Herausforderungen an die Planung und Durchführung von Diagnosestudien mit molekularen Biomarkern

Challenges in planning and conducting diagnostic studies with molecular biomarkers

Autoren **A. Ziegler**[1,2] **I.R. König**[1] **P. Schulz-Knappe**[3]

Institut
[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck
[2] Zentrum für Klinische Studien Lübeck, Universität zu Lübeck, Lübeck
[3] Protagen AG, Dortmund

Medizinisches Publizieren

Schlüsselwörter
- Biomarker
- Diagnosestudie
- Molekulare Diagnostik
- Stichprobenverzerrung
- Validierung
- Variationskoeffizient
- Verifikationsverzerrung

Keywords
- biomarker
- coefficient of variation
- diagnosis study
- molecular diagnostics
- sample selection bias
- validation
- verification bias

eingereicht 05.10.2012
akzeptiert 11.10.2012

Bibliografie
DOI 10.1055/s-0032-1327406
Dtsch Med Wochenschr 2013; 138: e2–e13 · © 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, Lübeck
Maria-Goeppert-Str. 1
23562 Lübeck
Tel. 0451/500-2789
Fax 0451/500-2999
eMail ziegler@imbs.uni-luebeck.de

Zusammenfassung

Die Bedeutung von Biomarkern für die personalisierte Medizin wächst stetig, und Biomarker finden ihre Anwendung im Bereich der Diagnose, Prognose sowie der Auswahl zielgerichteter Therapien. In vielen Biomarkerstudien werden inzwischen molekulare Biomarker verwendet, die sich bei -omics Experimenten gegen eine Vielzahl anderer Kandidaten durchgesetzt haben. Die Intensitäten, z.B. Proteinkonzentrationen, werden typischerweise zwischen zwei oder mehr Gruppen verglichen, um die diagnostische Wertigkeit eines molekularen Biomarkers zu bestimmen. Verschiedene prospektive oder retrospektive Studiendesigns können für molekulare Biomarke: e2–e13rstudien gewählt werden, und der Biomarker kann entweder durch eine einzelne Messung oder eine Kombination verschiedener Messungen, also ein Biomarkerprofil, gemessen werden. In dieser Arbeit werden die methodischen Herausforderungen an die Planung und Durchführung von diagnostischen Studien mit molekularen Biomarkern betrachtet. Zunächst werden die Grade der Evidenz von Diagnosestudien betrachtet. Anschließend werden die damit eng verbundenen verschiedenen Phasen von Diagnosestudien für Biomarker skizziert und die unterschiedlichen Studiendesigns diskutiert. Insbesondere unterscheidet sich die Auswahl der Personen je nach Phase der molekularen Biomarkerstudie wesentlich. Anhand von Beispielen sowie zweier systematischer Übersichten aus der Literatur werden die typischen Verzerrungsquellen molekularer Diagnosestudien illustriert und ihre Relevanz für Anwendungen diskutiert. Insbesondere werden die extreme Auswahl von Patienten und Kontrollen sowie die Verifikationsverzerrung betrachtet. Bei der Validierung molekularer Biomarker spielt die Variabilität der Biomarker-Messung, üblicherweise ausgedrückt als Variationskoeffizient, eine große Rolle. Es wird abschließend aufgezeigt, dass die erforderliche Fallzahl zur Validierung von Biomarkern quadratisch mit dem Variationskoeffizienten, also der Variabilität der Biomarkermessung, steigt. Die Konsequenz dieser Eigenschaft wird anhand von Realdaten verschiedener Labortechniken erläutert.

Einleitung

Biomarker werden in der Wissenschaft und Medizin zurzeit sehr intensiv beforscht. Die Nutzung biologischer Marker als Konzept ist natürlich schon Jahrtausende alt. Schon in der Antike nutzten Ärzte Biomarker für die Diagnose von Erkrankungen. Das bekannteste Beispiel ist wohl die Urinuntersuchung, die bereits bei Galen im zweiten Jahrhundert nach Christus beschrieben wird. Im Mittelalter war dies die Harnschau (Uroskopie), die bei Ärzten als quasi unfehlbare diagnostische Methode fast aller Krankheiten galt. Ein Teil dieser Indikatoren wird auch heute noch verwendet, wie z.B. Glukose im Urin als Hinweis auf einen Diabetes mellitus. Gleichzeitig wird nicht mehr nur Urin als Biomarker genutzt, und entsprechend war die Entwicklung sehr allgemeiner Definitionen notwendig [1, 2].

Definition

Die häufigste Definition für Biomarker ist nach Gallo et al. [2] die folgende: „Ein Biomarker ist eine Substanz oder biologische Struktur, die im menschlichen Körper gemessen werden kann, und die die Inzidenz oder das Resultat einer Krankheit beeinflussen, erklären oder vorhersagen kann." Allerdings ist fraglich, ob die Formulie-

rung, dass der Biomarker im Körper zu messen ist, eine sinnvolle Einschränkung darstellt. Eine alternative Definition haben Gallo et al. [2] ebenfalls gegeben, in der sie einen Biomarker definieren „als eine Substanz, eine Struktur oder einen Prozess, der in Körpermaterial gemessen werden kann und mit gesundheitsbezogenen Resultaten verbunden sein kann." Unserer Meinung nach ist diese Definition zu allgemein. Die Definition des Biomarkers sollte vielmehr eine spezifische Assoziation mit der Gesundheit aufweisen oder mit einem klinischen Resultat verbunden sein [1]. Daher präferieren wir die Definition der „Biomarkers Definitions Working Group" [3] der National Institutes of Health, die wir wie folgt übersetzen: „Ein Biomarker ist eine Charakteristik, die objektiv gemessen und evaluiert werden kann und als Indikator für normale biologische Prozesse, pathogene Prozesse oder für pharmakologische Reaktionen auf eine therapeutische Intervention dient." Molekulare Biomarker sind speziell solche Biomarker, die unter Verwendung molekularer Technologie, wie z.B. Genomik oder Proteomik, oder Bildgebungstechniken entdeckt werden können; für eine umfangreiche Definition s. Ziegler et al. [1].

Prognostische, diagnostische und prädiktive Biomarker

Biomarker werden in allen Patienten-relevanten Bereichen eingesetzt, also Diagnose, Prognose und Therapie. Während prognostische Biomarker den Krankheitsverlauf von Patienten vorhersagen, erlauben diagnostische Biomarker, die Erkrankung zu bestimmen. Prädiktive Biomarker sind grundsätzlich mit der Behandlung eines Patienten verbunden. So lässt sich mit prädiktiven Biomarkern zum Beispiel die Wahrscheinlichkeit für den Behandlungserfolg mit einer bestimmten Behandlung abschätzen oder aber beispielsweise die Wahrscheinlichkeit für eine bestimmte, schwere unerwünschte Wirkung einer Behandlung. Damit lässt sich die Auswahl der besten Behandlung eines Patienten steuern.

Häufig wird im Zusammenhang mit prädiktiven Biomarkern der Begriff „personalisierte Medizin", aber auch der Ausdruck stratifizierte Medizin verwendet. Dabei werden beide Begriffe in der Regel als die Identifikation der optimalen Behandlung sowie die optimale Dosierung und/oder das optimale Timing einer Behandlung in einer Subgruppe von Patienten verstanden. Doch ist es sinnvoll, den Term deutlich breiter zu fassen. Der Begriff sollte beispielsweise auch 1. die Nichtverabreichung einer Behandlung aufgrund von unerwünschten Wirkungen, 2. präventive Maßnahmen oder 3. tatsächlich die für eine einzelne Person maßgeschneiderte Behandlung umfassen [1]. So können z.B. beim Prostatakarzinom DNA-Biomarker verwendet werden, um zu entscheiden, ob eine Phase der intensiven Beobachtung ohne Nachteil für den Patienten anstelle einer unmittelbar zu beginnenden Tumortherapie gewählt werden kann. Möglicherweise ist dadurch nur dann eine radikale chirurgische Intervention mit anschließender Radiotherapie oder Chemotherapie indiziert, wenn der Patient eine aggressive Form des Tumors hat [4]. In anderen Fällen führt die Verwendung von Biomarkerprofilen möglicherweise zu präventiven Interventionen, wie zum Beispiel schon jetzt bei einigen Formen erblicher Tumoren. Dort kann das Ergebnis eines individuellen genetischen Tests die Entscheidungsgrundlage für eine spezifische, manchmal äußerst radikale Intervention, wie z.B. präventiver Chirurgie sein [4].

Tab. 1 Evidenzgrade für diagnostische Methoden entsprechend §11, Abs. 2 am 06.12.2012 [65].

Evidenzgrad	Kriterium
I a	Systematische Übersichtsarbeiten von Studien der Evidenzstufe I b
I b	Randomisierte kontrollierte Studien
I c	Andere Interventionsstudien
II a	Systematische Übersichtsarbeiten von Studien zur diagnostischen Testgenauigkeit der Evidenzstufe II b
II b	Querschnitts- und Kohortenstudien, aus denen sich alle diagnostischen Kenngrößen zur Testgenauigkeit (Sensitivität und Spezifität, Wahrscheinlichkeitsverhältnisse, positiver und negativer prädiktiver Wert) berechnen lassen
III	Andere Studien, aus denen sich die diagnostischen Kenngrößen zur Testgenauigkeit (Sensitivität und Spezifität, Wahrscheinlichkeitsverhältnisse) berechnen lassen
IV	Assoziationsbeobachtungen, pathophysiologische Überlegungen, deskriptive Darstellungen, Einzelfallberichte, u.ä.; nicht mit Studien belegte Meinungen anerkannter Expertinnen und Experten, Berichte von Expertenkomitees und Konsensuskonferenzen

Das ACCE-Modell

Doch wann ist ein Biomarker ein guter Biomarker? Entsprechend dem von dem US-amerikanischen Center of Disease Control (CDC) entwickelten ACCE-Modell wird ein Biomarker auf insgesamt vier Kriterien hin überprüft [5]:

1. **A**nalytical validity (analytische Validität),
2. **C**linical validity (klinische Validität),
3. **C**linical utility (klinischer Nutzen) und
4. **E**thical, legal, and social implications (ethische, rechtliche und soziale Aspekte, auf deutsch kurz: ELSA).

