität zu untersuchen und zu einer klinisch relevanten Interpretation von statistisch heterogenen Ergebnissen zu kommen. Diese Heterogenität reflektiert den klinischen Alltag, der berücksichtigt werden muss, wenn man versucht, diagnostische Genauigkeit auf der Grundlage von primärer diagnostischer Forschung zusammenzufassen. Heterogenität muss nicht nur als ärgerliche statistische Tatsache betrachtet werden, sondern kann auch als Phänomen gesehen werden, welches genutzt werden kann, um mehr über die Komplexität der täglichen Krankenversorgung zu lernen ¹⁸.

Die Untersuchung der Heterogenität im Rahmen der diagnostischen Forschung gibt dem Wissenschaftler wertvolle Hinweise über wichtige Quellen der Variabilität in der Genauigkeit diagnostischer Studien.

4.5 Methoden zur Abschätzung von Publikationsbias

Obwohl Publikationsbias mehr ein Problem von Studien zu diagnostischen Tests als von RCTs ist, haben empirische Studien zu Publikationsbias sich bislang vor allem mit therapeutischen Studien befasst. Viele Studien zur diagnostischen Genauigkeit nutzen Daten, die primär im Rahmen der klinischen Routine gesammelt wurden, so dass es keine eindeutige Dokumentation von versuchten Evaluationen gibt. Studien, die zur Veröffentlichung gelangen, sind häufiger verzerrt und überschätzen wahrscheinlich die Testgenauigkeit.

Irwig et al. ³³ stellen fest, dass nicht das Fehlen geeigneter statistischer Verfahren, sondern neben der schlechten Qualität der Primärstudien Publikationsbias der hauptsächlich limitierende Faktor hinsichtlich der Erstellung valider Metaanalysen ist. In Random-Effects-Modellen werden die Schätzer stärker von kleineren Studien beeinflusst und die Gefahr der Verzerrung der Ergebnisse durch Publikationsbias ist damit größer.

Die Wahrscheinlichkeit, Forschungsergebnisse zu identifizieren, hängt nicht nur davon ab, ob eine Studie veröffentlicht wird, sondern auch davon, wann und wo und in welchem Format dies geschieht. Disseminationsbias ist eine weiter gefasster Begriff als Publikationsbias und schließt sowohl Publikationsbias als auch verwandte Biasformen, z. B. aufgrund von der Zeit, der Art, der Sprache, dem multiplen Publizieren, dem selektiven Zitieren der Referenzen, dem Database Index Bias und der verzerrten Aufmerksamkeit der Medien ein ⁶⁹.

4.5.1 Verzerrungsquellen

Im folgenden werden zunächst die einzelnen Biasformen dargestellt:

- Publikationsbias: selektive Veröffentlichung signifikanter Ergebnisse, während nicht signifikante Ergebnisse unveröffentlicht bleiben. Publikationsbias kann aufgrund unterschiedlicher Faktoren entstehen ⁷⁶:

Kleine Studien haben nur eine geringe statistische Power und Ergebnisse können nur statistisch signifikant sein, wenn große Effekte beobachtet werden („small-study“-Effekte). Wahrscheinlicher ist es aber, dass aufgrund der kleinen Stichprobe, wahre Zusammenhänge nicht nachgewiesen werden können. Dies führt zu Publikationsbias, denn kleine Studien werden wahrscheinlich nur veröffentlicht, wenn sie statistisch signifikante Ergebnisse haben.

Publikationsbias entsteht, wenn Forschergruppen Ergebnisse nicht einreichen, wenn diese nicht statistisch signifikant sind. Auch Herausgeber und Gutachter sind an statistisch nicht signifikanten Ergebnissen weniger interessiert.

Auch Studien von guter Qualität mit negativen Ergebnissen werden weniger häufig veröffentlicht als positive Ergebnisse von qualitativ schlechteren Studien. Auch der „Peer review“ ist keine verzerrungsfreie Methode.

Selbst wenn in Übersichtsarbeiten und Metaanalysen die größtmögliche Vollständigkeit aller Studien erreicht werden soll und damit auch unveröffentlichte Studien mit eingeschlossen werden, muss doch berücksichtigt werden, dass unveröffentlichte Studien eventuell weniger reliable Ergebnisse liefern, da sie noch nicht einem Revierprozess unterliegen und damit potentiell qualitativ schlechter sein könnten.

- Location Bias

English Language Bias: Signifikante Ergebnisse werden englischsprachig, nicht signifikante Ergebnisse werden nicht englischsprachig veröffentlicht, was die Identifizierung

und den Einschluss in Metaanalysen weniger wahrscheinlich macht²⁴, vor allem wenn nur englischsprachige Publikationen in Metaanalysen eingeschlossen werden. Qualitativ wurden keine signifikanten Unterschiede zwischen englischsprachigen und nicht englischsprachigen Veröffentlichungen gefunden^{24;36;53;54}.

Citation Bias: Entsteht aus dem English Language Bias, da nicht englischsprachige Veröffentlichungen weniger häufig zitiert werden und deshalb häufiger übersehen werden.

Multiple Publikationsbias: Ergebnisse positiver Studien werden häufig mehrmals veröffentlicht, dies erhöht die Wahrscheinlichkeit, dass sie für Metaanalysen identifiziert werden. Es ist oftmals schwierig, multiple Publikationen einer Studie als solche zu identifizieren und nicht als vermeintliche Primärstudien in die Analyse einzuschließen⁷⁶.

Publikationsbias wird als häufigste Ursache von Verzerrungen in den Ergebnissen von Metaanalysen angesehen. Eine Reihe von mathematischen und graphischen Testverfahren wurde entwickelt, um zu testen, ob in einer Metaanalyse Publikationsbias vorliegt.

4.5.2 Funnelplot

Die bekannteste Methode ist das graphische Verfahren mittels Funnelplot. Diese Methode trägt in einem Scatterplot die Effektschätzer (x-Achse) gegen ein Maß der Studiengenauigkeit (die Stichprobengröße oder der Standardfehler, y-Achse) der jeweiligen Primärstudien auf. Sie geht von der Annahme aus, dass die Präzision des Schätzers, den wahren Wert abzubilden, mit der Stichprobengröße ansteigt. Schätzer von kleinen Studien werden daher „im unteren Bereich der Graphik“ größer um den wahren Wert streuen, während sich die Streuung der Schätzer verringert, je größer die Stichprobengröße (bei geringerer Anzahl derartiger Studien) wird. Eine inverse trichterartige, symmetrische Darstellung der Ergebnisse mit einer vertikalen Achse durch den wahren Wert ergibt sich, wenn keine Verzerrung durch Bias vorliegt. Eine asymmetrische Verteilungen entsteht, wenn kleinere Studien mit negativen Ergebnissen nicht veröffentlicht werden, während die Wahrscheinlichkeit von größeren Studien oder Studien mit signifikanten Ergebnisse höher ist, veröffentlicht werden. Dieses Phänomen führt zu einer Überschätzung der Testgüte.

Asymmetrie, dargestellt in Funnelplots, kann aber nicht als beweisend für eine Verzerrung durch Bias gelten, sondern auch eine Heterogenität zwischen den Primärstudien kann zu asymmetrischen Funnelplots führen. Wenn keine Heterogenität zwischen den Studien besteht, sollten 95% Studien innerhalb des 95%-Konfidenzintervalls um den Gesamteffektschätzer liegen⁷¹.

