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, 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. 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.

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 2x2 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.

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.

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.

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.
Explain the content of the image:

**Derivation of measures of discrimination**

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Likelihood ratio</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coronary stenosis</strong>^1^</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exercise electrocardiography^*</td>
<td>65</td>
<td>89</td>
<td>5.9</td>
<td>0.39</td>
</tr>
<tr>
<td>Stress thallium scintigraphy</td>
<td>85</td>
<td>85</td>
<td>5.7</td>
<td>0.18</td>
</tr>
<tr>
<td><strong>Pancreatic cancer</strong>^2^</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ultrasoundography</td>
<td>70</td>
<td>85</td>
<td>4.7</td>
<td>0.35</td>
</tr>
<tr>
<td>Computed tomography</td>
<td>85</td>
<td>90</td>
<td>8.5</td>
<td>0.17</td>
</tr>
<tr>
<td>Angiography</td>
<td>75</td>
<td>80</td>
<td>3.8</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>Peripheral arterial occlusive disease</strong>^3^</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermittent claudication</td>
<td>31</td>
<td>93</td>
<td>4.4</td>
<td>0.74</td>
</tr>
<tr>
<td>Posterior tibial or dorsalis pedis artery pulse</td>
<td>73</td>
<td>92</td>
<td>9.1</td>
<td>0.29</td>
</tr>
<tr>
<td><strong>Colorectal cancer</strong>^4^</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Change in bowel habit</td>
<td>88</td>
<td>72</td>
<td>3.1</td>
<td>0.17</td>
</tr>
<tr>
<td>Weight loss</td>
<td>44</td>
<td>85</td>
<td>2.9</td>
<td>0.66</td>
</tr>
<tr>
<td>Erythrocyte sedimentation rate &gt;30 mm in first hour</td>
<td>40</td>
<td>96</td>
<td>10.0</td>
<td>0.42</td>
</tr>
<tr>
<td>White blood cell count &lt;$10^9$/mm$^3$</td>
<td>75</td>
<td>90</td>
<td>7.5</td>
<td>0.28</td>
</tr>
<tr>
<td>Occult blood test &gt;1 positive out of 3</td>
<td>50</td>
<td>82</td>
<td>2.7</td>
<td>0.16</td>
</tr>
</tbody>
</table>

^*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. Evaluation studies should measure these factors as reproducibly as possible, recognising that interindividual and intrindividual 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, they should not know the x-ray results; pathologists establishing an independent diagnosis must not know the clinical conclusion already. 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, 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.

**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.29 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 (see also the next article in this series)28, 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, the usual procedure, or instead of 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, 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.21 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.20 Comparability of tested and not tested subjects at baseline is, again, important.
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.12

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.13 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, 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 arc communication technology are indispensable.27 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.28

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.

---