Valide Evidenz aus klinischen Studien
Diagnostikstudien

Dr. Dirk Hasenclever

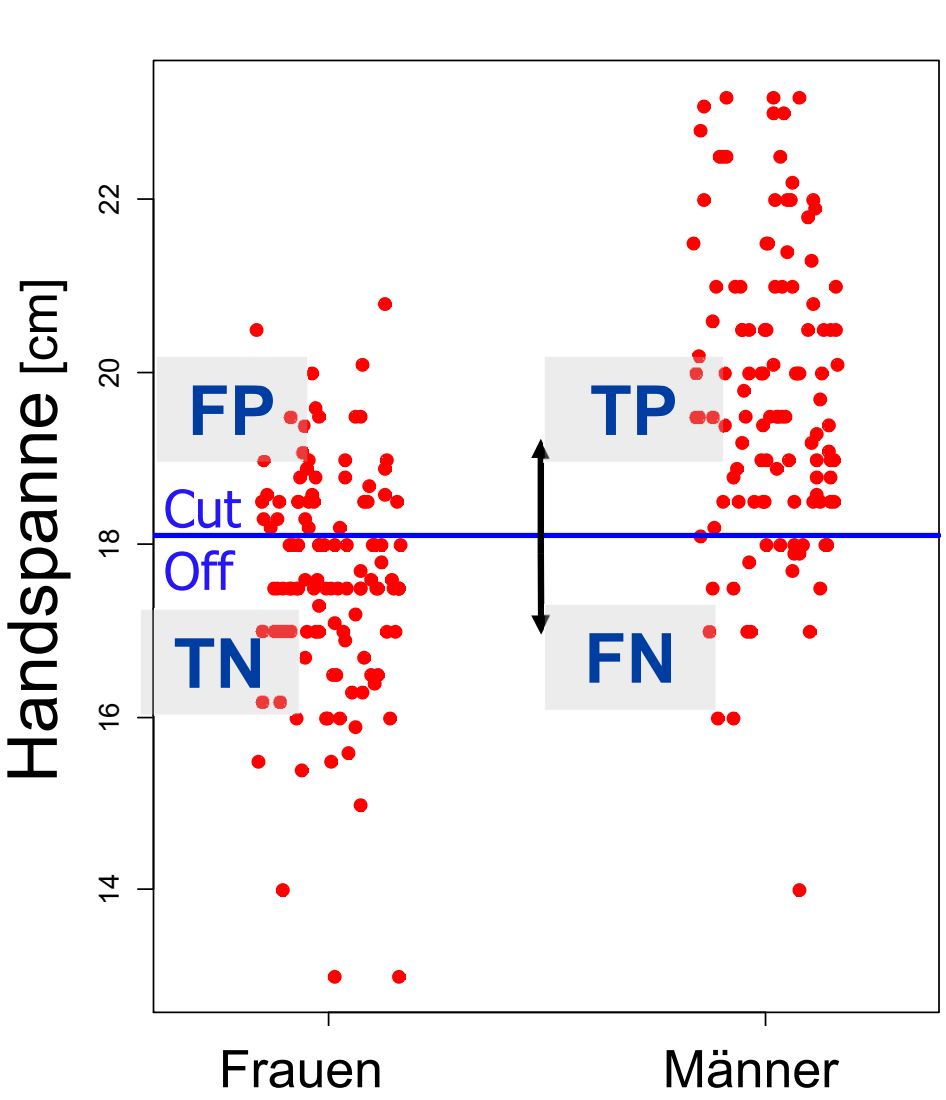
Institut für Medizinische Informatik, Statistik & Epidemiologie (IMISE),
& Zentrum für klinische Studien Leipzig, Universität Leipzig
dirk.hasenclever@imise.uni-leipzig.de

