

# Vollständiges und präzises Berichten von Studien zur diagnostischen Genauigkeit: Die STARD-Initiative

P. M. Bossuyt<sup>1</sup>  
 J. B. Reitsma<sup>1</sup>  
 D. E. Bruns<sup>2</sup>  
 C. A. Gatsonis<sup>3</sup>  
 P. P. Glasziou<sup>4</sup>  
 L. M. Irwig<sup>5</sup>  
 J. G. Lijmer<sup>1</sup>  
 D. Moher<sup>6</sup>  
 D. Rennie<sup>7</sup>  
 H. C. W. de Vet<sup>8</sup>  
 für die STARD Gruppe

## Medizinisches Publizieren

### Schlüsselwörter

- ▶ Berichten
- ▶ Checkliste
- ▶ diagnostische Genauigkeit
- ▶ Flussdiagramm
- ▶ Sensitivität und Spezifität

### Keywords

- ▶ checklist
- ▶ diagnostic accuracy
- ▶ flow diagram
- ▶ reporting
- ▶ sensitivity and specificity

### Institut

<sup>1</sup>Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, Niederlande

<sup>2</sup>Clinical Chemistry, Charlottesville, VA, USA

<sup>3</sup>Brown University, Center for Statistical Sciences, Providence, RI, USA

<sup>4</sup>Mayne Medical School, Department of Social and Preventive Medicine, Herston, Australien

<sup>5</sup>Screening and Test Evaluation Program, School of Public Health, University of Sydney, Australien

<sup>6</sup>Chalmers Research Group, Ottawa, ON, Kanada

<sup>7</sup>Journal of the American Medical Association, Chicago, IL, USA

<sup>8</sup>Free University, Institute for Research in Extramural Medicine, Amsterdam, Niederlande

### Bibliografie

DOI 10.1055/s-0031-1272979  
 Dtsch Med Wochenschr 2011;  
 136: e16–e19 · © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0012-0472

### Korrespondenz

**Univ.-Prof. Dr. Andreas Ziegler**  
 Institut für Medizinische, Biometrie und Statistik,  
 Universität zu Lübeck,  
 Universitätsklinikum Schleswig-Holstein, Campus Lübeck  
 Maria-Goeppert-Str. 1  
 23562 Lübeck  
 Tel. 0451/500 2789  
 Fax 0451/500 2999  
 eMail ziegler@imbs.uni-luebeck.de

## Wörtliche Übersetzung aus dem Englischen

A. Ziegler<sup>1,2</sup>

I. R. König<sup>1,2</sup>

<sup>1</sup>Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck

<sup>2</sup>Zentrum für Klinische Studien, Universität zu Lübeck, Lübeck

## Einleitung

Die Welt der diagnostischen Tests ist äußerst dynamisch. Neue Test werden mit hoher Geschwindigkeit entwickelt, die Technologie von existierenden Tests kontinuierlich verbessert. Übertriebene und verzerrte Ergebnisse schlecht designter und berichteter diagnostischer Studien können eine voreilige Verbreitung des diagnostischen Tests veranlassen und zu falschen Behandlungsentscheidungen von Klinikern führen. Ein konsequenter Evaluationsprozess diagnostischer Tests vor ihrer Einführung in die klinische Praxis könnte nicht nur die Anzahl der unerwünschten klinischen Konsequenzen reduzieren, die aus irreführenden Schätzungen der Testgenauigkeit resultieren. Vielmehr ließen sich auch die Kosten für die Gesundheitsversorgung durch das Vermeiden unnötiger diagnostischer Tests begrenzen. Ein essentieller Bestandteil dieses Evaluationsprozesses sind Studien zur diagnostischen Genauigkeit [6, 8, 10].

In Studien zur diagnostischen Genauigkeit wird das Ergebnis eines oder mehrerer Tests von Interesse mit dem Ergebnis des Referenzstandards verglichen. Alle Tests werden bei Personen angewendet, bei denen der Verdacht auf den Zustand von Interesse besteht. Der Begriff Test bezieht sich dabei auf eine beliebige Methode, mit der zusätzliche Informationen über den Gesundheitszustand eines Patienten erhalten werden sollen. Dies beinhaltet Informationen zur Krankengeschichte, körperlichen Untersuchungen, Labortests, bildgebenden Verfahren, Funktionstests und Histopathologie. Der Zustand von Interesse, Ziel-Zustand (engl.: target condition) genannt, kann sich auf eine bestimmte Krankheit oder irgendeinen anderen Zustand beziehen, der zu klinischen Handlungen führt, wie z.B. zu weiteren diagnostischen Tests oder dem Veranlassen, Modifizieren oder Beenden von Behandlungen. In diesem Zusammenhang ist der Referenzstandard (engl.: reference standard) das beste