Die analytische Validität gibt an, wie hoch die technische Zuverlässigkeit eines Biomarkers ist. In der entsprechenden Norm [6] wird dabei zusätzlich zwischen Genauigkeit, Richtigkeit und Präzision unterschieden. Hier misst die Richtigkeit, ob der Mittelwert einer großen Serie von Versuchen nahe an dem theoretisch erwarteten Wert liegt. Präzision hingegen betrachtet die Variabilität der Messungen. Genauigkeit schließlich ist die Kombination von Richtigkeit und Präzision.

Die klinische Validität gibt hingegen an, wie gut der Biomarker eine Krankheit erkennen oder vorhersagen kann. In der Praxis lässt sich nicht für alle Biomarker ein einheitliches Maß festlegen, ab wann ein Biomarker als klinisch valide zu bezeichnen ist. Dieses hängt unter anderem davon ab, ob alternative Vorhersagemodelle verfügbar sind, welches Ziel mit dem Biomarker verfolgt wird und wie stark die Belastung durch das Krankheitsbild wäre. Ohne adäquate Behandlungsmöglichkeit lässt sich der Einsatz eines diagnostischen Biomarkers aber, unabhängig von seiner Güte, nur selten rechtfertigen.

Die Beurteilung des klinischen Nutzens eines Biomarkers wird ausführlich im nächsten Abschnitt beschrieben. Das letzte Kriterium des ACCE-Modells schließlich bezieht sich auf ethische,

Tab.2 Phasen diagnostischer oder prognostischer Biomarkerstudien. Nach [1].

Phase	Beschreibung	Studienziel	Übliche Stichprobenumfänge
Ia	Entdeckung	Identifikation vielversprechender Biomarker	10–100
Ib	Entwicklung und Validierung des Assays	Definition und Optimierung des analytischen Prozesses in einen genauen, richtigen und präzisen Test	10–100
Ic	Retrospektive Validierung	Klinischer Assay entdeckt die Krankheit; Entwicklung eines ersten Algorithmus für einen Multimarker-Test	50–500
II	Retrospektive Verfeinerung	Validierung von Früherkennungseigenschaften des Biomarkers (des Biomarker-Sets); Entwicklung und/oder Verfeinerung des Algorithmus für einen Multimarker-Test	100–1000
III	Prospektive Untersuchung	Bestimmen der diagnostischen Genauigkeit (Sensitivität, Spezifität) in der Situation der klinischen Routine	200–1000
IVa	Randomisierte kontrollierte klinische Studie	Untersuchung der Wirksamkeit des Biomarkers im therapeutischen Setting	200–1000
IVb	Gesundheitsökonomische Studie	Quantifizierung der Kostenwirksamkeit	Hängt stark von der Größe des Risikos ab

rechtliche und soziale Implikationen, die im Kontext mit dem Einsatz eines Biomarkers auftreten können, und wird detailliert in der Literatur beschrieben, z.B. [7].

Evidenzgrade und Phasen diagnostischer Studien

Um die Anwendung eines Biomarkers in der Praxis rechtfertigen zu können, muss die Beurteilung des klinischen Nutzens positiv ausfallen, und dieses setzt sowohl eine hohe klinische Validität als auch eine hohe analytische Validität voraus. Die Güte der klinischen Validität wird dabei entscheidend von der Qualität der klinischen Studien beeinflusst.

Für Diagnosestudien hat der Gemeinsame Bundesausschuss die in **Tab.1** dargelegten Evidenzstufen in seiner Verfahrensordnung zur Nutzenbewertung verbindlich festgelegt. Diese Evidenzgrade gelten gleichermaßen auch für Methoden der Früherkennung. Den höchsten Evidenzgrad haben randomisierte Therapiestudien bzw. Meta-Analysen randomisierter Therapiestudien. Doch wie passen Biomarker zum Zwecke der Diagnose mit randomisierten Therapiestudien zusammen? Weiter oben wurde dargelegt, dass das Resultat eines diagnostischen Biomarkers einen Effekt auf die nachfolgenden Behandlungen haben sollte – der Biomarker also prädiktiv sein sollte. Entsprechend haben diagnostische Biomarker, die ihren Wert für den praktischen Einsatz in einer oder mehreren randomisierten Therapiestudien gezeigt haben und damit vom diagnostischen zum prädiktiven Biomarker werden, den größten klinischen Nutzen.

Rein diagnostische Biomarker ab Grad II und niedriger, deren Nutzen nicht im Zusammenhang mit Behandlungskonzepten untersucht wurde, haben einen geringeren Evidenzgrad. Allerdings beeinflussen das verwendete Studiendesign und die methodische Qualität der diagnostische(n) Studie(n) den Evidenzgrad diagnostischer Studien erheblich. Die Studiendesigns, die zum Einsatz kommen, hängen wiederum von der jeweiligen Phase der Biomarker-Studie ab.

Im Wesentlichen können vier Phasen für diagnostische und prognostische Biomarker voneinander unterschieden werden (**Tab.2**) [1, 8–12]. In Phase I geht es um technische und methodische Voruntersuchungen. Hier wird die Frage gestellt, ob der Biomarker prinzipiell als diagnostischer Test geeignet ist. Im Regelfall wird in Phase I eine Fall-Kontroll-Studie mit Kranken und Gesunden, häufig mit extrem Kranken und eindeutig Gesunden oder gar „hypernormalen Personen“ gewählt.

Diese Phase lässt sich noch deutlich feiner unterteilen, da heute mit den modernen Technologien an einer einzelnen Probe nicht nur ein einzelner Biomarker, sondern gegebenenfalls mehrere Millionen Biomarker gleichzeitig gemessen werden können (Phase Ia). Aus diesen vielen Messungen muss dann zunächst der richtige bzw. die richtigen Biomarker ausgewählt werden. Diese Hochdurchsatz-Technologien sind im Regelfall pro Einzelmolekül nicht so genau wie Messtechnologien, die auf einen einzelnen Biomarker speziell zugeschnitten sind. Daher fällt in diese Phase im Regelfall auch die analytische Validität des Biomarkers, zum Beispiel die Entwicklung eines speziellen Assays (Phase Ib).

Da die Hochdurchsatz-Technologien oft teuer sind, sind die Stichprobenumfänge dieser initialen Studien häufig gering. Bevor nun deutlich kostenintensivere größere Validierungsstudien durchgeführt werden, wird in der Regel mit Proben aus Biobanken der diagnostische Wert eines neuen Biomarkers bzw. mehrerer diagnostischer Biomarker ein erstes Mal überprüft (Phase Ic). Möglicherweise ist eine Kombination von mehreren Biomarkern vielversprechender als eine einzige Biomarkermessung. Wie sich die verschiedenen Biomarker am besten in einer Multimarkerregel kombinieren lassen, wird häufig in derselben Phase mit aufwendigen biostatistischen Verfahren, z.B. Verfahren des maschinellen Lernens untersucht.

In Phase II wird die Validität retrospektiv bei ausgewählten Personen überprüft. Hier wird der Frage nachgegangen, ob der Test seinen Zweck erfüllt, z.B. die Erkrankung erkennen kann.

Phase III ist die kontrollierte diagnostische Studie, bei der untersucht wird, wie präzise der Test im klinischen Alltag ist. Das Studiendesign der Wahl ist hierfür eine Kohortenstudie an symptomatischen Patienten. Ziel der Phase IV ist es schließlich, die Wirksamkeit des Biomarkers zu zeigen. Im ersten Teil dieser Phase wird dabei der Frage nachgegangen, wie der Test den klinischen Verlauf beeinflusst. Anschließend werden Kosten-Nutzen-Untersuchungen angestellt. In unserer eigenen Praxis hat sich gezeigt, dass eine etwas feingliedrigere Aufteilung der vier Phasen, insbesondere in der frühen Phase I hilfreich ist [1].

Tab.3 Grundprinzipien von Validierungsstudien diagnostischer Biomarker. Nach Weinstein et al. [21].

Prinzip	Erläuterung
Zwei Stichproben von Personen	Personen mit der Krankheit zur Schätzung der Sensitivität; Gruppe von Personen ohne die Krankheit zur Schätzung der Spezifität
Wohldefinierte Stichproben	Unabhängig vom Rekrutierungsschema Beschreibung der Charakteristika der Personen (z.B. Alter, Geschlecht, Krankheitsstadium, Komorbiditäten)
Wohldefinierter diagnostischer Test	Präzise eindeutige Definition des diagnostischen Tests; Anwendung in gleicher Weise bei allen Studienteilnehmern
Goldstandard / Referenzstandard	Bestimmung des wahren Krankheitsstatus jedes Studienteilnehmers durch einen perfekten oder fast perfekten Standard
Stichprobe von Untersuchern	Benötigt der Test geübte Beurteiler, werden zwei oder mehr Beurteiler benötigt
Verblindete Untersuchungen	Erhebung von Goldstandard/Referenzstandard und diagnostischem Test unabhängig voneinander und verblindet
Standardisiertes Berichten der Ergebnisse	Studienergebnisse sollten entsprechend Empfehlungen für Studien zur diagnostischen Genauigkeit publiziert werden

Methodische Grundprinzipien von Validierungsstudien diagnostischer Biomarker

Phase-III-Biomarkervalidierungsstudien haben als zentrales Kennzeichen die prospektive konsekutive Rekrutierung von symptomatischen Patienten, die dem üblichen Patientenspektrum entsprechen. Phase-IV-Validierungsstudien sind randomisierte Therapiestudien. Häufig werden ganz spezielle Studiendesigns [1, 13–19] verwendet, doch das wichtigste methodische Element dieser Studien ist die Randomisierung. Doch weder die Randomisierung noch die prospektive Rekrutierung allein garantieren eine hohe methodische Qualität und damit letztendlich die Gültigkeit einer Diagnosestudie. Viel mehr gilt für beide Studientypen eine ganze Reihe methodischer Grundprinzipien. Bei Phase-IV-Validierungsstudien sind dieses dieselben Grundprinzipien wie für alle klinisch-therapeutischen Studien [20].