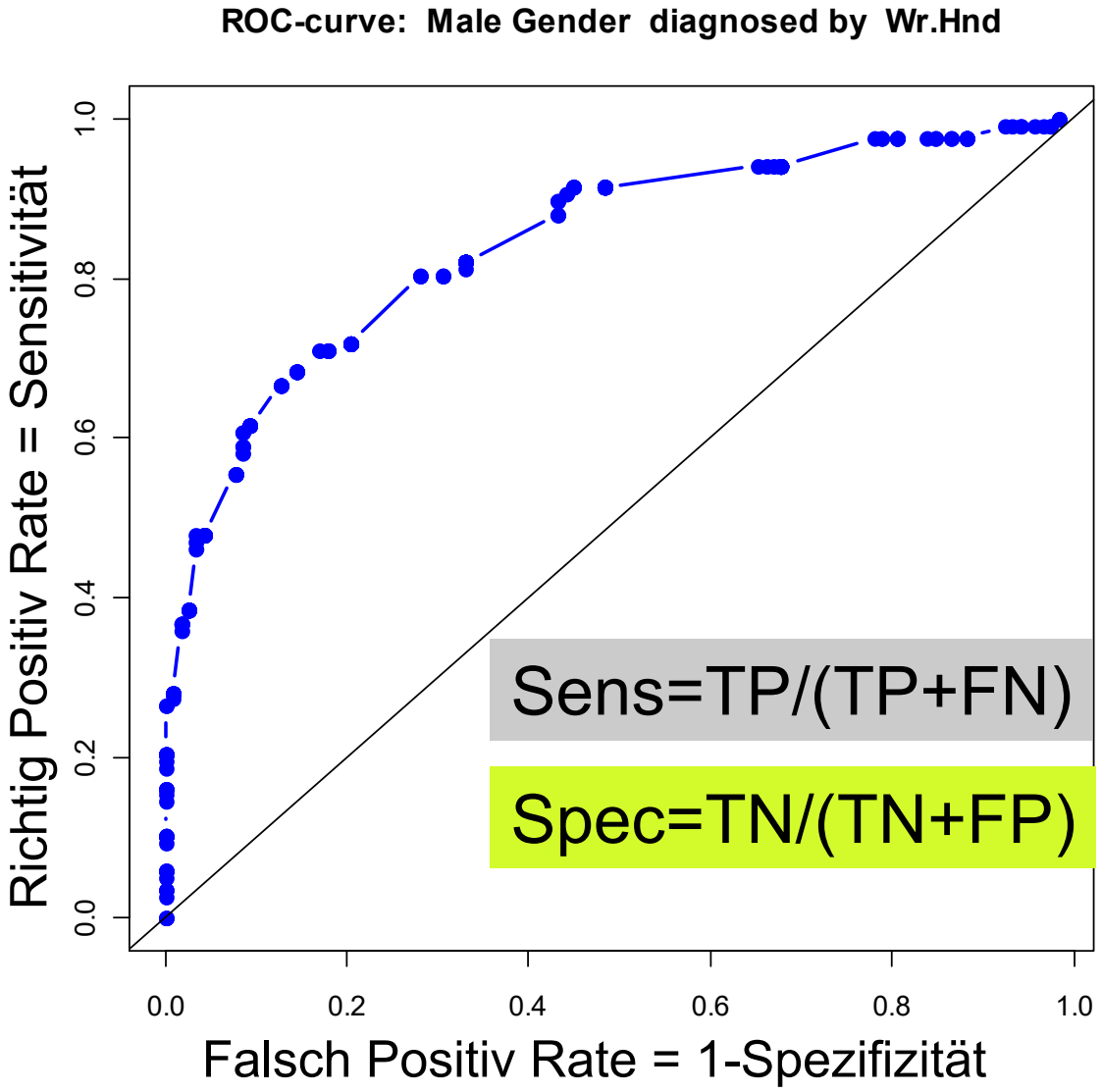
Phasen diagnostischer Studien

- **Phase I**: Unterscheiden sich Testresultate in Kranken und Gesunden?
- **Phase II**: Haben Patienten mit positivem Testresultat eine höherer Wahrscheinlichkeit, die Erkrankung zu haben?
- **Phase III**: Unterscheidet der Test Patienten mit und ohne die Zielerkrankung auch unter Patienten mit klinischem Verdacht auf die Erkrankung?
- **Phase IV**: Haben Patienten vom Einsatz des Tests einen klinischen Vorteil?

Sackett-2002-539

Phase 1 -> Phase II
ROC-Kurven – Wahl eines Schwellwerts
Geschlechtsdiagnose aufgrund der Handspanne





Kontinuierliche Testgröße

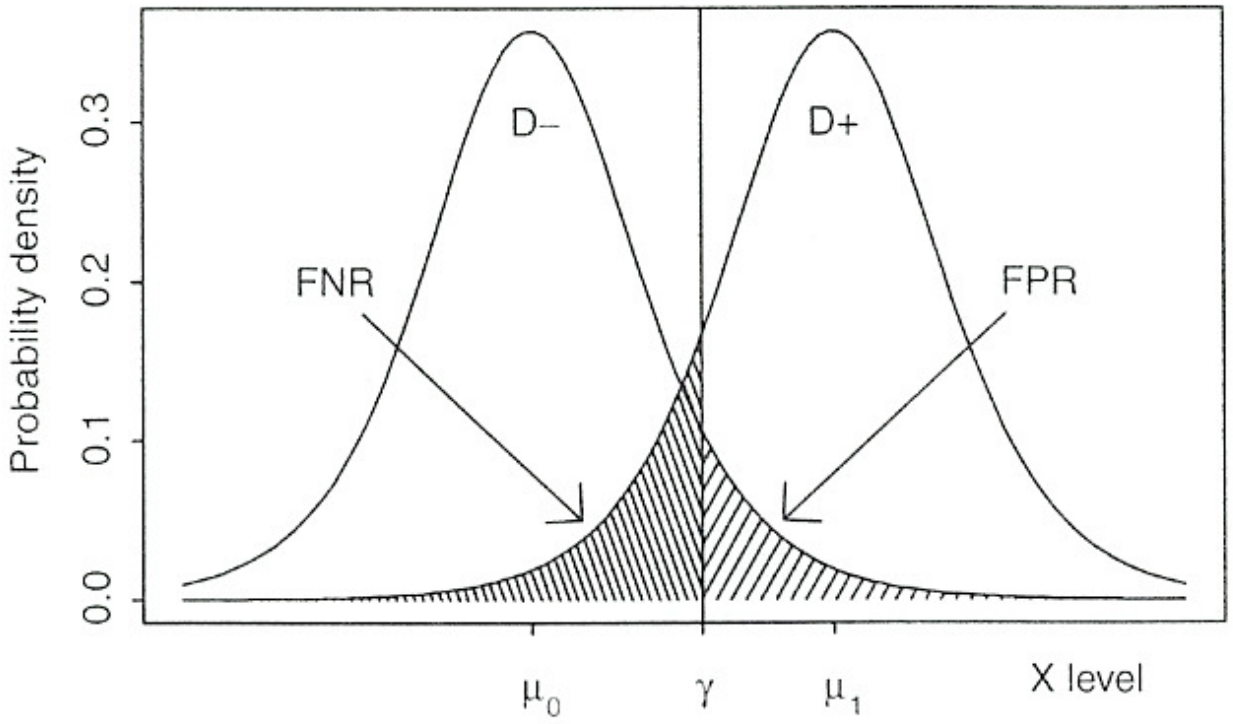
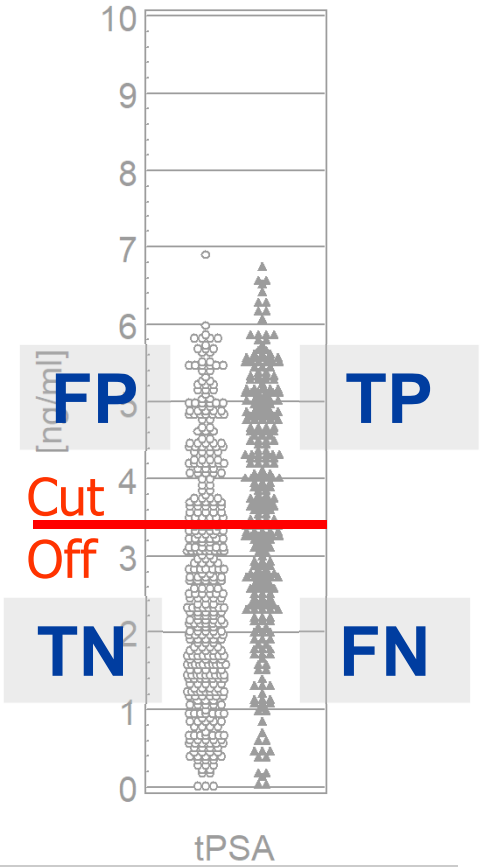