verfügbare Verfahren, um einwandfrei festzustellen, ob der Ziel-Zustand vorliegt oder nicht. Der Referenzstandard kann ein einzelnes Verfahren oder eine Kombination von Methoden sein. Er kann Labortests, bildgebende Verfahren und Pathologie beinhalten, aber genauso auch die klinische Nachuntersuchung. Der Begriff Genauigkeit (engl.: accuracy) bezieht sich auf den Grad der Übereinstimmung zwischen der Information des Tests von Interesse, genannt Indextest, und dem Referenzstandard. Diagnostische Genauigkeit kann auf viele Arten ausgedrückt werden, einschließlich Sensitivität, Spezifität, Likelihoodquotienten, diagnostische Chancenverhältnisse (engl.: odds ratios) und die Fläche unter der ROC-Kurve (engl.: receiver operator characteristic curve) [7, 13, 19].

Eine Reihe von Faktoren kann die interne und externe Validität einer Studie zur diagnostischen Genauigkeit gefährden. Eine Untersuchung der Studien zur diagnostischen Genauigkeit, die in vier wichtigen medizinischen Zeitschriften zwischen 1978 und 1993 publiziert wurden, ergab, dass die methodische Qualität bestenfalls mittelmäßig war [18]. Dabei wurde die Beurteilung dadurch erschwert, dass vielen Berichte wesentliche Elemente zum Design, der Durchführung und der Analyse der diagnostischen Studien fehlten [18]. Das Fehlen wesentlicher Informationen zum Design und der Durchführung der diagnostischen Studien wurde durch viele Autoren von Meta-Analysen bestätigt [4, 16]. Wie bei anderer Forschung auch, können Fehler im Studiendesign zu verzerrten Ergebnissen führen. Ein Publikation zeigte, dass diagnostische Studien mit bestimmten Designmerkmalen zu verzerrten, optimistischen Schätzungen der diagnostischen Genauigkeit im Vergleich zu Studien ohne diese Fehler führten [12].

Während des Cochrane-Kolloquiums 1999 in Rom diskutierte die Cochrane-Arbeitsgruppe für Diagnose- und Screening-Test-Methoden die geringe methodologische Qualität und das minderwertige Berichten von Bewertungen diagnostischer Tests. Die Arbeitsgruppe kam zu der Einschätzung, dass der erste Schritt zur Korrektur dieses Problems eine Verbesserung der Qualität des Berichtens diagnostischer Studien darstellt. In Anlehnung an die erfolgreiche CONSORT („consolidated standards of reporting trials“) -Initiative [1, 14, 15] war das Ziel der Arbeitsgruppe die Entwicklung einer Checkliste von Punkten, die im Bericht einer Studie zur diagnostischen Genauigkeit verwendet werden sollten.

Das Ziel der Initiative zum standardisierten Berichten von Studien zur diagnostischen Genauigkeit (engl.: **ST**Andards for **R**eporting of **D**iagnostic accuracy, STARD) ist, die Qualität des Berichtens von Studien zur diagnostischen Genauigkeit zu verbessern. Vollständiges und präzises Berichten ermöglicht dem Leser, potentielle Verzerrungen in einer Studie (interne Validität) zu identifizieren und die Generalisierbarkeit und Anwendbarkeit der Ergebnisse (externe Validität) zu beurteilen.

## Methoden

Das STARD-Lenkungsgremium (Teilnehmerliste im Anhang der Originalpublikation) begann mit einer ausführlichen Suche, um Veröffentlichungen zur Durchführung und dem Berichten diagnostischer Studien zu identifizieren. Die Suche schloss Medline, Embase, Biosis und die methodologische Datenbank der Cochrane Collaboration bis Juli 2000 ein. Weitere Referenzen wurden von Mitgliedern des Lenkungsgremiums aus den Literaturlisten der gefundenen Artikel, eigenen Dateien und dem Kontaktieren anderer Experten für diagnostische Forschung identifiziert. Sie überprüften alle relevanten Publikationen und extrahierten eine erweiterte Liste potentieller Items für die Checkliste.

