Evidence base of clinical diagnosis

Evaluation of diagnostic procedures

J André Knottnerus, Chris van Weel, Jean W M Muris

Development and introduction of new diagnostic techniques have greatly accelerated over the past decades. The evaluation of diagnostic techniques, however, is less advanced than that of treatments. Unlike with drugs, there are generally no formal requirements for adoption of diagnostic tests in routine care. In spite of important contributions,[1 2] the methodology of diagnostic research is poorly defined compared with study designs on treatment effectiveness, or on aetiology, so it is not surprising that methodological flaws are common in diagnostic studies.[3–5] Furthermore, research funds rarely cover diagnostic research starting from symptoms or tests.

Since quality of the diagnostic process largely determines quality of care, overcoming deficiencies in standards, methodology, and funding deserves high priority. This article summarises objectives of diagnostic testing and research, methodological challenges, and options for design of studies.

Summary points

Development of diagnostic techniques has greatly accelerated but the methodology of diagnostic research lags far behind that for evaluating treatments

Objectives of diagnostic investigations include detection or exclusion of disease; contributing to management; assessment of prognosis; monitoring clinical course; and measurement of general health or fitness

Methodological challenges include the "gold standard" problem; spectrum and selection biases; "soft" measures (subjective phenomena); observer variability and bias; complex relations; clinical impact; sample size; and rapid progress of knowledge

Objectives of testing

Diagnostic investigations collect information to clarify patients' health status, using personal characteristics, symptoms, signs, history, physical examination, laboratory tests, and additional facilities. Objectives include the following.

- *Increasing certainty of the presence or absence of disease*—This requires sufficient discriminative power. Measures of discrimination are commonly derived from a 2×2 table relating test outcome to a reference standard (figure), thus allowing tests to be compared. Tests for similar purposes may vary in accuracy, invasiveness, and risk, and, for example, history may be no less valuable than laboratory tests (table). To be useful, additional investigations should add relevant information to less invasive and cheaper tests performed earlier.
- *Supporting clinical management*—For example, determining presence, localisation, and shape of arterial lesions is necessary for treatment decisions.
- *Assessing prognosis*—As the starting point for clinical follow up and informing patients.
- *Monitoring clinical course*—When a disease is untreated, or during or after treatment.
- *Measuring fitness*—For example, for sporting activity or for employment.

Tests must be evaluated in accordance with their intended objectives, also taking into consideration possible inconvenience and complications, such as intestinal perforation during endoscopy. Using and not using a test, or using alternative tests, should therefore be compared.

If a test is evaluated before introduction into routine care, using or not using it can still be freely compared to study the effect on prognosis. Early evaluation helps decisions on whether to introduce a test and on planning its postmarketing surveillance.

Methodological challenges

The "gold standard" problem

To evaluate discriminatory power (accuracy), the outcome of a test is compared with an independently established standard diagnosis. "Gold standards" providing full certainty are rare. Even biopsies can fail to do so. Generally the challenge is to find a standard as close as possible to the theoretical gold standard.

Sometimes no suitable reference standard at all is available—in determining the accuracy of liver tests, neither imaging techniques nor biopsies will detect all liver abnormalities. Moreover, invasive procedures cannot easily be made the standard in a study. An independent standard may not even conceptually exist, as for example when evaluating symptoms incorporated in the definition of a disease (as in migraine), or when the symptoms are more important than anatomical status, as with prostatism. In studying the value of physical examination to detect severe disease in non-acute abdominal pain, comprehensive screening, including invasive procedures (if ethically allowable), might yield many irrelevant findings but still fail to exclude relevant pathology. An appropriate clinical follow up—a "delayed type cross sectional study," with a final assessment by independent experts—is then the best approach.[1 9]

New diagnostic tests superior to prevailing reference standards may be developed. If research into accuracy of test procedures were to consist only of comparing tests with standards, possible new standards would be ignored as they are not in agreement with prevailing standards. Up to date pathophysiological expertise is therefore required to be able to change a reference standard.