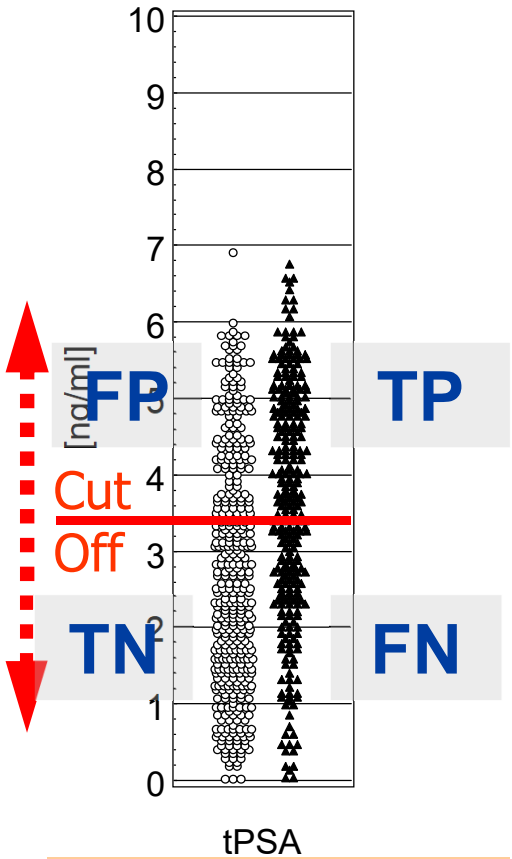
Figure 2 Hypothetical distributions of diagnostic test results X for diseased and nondiseased subjects. The vertical line at $X = \gamma$ indicates the decision limit for a positive test. The shaded area to the right of γ is the FPR; the shaded area to the left of γ is the FNR.

Shapiro-1999-113

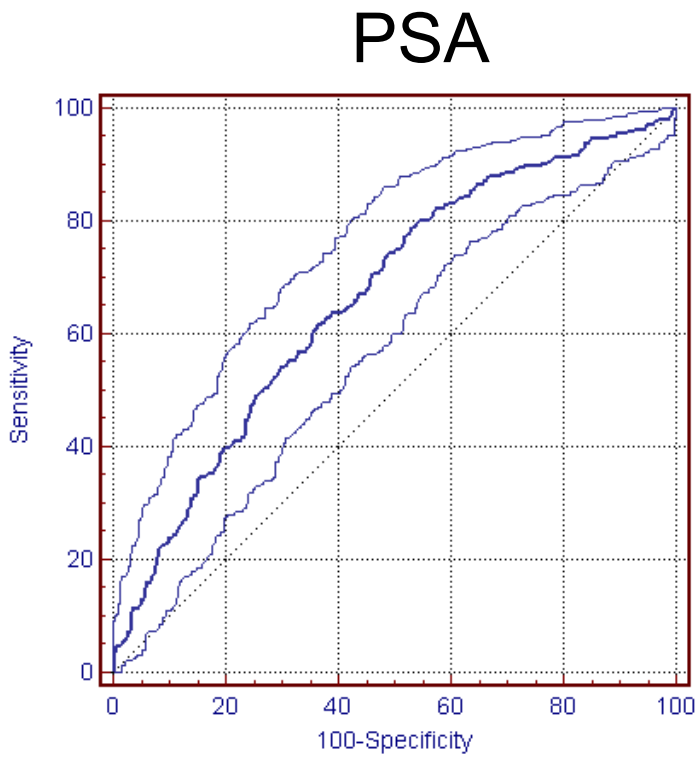
ROC-Kurve: Variation des Cut-Off



Sens=tp/(tp+fn)



Sens=f(Spec)



Spec=tn/(tn+fp)

ROC-Kurve

Diagnostische ROC Kurve

Serum creatinine in relation to renal artery stenosis

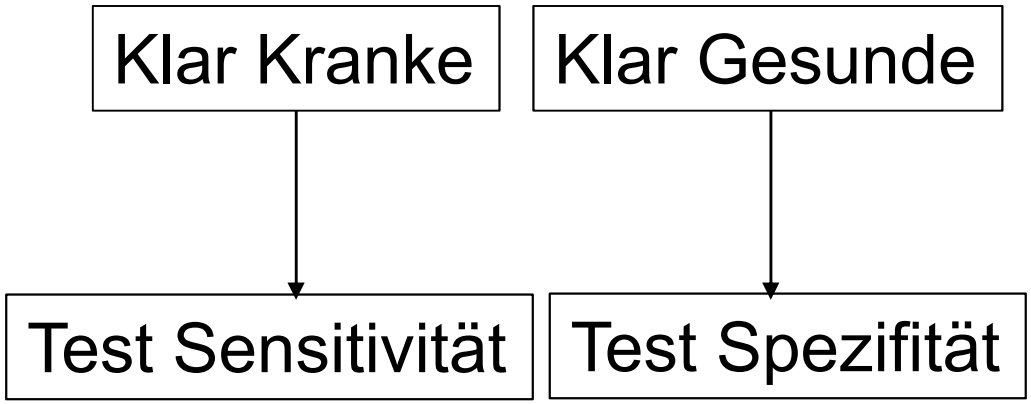
THE EVIDENCE BASE OF CLINICAL DIAGNOSIS

Table 7.7 Probability of test results and diagnostic information of dichotomised serum creatinine concentration values for nine possible cut-offs between high and low values.