Danach trafen sich die Mitglieder des STARD-Lenkungsgremiums für zwei Tage mit eingeladenen Experten der folgenden Interessensgruppen: Forscher, Herausgeber, Methodiker und Berufsverbände. Das Ziel der Konferenz war, die erweiterte Liste potentieller Items wo möglich zu reduzieren und das optimale Format und die Formulierung der Checkliste zu diskutieren. Wenn möglich war die Auswahl der verbliebenen Items evidenzbasiert.

Das Format des Treffens bestand in einer Mischung kleinerer Gruppensitzungen und Vollversammlungen. Jede Kleingruppe bearbeitete eine Gruppe zusammenhängender Items der Liste. Die Vorschläge der Kleingruppen wurden anschließend in Plenarsitzungen diskutiert. Über Nacht wurde ein erster Entwurf der STARD-Checkliste unter Verwendung der Vorschläge der Kleingruppen sowie Anmerkungen der Vollversammlung erstellt. Alle Teilnehmer der Konferenz diskutierten am folgenden Tag diesen Erst-

entwurf und nahmen weitere Änderungen vor. In einer späteren Runde konnten Mitglieder der STARD-Gruppe weitere Änderungen durch E-Mail-Kommentare vorschlagen.

Mögliche Nutzer führten einen Feldtest der Konferenzfassung der Checkliste und des Flussdiagramms durch, und zusätzliche Anmerkungen wurden gesammelt. Diese Version wurde auf der CONSORT-Webseite eingestellt und es gab einen Aufruf zur Kommentierung. Das Lenkungsgremium diskutierte alle Kommentare und erstellte abschließend die endgültige Checkliste.

## Ergebnisse

Die Suche nach publizierten Empfehlungen für diagnostische Forschung ergab 33 Checklisten. Ausgehend von diesen publizierten Leitlinien und Hinweisen des Lenkungsgremiums sowie den Mitgliedern der STARD-Gruppe erstellte das Lenkungsgremium eine Liste von 75 Items. Während des Konsensustreffens am 16. und 17. September 2000 bestätigten und entfernten Tagungsteilnehmer Items, um die endgültige 25-Punkte-Checkliste zu bilden. Teilnehmer der Tagung nahmen wesentliche Überarbeitungen in den Formulierungen und dem Format der Checkliste vor.

Die STARD-Gruppe erhielt wertvolle Kommentare und Anmerkungen in den verschiedenen Phasen der Evaluation nach der Konferenz, die dann zu der Fassung der STARD-Checkliste in **Tab. 1** führte.

► **Abb. 1** zeigt das Flussdiagramm, das Informationen über die Methode der Patientenrekrutierung (z.B. Verwendung einer konsekutiven Folge von Patienten mit spezifischen Symptomen oder von Fällen und Kontrollen), die Reihenfolge der Durchführung der Tests sowie die Anzahl der Patienten, die den Indextest und den Referenzstandard durchlaufen haben. Wir zeigen ein prototypisches Flussdiagramm, das die häufigsten Studiendesigns im Bereich der diagnostischen Forschung widerspiegelt. Beispiele für andere Designs stehen auf der STARD-Webseite zur Verfügung ([www.consort-statement.org/stardstatement.htm](http://www.consort-statement.org/stardstatement.htm)).

## Diskussion

Ziel der STARD-Initiative ist es, die Qualität des Berichtens diagnostischer Studien zu verbessern. Die Punkte der Checkliste und das Flussdiagramm können Autoren helfen, wesentliche Elemente des Designs und der Durchführung ihrer Studie, der Durchführung der Tests und ihrer Ergebnisse zu beschreiben.

Wir haben die Punkte unter Verwendung der üblichen Überschriften eines medizinischen Forschungsartikels angeordnet, doch soll die Reihenfolge der Anordnung in einem Artikel nicht verbindlich sein.

Das Leitprinzip bei der Entwicklung der STARD-Checkliste war es, Items auszuwählen, die den Lesern helfen, die mögliche Verzerrungen der Studie zu beurteilen und die Anwendbarkeit der Studienergebnisse einzuschätzen. Zwei weitere allgemeine Überlegungen formten den Inhalt und das Format der Checkliste. Zum einen ist die STARD-Gruppe davon überzeugt, dass eine allgemeine Checkliste für Studien zur diagnostischen Genauigkeit vermutlich weitere Verbreitung findet als verschiedene Checklisten für jedes einzelne Spezialgebiet. Möglicherweise wird diese auch eher von Autoren, Gutachtern und Herausgebern von Zeitschriften akzeptiert. Obwohl sich die Bewertung eines bildgebenden Verfahrens sich von der eines Labortests unterscheidet, sind diese Unterschiede eher gradueller denn prinzipieller Art. Zum anderen zielte die Entwicklung der Checkliste spezifisch auf Studien zur diagnostischen Genauigkeit ab. Wir schlossen nicht allgemeine Aspekte von Forschungsergebnissen ein, wie sie z.B. in den einheitlichen Anforderungen von Manuskripten für die Einreichung biomedizinischer Zeitschriften („uniform requirements for manuscripts submitted to biomedical journals“) beschrieben sind [9].