This is the first of a series of five articles

Netherlands School of Primary Care Research, PO Box 616, 6200 MD Maastricht, Netherlands
J André Knottnerus *professor of general practice*

Department of General Practice, University Medical Centre St Raboud, 6500 HB Nijmegen, Netherlands
Chris van Weel *professor of general practice*

Department of General Practice, Maastricht University, 6200 MD Maastricht, Netherlands
Jean W M Muris *senior lecturer in general practice*

Correspondence to:
J A Knottnerus
andre.knottnerus@hag.unimaas.nl

Series editor: J A Knottnerus

BMJ 2002;324:477–80

	x-Ray result (D)		
Physical examination result (T)	Positive (fracture)	Negative (no fracture)	Total
Positive (fracture)	190	80	270
Negative (no fracture)	10	720	730
Total	200	800	1000

The **SENSITIVITY** of T is the probability of a positive test result in people with D: 190/200 = **0.95**
The **SPECIFICITY** of T is the probability of a negative test result in people without D: 720/800 = **0.90**

The **PREDICTIVE VALUE** of a test result T_x is:
for a **POSITIVE** result, the probability of D in people with a positive test result: 190/270 = **0.70**
for a **NEGATIVE** result, the probability of absence of D in people with a negative test result: 720/730 = **0.99**
[for good discrimination, the difference between the predictive value (posterior or post-test probability of disease) and the prior or pre-test probability of disease should be large. (The prior probability of disease is the prevalence of D in the population to be tested: 200/800 = 0.20)]

The **LIKELIHOOD RATIO (LR)** of a test result T_x is the probability of the test result T_x in people with D, divided by the probability of T_x in people without D.
For a positive result, LR+ is: (190/200)/(80/800) = 9.5; for a negative result,LR- is: (10/200)/(720/800) = **0.06**
[a test is useless if LR = 1. The test is better the more LR differs from 1, that is, greater than 1 for LR+ and lower than 1 for LR-. For tests with multiple outcome categories, LR_x can be calculated for every separate category x]

The **ODDS RATIO (OR)** summarises the overall discrimination of a dichotomous test T: (190/10)/(80/720) = **171**
[a test is useless if OR = 1. T is better the more OR differs from 1]

The **RECEIVER OPERATING CHARACTERISTIC (ROC)** curve represents the relation between sensitivity and specificity for tests with a variable cut-off point, on an ordinal scale (eg, in case of 5 categories of suspicion of ankle fracture) or interval scale (eg, if degree of suspicion of ankle fracture is expressed as a percentage).
[a test is useless if the area under the curve (AUC) = 0.5. For a perfect test the AUC = 1.0]

Derivation of measures of discrimination

Spectrum and selection bias

Spectrum bias may occur when the study population has a different clinical spectrum (more advanced cases, for instance) than the population in whom the test is to be applied.[1 10 11] If sensitivity is determined in seriously diseased subjects and specificity in clearly healthy subjects, both will be grossly overestimated relative to practical situations where diseased and healthy subjects cannot be clinically distinguished in advance.

Selection bias is likely if inclusion in a study is related to test results. As subjects with abnormal exercise electrocardiograms are more often referred for coronary angiography, calibration of this investigation among preselected subjects will show higher sensitivity and lower specificity than if there had been no preselection.

Spectrum and selection bias often occur together—for example, when tests calibrated in hospital are introduced in primary care; all measures of accuracy may then be affected.[12]

Discrimination of some diagnostic tests (estimates based on on several sources)

Test	Sensitivity (%)	Specificity (%)	Likelihood ratio		Odds ratio
			Positive result	Negative result	
Coronary stenosis[6]					
Exercise electrocardiography*	65	89	5.9	0.39	15.0
Stress thallium scintigraphy	85	85	5.7	0.18	32.1
Pancreatic cancer[6]					
Ultrasonography	70	85	4.7	0.35	13.2
Computed tomography	85	90	8.5	0.17	51.0
Angiography	75	80	3.8	0.31	12.0
Peripheral arterial occlusive disease[7]					
Intermittent claudication	31	93	4.4	0.74	5.6
Posterior tibial or dorsalis pedis artery pulse	73	92	9.1	0.29	30.4
Colorectal cancer[8]					
Change in bowel habit	88	72	3.1	0.17	18.4
Weight loss	44	85	2.9	0.66	4.6
Erythrocyte sedimentation rate ≥30 mm in first hour	40	96	10.0	0.42	14.0
White blood cell count $<10^8/mm^3$	75	90	7.5	0.28	26.3
Occult blood test ≥1 positive out of 3	50	82	2.7	0.16	4.6

*Cut-off point: ST depression ≥1 mm.

"Soft" measures

Subjective phenomena such as pain and feeling unwell often evoke diagnostic and therapeutic actions and thus may themselves be "tests." Also, they are indispensable for assessment of clinical outcome.[13] Evaluation studies should measure these factors as reproducibly as possible, recognising that interindividual and intraindividual differences always have a role.