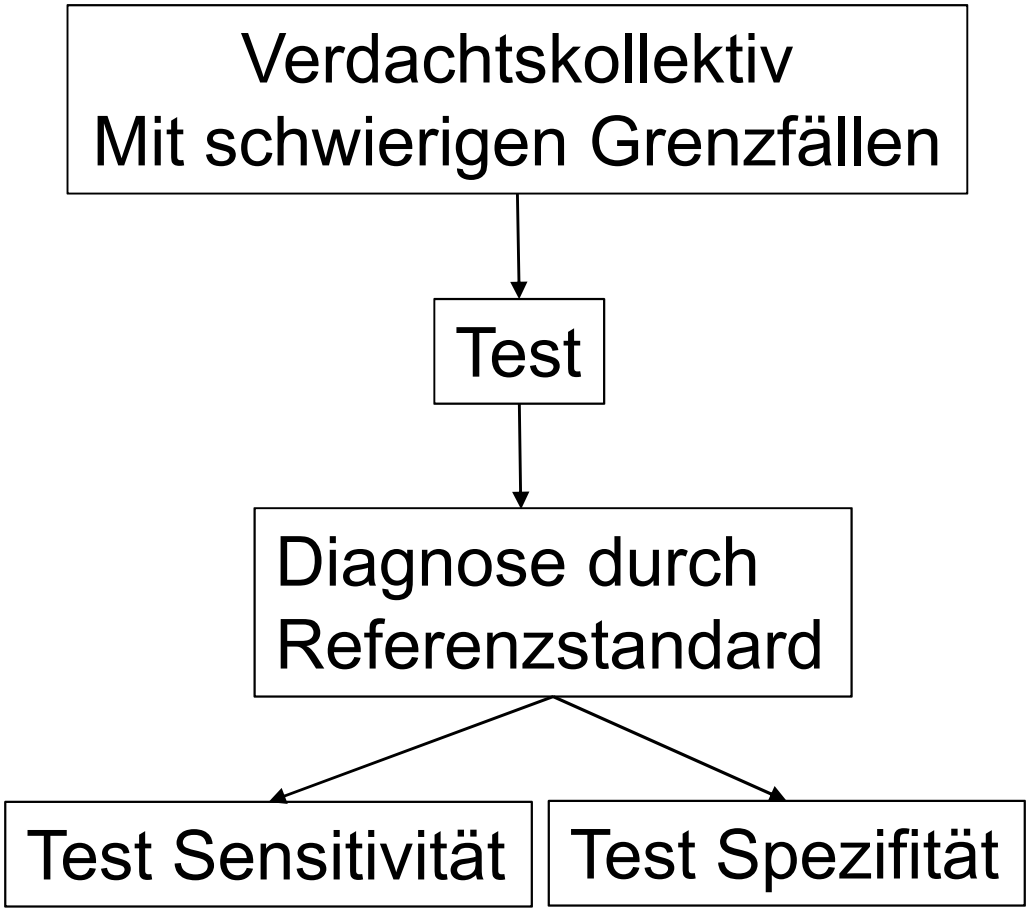
Serum creatinine (Micromol)	Stenosis Se	No stenosis 1 − Sp	All	LR+	LR−
>60	99 (99%)	318 (94%)	417 (95%)	1.05	0.18
>70	95 (95%)	282 (84%)	377 (81%)	1.14	0.31
>80	82 (82%)	215 (64%)	297 (68%)	1.29	0.50
>90	70 (70%)	144 (43%)	114 (49%)	1.64	0.52
>100	53 (53%)	73 (22%)	126 (29%)	2.44	0.60
>110	38 (38%)	32 (9%)	70 (16%)	4.00	0.69
>120	31 (31%)	22 (7%)	53 (12%)	4.77	0.74
>130	22 (22%)	13 (4%)	35 (8%)	5.64	0.81
>150	11 (11%)	5 (1%)	16 (4%)	7.33	0.90
Total	100 (100%)	337 (100%)	437 (100%)		

Aus Knotternerus BMJ 2002

Phase II versus Phase III
Zwei Populationen klarer Fälle oder eine Verdachtspopulation?



Zwei Populationen Studien sind **oft stark überoptimistisch!** (Spektrum/Case-Mix Bias)



Phasen diagnostischer Studien
Eingeteilt nach Fragestellung!

- **Phase I**: Unterscheiden sich Testresultate in Kranken und Gesunden?
- **Phase II**: Haben Patienten mit positivem Testresultat eine höherer Wahrscheinlichkeit, die Erkrankung zu haben?
- **Phase III**: Unterscheidet der Test Patienten mit und ohne die Zielerkrankung auch unter Patienten mit klinischem Verdacht auf die Erkrankung?
- **Phase IV**: Haben Patienten vom Einsatz des Tests einen klinischen Vorteil?

Sackett-2002-539

Bsp.: Phase I

Table 1 Answering a phase I question: do patients with left ventricular dysfunction have higher concentrations of B-type natriuretic peptide (BNP) precursor than normal individuals?

	Patients known to have disorder	Normal controls
Median (range) concentration of BNP precursor (pg/ml)	493.5 (248.9-909.0)	129.4 (53.6-159.7)

Sackett-2002-539

Bsp.: Phase II

Table 2 Answering a phase II question: are patients with higher concentrations of B-type natriuretic peptide (BNP) more likely to have left ventricular dysfunction than patients with lower concentrations?

Wahl eines Cutpoints

	Patients known to have target disorder	Normal controls
High BNP concentration	39	2
Normal BNP concentration	1	25

Test characteristics (95% CI):
- Sensitivity=98% (87% to 100%)
- Specificity=92% (77% to 98%)
- Positive predictive value=95% (84% to 99%)
- Negative predictive value=96% (81% to 100%)
- Likelihood ratio for an abnormal test result=13 (3.5 to 50.0)
- Likelihood ratio for a normal test result=0.03 (0.0003 to 0.19)

Sackett-2002-539

Bsp.: Phase III

Table 3 Answering a phase III question: among patients in whom it is clinically sensible to suspect left ventricular dysfunction (LVD), does the concentration of B-type natriuretic peptide (BNP) distinguish patients with and without left ventricular dysfunction?

	Patients with LVD on echocardiography	Patients with normal results on echocardiography
Concentration of BNP:		
High (>17.9 pg/ml)	35	57
Normal (<18 pg/ml)	5	29
Prevalence (pretest probability) of LVD	40/126=32%	

Test characteristics (95% CI):
- Sensitivity=88% (74% to 94%)
- Specificity=34% (25% to 44%)
- Positive predictive value=38% (29% to 48%)
- Negative predictive value=85% (70% to 94%)
- Likelihood ratio for an abnormal test result=1.3 (1.1 to 1.6)
- Likelihood ratio for a normal test result=0.4 (0.2 to 0.9)

Sackett-2002-539

Weiter Verzerrungsquellen (Überblick)

- **Verifiktionbias**
  - (Referenzstandard nicht in allen Fällen angewandt)
- **Nicht Interpretierbare Testergebnisse**
  - werden gerne einfach nicht erwähnt
- **Referenzstandard zeigt Beurteiler-Diskrepanzen**
- **Sensitivität und Spezifizität nicht konstant** bei sehr unterschiedlichen Prävalenzen
  - Selektive Anreicherung von Fällen reichert auch potentiell falsch positive an.