Wenn möglich basierte die STARD-Gruppe ihre Entscheidung, ein Item in die Checkliste aufzunehmen, auf der Evidenz dafür, dass dieses Item mit verzerrten Schätzungen (interne Validität) oder Veränderungen in den Maßen der diagnostischen Genauigkeit (externe Validität) einhergeht. Diese Evidenz reichte von beschreibenden Artikeln, in denen theoretische Prinzipien erläutert wurden, bis hin zu Arbeiten, die Ergebnisse der statistischen Modellierung oder empirische Evi-

Tab. 1 STARD Checkliste für das Berichten von Studien zur diagnostischen Genauigkeit.

Publikationsabschnitt und Thema	#	Beschreibung
<b>Titel/Zusammenfassung</b>	1	Identifikation als Studie zur diagnostischen Genauigkeit.
<b>Schlüsselwörter</b>		Empfehlung MeSH Überschrift „Sensitivität und Spezifität“.
<b>Einleitung</b>	2	Formulierung der Forschungsfragen oder der Studienziele, wie z. B. Schätzung diagnostischer Genauigkeit oder Vergleich der Genauigkeit zwischen Tests oder Teilnehmergruppen
<b>Methoden</b>		
<b>Teilnehmer</b>	3	Beschreibung der Studienpopulation: Ein- und Ausschlusskriterien, Setting und Orte, an denen die Daten erhoben wurden.
	4	Beschreibung der Rekrutierung der Teilnehmer: Basierte die Rekrutierung auf dem Vorhandensein von Symptomen, Ergebnissen früherer Tests oder dem Umstand, dass die Teilnehmer den Index- oder Referenztest erhalten hatten?
	5	Beschreibung der Stichprobenerhebung: War die Studienpopulation eine konsekutive Reihe von Teilnehmern, die durch Auswahlkriterien der Items 3 und 4 definiert waren? Wenn nein, Spezifizierung, wie Teilnehmer weiter ausgewählt wurden.
	6	Beschreibung der Datenerhebung: Wurde die Datenerhebung geplant, bevor der Indextest und der Referenzstandard durchgeführt wurden (prospektive Studie) oder nachher (retrospektive Studie)?
<b>Testverfahren</b>	7	Beschreibung des Referenzstandards und seiner Rationale.
	8	Beschreibung der technischen Spezifikationen von beteiligtem Material und Methoden, einschließlich der Beschreibung, wann und wie die Messungen durchgeführt wurden, und/oder Zitierung von Referenzen für Indextest und Referenzstandard.
	9	Beschreibung der Definition und der Rationale der Einheiten, Schwellenwerte und/oder Kategorien der Ergebnisse des Indextests und des Referenzstandards.
	10	Beschreibung der Anzahl, des Trainings und der Expertise der Personen, die den Indextest und den Referenzstandard durchgeführt und beurteilt haben.
	11	Beschreibung, ob die Beurteiler des Indextests und des Referenzstandards verblindet (maskiert) oder nicht verblindet waren im Hinblick auf die Ergebnisse des anderen Tests, und Beschreibung anderer verfügbarer klinischer Informationen für den Leser.
<b>Statistische Verfahren</b>	12	Beschreibung der Verfahren zur Berechnung oder dem Vergleich von Maßen der diagnostischen Genauigkeit sowie die statistischen Methoden zur Quantifizierung der Unsicherheit (z. B. 95% Konfidenzintervalle).
	13	Beschreibung der Verfahren zur Berechnung der Reproduzierbarkeit des Tests, falls durchgeführt.
<b>Ergebnisse</b>		
<b>Teilnehmer</b>	14	Berichten, wann die Studie durchgeführt wurde, einschließlich der Daten des Rekrutierungsbeginns und -endes.
	15	Berichten klinischer und demographischer Charakteristika der Studienpopulation (z. B. Alter, Geschlecht, Spektrum der beobachteten Symptome, Komorbiditäten, derzeitige Behandlungen, rekrutierende Zentren).
	16	Berichten der Anzahl der Teilnehmer, die die Einschlusskriterien erfüllten, und an der Indextest und/oder der Referenzstandard durchgeführt wurde oder nicht; Beschreibung, warum der jeweilige Test bei den Teilnehmern nicht durchgeführt wurde (ein Flussdiagramm wird dringend empfohlen).
<b>Testergebnisse</b>	17	Berichten von Zeitintervallen zwischen Indextest und Referenzstandard sowie von allen Therapiemaßnahmen zwischen der Durchführung der Tests.
	18	Berichten der Verteilung der Schwere der Erkrankung (Definition der Kriterien) bei Teilnehmern mit dem Ziel-Zustand; andere Diagnosen bei Teilnehmern ohne den Ziel-Zustand.
	19	Berichten einer Kreuztabelle der Ergebnisse des Indextests (einschließlich unbestimmte und fehlende Ergebnisse) gegen die Ergebnisse des Referenzstandards; für Tests mit kontinuierlichen Ergebnissen, die Verteilung der Testergebnisse gegen die Ergebnisse des Referenztests.
	20	Berichten des Auftretens aller unerwünschten Ereignisse aus der Durchführung des Indextests oder des Referenzstandards.
<b>Schätzwerte</b>	21	Berichten von Schätzwerten der diagnostischen Genauigkeit und Maße der statistischen Unsicherheit (z. B. 95% Konfidenzintervalle).
	22	Berichten, wie mit unbestimmten und fehlenden Ergebnissen sowie Ausreißern des Indextests umgegangen wurde.
	23	Berichten von Schätzungen der Variabilität diagnostischer Genauigkeit zwischen Subgruppen von Teilnehmern, Beurteilern oder Zentren, falls durchgeführt.
	24	Berichten von Schätzwerten der Testreproduzierbarkeit, falls durchgeführt.
<b>Diskussion</b>	25	Diskussion der klinischen Anwendbarkeit der Studienergebnisse.