4.5.3 Tests auf Bias

Es gibt kein universell akzeptiertes Maß, um die Testgenauigkeit in Metaanalysen von Screening- oder diagnostischen Daten zu messen. Sensitivität, Spezifität und prädiktive Werte stehen reziprok zueinander in Beziehung und sind von den gewählten Grenzwerten abhängig. Daher ist es nicht angemessen, die Schätzer der Sensitivität und Spezifität separat in Metaanalysen zu kombinieren³⁰. Um diesem auch bei der Untersuchung auf Publikationsbias Rechnung zu tragen, schlagen Song et al.⁷⁰ vor, als Maß für die Testgenauigkeit und deren Varianz, die Anzahl richtig positiver (TP), richtig negativer (TN), falsch positiver (FP) und falsch negativer (FN) Testergebnisse in einer einzigen Statistik, d , zusammenzufassen (siehe Hasselblad et al.³⁰):

$$d = \sqrt{3} (\log TP + \log TN - \log FP - \log FN) / \pi$$

$$\text{Varianz}(d) = 3 (1/TP + 1/FP + 1/FN + 1/TN) / \pi^2$$

Song et al.⁷⁰ setzen diesen Schätzer der diagnostischen Genauigkeit d ein, um das Vorhandensein von Publikationsbias in Metaanalysen diagnostischer Tests zu untersuchen.

Zwei einfache Methoden quantifizieren die Asymmetrie im Funnelplot und werden von Song et al.⁷⁰ zur Untersuchung eingesetzt:

Begg and Mazumdar entwickelten einen adjustierten Rang-Korrelations Test³, um die Assoziation zwischen den Schätzern der Testgenauigkeit und ihrer Varianz untersuchen. Die Abweichung von Spearmans *rho*-Wert (Kendalls Tau) von Null liefert ein Maß für die Funnelplotasymmetrie⁷⁰. Positive Werte legen einen Trend zu höherer Testgenauigkeit in Studien mit kleineren Stichprobengrößen nahe. Da die Varianz umgekehrt proportional zur Stichprobengröße ist, wird im eigentliche Sinne die Stichprobengröße mit der Effektgröße korreliert.

Egger et al.²² passen ein lineares Regressionsmodell basierend auf dem natürlichen Logarithmus des Effektschätzers an. Die Standardnormalabweichung (SND), definiert als der Effektschätzer d geteilt durch seinen Standardfehler $SE(d)$, wird als Funktion der Präzision des Schätzers, definiert als der Kehrwert des Standardfehlers $SE(d)$, dargestellt:

$$SND = \alpha + \beta * [SE(d)]^{-1}$$

α ist der Schnittpunkt mit der y-Achse und β die Steigung der Geraden.

Da die Genauigkeit vor allem von der Stichprobengröße abhängt, werden kleine Studien vor allem nahe der 0 abgebildet werden. Die Punkte homogener Primärstudien, die nicht von Se-

lektionsbias verzerrt werden, streuen demnach um eine Gerade, die durch den Ursprung verläuft und eine Steigung β besitzt, die ein Maß für die Größe und die Richtung des zugrundeliegenden wahren Effekts ist. Falls eine Asymmetrie besteht, d. h. wenn kleinere Studien systematisch zu anderen Ergebnissen als große Studien kommen, wird die Regressionsgerade nicht durch 0 verlaufen.

Das Intercept α ist ein Maß für die Asymmetrie des Funnelplots. Je größer die Abweichung vom Ursprung ist, um so ausgeprägter ist die Asymmetrie. Positive Werte ($\alpha > 0$) weisen auf den Trend hin, dass kleinere Studien eine größere diagnostische Genauigkeit aufweisen⁷⁰.

Beide Methoden haben ihre Schwachpunkte: Sie haben nur eine geringe statistische Power, wenn Metaanalysen auf 10 oder weniger Primärstudien basieren⁷². Die Methode nach Beggs besitzt nur eine geringe Power bei kontinuierlichen Daten, besonders wenn es nur eine geringe Anzahl von Primärstudien gibt, kann bei einem negativen Ergebnis die Abwesenheit von Publikationsbias nicht ausgeschlossen werden. Die Ergebnisse nach der Methode von Egger können in sich selbst verzerrt sein⁴⁹ und der Nachweis der Asymmetrie lässt noch keinen Rückschluss auf die Gründe zu.

Bei der „Trim and Fill“- Methode²¹ wird eine iterative Rang-basierte Datenaugmentation verwendet, um die Anzahl der fehlenden Studien abzuschätzen und um adjustierte Schätzer der Testgenauigkeit zu generieren, dadurch dass fehlende Studien „imputiert“ werden. Die grundsätzliche Annahme, die getroffen wird, ist, dass es neben einer Anzahl beobachteter Studien n , eine zusätzliche Anzahl von nicht beobachteten Studien k gibt. k soll mit dieser Methode abgeschätzt werden. Dabei wird zuerst die Anzahl der „asymmetrischen“ Studien auf der einen Seite des Trichters geschätzt. Dies sind die Studien, die auf der anderen Seite kein „Gegenüber“ haben. Diese Studien werden aus dem Trichter entfernt („getrimmt“). So entsteht ein symmetrischer Trichter. Von diesem wird das wahre Zentrum des Trichters mit Standard-Metaanalyse-Methoden geschätzt wird. Diese Schritte werden iterativ durchgeführt, bis die finalen Schätzer für den Effekt und für k gefunden sind. Die entfernten Studien werden wieder zurückgesetzt und die fehlenden „Spiegelbilder“ symmetrisch um die Achse des vorher gepoolten Schätzers platziert. Die finale Schätzer und dessen Varianz werden dann mit dem wieder aufgefüllten Funnelplot (beobachtete und aufgefüllte Studien) erneut angepasst. Sowohl Fixed-Effects als auch Random-Effects Modelle können genutzt werden, um den Impact der Modellwahl auf den Publikationsbias zu untersuchen. Positive Werte ($k > 0$) zeigen an, dass ein Trend besteht, die diagnostische Genauigkeit aufgrund der fehlenden Studien zu

überschätzen und damit Publikationsbias vorliegt. Eine geschätzte Anzahl fehlender Studien $k > 3$ wird als signifikantes Ergebnis gewertet. Mit dieser Methode können für den Bias adjustierte Schätzer der diagnostischen Genauigkeit berechnet werden und damit auch die Größe der Überschätzung.

Beispiel: Asymmetrische Funnelplots und Publikationsbias in Metaanalysen zur diagnostischen Genauigkeit (nach Song et al., 2001⁷⁰)

Song et al. fanden in dieser Untersuchung von 20 systematischen Reviews mit 28 durchgeführten Metaanalysen zu diagnostischen Tests, dass alle identifizierten Metaanalysen die MEDLINE Datenbank zur Literatursuche nutzten. Nur 6 Reviews schlossen noch weitere Datenbanken ein, vier kontaktierten Autoren, 2 unternahmen auch eine Handsuche, um relevante Studien zu identifizieren. Nur 2 Reviews gaben an, keine Restriktionen hinsichtlich der Sprache getroffen zu haben, während 12 solche spezifizierten. Keiner der Reviews hat als Teil der Datensynthese die Primärstudien hinsichtlich des Vorhandenseins von Publikationsbias untersucht. 23 Metaanalysen wiesen eine positive Korrelation nach der Rang-Korrelations Methode auf (6 davon signifikant). 25 von 28 Metaanalysen wiesen ein positives Intercept nach der Methode von Egger et al. auf als Hinweis für eine Asymmetrie, 12 davon waren signifikant. Nach der „Trim and Fill“-Methode mit Random-Effects-Modellen hatten 17 Studien fehlende Studien, in 7 Studien war dabei die Anzahl der fehlenden Studien größer als 3, was ein signifikantes Ergebnis nahe legt. Fixed-Effects-Modelle kamen zu ähnlichen Ergebnissen.