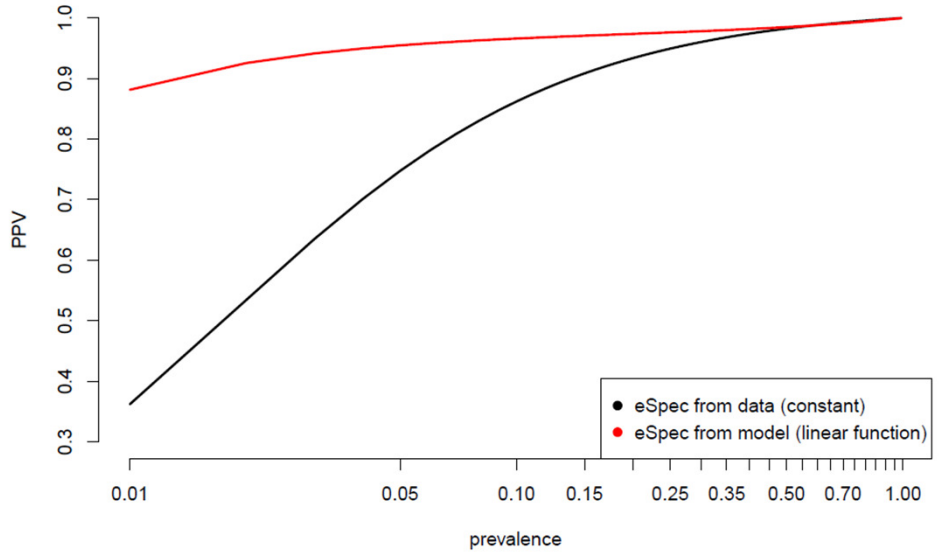


Figure 2 Estimates for positive predictive value (PPV) as a function of prevalence. The black curve is found by using standard methods and the red one when taking into account selection as proposed here.

Verifikations (Work-up)-Bias

- Goldstandard wird (da invasiv) hauptsächlich an Patienten mit positivem Testergebnis erhoben. (partielle Verifikation)

- Oder Surogat Gold-Standard in Fällen mit negativem Testergebnis (z.B. Follow up) (differentielle Verifikation)

Beispiel Verifikationsbias

Table II. Exercise radionuclide ventriculography for coronary artery disease (from Diamond *et al.*[31])

Source population

		Test result: Positive	Test result: Negative	
Disease	present	(302)*	(179)	→ Sensitivity = (63 %)
	absent	(80)	(372)	→ Specificity = (82 %)
		382	551	
		↓	↓	
		0·31	0·14	sampling fractions

E.g. 372=52 / 0.14

Verified sample

Disease	present	94	25	→ Sensitivity = 79 %
	absent	25	52	→ Specificity = 67 %

* The numbers in parentheses are model-based estimates (see text)

Assuming non-informed sampling within T+ and T-

Begg-1987-411

Bias durch uninterpretierbare Tests

- RelevanterAnteil uninterpretierbarer Tests
  - Bsp: Ultraschall Abdomen und Gasansammlung
  - Schilddrüsenknoten Feinnadelbiopsie

- Oft stillschweigend aus Bericht entfernt.

- Wichtig für klinischen Einsatz (zu erwartende Information) und Kosten

- Einfluß auf SENS/SPEC

Comparison of Diagnostic Cytologic Categories in Seven Series of biopsies of thyroid nodules

Series (Reference)	Country	Year	Total Cases	Cytologic Diagnosis			
				Benign	Malignant	Suspicious	Nondiagnostic
			n	← *n*(%) →			
Gardiner et al. (19)	Canada	1986	1465*	1085 (74)	16 (1)	146 (10)	218 (15)
Hawkins et al. (24)	Spain	1987	1399	1253 (90)	56 (4)	68 (5)	22 (2)
Khafagi et al. (30)	Australia	1988	618	404 (65)	29 (5)	68 (11)	117 (19)
Hall et al. (46)	United States	1989	795	509 (64)	81 (10)	75 (9)	130 (16)
Altavilla et al. (35)	Italy	1990	2433*	1890 (78)	32 (1)	119 (5)	392 (16)
Caplan et al. (40)	United States	1991	502	268 (53)	24 (5)	116 (23)	94 (19)
Gharib and Goellner†	United States	1991	10 971*	7071 (64)	416 (4)	1192 (11)	2292 (21)
Total			18 183	12 480 (69)	654 (3.5)	1784 (10)	3265 (17)

* Number of aspirates; multifocal aspirations on the same patient were counted as separate aspirations.
† Unpublished data.

Gharib, H. et. al. Ann Intern Med 1993;118:282-289

Annals of Internal Medicine

Yield, False-Negative and False-Positive Rates, Sensitivity, and Specificity of Thyroid Fine-Needle Aspiration Biopsy

Series	Cases			False-negative Rate	False-positive Rate	Sensitivity	Specificity
	Total	Operated	Malignant				
	n		*n*(%)	←	%	→	
Gardiner et al. (19)	1465†	207	46 (22)	11.5	0	65	91
Hawkins et al. (24)	1399	415	73 (18)	2.4	4.6	86	95
Khafagi et al. (30)	618	258	44 (17)	4.1	7.7	87	72
Hall et al. (46)	795	72	37 (51)	1.3	3.0	84	90
Altavilla et al. (35)	2433†	257	49 (19)	6.0	0	71	100
Caplan et al. (40)	502	185	64 (35)	9.3	4.0	91	99
Gharib and Goellner‡	10 971†	1750	682 (39)	2.0	0.7	98	99
Total	18 183	3144	995 (32)	5.2	2.9	83	92

* For definitions of false-negative rate, false-positive rate, sensitivity, and specificity, see text.
† Number of aspirates; multifocal aspirations on the same patient were counted as separate aspirations.
‡ Unpublished data.

Gharib, H. et. al. Ann Intern Med 1993;118:282-289

Annals of Internal Medicine

Inter-Observer Variation

- **Inter-Observer Variation:**
  - Oft klinische Urteile erforderlich
    - (Bildgebende Verfahren, Histologien etc.)
  - Unterschiedliche Sensitivität
  - Unterschiedliche Kompetenz

Bsp: Inter-Observer Variation

Table IV. Endoscopic diagnosis of oesophageal varices by two endoscopists*

Results	Definitive radiologic diagnosis	
	Positive	Negative
Both negative	3	12
1st negative/2nd positive	2	3
1st positive/2nd negative	1	7
Both positive	8	3
Inadequate examination	0	1
	Sensitivity	Specificity
1st endoscopist	64% (9/14)	60% (15/25)
2nd endoscopist	71% (10/14)	76% (19/25)

* Data from Conn *et al.*[41]

Begg-1987-411

Inter-Reader Reliability of Early FDG-PET/CT Response Assessment Using the Deauville Scale after 2 Cycles of Intensive Chemotherapy (OEPA) in Hodgkin's Lymphoma

PLoS One. 2016 Mar 10;11(3):e0149072.