Die Grundprinzipien diagnostischer Studien der Phasen I bis III sind in **Tab.3** zusammengestellt [8, 21, 22]. Werden diese methodischen Prinzipien nicht eingehalten, können die Schätzungen von Sensitivität und Spezifität, also der klinischen Validität des Biomarkers, erheblich verzerrt sein.

Wichtige Verzerrungsquellen von Validierungsstudien diagnostischer Biomarker

Die Fehler in Studiendesigns, die zu Verzerrungen bei diagnostischen Studien führen können, sind vielfältig. Betrachten wir, im Sinne eines Kochrezepts, die „Zutaten" einer Diagnosestudie, dann werden zunächst einmal Personen benötigt, die an der Studie teilnehmen. Entsprechend ist eine der wichtigsten Störquellen die Verzerrung, die durch eine selektive Auswahl oder die Teilnahmebereitschaft entstehen kann.

Als nächstes ist ein Indextest (= der ausgewählte zu testende Biomarker) erforderlich, und im Labor können zufällig oder auch systematisch Fehler auftreten. Darüber hinaus wird ein Referenzstandard zum Vergleich mit dem neuen Test benötigt. Und auch hier gibt es eine Reihe von Fehlerquellen; die relevanten werden weiter unten beschrieben. Ein Beispiel für das Zusammenspiel von Index- und Referenztest ist der Nachweis von zyklisch citrullinierten Peptiden (Anti-CCP-Ak) bei der rheumatoiden Arthritis [23]. Als Indextest zum Nachweis von Anti-CCP-Antikörpern dient ein ELISA der dritten Generation; als Referenzstandard werden in der Regel die Klassifikationskriterien des American College of Rheumatology verwendet.

Wichtig für die Beurteilung von Tests zur diagnostischen Genauigkeit ist, dass es möglicherweise eine Wechselwirkung zwischen dem Indextest und dem Referenzstandard geben kann. So können die beiden Tests z.B. zu so verschiedenen Zeitpunkten gemessen werden, dass sich in der Zwischenzeit der wahre Zustand verändert haben kann. Oder die Kenntnis des Ergebnisses des einen Tests beeinflusst in irgendeiner Weise das Vorgehen für den anderen Test oder, selbst bei identischer Vorgehensweise, dessen Ergebnis. Auch bei der Beurteilung der Tests selbst besteht die Möglichkeit, Fehler zu begehen.

Am Ende der Studiendurchführung werden die Daten analysiert. Hier stellt sich beispielsweise die Frage, wie mit fehlenden oder nicht interpretierbaren Testergebnissen umgegangen wird. Schließlich sollte eine Diagnosestudie auch nachvollziehbar publiziert werden. Entsprechende Fehler lassen sich vermeiden, indem die Veröffentlichung gemäß dem STARD-Statement (Standards for the Reporting of Diagnostic Accuracy Studies) erstellt wird [24].

Die oben beschriebenen wichtigen Verzerrungsquellen sind in **Tab.4** zusammengefasst, und eine ausführliche Diskussion findet sich in den Referenzen [21, 25–29].

Verzerrung auf der Ebene von Personen in der Studie

Tab.4 listet als erstes die Stichprobenverzerrung, weil die nicht repräsentative Auswahl die wichtigste Störquelle überhaupt ist. Die Möglichkeiten dieser sogenannten Auswahlverzerrung sind vielfältig [28, 30, 31], und eine längere Liste von Rekrutierungsproblemen, die zu dieser Verzerrung führen können, sind im Kasten dargestellt; die ersten drei dort dargestellten Punkte werden in der Regel unter dem englischen Begriff „spectrum composition bias" zusammengefasst.

Auswahlverzerrung

Dass die Auswahlverzerrung die Verzerrungsquelle mit dem größten Einfluss auf die Schätzungen von Sensitivität und Spezifität in Validierungsstudien mit Biomarkern ist, wurde in mehreren systematischen Übersichtsarbeiten gezeigt [25–27, 32]. **Abb.1** fasst die Resultate dieser Veröffentlichungen zusammen. Dort wird dargestellt, wie sich die Genauigkeitsschätzungen, ausgedrückt als Chancenverhältnis (Odds Ratio), ändern, wenn ein bestimmtes Element des methodischen Studiendesigns nicht vorhanden ist im Vergleich mit seinem Vorhandensein. Ist das relative diagnostische Chancenverhältnis größer

Tab.4 Wichtige Störquellen von Validierungsstudien diagnostischer Biomarker; in Anlehnung an Ref. [21].

Verzerrung	Erläuterung
Stichprobenverzerrung	Die Personen sind nicht repräsentativ für Personen, die den Test erhalten
Partielle Verifikationsverzerrung	Nur bei einem ausgewählten Teil der Personen, also nicht bei allen, wird der Referenzstandard erhoben
Differentielle Verifikationsverzerrung	Verwendung verschiedener Referenzstandards, und zwar in Abhängigkeit des Testergebnisses
Krankheitsverlaufsverzerrung	Die Zeit zwischen Referenzstandard und Indextest ist so lang, dass sich die Krankheitsbedingungen ändern
Einschlussverzerrung (Incorporation bias)	Referenzstandard und Indextest sind nicht unabhängig voneinander, z.B. wenn der Indextest Teil des Referenzstandards ist
Test-Beurteilungsverzerrung	Der Indextest wird in Kenntnis des Ergebnisses des Referenzstandards beurteilt
Referenzstandard-Beurteilungsverzerrung	Der Referenzstandard wird in Kenntnis des Ergebnisses des Indextests beurteilt
Klinische Beurteilungsverzerrung	Der Indextest wird unter Verwendung klinischer Daten beurteilt, die in der Routine nicht verfügbar sind

Gründe für eine Auswahlverzerrung

- Die Kontrollgruppe besteht aus extrem Gesunden (hypernormal controls).
- Es werden Fälle mit eingeschränktem Krankheitsspektrum eingeschlossen, z.B. durchweg schwere Fälle (selection for symptoms; severe cases);
- in der gleichen Weise kann sich die Zuweisung von Personen in die Studie von der Zuweisung von Personen in der Praxis unterscheiden; beispielsweise unterscheidet sich das Patientenspektrum in der Notaufnahme von dem in einer Tagesklinik.
- Personen werden in Abhängigkeit des Ergebnisses des Indextests in die Studie eingeschlossen (referral for index test bias); diese Verzerrung ist nicht identisch mit der Verifikationsverzerrung; bei der Verifikationsverzerrung wird nicht bei allen Studienpatienten der Referenzstandard erhoben;
- Gesunde kommen nicht zur Nachbeobachtung, und die Daten fehlen entsprechend (loss to follow-up bias).
- Nur ein eingeschränkter Kreis von Personen nimmt an der Studie teil; beispielsweise nur Personen mit gesicherter Diagnose (participation bias, auch self-selection bias).
- Es werden nur Personen eingeschlossen, bei denen bestimmte Voruntersuchungen vorliegen (limited challenge bias).
- Es werden nur Personen eingeschlossen, bei denen bestimmte Vordiagnosen vorliegen (increased challenge bias).
- Es werden nur Personen eingeschlossen, die bestimmte Untersuchung „durchhalten“ bzw. „geeignet sind“ (study examination bias).

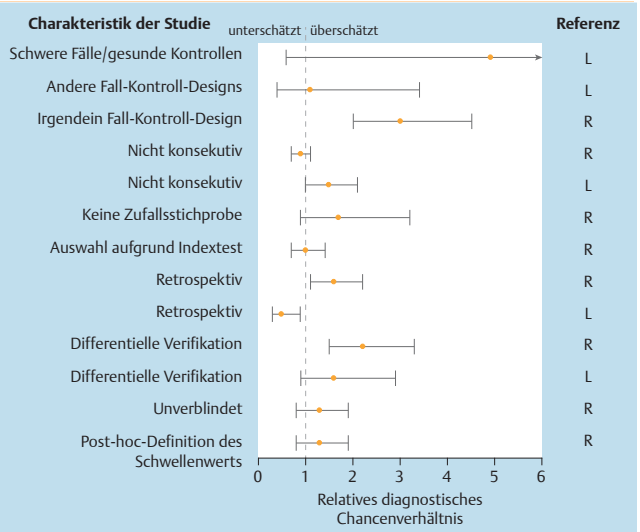

Abb.1 Einfluss verschiedener Charakteristika des Studiendesigns auf die Schätzungen der diagnostischen Genauigkeit, berichtet von Lijmer et al. (bezeichnet mit Referenz L) [26] bzw. Rutjes et al. (bezeichnet mit Referenz R) [27].

als 1, bedeutet dieses, dass Studien mit diesem Defizit im Studiendesign zu größeren Genauigkeitsschätzungen führen als Studien, die diesen Mangel nicht haben [27]. Entsprechend wird die Genauigkeit des Biomarkers durch die Verzerrung überschätzt.

Werden extrem kranke Fälle mit gesunden Kontrollen verglichen, wird die diagnostische Genauigkeit im Durchschnitt um etwa das Fünffache überschätzt. Das klassische Beispiel hierfür ist die Geschichte des carcinoembryonalen Antigens (CEA) [33] zur Diagnose des kolorektalen Karzinoms. In einer Fall-Kontroll-Studie, in die nur Patienten mit einem bekannten fortgeschrittenen kolorektalen Karzinom oder Rektumkarzinom eingeschlossen wurden, war der CEA-Wert bei 35 der 36 Patienten erhöht. Die CEA-Werte waren deutlich niedriger bei gesunden Kontrollen [34]. In Nachfolgestudien wurden Patienten mit deutlich niedrigeren Krankheitsstadien des kolorektalen Karzinoms eingeschlossen, und die Genauigkeit des CEA-Tests sank drastisch (z.B. [35]). Konsequenterweise wurde der CEA-Tests sowohl für die Neudiagnose als auch das Screening aus der klinischen Routine verbannt [33].