Darüber hinaus wurden weitere Verfahren beschrieben und diskutiert⁷⁶. Hierzu zählen neben graphischen Methoden auch formale Tests, die beispielsweise beobachtete Assoziationen erklären (Hackshaw-Methode) und komplexe Modellierungen (Given-Methode). Die Methode von Hackshaw beruht darauf, dass die Anzahl tatsächlich veröffentlichter Studien mit der Anzahl von Studien verglichen wird, die man bei Abwesenheit einer wahren Assoziation erwarten würde, bei gegebenem Signifikanzniveau und der Anzahl der Studien, die bei diesem Niveau signifikant sind. Der Effekt des Publikationsbias auf die Ergebnisse der Metaanalyse kann damit abgeschätzt werden⁷⁶. Die Methode nach Given beruht auf einem Bayes'schem Modell, das die beobachteten Daten durch die Simulation der Ergebnisse von fehlenden Studien vergrößert und dadurch einen „vollständigen“ Datensatz für die Metaanalyse erzeugt. In diesem Modell werden die beobachteten Daten Y als teilweise Realisation der Zufallsvariable $X = (YZ)$ betrachtet. Die vollständige Realisation X stellt den kompletten Datensatz dar, die Realisation Z sind die fehlenden oder latenten Daten. Im Falle von Publikationsbias werden die nicht veröffentlichten Daten als die latenten Daten behandelt, die die beobachteten Daten augmentieren. Random-Effects-Modell können dann ausgeweitet werden, um diesen Publikationsbias einzubeziehen. Die Anzahl der nicht publizierten Studien und deren Effekte sind unbekannt und müssen geschätzt werden, Unsicherheiten dieser Schätzer werden in der abschließenden Metaanalyse berücksichtigt, in dem sie als Parameter in der Bayes'schen Analyse integriert werden. Limitationen dieses

se integriert werden. Limitationen dieses komplexen Modells sind, dass es zahlreiche Annahmen für das Modell trifft und eine debattierbare Wahl an Prior-Verteilungen trifft ⁷⁶.

4.5.4 Zusammenfassung

Wenn es auch bisher keinen internationalen Konsens über das am besten geeignete Verfahren gibt, so kann dennoch festgestellt werden, dass die Tests von Beggs und Mazumdar sowie Egger zu den gebräuchlichsten gehören und dementsprechend verwendet werden sollten. Neben der Durchführung eines statistischen Tests auf Publikationsbias ist die Erstellung eines Funnelplots in jeder Metaanalyse sinnvoll.

Methoden, Publikationsbias zu vermeiden:

- Bemühungen, alle relevanten Studien zu identifizieren, nicht nur elektronische Datenbanken, sondern auch Handsuche, Kontakt mit Experten.....,
- Keine Restriktionen hinsichtlich der Sprache,
- formale Untersuchung bei der Analyse der Daten,
- Schaffung von Registern. Da die Registrierung üblicherweise erfolgt bevor Ergebnisse bekannt werden, würde dies mögliche Effekte von Publikationsbias minimieren.

4.6 Weitere Aspekte von Metaanalysen diagnostischer Studien

4.6.1 Sensitivitätsanalysen

Mittels Sensitivitätsanalysen kann die Robustheit einer Metaanalyse eingeschätzt werden. Sensitivitätsanalysen im Rahmen von Metaanalysen beinhalten die Wiederholung der Metaanalyse an Subsets der originalen Datensätze. Der Einfluss von Einzelstudien kann auch dadurch überprüft werden, dass für die Metaanalyse Studien sukzessive aus- bzw. eingeschlossen werden und die Änderung der Ergebnisse jeweils beobachtet wird. Dabei ist zu beachten, dass der Einfluss einer Einzelstudie auf das Gesamtergebnis viel stärker erscheint, wenn die Studie als eine der ersten in die kumulative Metaanalyse aufgenommen wird. Meistens werden die Studienqualität, Studiengröße, statistische Verfahren (FEM vs. REM) und eventuell andere Faktoren wie Publikationsjahr (zeitliche Robustheit ⁴⁶) oder Setting in die Sensitivitätsanalyse einbezogen. Sensitivitätsanalysen sind vor allem bedeutsam, wenn Annahmen notwendig waren (z. B. kann es unklar sein, ob eine Studie die Einschlusskriterien erfüllt, weil bestimmte Angaben in der Publikation fehlen) oder die Ergebnisse der Einzelstudien widersprüchlich sind ^{14,23}. In jedem Fall sollte sichergestellt werden, dass Sensitivitätsanalysen begründet werden und plausibel sind. Dies hängt von der jeweiligen

sen begründet werden und plausibel sind. Dies hängt von der jeweiligen Fragestellung der untersuchten Technologie ab.

4.6.2 Qualitative Auswertungen diagnostischer Studien

Die qualitative Auswertung von Studien zu diagnostischen Tests ist obligater Bestandteil von systematischen Übersichten im Rahmen von HTAs. Zum einen muss für Metaanalysen ohnehin eine Datenextraktion und Qualitätsbewertung durchgeführt werden. Zum anderen werden diagnostische Studien im Kontext zahlreicher Fragestellungen und deshalb auch Studiendesigns durchgeführt. Gemeinhin werden die bereits dargestellten sechs Evaluationsebenen unterschieden. Metaanalysen werden (bisher) nur für die diagnostische Genauigkeit durchgeführt. Dies reicht jedoch für eine umfassende Bewertung diagnostischer Studien nicht aus. Aus diesem Grund sollten noch weitere verfügbare Informationen herangezogen werden. Am häufigsten sind dabei Studien zum diagnostischen Impact zu finden. Die Auswertung sollte die in der Tabelle 7 aufgeführten Aspekte berücksichtigen ⁶¹.

Tabelle 7: Merkmale prospektiver und retrospektiver Studien zum diagnostisch-therapeutischen Impact diagnostischer Test

Berichtsteil	Merkmal
Design / Protokoll	<ul style="list-style-type: none"> • Setting • Untersuchungsabfolge, z. B. zeitliche Abfolge oder Stellung im Managementplan • Rekrutierung von Indexpatienten und Kontrollpatienten, Ein- und Ausschlusskriterien • Operationalisierung und Verifizierung der Befunde • Operationalisierung und Verifizierung der Outcomes • Festlegung der Beobachtungseinheit, z. B. Patienten, Organe, Proben
Beschreibung der Testverfahren	<ul style="list-style-type: none"> • technische Charakteristika, z. B. Gerätetyp, Hilfsmittel, Reagenzien, Test-Kits, vorbereitende Maßnahmen • Auswertungsalgorithmus bei computergestützten Verfahren
Patientenselektion	<ul style="list-style-type: none"> • Beschreibung der Studienpopulation, z. B. Stadienverteilung, Komorbidität, Alter, Geschlecht • Methode der Rekrutierung der Patienten und eventueller Kontrollen • Definition der Kohorte (bei retrospektiven Studien)
Auswertung / Interpretation der Daten	<ul style="list-style-type: none"> • Kenntnisstand des / der Auswerter(s) in Bezug auf Vor-Testergebnisse und Krankheitsstatus • blinde oder offene Auswertung • Definition / Klassifikation der Testergebnisse • Berücksichtigung ergänzender Informationen, z. B. Interviews (bei retrospektiven Studien)
Datenanalyse / statistische Auswertung	<ul style="list-style-type: none"> • Datenaufbereitung, Beschreibung und Begründung von Klassenbildung, z. B. Dichotomisierung kontinuierlicher Variablen • Berechnung von Effektschätzern, Angabe von statistischen Testverfahren • Umgang mit unklaren oder nicht interpretierbaren Befunden • Anzahl der korrekt und nicht korrekt durch den Test identifizierten Entitäten (Vierfeldertafel) • Angabe von abgeleiteten Effektschätzern, z. B. Sensitivität, Spezifität, prädiktive Werte sowie Konfidenzintervalle
Charakteristika	<ul style="list-style-type: none"> • Anzahl der untersuchten Patienten bzw. Organe