Regine Kluge[1]*, Lidia Chavdarova[2], Martha Hoffmann[3], Carsten Kobe[4], Bogdan Malkowski[5], Françoise Montravers[6], Lars Kurch[1], Thomas Georgi[1], Markus Dietlein[4], W. Hamish Wallace[7], Jonas Karlen[8], Ana Fernández-Teijeiro[9], Michaela Cepelova[10], Lorrain Wilson[11], Eva Bergstraesser[12], Osama Sabri[1], Christine Mauz-Körholz[13], Dieter Körholz[13], Dirk Hasenclever[14]

Purpose

The five point Deauville (D) scale is widely used to assess interim PET metabolic response to chemotherapy in Hodgkin lymphoma (HL) patients. An International Validation Study reported good concordance among reviewers in ABVD treated advanced stage HL patients for the binary discrimination between score D1,2,3 and score D4,5. Inter-reader reliability of the whole scale is not well characterised.

Methods

Five international expert readers scored 100 interim PET/CT scans from paediatric HL patients. Scans were acquired in 51 European hospitals after two courses of OEPA chemo-therapy (according to the EuroNet-PHL-C1 study). Images were interpreted in direct com-parison with staging PET/CTs.

Is it lymphoma residual?

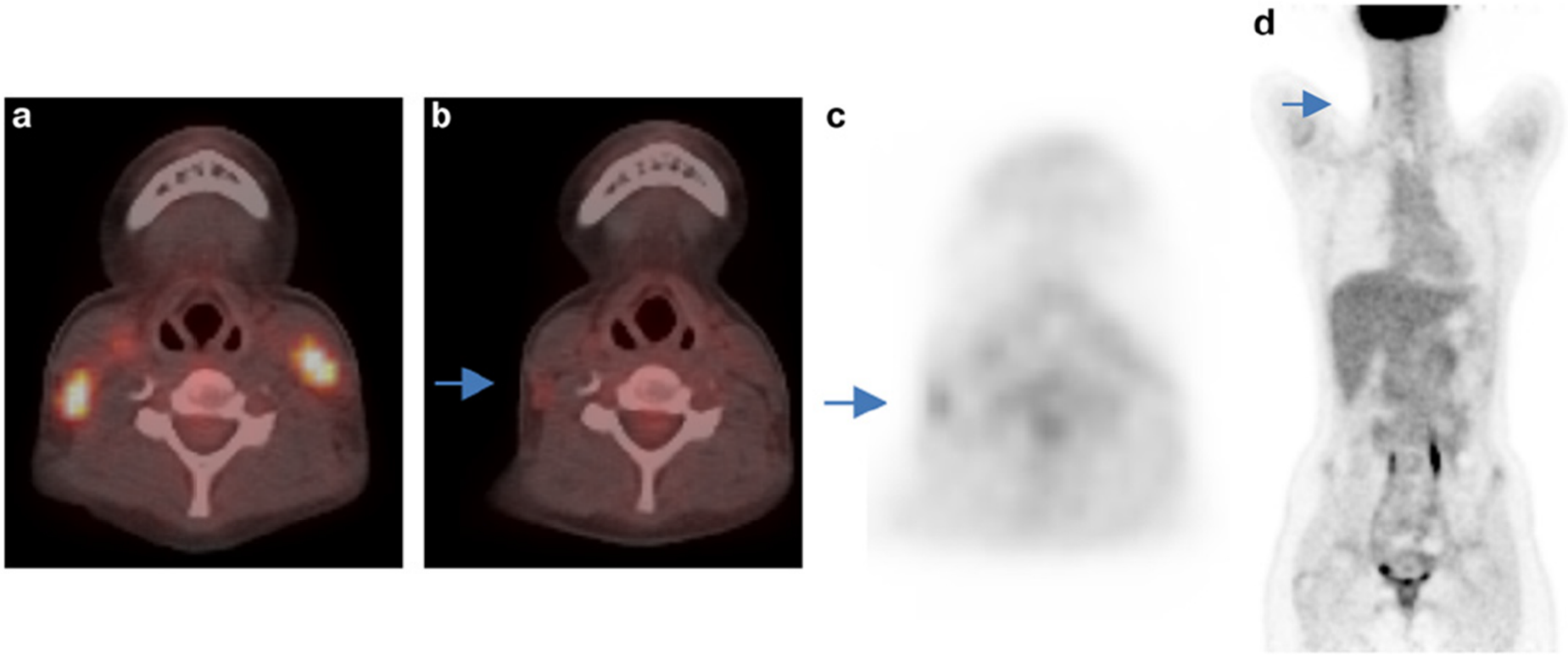


Fig 3. Discrepant D scoring of the neck region. Case 44: a) Lymphoma involvement at both sides of the neck before treatment. b) and c) residual lymph node on the right neck with slightly enhanced FDG uptake after to cycles ChT. d) The residual uptake is less than in the liver, but it is difficult to decide if it is above or below the MBP uptake. Deauville score of the readers R1-5, respectively: 2-3-3-1-3. Courtesy Dept. Clinical Physiology, Sahlgrenska University Hospital, Gothenburg Sweden.

Conclusion

Inter-reader reliability of the five point D-scale is poor in this interobserver analysis of paediatric patients who underwent OEPA. Inter-reader variability is maximal in cases assigned to D2 or D3. The binary distinction D1,2,3 versus D4,5 is the most reliable criterion for clinical decision making.

Table 1. Frequency of Deauville (D) scores by reader R1 –R5.

D score	R1	R2	R3	R4	R5
1	49	52	10	49	14
2	12	10	13	18	49
3	19	24	52	16	18
4	15	13	21	14	14
5	5	1	4	3	5

Table 2. Agreement statistics for Deauville (D) scoring among five readers.