Dass die Unterschiede in den Genauigkeitsschätzungen tatsächlich sehr groß sein können, zeigt eine Arbeit von Lachs et al. [36] zur Untersuchung von Harnwegsinfektionen durch Leukozyten-Esterase und bakterielles Nitrit mit einem Teststreifen. In die Studie wurden konsekutiv 366 erwachsene Patienten eingeschlossen. Bei den 107 Patienten, die aufgrund klinischer Zeichen und Symptome eine hohe a priori Wahrscheinlichkeit für eine Harnwegsinfektion hatten, betrug die Sensitivität des Tests mit dem Teststreifen 92% bei einer Spezifität von 42%. Hingegen war die Sensitivität bei den 259 Patienten mit niedriger a priori Wahrscheinlichkeit für eine Harnwegsinfektion mit 56% gering, und die Spezifität betrug hier 78%. Die Zusammensetzung der Patientengruppen hat hier also eine vollständige Veränderung der Eigenschaften des Biomarker-Tests zur Folge.

Werden Fall-Kontroll-Studiendesigns verwendet, bei denen nicht extrem kranke Patienten eingeschlossen werden, ist die Verzerrung nicht so ausgeprägt wie bei der Verwendung extremer Patienten- und Probandengruppen. Insgesamt aber muss

man dennoch bei der Verwendung von Fall-Kontroll-Studien mit sehr optimistischen Ergebnissen rechnen, und zwar mit einer Überschätzung der diagnostischen Genauigkeit um etwa den Faktor drei.

In der Literatur gibt es viele weitere Beispiele, wie die Auswahl der Studiengruppe zu einer Verzerrung der Ergebnisse führen kann, selbst wenn es sich nicht um Fall-Kontroll-Studien, sondern um prospektive Studien handelt. So wurde z.B. in einer prospektiven Studie, bei der Föten zwischen der 11. und der 14. Schwangerschaftswoche mit Ultraschall untersucht wurden, der diagnostische Wert des Fehlens des Nasenbeins als ein Hinweis auf eine chromosomale Anomalie analysiert, und die Sensitivität des Tests wurde mit 69% angegeben [37]. Allerdings wurden die Auswertungen auf die Föten beschränkt, die eine Trisomie 21 aufwiesen (**Abb.2**) – auch der Titel der Arbeit beschränkte sich auf diese Fehlbildung. Dabei lassen sich mit der Chorionzottenbiopsie – diese war der Referenzstandard in [37] – oder der Amniozentese auch andere chromosomale Anomalien entdecken. Insgesamt wurden so 295 Föten ausgeschlossen; bei den ausgeschlossenen Föten betrug die Sensitivität aber nur 32%. Und damit ist der Test zum Entdecken chromosomaler Anomalien genau so gut wie ein Münzwurf, denn die Sensitivität bei Berücksichtigung aller Föten beträgt nur 52% (**Abb.2**).

a

Chromosomale Anomalie		
Nasenbein	Ja	Nein
nicht vorhanden	229	129
vorhanden	104	5094

b

Chromosomale Anomalie		
Nasenbein	Ja	Nein
nicht vorhanden	324	129
vorhanden	304	5094

Abb.2 Auswahlverzerrung. **a**) Beobachtete Häufigkeiten, wenn nur Föten mit Trisomie 21 eingeschlossen werden. **b**) Beobachtete Häufigkeiten, wenn alle Föten mit chromosomalen Anomalien berücksichtigt werden.

Konsekutive Rekrutierung von Studienteilnehmern

Werden Personen nicht konsekutiv, also unmittelbar aufeinanderfolgend, in eine Studie eingeschlossen, ist nicht eindeutig klar, ob eine Verzerrung zu erwarten ist (**Abb.1**). Wenn mit der Rekrutierung eine selektive Auswahl in die Studie verbunden ist, wird die nicht konsekutive Rekrutierung zu einer Verzerrung führen. In anderen Fällen ist dieses nicht unbedingt zu erwarten. Wichtig ist allerdings festzuhalten, dass bei einer nicht konsekutiven Rekrutierung eine Verzerrung nicht ausgeschlossen werden kann und diese möglicherweise sogar recht groß ist. Überraschenderweise kann diese Verzerrung auch dann auftreten, wenn die Personen zufällig ausgewählt werden (**Abb.1**), da auch hier keine konsekutive Reihe von Personen vorliegt.

Retrospektive Studie versus prospektive Studie

Bei Biomarkerstudien sind die Begriffe „prospektiv“ und „retrospektiv“ nicht einheitlich definiert. Allgemein erwartet man bei einer prospektiven Studie, dass sämtliche Untersuchungen zunächst geplant und dann durchgeführt werden. Dieses schließt insbesondere die Rekrutierung von Patienten erst nach Abschluss der Planung ein. Bei Studien wird jedoch häufig zum einen auf bereits durchgeführte randomisierte Therapiestudien, zum anderen auf prospektiv gesammelte Bioproben zurückgegriffen. Hier werden dann prospektiv Hypothesen zu spezifischen Biomarkern gestellt und auch prospektiv an vorhandenem Biomaterial überprüft. Ein Vorteil dieser Form der Studiendurchführung ist, dass das Probenmaterial im Regelfall entsprechend einem hohen Qualitätsstandard mit einem strikten Protokoll gewonnen werden kann. Allerdings ist der Zeitpunkt, wann das Biomaterial gewonnen wird und welche Art des Biomaterials vorliegt, in einer Biobank sehr eingeschränkt. So kann eine Biobank ungeeignet sein, wenn Proteine oder Metaboliten betrachtet werden. Deren Konzentrationen kann sich sehr schnell ändern, so dass ein spezifischer Abnahmepunkt der Probe sowie ein definierter Messpunkt relevant sein kann; für eine ausführliche Diskussion siehe [17, 38]. Das Biomaterial muss hier also auf einen Zeitpunkt genau verfügbar sein, und dieses Problem tritt bei einer prospektiven Rekrutierung nicht auf. Allerdings ist der Zeitpunkt der Biomarkermessung in anderen Fällen unproblematisch, wie z.B. bei der Bestimmung von genetischen Markern, also DNA-Markern, da diese lebenslang konstant sind.

Im Gegensatz zu prospektiven Studien sind retrospektive Studien solche, bei denen sowohl die Personen schon rekrutiert als auch die Biomarkermessungen schon durchgeführt sind. Dieser Ansatz wird häufiger im Bereich der Genetik verwendet, wo die Daten verschiedener sogenannter genomweiter Assoziationsstudien für die Bestätigung eines neuen Biomarkerbefundes verwendet werden. Diese Form der retrospektiven Studie wird dann in der Regel als „In-silico-Replikation“ bezeichnet.

In der Regel werden prospektive Studien zu einer geringeren Verzerrung führen als retrospektive Studien. Doch das gilt nicht immer [38], und es gibt wenige Situationen, in denen eine retrospektive Studie besser sein kann als eine prospektive Studie. Zwei offensichtliche Aspekte, die zur Verwendung retrospektiver Studien führen, sind der hohe Zeitaufwand und die Kosten für eine prospektiv geplante Studie. Da auf der anderen Seite die prospektive Studie standardisiert durchgeführt wird, kann eine einheitliche Qualitätskontrolle aller Daten erfolgen, was im Normalfall zu einer höheren Validität der Studie führt. Bei einer äußerst gut geführten Biobank sind diese Aspekte in der Regel allerdings auch erfüllt. Hingegen kann es bei prospektiven Studien auch zu Verzerrungen kommen. So ist das Prinzip der Beobachtungs- und Behandlungsgleichheit verletzt, wenn die behandelnden oder anderweitig beteiligten Ärzte das Ergebnis eines Tests kennen und die Kenntnis des Testergebnisses ihr Handeln in irgendeiner Form beeinflusst. Dennoch muss bei retrospektiven Studien immer wieder hinterfragt werden, ob die Daten überhaupt vollständig sind, standardisiert erfasst wurden, die Messungen an den richtigen Zeitpunkten stattgefunden haben und dass es nicht doch zu einer Auswahlverzerrung gekommen ist.

Zusammengefasst sind prospektive Studien fast immer retrospektiven Studien überlegen. Dieses wird auch durch die systematische Übersicht von Rutjes et al. [27] bestätigt, die zeigen, dass Studien, in denen die Datensammlung retrospektiv erfolgte, die diagnostische Genauigkeit der Tests um 60% überschätzten.

Auswahl aufgrund des Indextests

Eine überraschende Tendenz für eine Verzerrung zeigt sich, wenn bewusst Personen in Abhängigkeit des Ergebnisses des Indextests in die Studie rekrutiert werden. Knotterus und Muris [38] haben diese Verzerrungsmöglichkeit detailliert beschrie-

a	Referenzstandard	
Indextest	Positiv	Negativ
Positiv	80	10
Negativ	20	40

Korrektur für Verifikations-Verzerrung →

b	Referenzstandard	
Indextest	Positiv	Negativ
Positiv	80	10
Negativ	20 + 60 = 80	40 + 120 = 160

Abb.3 Partielle Verifikationsverzerrung **a)** Beobachtete Häufigkeit **b)** Korrigierte Häufigkeit

ben. So können beispielsweise bevorzugt Personen mit klarer Symptomatik oder gar einer Reihe Ergebnissen wenig zuverlässiger Tests, mit widersprüchlichen Testergebnissen, oder gar mit einem positiven Ergebnis im Indextest eingeschlossen werden. Diese Patienten mit klaren Symptomen, aber möglicherweise einer anderen Differentialdiagnose, sind im Allgemeinen sehr schwierig zu beurteilen, so dass es zu einer Abnahme der Richtigkeit der Einschätzung kommen kann. Entsprechend führt dieses Vorgehen tendenziell zu einer Unterschätzung der Genauigkeit (**Abb.1**) [27].

Verzerrung auf der Ebene der Tests

Die Entscheidung, welches Verfahren als Referenzstandard geeignet ist, ist in manchen Anwendungen äußerst schwierig. Selbstverständlich sollte der Referenzstandard perfekt oder fast perfekt sein. Doch selbst erfahrene Pathologen oder Radiologen sind nicht unfehlbar. In einigen diagnostischen Problemen ist ein Referenzstandard nicht einmal verfügbar, wie bei der Epilepsie, oder die Anwendung des Referenzstandards ist aufgrund seines hohen Risikopotentials unethisch. Doch auch in diesen Situationen kann die Ähnlichkeit der Biomarkermessung mit den Resultaten anderer Tests berichtet werden, und es können auch hier Sensitivität und Spezifität bestimmt werden. Dieses ist immer noch besser, als Patienten einfach von Studien auszuschließen.