Observer variability and observer bias

Interobserver and intraobserver variability in reading and interpreting diagnostic data not only influence "soft" diagnostic aspects, but also results of "harder" investigations like *x* rays and biopsies. Even without human interpretation, interinstrument and intrainstrument variations occur. Variability should be limited in order to assure utility of information.

Prior knowledge may evoke observer bias. If doctors' accuracy in diagnosing ankle fractures on the basis of physical examination is being evaluated, theyshould not know the *x* ray results; pathologists establishing an independent diagnosis must not know the clinical conclusion already.[14] Bias can also occur if, in comparing two techniques, observers are prejudiced and perform one more carefully than the other. And since, for a fair assessment, diagnostic skills should be at a similar level for each technique, new tests can be at a disadvantage shortly after being introduced.

Complex relations

Ideally evidence reflects the clinical context,[15] where tests are often not applied in isolation but in combinations, as, for instance, in the context of protocols. Moreover, tests can be used to differentiate between a number of diseases, rather than just checking for one. Multivariate statistical techniques then help to evaluate the (added) value of diagnostic items separately and in combination. While analysis of data to determine aetiology generally addresses the overall impact of factors adjusted for covariables, analysis of diagnostic data focuses on the best individual prediction. Accordingly, diagnostic data analysis needs specific methodological development.[16–18]

Sample size

Whether sample size is adequate to provide the desired information with sufficient precision is often ignored in diagnostic studies. Progress in diagnostic performance consists of a series of small steps that gradually increase certainty rather than by one big breakthrough. Evaluating small steps, however, requires large study populations.

Clinical impact

More accurate tests do not necessarily improve management. They may add little to what is known already, or to the results of earlier, perhaps less invasive or cheaper, investigations. Also, clinicians may not make full use of information from results. In a classic

study of the value of upper gastrointestinal endoscopy, management changed in 23% of cases without a change in diagnosis, while in 30% of those with changes in diagnosis management was not altered.[19] Also, tests may have no practical benefit; brain scans showing details of untreatable brain conditions would be an example. Therefore, diagnostic research should consider not only the accuracy of diagnostic tests but also their practical clinical value.

If the probability of disease is extremely low or high, the outcome of subsequent investigations rarely influences management and false positive or false negative results, respectively, are common.[2] Generally, investigations are indicated when the probability of disease is somewhere between the two extremes. Evaluation studies must take place in populations with prior probabilities for which the test is particularly suitable. For example, tests with moderate specificity are inappropriate for population screening (with low probability of disease) because of the high risk of false positive results.

Changes over time and the mosaic of evidence

Thorough evaluation may take longer than developing better techniques. The position of computer assisted tomography was not yet defined when magnetic resonance imaging and positron emission tomography appeared; evaluation studies can thus be outdated before they are completed. Progress is especially rapid where information technology and molecular genetics are important. Therefore, we need comprehensive scenarios with relatively stable overall frameworks into which new data are inserted like pieces of a puzzle. For example, evaluation of the impact of new imaging techniques on the effectiveness of breast screening can be based on data on the accuracy of the techniques being compared if other "mosaic" pieces are already available and unchanged. Since accuracy can often be assessed cross sectionally, lengthy new prospective studies may then be avoided.

Options in diagnostic research

Clinical studies

Methodological approaches must be relevant to the type of study objective (box). Diagnostic accuracy—that is, the relation between the test under study and the disorder as expressed in measures of discrimination (see table 1), can be assessed cross sectionally if the results of the test and the reference standard procedure are known for all subjects in the study population. Possible designs are comparing test distributions in samples already known to have the disorder (cases) and known to be free of it (controls); or comparing disease distributions in samples with already known test results; and a survey in an "indicated" population (a target population in which testing would be relevant). Case-control sampling or sampling based on test results is efficient as a phase I study[1] (see also the next article in this series[20]), and it should be considered before any extensive study in a population where neither the distribution of a disease nor the test results are known. If tests already adopted are also applied to all study subjects, the added value of a new test can be directly estimated. Furthermore, and very importantly, the clinical diagnostic contribution of the test being evaluated can also be assessed if tests that have been performed earlier or are less invasive—for example, from history or physical examination—are also included in the design. Invasiveness and possible adverse effects of the approaches being compared can then be measured.