	Overall probability of concordance	Specific probability of concordance for categories					Global kappa	Range of pairwise kappas
Scale		D1	D2	D3	D4	D5		
Five categories: D1, D2, D3, D4, D5	0.422	0.520	0.255	0.360	0.494	0.556	0.244	0.144–0.387
Three categories: D1,2, D3, D4,5	0.604	0.705		0.360	0.642		0.344	0.222–0.485
Two categories: D1,2, D3,4,5	0.674	0.705			0.636		0.358	0.189–0.514
Two categories: D1,2,3, D4,5	0.864	0.916			0.642		0.559	0.428–0.710

Biasquellen bei Diagnostikstudien V

- **Zeittrends**:
  - Ständige technologische Verbesserungen machen Diagnostik-Studien mit Bildgebung schnell obsolete

Biasquellen bei Diagnostikstudien VI

- **Einfluß klinischer Faktoren auf Testinterpretation (Clinical Review Bias)**
  - Unterscheide:
    - Beurteilung in Kenntnis / Unkenntnis der Klinik
    - Absolute vs. inkrementelle Information des Tests
  - Verblindung
- Extremfall (**Review Bias**):
  - Testinterpretation in Kenntnis des Goldstandards

Klinischer Review-Bias

Table VI. Influence of clinical history (a) Chest radiographs* (b) ECG stress test†

(a)	Suggestive history	Non-suggestive history
Sensitivity	72% (23/32)	16% (5/32)

(b)	Blinded study†		Unblinded study§	
History	Sensitivity	Specificity	Sensitivity	Specificity
Typical angina	71%	81%	74%	57%
Atypical angina	75%	80%	60%	94%

* from Doubilet and Herman[44]
† from Weintraub *et al.*[45]
§ from Melin *et al.*[46]

Same radiologist looking twice at radiographs with differing accompanying suggestive clinical history.

Begg-1987-411

Empirische Evidence für Relevanz dieser Biasquellen

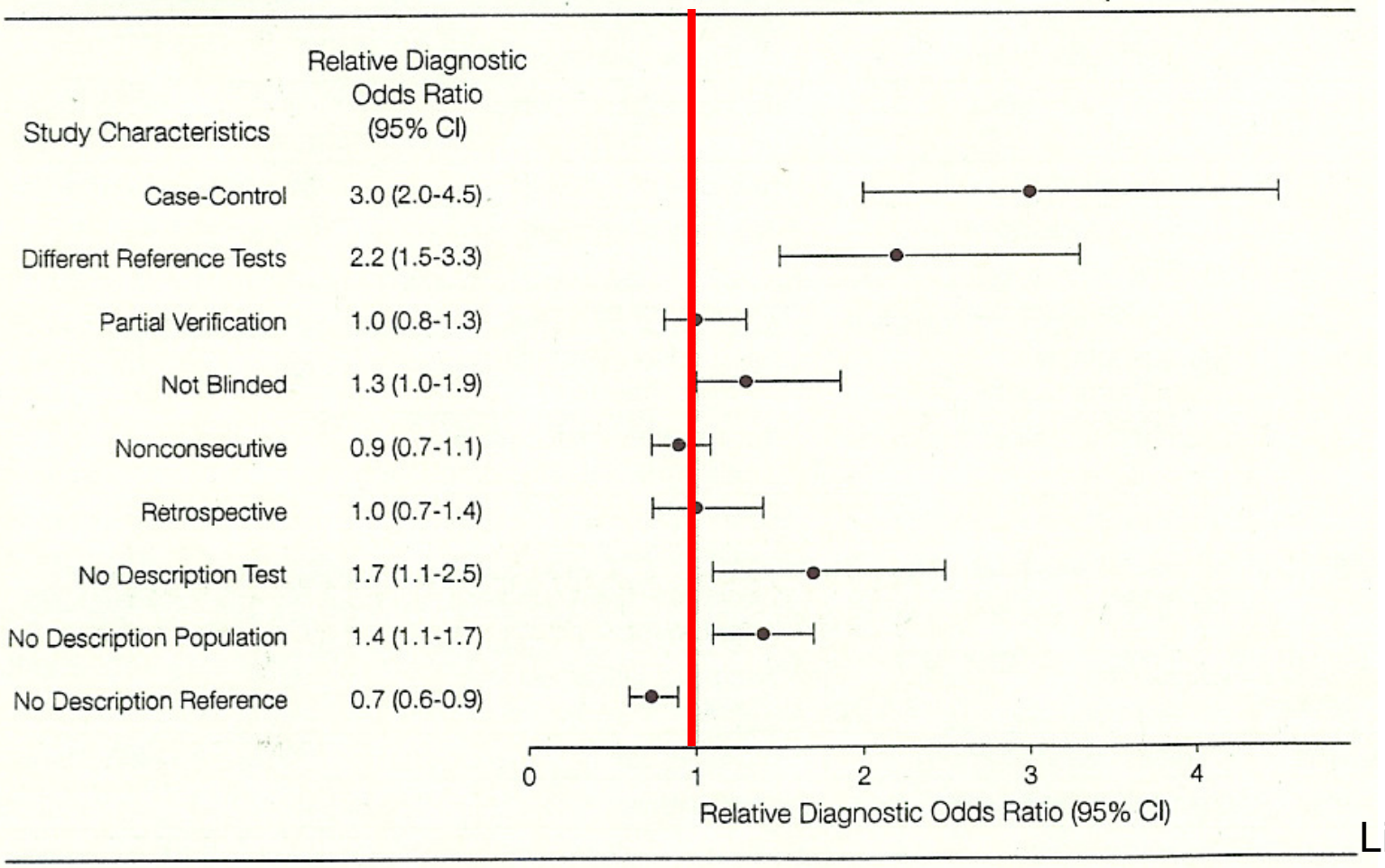
Table 2. Results of the Scoring of Study Quality (N = 218)

Study Characteristic	Score	No. of Subjects
Spectrum	Clinical population	213
	Case-control	5
Verification	Complete	106
	Different reference tests	48
	Partial	54
Interpretation of test results	Blinded	70
	Not blinded	148
Patient selection	Consecutive	97
	Nonconsecutive	121
Data collection	Prospective	98
	Unknown	92
	Retrospective	28
Details test	Sufficient	195
	Insufficient	23
Details reference test	Sufficient	138
	Insufficient	80
Details population	Sufficient	132
	Insufficient	86

Lijmer-1999-1061

Empirische Evidence für Relevanz dieser Biasquellen II

Figure. Relative Diagnostic Odds Ratios and 95% Confidence Intervals (CIs) of the 9 Study Characteristics Examined With a Multivariate Regression Analysis