Verifikationsverzerrung

Neben der selektiven Auswahl von Kranken und Gesunden für die Studie ist die Verifikationsverzerrung die zweite wichtige Störquelle. Im Englischen werden hierfür synonym neben „verification bias“ die Begriffe „work-up bias“, „referral bias“ oder „ascertainment bias“ verwendet. Speziell unterscheidet man zwischen partieller Verifikationsverzerrung, bei der der Referenzstandard nur bei einem Teil der Studienteilnehmer angewendet wird, und differentieller Verifikationsverzerrung. Bei der differentiellen Verifikationsverzerrung werden verschiedene Referenzstandards verwendet, und zwar in Abhängigkeit des Ergebnisses des Indextests, hier also des Biomarkers. Diese Form der Verzerrung ist im Englischen nicht nur unter dem Begriff „differential verification bias“ bekannt, sondern auch unter der Bezeichnung „double gold standard bias“ bzw. „double reference standard bias“.

Differentielle Verifikations-Verzerrung führt zu einer deutlichen Überschätzung der Studienergebnisse (**Abb.1**). Häufig tritt dieses Phänomen auf, wenn der Referenzstandard in irgendeiner Form invasiv, z.B. mit einer Operation verbunden ist. Die invasive Diagnostik wird dann nur bei Test-Positiven durchgeführt. Ein anderer Referenzstandard, wie z.B. die klinische Nachbeobachtung bei Personen mit negativem Testergebnis wird dann bei Test-Negativen verwendet.

Ein Beispiel hierfür ist die Diagnosestudie zur Lungenembolie [39], bei der die Lungenperfusions-/-ventilationsszintigraphie als diagnostischer Test eingesetzt wurde. Der Referenzstandard ist die radiologische Betrachtung der Lungenarterie, die allerdings mit größerer Wahrscheinlichkeit durchgeführt wird, wenn die Szintigraphie positiv ausfällt. Hingegen wird bei Personen mit negativem Ergebnis der Szintigraphie einfach eine klinische Nachbeobachtung durchgeführt.

Genau wie die differentielle Verifikationsverzerrung kann auch die partielle Verifikationsverzerrung zu erheblichen Verzerrungen führen. Allerdings ist dieses Phänomen in der Praxis nicht so stark ausgeprägt; detaillierte Beschreibungen des Phänomens geben Referenzen [40, 41].

Zur Illustration dient hier das fiktive in **Abb.3** gegebene Beispiel; reale Beispiele geben z.B. [40–44]. Gehen wir davon aus, dass bei negativem Indextest nur in einem Viertel der Fälle auch der Referenztest durchgeführt wird, hingegen alle Personen mit positivem Indextest den Referenztest erhalten. Weiterhin gehen wir davon aus, dass in der Studie die in **Abb.3a** dargestellten Häufigkeiten beobachtet wurden. Zur Berechnung der Sensitivität werden nur die Personen mit positivem Referenzstandard benötigt. Werden die beobachteten Häufigkeiten hierfür herangezogen, beträgt die Sensitivität 80% [= 80/(80 + 20)]. Gleichzeitig beträgt die Spezifität, für die nur die Studienteilnehmer mit negativem Ergebnis des Referenzstandards verwendet werden, ebenfalls 80% [= 40/(40 + 10)].

Da allerdings der Referenzstandard nur bei einem Viertel aller Biomarker-Test-Negativen, also Indextest-Negativen angewendet wurde, müssten für eine korrekte Berechnung von Sensitivität und Spezifität die Häufigkeiten korrigiert werden. Die einfache Hochrechnung ist in **Abb.3b** dargestellt. Diese Tabelle entsteht aus **Abb.3a,** indem die Häufigkeiten der Test-negativen um den Faktor drei erhöht werden. Mit diesen korrigierten Zahlen erhält man nun für die Sensitivität nur noch einen Wert von 50% [= 80/(80 + 80)] sowie für die Spezifität einen Wert von 94,11% [= 160/(160 + 10)]. Das bedeutet, dass durch die partielle Verifikationsverzerrung in der fiktiven Studie die Sensitivität erheblich überschätzt, die Spezifität hingegen deutlich unterschätzt wird. Gleichzeitig sinkt der Anteil der korrekt diagnostizierten Studienteilnehmer von 80% [= (80 + 40)/(80 + 10 + 20 + 40)] auf 72,72% [= (80 + 160)/(80 + 10 + 80 + 160)]. Zur Korrektur der partiellen Verifikationsverzerrung sind in der Literatur zwei bekannte statistische Verfahren vorgeschlagen worden, und sie firmieren unter den Bezeichnungen Begg-Greenes-Methode [45] sowie Diamond-Methode [46].

In gleicher Weise führt die Verwendung verschiedener Referenzstandards, also die differentielle Verifikationsverzerrung, zu einer 60%igen Überschätzung der diagnostischen Genauigkeit eines Tests im Vergleich zu Studien, die nur einen einzigen Referenzstandard verwendet haben.

Verblindung

Die offensichtlichste mögliche Fehlerquelle auf der Ebene der Tests ist die fehlende Verblindung. Fehlt sie, muss mit einer Überschätzung der diagnostischen Genauigkeit gerechnet werden. Die Verzerrung, die aus der fehlenden Verblindung entsteht, wird im Englischen „review bias“ genannt. Selbstverständlich ist die Verblindung bei „weichen“ Zielkriterien, wie klinischen Symptomen, relevanter als bei „harten“ Endpunkten, wie Biomarkermessungen im Labor, obwohl sich gerade Laboruntersuchungen auf einfache Weise durch die geeignete Kodierung der Proben verblinden lassen. Der Preis für die fehlende Verblindung liegt im Durchschnitt bei einer etwa 30%igen Überschätzung der diagnostischen Genauigkeit [27] (**Abb.1**).

Goldstandard versus Referenzstandard

In den vorherigen Ausführungen wurde immer direkt der Begriff Referenzstandard verwendet; dieses ist auch der Begriff, der sich in einigen Richtlinien wiederfindet, zum Beispiel im STARD-Statement [24]. In der Umgangssprache sowie der Literatur findet sich darüber hinaus der Begriff des Goldstandards. Der Goldstandard soll den *wahren Krankheitszustand* einer Person widerspiegeln. Hingegen ist der Referenzstandard die beste *verfügbare* Methode, um den Krankheitszustand der Person zu bestimmen [47]. Wichtiger ist jedoch, dass sich Goldstandard und Referenzstandard voneinander unterscheiden können. Insbesondere könnte der Referenzstandard fehlerhaft sein und eher mit dem Indextest übereinstimmen, also dem Biomarkertest von Interesse. Auf alle Fälle muss man mit unterschiedlichen Schätzungen von Sensitivität und Spezifität rechnen, wenn sich Goldstandard und Referenzstandard voneinander unterscheiden.

Obwohl der Referenzstandard möglichst perfekt sein sollte, ist es in der Praxis häufig schwierig, einen guten Referenzstandard zu wählen. In bildgebenden Studien werden häufig Operation, pathologische Befunde oder die klinische Nachbeobachtung als Standard gewählt [21]. Wird ein Referenzstandard verwendet, der nicht dem üblichen Standard entspricht, wird, wie das Beispiel [48] zeigt, die Validität der ganzen Studie in Frage gestellt [49]. In ihrer Arbeit haben Dehdashti und Kollegen [48] Bariumbrei als Referenzstandard zur Diagnose der Refluxerkrankung verwendet, obwohl die Nordamerikanische Gesellschaft für pädiatrische Gastroenterologie, Hepatologie und Ernährung dieses nicht unterstützt. Der gegenwärtige Referenzstandard ist hier das Monitorieren des pH-Werts in der Speiseröhre. Zusammen mit anderen methodischen Schwächen hat der Autor des Leserbriefs [49] geschrieben: „…, this study has several critical methodological flaws, …“

Einschlussverzerrung: Referenzstandard und Indextest hängen voneinander ab

In einigen Fällen ist der Indextest Teil des Referenzstandards. Entsprechend sind die beiden Tests nicht unabhängig voneinander. Das bekannteste Beispiel hierfür wird von Guyatt et al. [33] gegeben. In einer Studie zu Screening-Instrumenten auf Depression bei Personen im terminalen Krankheitsstadium wurden 100% Sensitivität und 100% Spezifität beobachtet. Der Indextest bestand dabei in neun Fragen, von denen eine Frage lautete: „Sind Sie depressiv?“

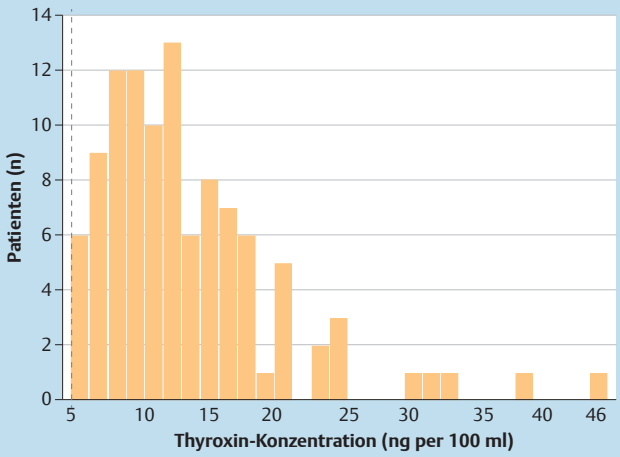

Abb.4 Anzahl der Personen gegen freie Thyroxin-Konzentration (ng per 100 ml) für 105 Personen nach [50] mit Schilddrüsenhormonvergiftung. Die gestrichelte Linie gibt den oberen Wert des Referenzbereichs des Tests an.