von Indexpatienten und Kontrollpatienten / Patientenfluss	<ul style="list-style-type: none"> • Beschreibung der Studienpopulation, z. B. Stadienverteilung, Komorbidität, Alter, Geschlecht • Details zu Ein- und Ausschluss von Patienten • Vollständigkeit der Testdurchführung
Diskussion designtypischer Biasformen	<ul style="list-style-type: none"> • Sepktrum-Bias • Patientenselektionsbias • Diagnostic / Review Bias • Verifikationsbias (Work-up Bias) • Bias durch selektive Vollständigkeit der Daten
Generalisierbarkeit (externe Validität)	<ul style="list-style-type: none"> • Reproduzierbarkeit der Testergebnisse in anderen Settings bzw. Abhängigkeit von der Interpretation • Abhängigkeit bzw. Änderung der Richtung der Ergebnisse z. B. von Krankheitsstadium, Komorbidität, Alter Geschlecht • Vergleichbarkeit diagnostisch-therapeutischer Strategien (bei retrospektiven Studien)

5 Weitere Überlegungen zur Bewertung diagnostischer Tests und Schlussfolgerungen

Der Stand der gegenwärtigen methodischen Ansätze zu Metaanalysen von diagnostischen Genauigkeitsstudien lässt sich folgendermaßen zusammenfassen:

Der Hauptunterschied zwischen systematischen Reviews von Studien zur diagnostischen Genauigkeit und systematischen Reviews von RCTs entsteht in der Identifikation der Studien, der Abschätzung von potentiell Bias und den Methoden, die benutzt werden, um die Ergebnisse statistisch zusammenzufassen ¹⁵.

Eine systematische Übersichtsarbeit oder eine Metaanalyse sollte alle verfügbare Evidenz beinhalten. Um diese Evidenz zusammenfassen zu können, müssen zunächst die relevanten Veröffentlichungen identifiziert werden. Die elektronische Literatursuche für Reviews zur diagnostischen Genauigkeit kann sich schwierig gestalten, da geeignete designbezogene Indexwörter fehlen. Eine elementare Erkenntnis aus zahlreichen Untersuchungen ist, dass eine Recherche in biomedizinischen Datenbanken alleine nicht ausreichend ist, um alle potentiell relevanten Studien zu identifizieren. Datenquellen, die berücksichtigt werden sollten, beinhalten: biomedizinische Datenbanken, Literaturverzeichnisse bereits identifizierter Studien, Handsuche von Kongressbänden, Bücher, nicht in Datenbanken gelistete Zeitschriften und Referenzlisten, Studienregister.

Die empirische Forschung legt nahe, dass die wichtigsten Aspekte der Studienqualität die Selektion einer klinisch relevanten Kohorte, den konsistenten Gebrauch eines einzigen guten Referenztests und die Verblindung des experimentellen und Referenztestergebnisses sind. Unvollständige Berichterstattung ist mit Bias assoziiert. Interne und externe Validität greifen

dabei häufig ineinander und beschreiben beide Aspekte zur Beurteilung der Qualität der identifizierten Primärstudien. Ähnlich wie für RCTs (CONSORT-Satement¹), werden daher zur Zeit für Studien zur diagnostischen Genauigkeit „Standards for Reporting of Diagnostic Accuracy“ (STARD Statement) erarbeitet⁵⁰. Der Umgang mit der anhand dieser Kriterien ermittelten Qualität der Primärstudien im Rahmen von Metaanalysen wird nicht einheitlich diskutiert.

Die Wahl der statistischen Methode, um die Studienergebnisse zu poolen, ist abhängig von der Art der Heterogenität, von der Variabilität des diagnostischen Grenzwertes und der diagnostischen Genauigkeit. Sensitivität, Spezifität und Likelihood Ratios können direkt kombiniert werden, wenn die Ergebnisse homogen sind. Wenn ein Grenzwerteffekt existiert, werden die Studienergebnisse am besten in SROC – Kurven zusammengefasst. Die Dichotomisierung von Testergebnissen, um Sensitivität und Spezifität zu erhalten, führt zu einem Informationsverlust. Methoden zur Berücksichtigung von kontinuierlichen oder wenigstens ordinal skalierten Daten sollten für Metaanalysen eingesetzt werden, wenn in den Primärstudien derartige Daten präsentiert werden. Bislang gibt es wenige Beispiele für diese Vorgehensweise in der Literatur. Das könnte darauf zurückzuführen sein, dass in der Praxis meist doch Testergebnisse dichotomisiert werden, etwa um die Entscheidungsfindung in der Klinik zu simulieren.

Die Analyse der Heterogenität bietet viele Stolpersteine. Es bleibt eine Herausforderung die Heterogenität zu untersuchen und zu einer klinisch relevanten Interpretation von statistisch heterogenen Ergebnissen zu kommen. Wenn die Studienergebnisse sehr heterogen sind, kann es am besten sein, von einer statistischen Zusammenfassung abzusehen. Heterogenität muss nicht nur als ärgerliche statistische Tatsache betrachtet werden, sondern kann auch als Phänomen gesehen werden, welches benutzt werden kann, um mehr über die Komplexität der täglichen Krankenversorgung zu lernen¹⁸. Die Untersuchung der Heterogenität im Rahmen der diagnostischen Forschung gibt dem Wissenschaftler wertvolle Hinweise über wichtige Quellen der Variabilität in der Genauigkeit diagnostischer Studien.

Viele Studien zur diagnostischen Genauigkeit nutzen Daten, die primär im Rahmen der klinischen Routine gesammelt wurden, so dass es keine eindeutige Dokumentation von versuchten Evaluationen gibt. Studien, die zur Veröffentlichung gelangen, sind häufiger verzerrt und überschätzen wahrscheinlich die Testgenauigkeit. Publikationsbias ist daher mehr ein Problem von Studien zu diagnostischen Tests als von RCTs ist. Trotz dieser Tatsache haben sich

empirische Studien zu Publikationsbias bislang vor allem mit therapeutischen Studien befasst. Wenn es auch bisher keinen internationalen Konsens über das am besten geeignete Verfahren gibt, so ist neben der Durchführung eines statistischen Tests auf Publikationsbias die Erstellung eines Funnelplots in jeder Metaanalyse sinnvoll.

Die gesamte Evaluation der Performance eines diagnostischen Tests beinhaltet, die Reliabilität des Test zu untersuchen, die diagnostische Genauigkeit, den diagnostischen und therapeutischen Impact und den Nettoeffekt, den der Test auf den medizinischen Outcome hat, zu bestimmen. Getrennte systematische Reviews können für diese einzelnen Aspekte durchgeführt werden¹⁵.