For studying the impact of a test on clinical decision making and prognosis the randomised controlled trial is the standard method. The experimental group undergoes the index test and the control group the usual test or no test. The value of the index test in addition to or as a replacement for the usual procedure, or instead of no test, can be assessed as (possible) gain in correct diagnoses, management, and prognosis. A variant is to apply the index test to all subjects but randomise disclosure of its results to the care givers, if this is ethically permissible. This constitutes an ideal placebo procedure for the patient. Studies on breast cancer screening, with a treatment protocol linked to the screening result, were classic examples of randomised controlled trials of diagnostic methods.[21] If such a trial is not feasible, observational approaches can be considered. The cohort design compares the clinical outcome of previously tested and untested groups, without the diagnostic information being randomised.[22] A point of concern is whether both groups have a similar clinical spectrum at baseline, especially regarding unmeasured factors. The case-control design is efficient if patient outcome among indicated subjects is already known: were fewer cases than controls tested? Examples are studies on the relation between breast cancer mortality and previous mammographic screening.[23] Comparability of tested and not tested subjects at baseline is, again, important.

Options in diagnostic research in relation to study objectives

Clinical studies
Objective—Diagnostic accuracy
Options: Cross sectional studies
Case-control sampling
Sampling based on test results
Surveys in indicated population
Objective—Impact of (additional or replacing) diagnostic testing on prognosis or management
Options: Randomised controlled trial
Cohort study
Case-control study
Before and after study

Synthesising findings and expertise
Objective—Synthesising results of multiple studies
Options: Systematic review
Meta-analysis
Objective—Determining the most (cost) effective diagnostic strategy
Options: Clinical decision analysis
Cost effectiveness analysis
Objective—Translating findings for practice
Options: Integrating results of the above approaches
Expert consensus methods
Developing guidelines

Integrating information in clinical practice
Options: ICT support studies
Studying diagnostic problem solving
Evaluation of implementation in practice

The impact on clinical management can be also investigated by comparing the (intended) management before and after test results are available, as was done early in evaluation of computer assisted tomography of the brain. Such before and after comparisons have specific potentials and limitations.[24]

Appropriate inclusion and exclusion criteria are indispensable for focusing on the relevant clinical question, target population, clinical spectrum, and setting (primary care or a population referred to hospital, for instance).

Synthesising findings and expertise

If results from a number of studies are available, a systematic review of diagnostic methods and meta-analysis of pooled data can provide a comprehensive synthesis of present knowledge. Diagnostic accuracy can be assessed overall and for subgroups. Much effort is being invested to make systematic reviews of diagnostic methods as solid as the methodologically more established systematic reviews of treatment methods.[25 26]

If the diagnostic problem is well structured, and if estimates are available for accuracy and risks of testing, occurrence and prognosis of the suspected disorder, and "values" of clinical outcomes, quantitative decision analysis can identify the most effective/cost effective strategy. A combined analysis of diagnostic and treatment aspects is essential. Often qualitative analysis can be already very useful. For example, non-invasive techniques can nowadays detect carotid stenoses reasonably well in asymptomatic patients. This allows preselection of patients for the more invasive investigation, carotid angiography, to decide about surgical intervention; it would yield quite a complex "decision tree." But if surgery of asymptomatic stenosis is not shown to improve prognosis,[27] the decision tree is greatly simplified: it would no longer include angiography nor surgery, and maybe not even non-invasive testing.

Decision analysis cannot always provide an answer. Problems may be too complex to be summarised in a tree; data may be missing; and there can be disagreement over valuing outcomes. Consensus procedures are then essential to translate research into practice guidelines. Clinical experts can integrate current knowledge with experience to achieve agreement on clinical guidelines for diagnostic approaches to particular medical problems.

Integrating information in practice

To help clinical investigators harvest data from clinical databases to support clinicians in improving diagnostic decisions, innovations in information anc communication technology are indispensable.[28] For utilising the potentials in this field, specific methodological requirements apply, such as avoiding confounding by indications or contraindications.

Ensuring that information providing approaches have optimal impact on the diagnostic decision making of individual clinicians is far from simple. The growing cognitive efforts associated with diagnostic management make insight into diagnostic problem solving increasingly important.[29]

Clinical studies, systematic reviews, and guideline construction are all necessary but not alone sufficient to improve practice. Implementation research has been developed to bridge the gap from clinical science to routine diagnostic management.