Ein zweites klares Beispiel ist die Arbeit von Harvey [50], in der 107 Patienten mit einer Schilddrüsenvergiftung untersucht wurden. Die endgültige Diagnose wurde auf der Basis aller verfügbaren Informationen, einschließlich der Ergebnisse der Schilddrüsenfunktionstests gestellt. Die Schlussfolgerung war, dass der klinische Schweregrad der Erkrankung stärker mit der freien Thyroxinkonzentration assoziiert war als mit irgendeinem anderen der untersuchen Indizes. Allerdings muss hier beachtet werden, dass die primäre Diagnose unter Verwendung der freien Thyroxinkonzentration erfolgte [51]. So haben selbstverständlich alle Fälle der Studie gerade freie Thyroxinkonzentrationen außerhalb des Referenzbereichs (**Abb.4**).

Verzerrung auf der Ebene der Beurteilung der Testergebnisse

Fehlende Werte

Bei vielen molekularen Tests sind die Ergebnisse nicht für alle Personen eindeutig, d.h., unklar oder unsicher oder gar nicht bestimmt worden. Können aber nicht alle als Test positiv oder Test negativ klassifiziert werden, sollten die Werte nicht einfach weggelassen werden. Denn die Häufigkeiten der einzelnen Kategorien sind ein wichtiger Hinweis auf den Nutzen des Tests.

Werden die Ergebnisse weggelassen, können die Schätzer für Sensitivität und Spezifität verzerrt sein. Dieses wird methodisch in einer Reihe von Arbeiten betrachtet, z.B. [52–54]. Illustriert wird dieses Phänomen hier in **Abb.5** anhand der kürzlich publizierten Daten von Ramos et al. [55]. In dieser Arbeit wurden die Wertigkeit des Interferon Gamma Release Assays (IGRA) zur Diagnose der Tuberkulose untersucht und nach unserer Einschätzung die Daten vollständig berichtet und korrekt analysiert. Die vollständigen Daten sind in **Abb.5a** dargestellt. Werden die unklaren Testergebnisse entsprechend der ungünstigsten Kategorien eingeordnet (**Abb.5b**), ergeben sich für die Sensitivität 27/71 = 38,00% und für die Spezifität 238/302 = 78,81%. Werden die fehlenden Werte hingegen ignoriert (**Abb.5c**), erhält man als Schätzer für die Sensitivität 27/67 = 40,30%, und die Spezifität ergibt sich zu 238/280 = 85,00%. Die

Unterschiede zwischen den Schätzungen sind also auch in diesem jüngeren Beispiel deutlich ausgeprägt. Dabei beträgt der Anteil der fehlenden Werte hier nur etwa 7%; in der Literatur finden sich Angaben bis zu 40% fehlende Werte [56].

a	Referenzstandard		b	Referenzstandard		c	Referenzstandard	
Indextest	Positiv	Negativ		Positiv	Negativ		Positiv	Negativ
Positiv	27	42		27	42		27	42
Unklar	4	22		4	22		–	–
Negativ	40	238		40	238		40	238

Abb.5 Fehlende Daten anhand des Beispiels von Ramos et al. [55] zur Diagnose des Interferon-Gamma-Release-Assays zur Diagnose der Tuberkulose. **a**) Vollständige Daten, **b**) Gruppierung der unklaren Befunde in die schlechtere Kategorie, **c**) Ignorieren der fehlenden Werte.

Systematisch wurde der Effekt des Weglassens nicht interpretierbarer Testergebnisse in zwei Studien untersucht [32]. Allerdings findet sich in beiden Arbeiten kein Hinweis auf die Richtung und Stärke der möglichen Verzerrung [57, 58].

Schließlich merken wir an, dass Testergebnisse, die weder eindeutig positiv noch eindeutig negativ sind, von eigenem diagnostischem Wert sein oder auf eine andere Erkrankung hinweisen können [59].

Post-Hoc-Definition des Schwellenwerts

Viele molekulare Biomarker liefern nicht nur ein positives oder negatives Testergebnis, wie z.B. Mutation vorhanden oder nicht vorhanden, sondern ein quantitatives Testergebnis. Erst unter Verwendung eines Schwellenwerts, der in der Regel über Referenzbereiche oder Normwerte bestimmt wird, wird hieraus ein auffälliges oder unauffälliges Testergebnis definiert. Wird der Schwellenwert erst mit den Daten der aktuell durchgeführten Diagnosestudie ermittelt, wird dieser in der Regel so bestimmt, dass Sensitivität und/oder Spezifität in gewisser Weise optimiert werden. Damit resultiert in der Regel eine Überschätzung der diagnostischen Genauigkeit von etwa 30% (**Abb.1**). Entsprechend ist die Definition des Schwellenwerts bzw. bei Verwendung mehrerer Biomarker die Definition und Offenlegung der Multimarker-Regel vor Beginn der Studie von großer Bedeutung.

Variationskoeffizient und Fallzahlplanung

Eine zentrale Frage während der Studienplanung ist die nach der erforderlichen Fallzahl, um Unterschiede in den Mittelwerten der Biomarker zwischen zwei Gruppen entdecken zu können. Es ist intuitiv klar, dass diese sowohl von der Präzision der Biomarkermessungen abhängt als auch vom Unterschied zwischen den Gruppen. Dabei wird die Präzision als Variationskoeffizient (VK) $v = \sigma/\mu$ angegeben, der relativen Streuung der Biomarkermessungen. Gute Labortests haben einen VK von unter 10% = 0,1. Der Unterschied wird in diesem Zusammenhang angegeben als Fold Change $f = \mu_2/\mu_1$ bzw. $100 \times f$. Dieser drückt aus, um welchen Faktor bzw. um wie viel Prozent die durchschnittlichen Biomarkerwerte in Gruppe 2 sich von den durchschnittlichen Biomarkerwerten in Gruppe 1 unterscheiden. Sind z.B. die durchschnittlichen Werte des Biomarkers in Gruppe 2 doppelt so groß wie die durchschnittlichen Werte in Gruppe 1, ist der Fold Change 2.

Wird davon ausgegangen, dass die beiden Gruppen gleich groß sind, kann mit etwa 90%iger Sicherheit und bei dem üblichen Signifikanzniveau von 5% ein Unterschied zwischen den beiden Gruppen bei einer Fallzahl von

$$n = 20 \times \frac{v^2}{(1-f)^2}$$ **Gleichung 0.1.**

je Gruppe entdeckt werden. Diese Fallzahlformel wird im Anhang aus der Standardfallzahlplanungsformel für Mittelwertsdifferenzen hergeleitet.

Gleichung 0.1 zeigt dabei, dass die Fallzahl quadratisch mit dem VK steigt. Anschaulich gesprochen bedeutet dieses, dass die Fallzahl viermal so groß sein muss, wenn sich der VK aufgrund von Messungenauigkeiten verdoppelt. Der quadratische Zusammenhang zwischen Fallzahl und VK wird in **Abb.6** für verschiedene Fold Changes dargestellt.

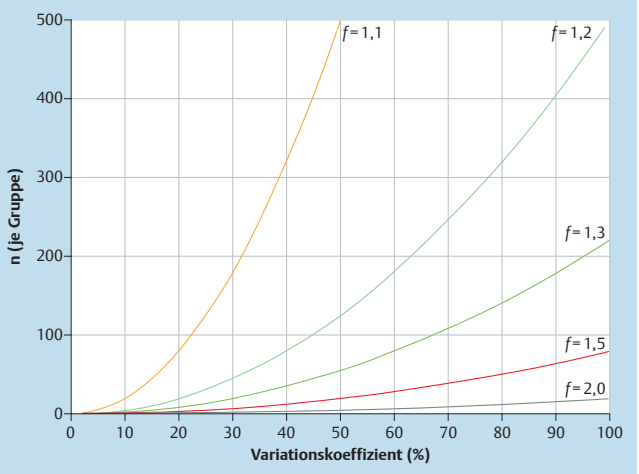

Abb.6 Anzahl der Personen n, die je Gruppe zum Entdecken eines vorgegebenen Fold Changes *f* benötigt werden in Abhängigkeit vom Variationskoeffizienten (VK). Die benötigte Fallzahl steigt quadratisch mit dem VK.

Im Laborbereich lässt sich dieser Zusammenhang bei einer ganzen Reihe von Anwendungen ausnutzen, was im Folgenden anhand zweier Beispiele gezeigt wird.

Abb.7 zeigt für ein Experiment mit Genexpressions-Chips, dass der VK erheblich mit steigender Expressionsstärke fällt. So beträgt der VK bei Expressionen oberhalb der Nachweisgrenze (4,5) etwa 4,5%. Hingegen ist dieser um den Faktor 2 bis 4 geringer für stark exprimierte Transkripte (normalisierte Expressionstärke ab 11). Stehen also zwei Transkripte mit unterschiedlichen Expressionsstärken für eine Validierung zur Auswahl, sollte das Transkript mit dem geringeren VK bevorzugt für die Validierung berücksichtigt werden.

Als zweites Beispiel betrachten wir die Hochleistungsflüssigkeitschromatographie (HPLC), bei der die Präzision von der Zeitkonstante abhängt (**Abb.8**). Liegt die Zeitkonstante bei 0,5

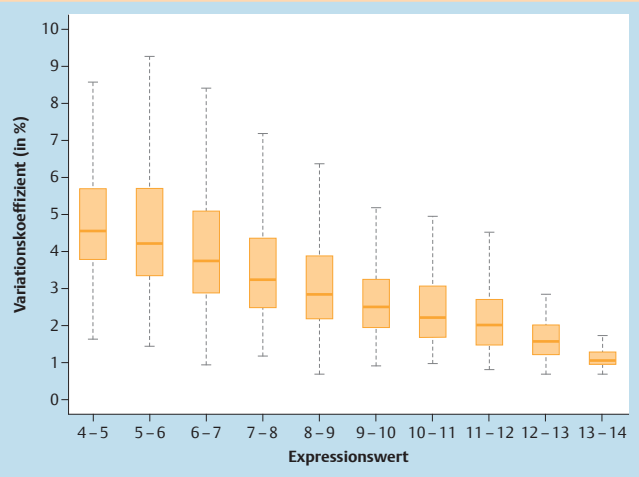

Abb.7 Variationskoeffizienten (VK) gegen Signalintensität von technischen Replikaten einer Genexpressionsstudie unter Verwendung des Affymetrix u133a 2.0 Microarrays. Die normalisierten Expressionsstärken sind in Intervallen von einer Einheit angegeben. Der Variationskoeffizient ist als Boxplot mit Median, Quartilen und kleinstem bzw. größtem Nichtausreißer dargestellt.