Der Outcome der medizinischen Versorgung ist das Ergebnis eines Entscheidungsfindungsprozesses, der aus zwei hauptsächlichen Typen von Entscheidungen besteht: die diagnostische Entscheidung und die therapeutische Entscheidung. Beide Typen der Entscheidung basieren auf dem aktuellen Wissen um die Manifestationen der unterschiedlichen Krankheiten und über die Effekte der unterschiedlichen Interventionen. Das aktuelle Wissen ist das Ergebnis der Basisausbildung, der klinischen Erfahrung, der weiteren Fortbildung und der Information, die durch alle möglichen Kommunikationskanäle verfügbar ist. Früher basierte die Kunst der Medizin auf der Evidenz, die sich aus der täglichen Erfahrung und der Intuition ergab. Heute besteht die Kunst darin, die Stücke wissenschaftlicher Evidenz in die tägliche Praxis zu integrieren. Medizinische Praxis versucht ständig effektiver, effizienter und nebenwirkungsärmer zu werden, in dem sie wissenschaftliche Evidenz standardisiert zusammenfasst, in dem sie diese Evidenz in standardisierter Weise in praktische Leitlinien übersetzt und in dem sie diese wissenschaftliche Evidenz für alle Beteiligten im Gesundheitswesen verfügbar macht.

Systematische Reviews und Metaanalysen zu diagnostischen Test können dabei in vieler Hinsicht wertvolle Beiträge leisten. Sie ermöglichen es, schlechte oder nutzlose Tests zu eliminieren bevor sie weitverbreitet Anwendung finden. Sie stellen eine verbesserte Qualität der Information über diagnostische Tests zur Verfügung, in dem sie das Patientenspektrum darstellen oder relevante Subgruppen mit dem Ziel einer verbesserten Patientenversorgung analysieren⁶⁵.

Mit dem Wissen um diese Limitationen können nach Irwig³⁴ Leser die Ergebnisse von Metaanalysen dann akzeptieren und für ihre eigenen Bedürfnisse nutzen, wenn:

1. sie Patienten behandeln, die die gleichen Charakteristika aufweisen, wie die Patienten in der Metaanalyse,

2. das Profil der Patienten sich zwar unterscheidet, aber die Metaanalyse auch Ergebnisse für Subgruppen liefert,
3. das Profil der Patienten sich zwar unterscheidet, in der Metaanalyse aber deutlich wird, dass die diagnostische Genauigkeit unabhängig von Patientencharakteristika ist.

Eine entscheidende Limitation der meisten systematischen Übersichtsarbeiten ist aber immer noch die Tatsache, dass meistens die diagnostische Genauigkeit nur eines Tests untersucht wird und nicht die diagnostische Genauigkeit des gesamten diagnostischen Prozesses. In der Realität ist der untersuchte Test normalerweise nur ein Glied in der Kette von verschiedenen diagnostischen Tests. Bestehendes Wissen - z. B. klinische Informationen - und Überweisungsmuster beeinflussen nicht nur die „prior probability“, sondern auch die diagnostische Genauigkeit. Aspekte, die in Metaanalysen diagnostischer Test häufig nicht adressiert werden, sind u. a. Fragen, inwieweit Studien zur diagnostischen Tests auch den klinischen Outcome verbessern ⁴¹, oder Aspekte wie Kosten-Effektivität, die Priorisierung von Ressourcen, ethische und juristische Erwägungen, Erwartungen von Patienten oder der Gesellschaft ²⁹. So fanden Walter et al. ⁸¹, in einer Untersuchung zu Metaanalysen (1996 – 1997) von Screeningtests, dass 54% der Metaanalysen sich nur mit der Testgüte beschäftigten, 23% setzen klinische Outcome-Parameter, wie Mortalität ein. Nur ein sehr kleiner Anteil erhob beides oder schloss auch noch die Kosten mit ein.

Daher ist eine umfassende Bewertung von diagnostischen Testverfahren im Rahmen von HTAs wünschenswert, insbesondere im Hinblick darauf, dass viele Tests mit teilweise umfangreichen Folgeuntersuchungen verbunden sind und zu hohen Kosten führen können. Eine in diesem Sinne vollständige Evaluation von Testverfahren unter Berücksichtigung der Kosten aus gesellschaftlicher Perspektive findet sich indessen so gut wie nie in der Literatur. In Abbildung 7 ist ein solcher vollständiger Vergleich zwischen zwei Testverfahren stark vereinfacht dargestellt. Eine „Zerlegung“ der Testevaluation in verschiedene Testphasen, die von der technischen Wirksamkeit bis hin zur Erfassung von patientenrelevanten Endpunkten und Kosten führt, wäre wünschenswert. ²⁹

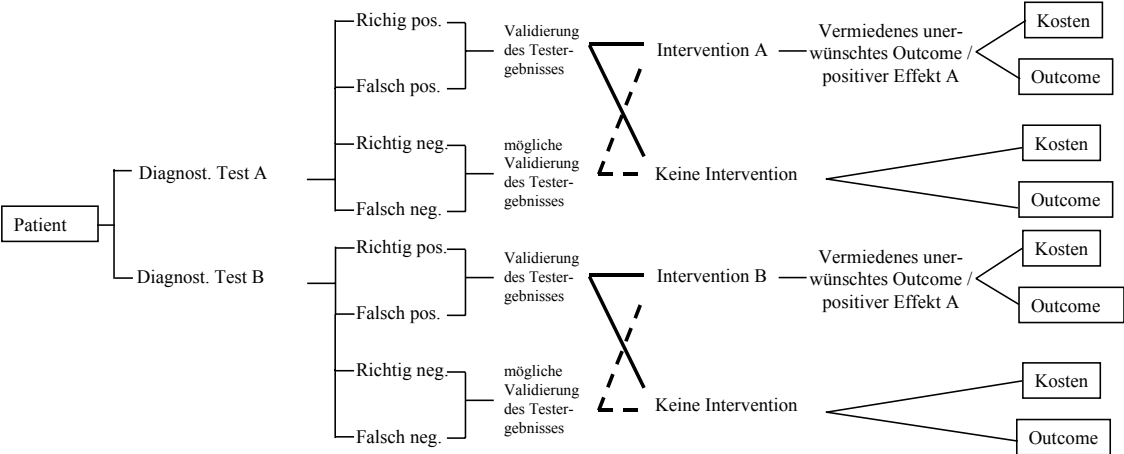


Abbildung 7: Vereinfachtes Schema einer umfassenden Evaluation diagnostischer Tests

Anlage

1 Literaturverzeichnis

1. Altman DG. Better reporting of randomized controlled trials: the CONSORT statement. Authors must provide enough information for readers to know how the trial was performed. *BMJ* 1996;**313**:570-1.
2. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subjected to selection bias. *Biometrics* 1983;**39**:207-15.
3. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994;**50**:1088-99.
4. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342**:1878-86.
5. Blakeley DD, Oddone EZ, Hasselblad V, Simel DL, Matchar DB. Noninvasive carotid artery testing. *Ann Intern Med* 1995;**122**:360-7.
6. Boissel JP, Cucherat M. The meta-analysis of diagnostic test studies. *Eur Radiol* 1998;**8**:484-7.
7. Brenner H, Gefeller O. Variation of sensitivity and specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 1997;**16**:981-91.
8. Breslow NE, Day NE. Statistical methods in cancer research, Volume I: The analysis of case-control studies, IARC Scientific Publications, No 32. New York: Oxford University Press, 1993.
9. Breslow NE, Day NE. Statistical methods in cancer research, Volume II: The design and analysis of cohort studies, IARC Scientific Publications, No 82. New York: Oxford University Press, 1994.
10. Britton A, McKnee M, Black N, McPerson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technology Assessment* 1998;**2**:1-119.
11. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *Can Med Assoc J* 1979;**121**:1193-254.

12. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954;**10**:101-29.
13. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887-92.
14. Davey Smith G, Egger M, Phillips AN. Meta-analysis and data synthesis in medical research. In Detels R, Holland WW, McEwen J, Owens DK, eds. *Oxford Textbook of Public Health*, pp 631-49. Oxford - New York - Tokio: Oxford University Press, 1997.
15. Deeks JJ. Systemic reviews of evaluations of diagnostic and screening tests. In Egger M, Smith GD, Altman DG, eds. *Systematic reviews in health care. Meta-analysis in context*, pp 248-82. London: BMJ, 2001.
16. DerSimonian R, Laird N. Meta analysis in clinical trials. *Control Clin Trials* 1986;**7**:177-88.
17. Devillé WLJM, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000;**53**:65-9.
18. Devillé WLJM, Bouter LM, van der Windt DAWM, Yzermans JC, Bezemer PD. Heterogeneity in systematic reviews of studies on diagnostic accuracy. In Devillé WLJM, ed. *Evidence in diagnostic research. Reviewing diagnostic accuracy: from research to guidelines*, pp 75-91. Wageningen: Ponsen&Looijen, 2001.
19. Devillé WLJM, Buntinx F, van der Windt DAWM, de Vet HCW, Montori V, Bezemer PD *et al.* Didactic guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In Devillé WLJM, ed. *Evidence in diagnostic research. Reviewing diagnostic accuracy: from search to guidelines*, pp 93-112. Wageningen: Ponsen & Looijen, 2001.
20. Devillé WLJM, Yzermans JC, van Duijn NP, Bezemer PD, van der Windt DAWM, Bouter LM. Which factors affect the accuracy of the urine dipstick test for the detection of bacteruria or urinary tract infections? A meta-analysis. In Devillé WLJM, ed. *Evidence in diagnostic research. Reviewing diagnostic accuracy: from research to guidelines*, pp 39-73. Wageningen: Ponsen & Looijen bv, 2001.
21. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method for testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000;**56**:455-63.

22. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;629-34.
23. Egger M, Smith GD, Altman DG. Systematic reviews in health care. Meta-analysis in context. London: BMJ, 2001.
24. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet* 1997;350:326-9.
25. Flynn K, Adams E. Assessing diagnostic technologies. 1, 1-15. 1996. Boston, HSR&D. Technology Assessment Program.
26. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991;11:88-94.
27. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;7:889-94.
28. Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Stat Med* 2000;19:3325-36.
29. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM *et al.* Current methods of the U.S. Preventive Services Task Force. *Am J Prev Med* 2001;20:21-35.
30. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;117:167-78.
31. Hawkins DM, Garrett JA, Stephenson B. Some issues in resolution of diagnostic tests using an imperfect gold standard. *Stat Med* 2001;20:1987-2001.
32. Hellmich M, Abrams KR, Sutton AJ. Bayesian approaches to meta-analysis of ROC Curves. *Med Decis Making* 1999;19:252-64.
33. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;48:119-30.
34. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers C *et al.* Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120:667-76.
35. Jadad AR, McQuay HJ. Be systematic in your searching. *BMJ* 1993;307:66.

36. Jadad AR, Moore A, Carroll D, Jenkinson C, Reynolds DJM, Gavaghan DJ *et al.* Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996;**17**:1-12.
37. Jovell AJ, Navarro-Rubio MD. Evaluation of scientific evidence. *Medicina Clinica (Barcelona)* 1995;**105**:740-3.
38. Kardaun JWPF, Kardaun OJWF. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. *Meth Info Med* 1990;**29**:12-22.
39. Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol* 1992;**27**:245-54.
40. Kester AD, Buntinx F. Meta-analysis of ROC-Curves. *Med Decis Making* 2000;**20**:430-9.
41. Khan KS, Dinnes J, Kleijnen J. Systematic reviews to evaluate diagnostic tests. *Eur J Obst Gyn Rep Biol* 2001;**95**:6-11.
42. Knipschild P. Systematic reviews: Some examples. *BMJ* 1994;**309**:719-21.
43. Köbberling J, Trampisch HJ, Windeler J. Memorandum zur Evaluierung diagnostischer Massnahmen. Stuttgart - New York: Schattauer, 1989.
44. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;**317**:1185-90.
45. Lau J, Ioannidis JPA, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med* 1997;**127**:820-6.
46. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JHP *et al.* Empirical evidence of design-related bias in studies of diagnostic test. *JAMA* 1999;**282**:1061-6.
47. Lloyd CJ. Regression models for convex ROC curves. *Biometrics* 2000;**56**:862-7.
48. Lowe HJ, Barnett GO. Understanding and using the medical subject headings (Mesh) vocabulary to perform literature searches. *JAMA* 1994;**271**:1103-8.

49. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med* 2001;**20**:641-54.
50. McQueen MJ. Overview of evidence-based medicine: challenges for evidence-based laboratory medicine. *Clinical Chemistry* 2001;**47**:1536-46.
51. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. *Med Decis Making* 1993;**13**:253-7.
52. Mijnhout GS, Hooft L, van Tulder MW, Devillé WLJM, Teule GJJ, Hoekstra OS. How to perform a comprehensive search for FDG-PET literature. *Eur J Nucl Med* 2000;**27**:91-7.
53. Moher D, Fortin P, Jadad AR, Kiassen T, Leloirier J, Liverati A *et al.* Completeness of reporting of trials published in languages other than English: Implications for the conduct and reporting of systematic reviews. *Lancet* 1996;**347**:363-6.
54. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR *et al.* What contributions do languages other than English make on the results of meta-analyses? *J Clin Epidemiol* 2000;**53**:964-72.
55. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;**12**:1293-316.
56. Mulrow C, Cook D. Systematic reviews. Synthesis fo best evidence for health care decisions. Philadelphia: American College of Physicians, 1998.
57. Mulrow CD. The medical review article: State of the science. *Ann Intern Med* 1987;**106**:485-8.
58. Oosterhuis WP, Niessen RWLM, Bossuyt PMM. The science of systematic reviewing studies of diagnostic tests. *Clin Chem Lab Med* 2000;**38**:577-88.
59. Perleth M, Antes G. Evidenz-basierte Medizin. Wissenschaft im Praxisalltag. München: MMV Medizin Verlag, 1998.
60. Perleth M, Antes G. Evidenz-basierte Medizin. Wissenschaft im Praxisalltag. München: MMV Medizin Verlag, 1999.

61. Perleth M, Bitzer E. Brauchen wir ein "CONSORT-Statement" für diagnostische Studien. *Gesundh Wes* 2000;**62**:114-5.
62. Perleth M, Jakubowski E, Busse R. Bewertung von Verfahren zur Diagnostik der akuten Sinusitis maxilaris bei Erwachsenen. Baden-Baden: Nomos-Verlag, 1999.
63. Perleth M, Raspe H. Levels of Evidence. Was sagen Sie wirklich aus? *Z ärztl Fortbild Quallsich* 2000;**94**:699-700.
64. Raspe H, Ollenschläger G. EBM braucht zur Literaturbewertung methodische und klinische Kriterien! *Z ärztl Fortbild Quallsich* 2000;**94**:131-2.
65. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;**274**:645-51.
66. Rutter CM, Gatsonis C. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995;**2**:S48-S56.
67. Rutter CM, Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;**20**:2865-84.
68. Shapiro DE. Issues in combining independent estimates of sensitivity and specificity of a diagnostic test. *Acad Radiol* 1995;**2**:S37-S47.
69. Song F, Eastwood S, Gilbody S, Duley L, Sutton A. Publication and related biases. *Health Technology Assessment* 2000;**4 (10)**.
70. Song F, Khan KS, Dinnes J and Sutton AJ. Asymmetric funnel plots and publication bias in meta-analysis of diagnostic accuracy. *Int J Epidemiol* 2002;**31**:88-95.
71. Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001;**54**:1046-55.
72. Sterne JAC, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000;**53**:1119-29.
73. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *Health Technology Assessment* 1998;**2**:1-272.
74. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**:1351-5.