denz aus diagnostischen Studien zeigten. Für einige Items war die vorhandene Evidenz allerdings limitiert.

Ein separates ergänzendes Dokument erläutert die Bedeutung und die Rationale jedes einzelnen Items und fasst die Arten und den Grad der Evidenz kurz zusammen [3]. Dieses Dokument sollte die Verwendung, das Verständnis und die Ver-

breitung der STARD-Checkliste unterstützen.

Die STARD-Gruppe unter erheblichem Aufwand ein Flussdiagramm für diagnostische Studien entwickelt. Ein Flussdiagramm hat das Potential, die wesentlichen Informationen des Studiendesigns und des Patientenflusses in transparenter Weise zu kommunizieren [16]. Ein ver-

gleichbares Flussdiagramm ist ein wesentliches Element des CONSORT Statements für das Berichten randomisierter Therapiestudien [5,15]. Das Flussdiagramm könnte wegen der Vielfalt von Studiendesigns in der diagnostischen Forschung noch viel bedeutsamer sein. In den Berichten von Studien zur diagnostischen Genauigkeit geben Flussdiagramme den Verlauf wieder: die Rekrutierung

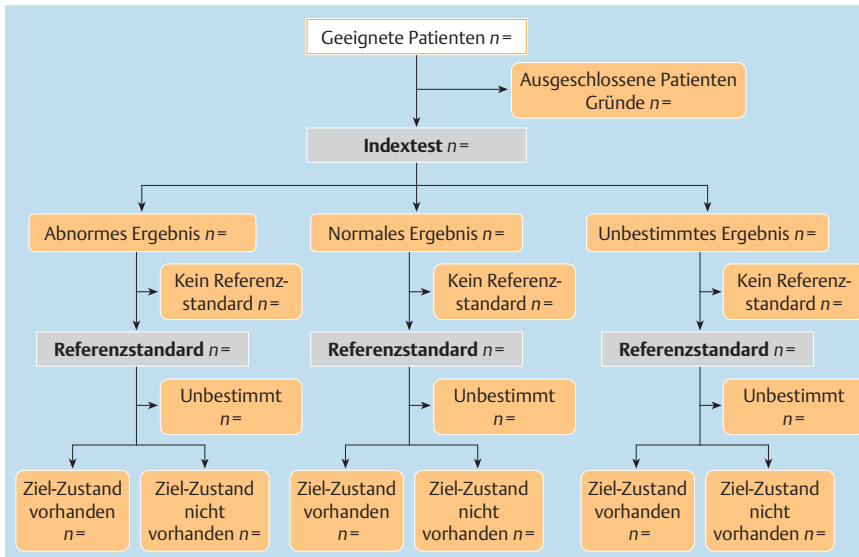


Abb. 1 Prototypisches Flussdiagramm einer Studie zur diagnostischen Genauigkeit.