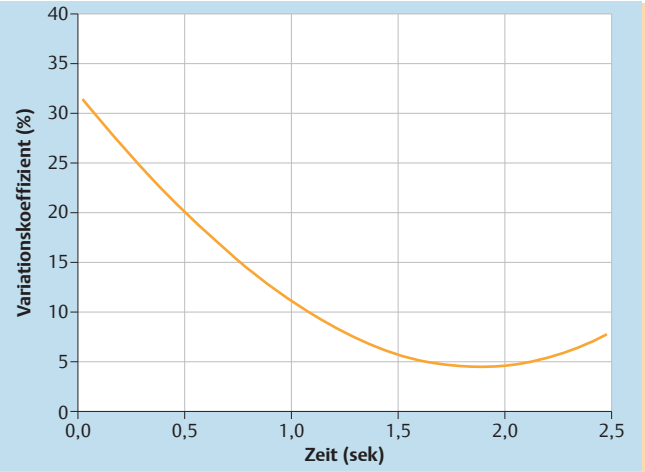

Abb.8 Variationskoeffizienten (VK) gegen Zeitkonstante bei der High Performance Liquid Chromatography (HPLC).

sek, beträgt der VK etwa 20%. Hingegen fällt der VK auf unter 5% bei 2 sek. Der Unterschied in der Variabilität beträgt in diesem Beispiel etwa den Faktor 4, und das bedeutet, dass bei einer Messung zu 0,5 sek etwa 16 mal so viele Personen in die Studie eingeschlossen werden müssten, als wenn bei 2 sek gemessen wird. Natürlich ist hier erforderlich, die Messtechnologie genau zu verstehen. Denn ein Messfenster von 2 sek bei der HPLC kann auch dazu führen, dass andere Analyten im Zeitfenster vermessen werden und dadurch den Messwert des Zielanalyten verfälschen. Bei sehr schwachen Signalen ist eine lange Messzeit auch mit mehr Rauschen belegt als eine kurze Messzeit. Das Verhältnis aus Signal zu Rauschen ist also gerade bei schwächeren Signalen sehr stark von der Messzeit abhängig.

Diskussion

Schon 1995 wurde in einer JAMA-Publikation gezeigt, dass die Forschung zu diagnostischen Methoden zwingend erforderlich ist [60]. In ihrer Übersichtsarbeit haben die Autoren 112 Arbeiten zu diagnostischen Tests betrachtet, die zwischen 1978 und 1993 in vier wichtigen medizinischen Zeitschriften veröffentlicht wurden. Insgesamt wiesen 80% der Publikationen methodische Fehler auf, die zur einer relevanten Verzerrung der Ergebnisse führten [61].

Doch auch vor kurzem fielen die Schlussfolgerungen von Fontela et al. [25] vergleichbar ernüchternd aus. In ihrer 2009 veröffentlichten Studie zur Qualität diagnostischer Studien für Tuberkulose, HIV und Malaria, in denen ein kommerzielles Test-Kit verwendet wurde, wurden 90 Arbeiten identifiziert, die die gewählten Einschlusskriterien erfüllten. Doch keine einzige dieser 90 Publikationen war frei von Mängeln. So wurde z.B. nur in 10% der Arbeiten der Referenzstandard adäquat beschrieben, und nur 16% der Studien hatten eine verblindete Nachbeobachtung.

Darüber hinaus kamen Fontela et al. [25], genau wie viele andere zur Schlussfolgerung, dass die Qualität des Berichtens diagnostischer Studien niedrig war [26, 59, 62]. Dabei lassen sich die Mängel in der Berichtsqualität sehr einfach durch die vollständige Umsetzung der STARD Leitlinie beim Schreiben der Veröffentlichung vermeiden [24], auch wenn dieser Prozess anscheinend nicht schnell voranschreitet [63, 64].

Die Berücksichtigung der fundamentalen Prinzipien diagnostischer Studien in Planungs-, Durchführungs- sowie Auswertungsphase mit anschließender Publikation entsprechend üblicher Leitlinien wie STARD zeigt die Perspektive für eine deutliche Verbesserung der gegenwärtigen Situation auf.

Konsequenz für Klinik und Praxis

- Diagnosestudien sind immer dann kritisch zu werten, wenn mindestens einer der folgenden Punkte nicht erfüllt ist:
  - Gibt es einen unabhängigen, verblindeten Vergleich mit einem Referenzstandard?
  - Schloss die Studie ein genügend breites Spektrum von Patienten ein, an welchen der Test in der klinischen Praxis angewendet wird?
  - Beeinflussten die Ergebnisse des zu evaluierenden Tests die Entscheidung, ob der Standardtest angewendet wurde?
  - Wurde der Test unabhängig von der vorliegenden Studie in einer zweiten Studiengruppe untersucht?
- Bei diagnostischen Biomarkern ist zusätzlich wichtig, dass diese valide und wiederholbar gemessen werden können.
- Damit Biomarkerstudien eine hohe Qualität erreichen, ist die detaillierte Planung der Studie unter Beteiligung von Experten verschiedener Fachdisziplinen erforderlich.

Anhang

Ausgangspunkt ist folgende Standardformel für die Fallzahl, die je Gruppe benötigt wird, wenn zwei gleich große Gruppen miteinander verglichen werden:

$$n = 2 \times (z_{1-\alpha/2} + z_{1-\beta})^2 \times \frac{\sigma^2}{(\mu_2 - \mu_1)^2}$$ **Gleichung 0.2.**

Dabei ist α das Signifikanzniveau, üblicherweise 0,05, 1–β die statistische Macht der Studie, üblicherweise 80%, 90% oder 95%. μ_1 und μ_2 sind die durchschnittlichen Biomarkerwerte der bei-

den Gruppen, und σ ist die Streuung der Biomarkermessung einer einzelnen Person. Wird der Einfachheit halber 1 – β = 0,9 gewählt, ergibt sich unter Verwendung von Werten der Normalverteilung $(z_{1-\alpha/2} + z_{1-\beta})^2 = 10{,}5074 \approx 10$. Entsprechend lässt sich **Gleichung 0.2** vereinfacht auch näherungsweise schreiben als

$$n = 20 \times \frac{\sigma^2}{(\mu_2 - \mu_1)^2} \qquad \textbf{Gleichung 0.3.}$$

Dieses ist die Fallzahl, die pro Gruppe benötigt wird, um bei einem Signifikanzniveau von 5% mit etwa 90%iger Sicherheit einen Unterschied in den durchschnittlichen Biomarkerwerten zwischen den beiden Gruppen entdecken zu können.

Im Bereich der Biomarker wird häufig die Darstellung im Fold Change $f = \mu_2/\mu_1$ und dem Variationskoeffizienten (VK) $v = \sigma/\mu$ gegenüber der Differenz der Mittelwerte $\mu_2 - \mu_1$ und der Standardabweichung σ bevorzugt. Unter Verwendung dieser Parameter lässt sich **Gleichung 0.3** auch schreiben als

$$n = 20 \times \frac{v^2}{(1-f)^2} \qquad \textbf{Gleichung 0.4.}$$

Danksagung: Diese Arbeit ist aus Vorträgen entstanden, die die Autoren bei der Tagung „Functional Genomics and Proteomics – Applications, Molecular Diagnostics & Next Generation Sequencing“ Anfang Februar 2012 in Frankfurt gehalten haben. Die Autoren danken dem Projektträger im Deutschen Zentrum für Luft- und Raumfahrt e.V., Gesundheitsforschung, für die Einladung zur entsprechenden Sitzung mit dem Arbeitstitel „Validierung in diagnostischen Studien“.

Die Arbeit von AZ zu Biomarkern wird finanziell unterstützt durch das Bundeministerium für Bildung und Forschung (01GI0916, 0315536F, 01ER0805, 01KU0908A, 01KU0908B, Exzellenzcluster Inflammation at Interfaces) und die Europäische Union (HEALTH-2011–278913).

Die Autoren danken Dr. Stavros Kromidas für die freundliche Erlaubnis, das Beispiel zur Hochleistungsflüssigkeitschromatographie (HPLC) verwenden zu dürfen.

Autorenerklärung: AZ ist Wissenschaftlicher Berater bei der Protagen AG, Dortmund, war bis zum 14.01.2011 Wissenschaftlicher Berater bei IntegraGen SA, Evry, Frankreich, und hat einen Kooperationsvertrag mit Affymetrix Inc., Santa Clara, USA. PSK ist wissenschaftlicher Leiter der Protagen AG, Dortmund.

Englische Version dieses Beitrages: DOI 10.1055/s-0033-1343172

Abstract

Biomarkers are of increasing importance for personalized medicine in many areas of application, such as diagnosis, prognosis, or the selection of targeted therapies. In many molecular biomarker studies, intensity values are obtained from large scale -omics experiments. These intensity values, such as protein concentrations, are often compared between at least two groups of subjects to determine the diagnostic ability of the molecular biomarker. Various prospective or retrospective study designs are available for molecular biomarker studies, and the biomarker used may be univariate or may even consist in a multimarker rule. In this work, several challenges are discussed for the planning and conduct of biomarker studies. The phases of diagnostic biomarker studies are closely related to levels of evidence in diagnosis, and they are therefore discussed upfront. Different study designs for molecular biomarker studies are discussed, and they primarily differ in the way subjects are selected. Using two systematic reviews from the literature, common sources of bias of molecular diagnostic studies are illustrated. The extreme selection of patients and controls and verification bias are specifically discussed. The pre-analytical and technical variability of biomarker measurements is usually expressed in terms of the coefficient of variation, and is of great importance for subsequent validation studies for molecular biomarkers. It is finally shown that the required sample size for biomarker validation quadratically increases with the coefficient of variation, and the effect is illustrated using real data from different laboratory technologies.