75. Thompson SG, Pocock SJ. Can meta-analyses be trusted. *Lancet* 1991;**338**:1127-30.
76. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *J Clin Epidemiol* 2000;**53**:207-16.
77. Tusch G, Heinrich M, Perleth M. Empirical comparison of meta-analytic methods for diagnostic accuracy studies. 47. Biometrisches Kolloquium der Deutschen Region der Internationalen Biometrischen Gesellschaft (IBS), Homburg /Saar 2001.
78. Vamvakas EC. Meta-analyses of studies of the diagnostic accuracy of laboratory tests. A review of the concepts and methods. *Arch Pathol Lab Med* 1998;**122**:675-86.
79. Walter SD, Irwig L, Glasziou P. Meta-analysis of diagnostic tests with imperfect reference standards. *J Clin Epidemiol* 1999;**52**:943-51.
80. Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *J Clin Epidemiol* 1988;**41**:923-37.
81. Walter SD, Jadad AR. Meta-analysis of screening data: a survey of the literature. *Stat Med* 1999;**18**:3409-24.

2 Literaturrecherche

Zeitraumen: 1995 bzw. 2000 – 2001	
Datenbanken:	
<ul style="list-style-type: none"> • EMBASE • Current Contents • CINAHL • MEDLINE • BIOSIS Previews 	
weitere Datenquellen:	
Referenzlisten publizierter Artikel	
Handsuche in folgenden Zeitschriften:	
<ul style="list-style-type: none"> • Medical Decision Making • Journal of Clinical Epidemiology • Statistics in Medicine 	<p>Volume 20, 2000; 1 – 6, Volume 21, 2001; 1 – 5 (Sept./Oct)</p> <p>Volume 53, 2000; 1 – 12, Volume 54, 2001; 1- 10 (Oct.)</p> <p>Volume 19, 2000; 1 – 24; Volume 20, 2001; 1 – 20 (30.10.01)</p>
Reports, Bücher, Dissertationen, Habilitationsschriften	

Datenbank-Suchstrategie		
Datenbank: EMBASE (OVID)		
Zeitraum der Recherche: 1995 – 2001		
Rechercheschritt:	Suchbegriff:	Treffer:
#1	cancer diagnosis/ or exp qualitative diagnosis/ or computer assisted diagnosis/ or exp. quantitative diagnosis/ or exp diagnosis/ or radioisotope diagnosis/ or exp “diagnosis, measurement and analysis”/ or sex diagnosis/ or tumor diagnosis/ or differential diagnosis/ or virus diagnosis/ or early diagnosis/ or laboratory diagnosis/ or prenatal diagnosis/ or psychiatric diagnosis	2108747
#2	specificity.mp.	91455
#3	exp Meta-Analysis/ or meta-analysis.mp.	14340
#4	limit 3 to (human and yr=1995 – 2002)	11380
#5	exp Technique/ or method.mp.	344998
#6	methods.mp.	440373
#7	techniques.mp.	133466
#8	technic.mp.	339
#9	technics.mp.	371
#10	5 or 6 or 7 or 8 or 9	801767
#11	4 and 10	2763
#12	1 and 11	2490
#13	Accuracy/ or exp diagnostic accuracy	73297
#14	false negative.mp.	6234
#15	ROC curve.mp	1112
#16	2 or 13 or 14 or 15	156454
#17	1 and 16	115538
#18	11 and 17	159

Datenbank-Suchstrategie		
Datenbank: Current Contents (OVID)		
Zeitraum der Recherche: 2000 - 2001		
Rechercheschritt:	Suchbegriff	Treffer:
#1	diagnosis mp.*	189957
#2	diagnostic mp.*	83963
#3	false negative mp.*	4244
#4	accuracy mp.*	93228
#5	specificity mp.*	84744
#6	ROC curve mp.*	1172
#7	meta-analysis mp.*	6655
#8	metaanalysis mp.*	5215
#9	meta-analytic mp.*	719
#10	method mp.*	582812
#11	methods mp.*	547567
#12	technique mp.*	267560
#13	techniques mp.*	204008
#14	technic mp.*	220
#15	technics mp.*	408
#16	1 or 2	242678
#17	limit 16 to yr=1995 – 2002	203367
#18	3 or 4 or 5 or 6	175006
#19	7 or 8 or 9	10881
#20	10 or 11 or 12 or 13 or14 or 15	1309030
#21	19 and 20	3561
#22	17 and 18 and 21	105

* mp = in abstract, title author keywords, keywords plus

Datenbank-Suchstrategie		
Datenbank: CINAHL (OVID)		
Zeitraum der Recherche: 1995 – 2001		
Rechercheschritt:	Suchbegriff	Treffer:
#1	diagnosis mp.*	15513
#2	diagnostic mp.*	6704
#3	false negative mp.*	157
#4	accuracy mp.*	2335
#5	specificity mp.*	1300
#6	ROC curve mp.*	37
#7	meta-analysis mp.*	2070
#8	metaanalysis mp.*	20
#9	meta-analytic mp.*	69
#10	method mp.*	12884
#11	methods mp.*	26510
#12	technique mp.*	5578
#13	techniques mp.*	8200
#14	technic mp.*	4
#15	technics mp.*	2
#16	1 or 2	20421
#17	limit 16 to yr=1995 – 2002	14467
#18	3 or 4 or 5 or 6	3533
#19	7 or 8 or 9	2093
#20	10 or 11 or 12 or 13 or14 or 15	46457
#21	19 and 20	499
#22	17 and 18 and 21	3

* mp = in title, cinahl subject heading, abstract, instrumentation

Datenbank-Suchstrategie		
Datenbank: MEDLINE		
Zeitraum der Recherche: 1995 – 2001		
Rechercheschritt:	Suchbegriff	Treffer:
#1	'diagnosis' [†]	709171
#2	'diagnostic techniques and procedure' [†]	509914
#3	#1 or #2	752383
#4	'sensitivity-and specificity' [†]	76418
#5	'reproducibility-of-results' [†]	47497
#6	'false-negative-reactions' [†]	2535
#7	'false-positive-reactions' [†]	3854
#8	'regression-analysis' [†]	54079
#9	likelihood ratio	784
#10	ROC curve	3799
#11	accuracy	28502
#12	#4 or #5 or #6 or #7 or #8 or #9 or #10 or #11	178132
#13	#3 and #12	91313
#14	#13 and (human in TG)	84742
#15	#14 not (case report in TG)	82464
#16	meta-analy*	6127
#17	Meta-analysis in PT	4171
#18	'Meta-analysis' [†]	2200
#19	#16 or #17 or #18	7823
#20	#19 and (method* or technic* or technique* or concept*)	3921
#21	#13 and #20	347