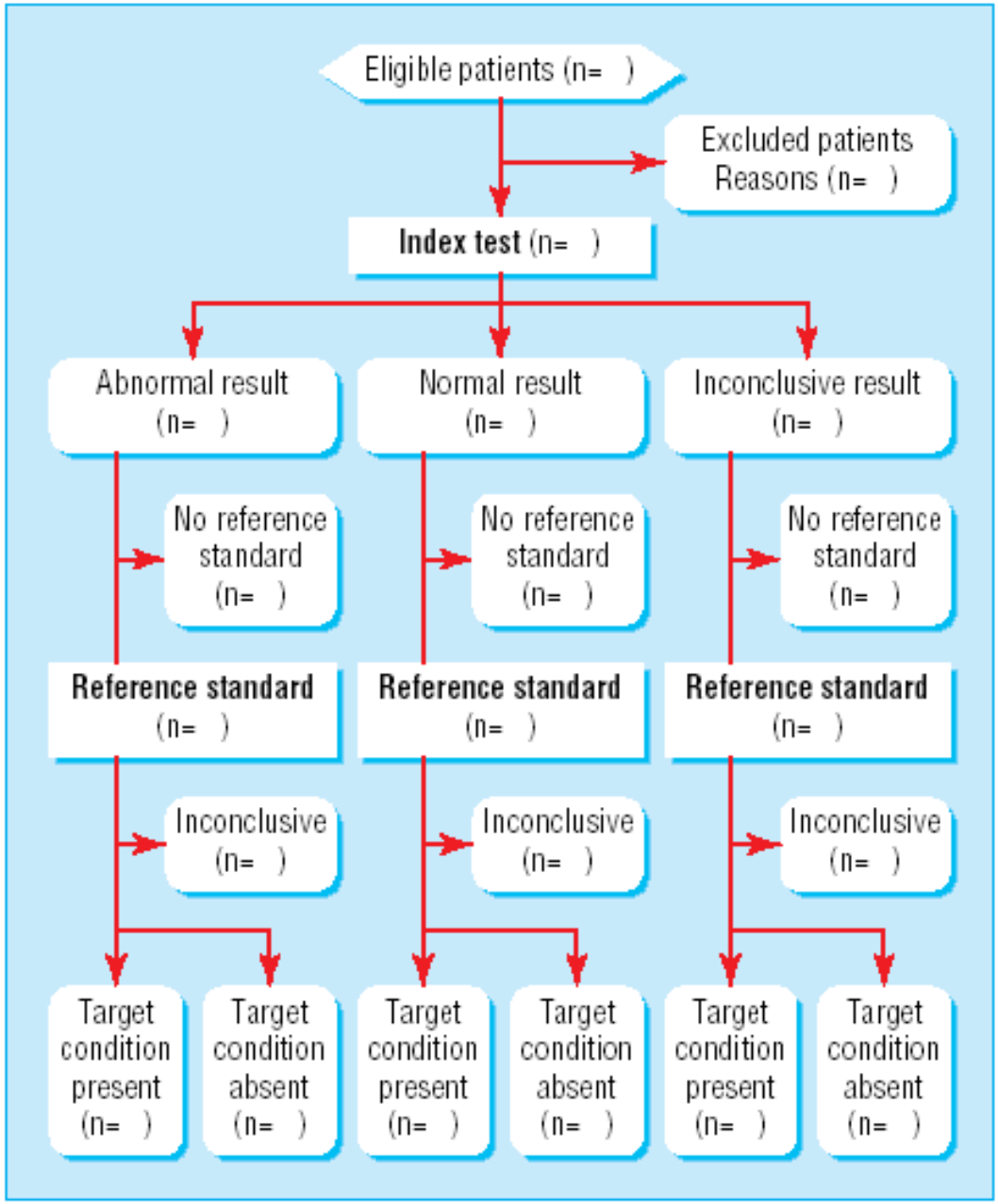
Lijmer-1999-1061

Qualitätsbewertung für Diagnostikstudien

Table 2: The QUADAS tool

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

Whiting 2003



Prototype of a flow diagram for a study on diagnostic accuracy

STARD Standard für Publikation von Diagnostikstudien

STARD checklist for reporting diagnostic accuracy studies

Section and topic	Item	Description
Title, abstract, and keywords	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity")
Introduction	2	State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups
Methods:		
Participants	3	Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected
	4	Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Describe participant sampling: was this a consecutive series of participants defined by selection criteria in items 3 and 4? If not, specify how participants were further selected
	6	Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?
Test methods	7	Describe the reference standard and its rationale
	8	Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for index tests or reference standard, or both
	9	Describe definition of and rationale for the units, cut-off points, or categories of the results of the index tests and the reference standard
	10	Describe the number, training, and expertise of the persons executing and reading the index tests and the reference standard
	11	Were the readers of the index tests and the reference standard blind (masked) to the results of the other test? Describe any other clinical information available to the readers.
Statistical methods	12	Describe methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty (eg 95% confidence intervals)
	13	Describe methods for calculating test reproducibility, if done
Results:		

Participants	14	Report when study was done, including beginning and ending dates of recruitment
	15	Report clinical and demographic characteristics (eg age, sex, spectrum of presenting symptoms, comorbidity, current treatments, and recruitment centre)
	16	Report how many participants satisfying the criteria for inclusion did or did not undergo the index tests or the reference standard, or both; describe why participants failed to receive either test (a flow diagram is strongly recommended)
Test results	17	Report time interval from index tests to reference standard, and any treatment administered between
	18	Report distribution of severity of disease (define criteria) in those with the target condition and other diagnoses in participants without the target condition
	19	Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, report the distribution of the test results by the results of the reference standard
	20	Report any adverse events from performing the index test or the reference standard
Estimates	21	Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)
	22	Report how indeterminate results, missing responses, and outliers of index tests were handled
	23	Report estimates of variability of diagnostic accuracy between readers, centres, or subgroups of participants, if done
	24	Report estimates of test reproducibility, if done
Discussion	25	Discuss the clinical applicability of the study findings

Weitere Diagnostikstudien-Typen

- Vergleich von Tests / Markern
- Kombination von Tests
- Klinischer Einsatz von Tests
- Vergleich von Diagnosestrategien
- ...

Lernziele

- Grundbegriffe diagnostischer Tests:
  - Sensitivität, Spezifität , NPW, PPW, diagnostische LR, ROC-Kurven
- Bayes-Formel
  - Abhängigkeit der Testaussage von der Prävalenz
- Biasquellen in Screeningstudien
- Phasen-Konzept für diagnostische Studien
- Verzerrungsquellen bei diagnostischen Studien
  - Case-mix, Verifikation, Interobserver-Variabilität, Review-Bias…

Literatur

- Sackett-D-2002-539-a: The architecture of diagnostic research (Bitte Lesen!)
- Lijmer-J-1999-1061-a: Empirical Evidence of Design-Related Bias in studies of Diagnostic Tests
- Begg-C-1987-411-a: Biases in the assessment of diagnostic tests.