Literatur

1 *Ziegler A, Koch A, Krockenberger K et al.* Personalized medicine using DNA biomarkers: a review. Hum Genet 2012; 131: 1627–1638
2 *Gallo V, Egger M, McCormack V et al.* STrengthening the Reporting of OBservational studies in Epidemiology – Molecular Epidemiology (STROBE-ME): an extension of the STROBE Statement. PLoS Med 2011; 8: e1001117
3 Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2011; 69: 89–95
4 *Kroll W.* Biomarkers – predictors, surrogate parameters – a concept definition. In: *Schmitz G, Endres S, Götte D* eds, Biomarker. Stuttgart, Schattauer 2008; 1–14
5 *Haddow JE, Palomaki GE.* ACCE: A model process for evaluating data on emerging genetic tests. In: *Khoury M, Little J, Burke W* eds, Human Genome Epidemiology: A Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease. Oxford, Oxford University Press 2003; 217–233
6 DIN Deutsches Institut für Normung e.V.. DIN ISO 5725-1:1997-11 Genauigkeit (Richtigkeit und Präzision) von Messverfahren und Messergebnissen – Teil 1: Allgemeine Grundlagen und Begriffe. In: DIN Deutsches Institut für Normung e.V., ed, DIN-Taschenbuch 355: Statistik – Genauigkeit von Messungen – Ringversuche. Berlin, Beuth 2004; 1–44
7 *Evans JP, Skrzynia C, Burke W.* The complexities of predictive genetic testing. BMJ 2001; 322: 1052–1056
8 *Jensen K, Abel U.* Methodik diagnostischer Validierungsstudien. Fehler in der Studienplanung und Auswertung. Med Klin (München) 1999; 94: 522–529
9 *Pepe MS.* The statistical evaluation of medical tests for classification and prediction. New York, Oxford University Press 2003
10 *Zhou X-H, Obuchowski NA, McClish DK.* Statistical methods in diagnostic medicine. New York, John Wiley & Sons 2001
11 *Pepe MS, Etzioni R, Feng Z et al.* Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001; 93: 1054–1061
12 *Abel U, Jensen K.* Klinische Studien außerhalb des Arzneimittelgesetzes: Diagnosestudien. Bundesgesundheitsbl 2009; 52: 425–432
13 *Buyse M, Michiels S, Sargent DJ et al.* Integrating biomarkers in clinical trials. Expert Rev Mol Diagn 2011; 11: 171–182
14 *Buyse M, Sargent DJ, Grothey A et al.* Biomarkers and surrogate end points – the challenge of statistical validation. Nat Rev Clin Oncol 2010; 7: 309–317
15 *Sargent DJ, Conley BA, Allegra C et al.* Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 2005; 23: 2020–2027
16 *Mandrekar SJ, Grothey A, Goetz MP et al.* Clinical trial designs for prospective validation of biomarkers. Am J Pharmacogenomics 2005; 5: 317–325
17 *Mandrekar SJ, Sargent DJ.* Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol 2009; 27: 4027–4034
18 *Mandrekar SJ, Sargent DJ.* Clinical trial designs for predictive biomarker validation: one size does not fit all. J Biopharm Stat 2009; 19: 530–542
19 *Mandrekar SJ, Sargent DJ.* Predictive biomarker validation in practice: lessons from real trials. Clin Trials 2010; 7: 567–573

20 *Schäfer H*. Anforderungen an einen patientenorientierte klinisch-therapeutische Forschung. Dtsch Med Wochenschr 1997; 122: 1531–1536
21 *Weinstein S, Obuchowski NA, Lieber ML*. Clinical evaluation of diagnostic tests. AJR Am J Roentgenol 2005; 184: 14–19
22 *Obuchowski NA*. How many observers are needed in clinical studies of medical imaging? AJR Am J Roentgenol 2004; 182: 867–869
23 *Egerer K, Feist E, Burmester GR*. The serological diagnosis of rheumatoid arthritis: antibodies to citrullinated antigens. Dtsch Arztebl Int 2009; 106: 159–163
24 *Ziegler A, König IR*. Leitlinien für Forschungsberichte: Deutschsprachige Übersetzungen von CONSORT 2010, PRISMA und STARD. Dtsch Med Wochenschr 2011; 136: 357–358
25 *Fontela PS, Pant Pai N, Schiller I et al.* Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. PLoS ONE 2009; 4: e7753
26 *Lijmer JG, Mol BW, Heisterkamp S et al.* Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282: 1061–1066
27 *Rutjes AW, Reitsma JB, Di Nisio M et al*. Evidence of bias and variation in diagnostic accuracy studies. Can Med Assoc J 2006; 174: 469–476
28 *Sica GT*. Bias in research studies. Radiology 2006; 238: 780–789
29 Centre for Review and Dissemination. Systematic reviews: CRD's guidance for undertaking reviews in health care. 2009; www.york.ac.uk/inst/crd/SysRev/!SSL!/WebHelp/TITLEPAGE.htm Zugriff: 26.04.2013.
30 *Blackmore CC*. The challenge of clinical radiology research. AJR Am J Roentgenol 2001; 176: 327–331
31 *Brealey S, Scally AJ*. Bias in plain film reading performance studies. Br J Radiol 2001; 74: 307–316
32 *Whiting P, Rutjes AW, Reitsma JB et al.* Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140: 189–202
33 *Guyatt G, Rennie D, Meade MO, Cook DJ eds*. Users' Guide to the Medical Literature: A Manual for Evidence-Based Clinical Practice. 2.: ed. Minion, McGraw-Hill 2008
34 *Thomson DM, Krupey J, Freedman SO et al.* The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. Proc Natl Acad Sci U S A 1969; 64: 161–167
35 *Zielinski C*. Aussagekraft des carcinoembryonalen Antigens. Dtsch Med Wochenschr 1995; 120: 893
36 *Lachs MS, Nachamkin I, Edelstein PH et al*. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 1992; 117: 135–140
37 *Cicero S, Rembouskos G, Vandecruys H et al.* Likelihood ratio for trisomy 21 in fetuses with absent nasal bone at the 11-14-week scan. Ultrasound Obstet Gynecol 2004; 23: 218–223
38 *Knottnerus JA, Muris JW*. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003; 56: 1118–1128
39 PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). The PIOPED Investigators. JAMA 1990; 263: 2753–2759
40 *Punglia RS, D'Amico AV, Catalona WJ et al.* Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. N Engl J Med 2003; 349: 335–342
41 *de Groot JA, Bossuyt PM, Reitsma JB et al.* Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ 2011; 343: d4770
42 *Martus P, Schueler S, Dewey M*. Fractional flow reserve estimation by coronary computed tomography angiography. J Am Coll Cardiol 2012; 59: 1410–1411 author reply 1411
43 *Hanrahan CF, Westreich D, Van Rie A*. Verification bias in a diagnostic accuracy study of symptom screening for tuberculosis in HIV-infected pregnant women. Clin Infect Dis 2012; 54: 1377–1378 author reply 1378-1379
44 *Richardson ML, Petscavage JM*. Verification bias: an under-recognized source of error in assessing the efficacy of MRI of the meniscii. Acad Radiol 2011; 18: 1376–1381
45 *Begg CB, Greenes RA*. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39: 207–215
46 *Diamond GA*. "Work-up bias". J Clin Epidemiol 1993; 46: 207–209
47 *Rutjes AW, Reitsma JB, Coomarasamy A et al.* Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007; 11: ix–51
48 *Dehdashti H, Dehdashtian M, Rahim F et al.* Sonographic measurement of abdominal esophageal length as a diagnostic tool in gastroesophageal reflux disease in infants. Saudi J Gastroenterol 2011; 17: 53–57
49 *Sarkhy AA*. Methodological issues in diagnostic studies. Saudi J Gastroenterol 2011; 17: 161–162
50 *Harvey RF*. Indices of thyroid function in thyrotoxicosis. Lancet 1971; 2: 230–233
51 *Andersen B*. Methodological errors in medical research. An incomplete catalogue. Oxford, Blackwell 1990
52 *Begg CB*. Biases in the assessment of diagnostic tests. Stat Med 1987; 6: 411–423
53 *Simel DL, Feussner JR, DeLong ER et al.* Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making 1987; 7: 107–114
54 *Ronco G, Montanari G, Aimone V et al.* Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. Cytopathology 1996; 7: 151–158
55 *Ramos JM, Robledano C, Masia M et al.* Contribution of Interferon gamma release assays testing to the diagnosis of latent tuberculosis infection in HIV-infected patients: A comparison of QuantiFERON-TB gold in tube, T-SPOT.TB and tuberculin skin test. BMC Infect Dis 2012; 12: 169
56 *Begg CB, Greenes RA, Iglewicz B*. The influence of uninterpretability on the assessment of diagnostic tests. J Chronic Dis 1986; 39: 575–584
57 *Philbrick JT, Horwitz RI, Feinstein AR et al.* The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. JAMA 1982; 248: 2467–2470
58 *Detrano R, Gianrossi R, Mulvihill D et al.* Exercise-induced ST segment depression in the diagnosis of multivessel coronary disease: a meta analysis. J Am Coll Cardiol 1989; 14: 1501–1508
59 *Bossuyt PM, Reitsma JB, Bruns DE et al.* The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138: W1–12
60 *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
61 *Sardanelli F, Di Leo G*. Biostatistics for Radiologists. Mailand, Springer 2009
62 *Rutjes AW, Reitsma JB, Di Nisio M et al.* Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006; 174: 469–476
63 *Wilczynski NL*. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication – before-and-after study. Radiology 2008; 248: 817–823
64 *Hollingworth W, Jarvik JG*. Technology assessment in radiology: putting the evidence in evidence-based radiology. Radiology 2007; 244: 31–38
65 Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses. http://www.g-ba.de/downloads/62-492-654/VerfO_2012-10-18.pdf. Fassung vom: 18.12.2008. BAnz. Nr. 84a (Beilage) vom 10.06.2009. Letzte Änderung: 18.10.2012. BAnz AT 05.12.2012 B3. In Kraft getreten am 06.12.2012. Letzter Zugriff 26.04.2013.