† all subheadings in MIME, MJME

Datenbank-Suchstrategie		
Datenbank: BIOSIS Previews (OVID)		
Zeitraum der Recherche: 2000 – 2001		
Rechercheschritt:	Suchbegriff	Treffer:
#1	diagnosis.mp.*	49656
#2	diagnostic mp.*	60457
#3	false negative mp.*	799
#4	accuracy mp.*	8539
#5	specificity mp.*	17032
#6	ROC curve mp.*	253
#7	meta-analysis mp.*	1274
#8	metaanalysis mp.*	45
#9	meta-analytic mp.*	71
#10	method mp.*	411227
#11	methods mp.*	248069
#12	technique mp.*	26011
#13	techniques mp.*	165423
#14	technic mp.*	26
#15	technics mp.*	39
#16	1 or 2	87836
#17	limit 16 to yr=1995 – 2002	87836
#18	3 or 4 or 5 or 6	24903
#19	7 or 8 or 9	1326
#20	10 or 11 or 12 or 13 or 14 or 15	479091
#21	19 and 20	989
#22	17 and 18 and 21	41

* mp = in title, book title, original language book title, abstract, subject headings, biosystematic codes/super taxa, heading words

3 Indexwörter oder Textwörter

- die hilfreich sind Studien zur diagnostischen Genauigkeit zu finden, nach Deeks ¹⁵:

Nützliche MeSH-Terms: explode „Sensitivity-and-Specificity“/all subheadings

explode „Mass-screening“

„Predictive-value-of-tests“

„ROC-Curve“

Textwörter: Specificit*, Sensitivit*; falsce negativ*, predictive value*, accuracy*, likelihood ratio*, screening

Sub-headings: Sub-headings werden dazu verwendet, um mit anderen subject headings die Art der Artikel zu begrenzen, nach denen gesucht wird. /*diagnostic use* zusätzlich zu agents, Untersuchungen usw. benutzt werden, um die Treffer auf solche zu begrenzen, die das Agent oder die Untersuchung für diagnostische Zwecke nutzen.

Der Indexterm „Sensitivity-and-Specifity“ scheint der geeignetste für diese Studien zu sein, aber er wird nur inconsistent genutzt und ist nicht sensitive genug. Der alternative MESH-Term „diagnosis“ schließt noch viele andere Terme mit ein und ist daher unspezifisch.

Levels of Evidence in der Fassung des Centre for Evidence-based medicine*

Empfehlungsstärke	Evidenz-Level	Therapie / Prävention, Ätiologie / Nebenwirkungen	Prognose	Diagnose	Differentialdiagnose / Symptom-Prävalenz-Studie	Ökonomische Evaluation / Entscheidungsanalyse
A	1a	Systematischer Review mit homogenen RCTs)	Systematische Review mit homogenen Inzeptionskohortenstudien	Systematischer Review mit homogenen Level 1 diagnostischen Studien	Systematischer Review mit homogenen prospektiven Kohortenstudien	Systematischer Review mit homogenen Level 1 ökonomischen Studien
	1b	Einzelner RCT mit schmalen Konfidenzintervallen	Einzelne Inzeptionskohortenstudie mit mindestens 80% Follow-up	Prospektive Studie mit unabhängigem, verblindetem Vergleich eines angemessenen Spektrums von Patienten, von denen alle mit dem diagnostischen Test und dem Referenzstandard getestet wurden	Prospektive Kohortenstudien mit mindestens 80% Follow-up und ausreichend langer Beobachtungszeit	Analyse basiert auf klinisch relevanten Alternativen bzw. Kosten, systematischem Review der Evidenz, Mehrweg-Sensitivitätsanalysen durchgeführt
	1c	"Alles oder Nichts-Studie"***	"Alles oder Nichts-Fallsrie"	Spezifität ist so hoch, daß ein positives Ergebnis die Diagnose sicherstellt oder Sensitivität so hoch, daß ein positives Ergebnis die Diagnose ausschließt	"Alles oder Nichts-Fallsrie"	absolute <i>better-value</i> oder <i>worse-value</i> -Analyse
B	2a	Systematischer Review mit homogenen Kohortenstudien	Systematischer Review mit homogenen retrospektiven Kohortenstudie oder unbehandelte Kontrollgruppen in RCTs	Systematischer Review mit homogenen ≥ 2 -Diagnosestudien	Systematischer Review mit homogenen Studien Level 2b oder besser	Systematischer Review mit homogenen ökonomischen Studien Level 2 oder besser
	2b	Einzelne Kohortenstudie (einschließlich RCT mit niedriger Qualität [bspw. <80% Follow-up])	Retrospektive Kohortenstudie oder Follow-up unbehandelter Kontrollpatienten in einem RCT	<ul style="list-style-type: none"> unabhängiger, blinder oder objektiver Vergleich Studie an nicht-konsekutiven Patienten und / oder mit schmalen klinischen Spektrum, wobei bei allen Patienten der Vergleichs- und der Referenztest durchgeführt wurde 	<ul style="list-style-type: none"> Retrospektive Kohortenstudie Prospektive Kohortenstudie mit schlechtem Follow-up 	Analyse basiert auf klinisch relevanten Alternativen bzw. Kosten, Evidenzlage nur unzureichend bestimmt bzw. nur Einzelstudien berücksichtigt, Mehrweg-Sensitivitätsanalysen durchgeführt
	2c	"Outcomes"-Forschung, ökologische Studien	"Outcomes"-Forschung, ökologische Studien		ökologische Studien	Audit oder „Outcome“-Forschung
	3a	Systematischer Review mit homogenen Fall-Kontroll-Studien			Systematischer Review mit homogenen Studien mit Level 3b oder besser	Systematischer Review mit homogenen Studien mit Level 3b oder besser
	3b	Einzelne Fall-Kontroll-Studie		Unabhängiger, verblindeter Vergleich anhand eines angemessenen Patientenspektrums, Referenztest nicht bei allen Patienten angewendet	nicht-konsekutive Kohortenstudie oder nicht repräsentatives, kleines Sample	Analyse basiert auf nur wenigen untersuchten Alternativen bzw. Kosten, unzureichende Qualität der Daten, aber klinisch relevante Sensitivitätsanalysen vorhanden
C	4	Fallsrie; Kohorten- oder Fall-Kontrollstudie mit niedriger methodischer Qualität	Fallsrie; prognostische Kohortenstudie mit niedriger methodischer Qualität	<ul style="list-style-type: none"> Referenzstandard nicht objektiv, nicht verblindet oder nicht unabhängig Positive und negative Tests mit unterschiedlichen Referenztests verifiziert Studie an nicht angemessenem Patientenspektrum durchgeführt 	Fallsrie oder überholter Referenztest	keine Sensitivitätsanalyse vorhanden
D	5	Expertenmeinung ohne expliziter kritischer Bewertung oder auf physiologischen Daten basiert, Laborforschung	Expertenmeinung ohne expliziter kritischer Bewertung oder auf physiologischen Daten basiert, Laborforschung	Expertenmeinung ohne expliziter kritischer Bewertung oder auf physiologischen Daten basiert, Laborforschung	Expertenmeinung ohne expliziter kritischer Bewertung oder auf physiologischen Daten basiert, Laborforschung	Expertenmeinung ohne expliziter kritischer Bewertung oder auf physiologischen Daten basiert, Laborforschung

* Quelle: <http://cebm.jr2.ox.ac.uk/docs/levels.html>.

** In diese Kategorie fallen Studien, wenn alle Patienten vor dem Verfügbarwerden der Therapie starben, danach aber wenigsten einige überleben oder wenn einige Patienten vor dem Verfügbarwerden starben, nun aber alle überleben.