und die Auswahl von Studienteilnehmern (externe Validität), den Fluss der Teilnehmer in Verbindung mit Zeitablauf sowie Testergebnissen, die Anzahl der Patienten, die entweder den Indextest oder den Referenzstandard oder beide Tests nicht erhalten haben (Möglichkeit der Verifikationsverzerrung) [2, 11, 17] sowie die Anzahl der Patienten in jeder Phase der Studie, so dass die korrekte Nennerinformationen für Anteile (interne Konsistenz) verfügbar ist.

Die STARD-Gruppe plant, die Auswirkung des Statements auf die Qualität publizierter Berichte von Studien zur diagnostischen Genauigkeit in einem „vorher-nachher“-Vergleich zu beurteilen [15]. Aktualisierungen der Dokumente der STARD-Initiative werden zur Verfügung gestellt, sobald sich neue Evidenz zu Quellen von Verzerrung oder Variabilität zeigt. Wir begrüßen Kommentare jedweder Art, bezogen auf Inhalt oder Form, um die aktuelle Version zu verbessern.

Die Übersetzung entstand während des Forschungsfreisemesters von AZ an dem Institut Montefiore der Universität Lüttich (Belgien). AZ dankt den Kollegen am Institut Montefiore sehr herzlich für die Möglichkeit, dort zu arbeiten. Die Übersetzung entstand während der Vorbereitung der Autoren auf den Kursteil Evidenzbasierte Medizin im Rahmen des Kurses Ärztliches Qualitätsmanagement an der Akademie für Fortbildung und Weiterbildung der Landesärztekammer Hessen (Sigrid Blehle, Dr. Roland Kaiser).

**Danksagung:** Diese Initiative zur Verbesserung des Berichtens von Studien zur diagnostischen Genauigkeit wurde durch eine große Anzahl von Personen rund um den Globus unterstützt, die Vorläuferversionen kommentiert haben. Finanzielle Unterstützung zum Einberufen der STARD-Gruppe erfolgte teilweise durch das Dutch Health Care Insurance Board, Amstelveen, Niederlande, die International Federation of Clinical Chemistry, Mailand, Italien; die Medical Research Council's Health Services Research Collaboration, Bristol, Großbritannien und das Academic Medical Center in Amsterdam, Niederlande.

[Veröffentlichung des Originals: Family Practice 2004; 21: 4–10]

**Autorenerklärung:** Die Autoren erhalten ein Honorar von der Akademie für Fortbildung und Weiterbildung der Landesärztekammer Hessen für das Erstellen von Unterlagen für den Kursteil Evidenzbasierte Medizin im Rahmen des Kurses Ärztliches Qualitätsmanagement.

**Literatur**

- 1 Begg C, Cho M, Eastwood S et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996; 276: 637–639
- 2 Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987; 6: 411–423
- 3 Bossuyt PM, Reitsma JB, Bruns DE et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. Clin Chem 2003; 49: 1–6

- 4 de Vries SO, Hunink MG, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. Acad Radiol 1996; 3: 361–369
- 5 Egger M, Jüni P, Bartlett C. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001; 285: 1996–1999
- 6 Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11: 88–94
- 7 Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981; 94: 557–592
- 8 Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986; 134: 587–594
- 9 International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. JAMA 1997; 277: 927–934
- 10 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: 245–254
- 11 Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making 1987; 7: 139–148
- 12 Lijmer JG, Mol BW, Heisterkamp S et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282: 1061–1066
- 13 Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8: 283–298
- 14 Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001; 285: 1992–1995
- 15 Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001; 285: 1987–1991
- 16 Nelemans PJ, Leiner T, de Vet HC, van Engelenhoven JM. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. Radiology 2000; 217: 105–114
- 17 Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. Med Decis Making 1987; 7: 115–119
- 18 Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995; 274: 645–651
- 19 Sackett DL, Haynes RB, Guyatt GH, Tugwell PX. The selection of diagnostic tests. In: Sackett DL ed (Hrsg). Clinical epidemiology. 2 ed. Boston: Little, Brown and Company, 1991: 47–57