Setting formal standards

Assessment of diagnostic technologies would be greatly stimulated if formal standards for acceptance of diagnostic procedures in routine care were adopted by health authorities. Professional organisations are responsible for setting, implementing, maintaining, and improving clinical standards. International cooperation is important, as has been proved in the field of quality control of drugs. Along these lines, governmental, industrial, and societal funding for assessments of diagnostic technologies should be intensified.

Competing interests: None declared.

*"The Evidence Base of Clinical Diagnosis," edited by J A Knottnerus, can be purchased through the BMJ Bookshop (*www.bmjbookshop.com*)*

1 Feinstein AR. *Clinical* epidemiology. *The architecture of clinical research.* Philadelphia: WB Saunders, 1985.
2 Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine.* Boston: Little, Brown, 1985.
3 Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. *JAMA* 1984;252:2418-22.
4 Reid ML, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic research. Getting better but still not good. *JAMA* 1995;274:645-51.
5 Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6
6 Panzer RJ, Black ER, Griner PF, eds. *Diagnostic strategies for common medical problems.* Philadelphia: American College of Physicians, 1991.
7 Stoffers HEJH, Kester ADM, Kaiser V, Rinkens PELM, Knottnerus JA. Diagnostic value of signs and symptoms associated with peripheral arterial obstructive disease seen in general practice: a multivariable approach. *Med Decis Making* 1997;17:61-70.
8 Fijten GHF. *Rectal bleeding, a danger signal?* Amsterdam: Thesis Publishers, 1993.
9 Knottnerus JA, Dinant GJ. Medicine based evidence, a prerequisite for evidence based medicine. *BMJ* 1997;315:1109-10.
10 Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926-30.
11 Begg CB. Biases in the assessment of diagnostic tests. *Med Stat* 1987;6:411-23.
12 Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol* 1992;45:1143-54.
13 Feinstein AR. *Clinimetrics.* New Haven: Yale University Press, 1987.
14 Schwartz WB, Wolfe HJ, Pauker SG. Pathology and probabilities, a new approach to interpreting and reporting biopsies. *N Engl J Med* 1981;305:917-23.
15 Van Weel C, Knottnerus JA. Evidence-based interventions and comprehensive treatment. *Lancet* 1999;353:916-8.
16 Spiegelhalter DJ, Crean GP, Holden R, Knill-Jones RP. Taking a calculated risk: predictive scoring systems in dyspepsia. *Scand J Gastroenterol* 1987;22(suppl 128):S152-60.
17 Knottnerus JA. Application of logistic regression to the analysis of diagnostic data. *Med Decis Making* 1992;12:93-108.
18 Moons KG, Stijnen T, Michel BC, Buller HR, Van Es GA, Grobbee DE, et al. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas un received operating characteristic curves. *Med Decis Making* 1997;17:447-54.
19 Liechtenstein JI, Feinstein AR, Suzio KD, DeLuca V, Spiro HM. The effectiveness of panendoscopy on diagnostic and therapeutic decisions about chronic abdominal pain. *J Clin Gastroenterol* 1980;2;31-6.
20 Sackett DL, Haynes RB. The architecture of clinical diagnosis. In press.
21 Shapiro S, Venet W, Strax Ph, Roeser R. Ten to fourteen year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982;69:349-55.
22 Harms LM, Schellevis FG, van Eijk JT, Donker AJ, Bouter LM. Cardiovascular morbidity and mortality among hypertensive patients in general practice: the evaluation of long-term systematic management. *J Clin Epidemiol* 1997;50:779-86.
23 Verbeek ALM, Hendriks JHCL, Holland R, Mravunac M, Sturmans F, Day NE. Reduction of breast cancer mortality through mass-screening with modern mammography. *Lancet* 1984;i:1222-4.
24 Guyatt GH, Tugwell P, Feeny DH, Drummond MF, Haynes RB. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. *J Chron Dis* 1986;39:295-304.
25 Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;48:119-30.
26 Buntinx F. Brouwers M. Relation between sampling device and detection of abnormality in cervical smears: meta-analysis of randomised and quasi-randomised studies. *BMJ* 1996;313:1285-90.
27 Benavente O, Moher D, Pham B. Carotid endarterectomy for asymptomatic carotid stenosis: a meta-analysis. *BMJ* 1998;317:1477-80.
28 Van Wijk MA, van der Lei J, Mosseveld M, Bohnen AM, van Bemmel JH. Assessment of decision support for blood test ordering in primary care. A randomized trial. *Ann Intern Med* 2001;74:274-81.
29 Elstein AS. Heuristics and biases: selected errors in clinical reasoning. *Acad Med* 1999;74:791-4.