Evidence base of clinical diagnosis

Designing studies to ensure that estimates of test accuracy are transferable

Les Irwig, Patrick Bossuyt, Paul Glasziou, Constantine Gatsonis, Jeroen Lijmer

Measures of test accuracy are often thought of as fixed characteristics determinable by research and then applicable in practice. Yet even when tests are evaluated in a study of adequate quality—one including such features as consecutive patients, a good reference standard, and independent, blinded assessments of tests and the reference standard[1]—performance of a diagnostic test in one setting may vary significantly from the results reported elsewhere.[2–8] In this paper, we explore the reasons for this variability and its implications for the design of studies of diagnostic tests.

Summary points

- Test accuracy may vary considerably from one setting to another
- This may be due to the target condition, the clinical problem, what other tests have been done, or how the test is carried out
- Larger studies than those usually done for diagnostic tests will be needed to assess transferability of results
- These studies should explore the extent to which variation in test accuracy between populations can be explained by patient and test features

True variability in test accuracy

To interpret a test's results in different setting requires an understanding of whether and why the test's accuracy varies. Broadly speaking, measures of accuracy fall into two broad categories: measures of discrimination between people who are and who are not diseased, and measures of prediction used to estimate post-test probability of disease.

Measures of discrimination

Global measures of test accuracy assess only the ability of the test to discriminate between people with and without a disease. Common examples are the area under the receiver operating characteristic curve (ROC), and the odds ratio (OR), sometimes also referred to as the diagnostic odds ratio. Such results may suffice for some broad health policy decisions—for example, to decide whether a new test is in general better than an existing test for the target condition.

Measures for prediction

The measures used to estimate the probabilities of the target condition in people who have a particular test result require both discrimination and calibration. The predictive value—the proportion of people with a particular test result who have the disease of interest—is an example. It is clumsy and difficult to estimate disease rates for categories of patients who may have different pretest probabilities of having the disease. Therefore, the estimation is often done indirectly using Bayes's theorem, based on the pretest probability and measures of test characteristics such as sensitivity and specificity or likelihood ratios in specific patients. These measures of test performance require more than discrimination. They require tests to be calibrated.

Transferability of test results

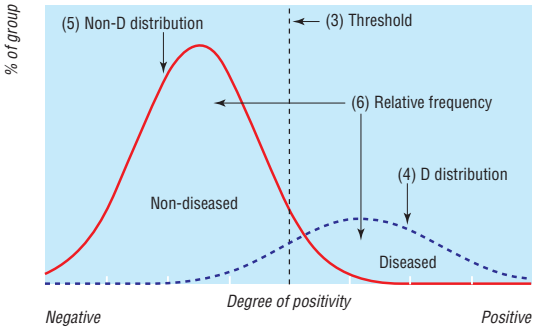
The transferability of measures of test performance from one setting to another depends on which indicator of test performance is used. The figure shows the assumptions involved in transferability. The table indicates the relation between these assumptions and the transferability of the different measures of test performance.

The main assumptions in transferring tests across settings fall into six categories.

The definition of disease is constant—Many diseases have ambiguous definitions. For example, there are no single reference standards for heart failure, Alzheimer's disease, or diabetes. Reference standards may differ because individual investigators' conceptual frameworks differ, or because it is difficult to apply the same framework in a standardised way.

The same test is used—Although based on the same principle, tests may differ—for example, over time or if made by different manufacturers.

The thresholds between categories of test result (for example, positive and negative) are constant—This is possible with a well standardised test that can be calibrated for different settings. However, there may be no accepted means of calibration—for example, different observers of imaging tests may have different thresholds for calling an image "positive." The effect of different cut-off points is classically studied by use of a receiver operating characteristic curve. In some cases calibration may be improved by using category specific likelihood ratios rather than a single cut-off point.



Distribution of test results in patients with and without the target disease. The numbers refer to assumptions for the transferability of test results (see text and table)

This is the third in a series of five articles

Screening and Test Evaluation Program, Department of Public Health and Community Medicine, University of Sydney, NSW 2006, Australia
Les Irwig *professor*
Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, PO Box 22700, 1100 DE Amsterdam, Netherlands
Patrick Bossuyt *professor of clinical epidemiology*
Jeroen Lijmer *clinical researcher*
School of Population Health, University of Queensland Medical School, Herston, QLD 4006, Australia
Paul Glasziou *professor of evidence based practice*
Center for Statistical Sciences, Brown University, Providence, RI 02192, USA
Constantine Gatsonis *professor*
Correspondence to: L Irwig lesi@health.usyd.edu.au

Series editor: J A Knottnerus

BMJ 2002;324:669–71

Assumptions for transferring different test performance characteristics (**X**=important; x=less important)

	Assumption*				
	3	4	5	6	Comment
Measures of test's discriminatory power					
Odds ratio	X	X	X		Used for global assessment of discriminatory power and are transferable if assumptions are met. Neither is concerned with calibration and therefore they cannot be used for assessing probability of disease in individuals
Area under ROC		X	X		
Measures of discriminatory power and calibration					
Predictive value	X	X	X	X	Directly estimates probability of disease in individuals
Sensitivity	X	X		x	Can be used to estimate probability of disease in individuals, using Bayes's theorem
Specificity	X		X	x	
Likelihood ratios for multicategory test	X	X	X		

ROC=receiver operating characteristic curve. *Numbered as described in text.

The distribution of test results in the disease group is constant in average (location) and spread (shape)—This assumption is not fulfilled if the spectrum of disease changes—if, for example, a screening setting is likely to include earlier disease, for which test results will be closer to those for a group without the disease (hence reducing sensitivity).

The distribution of test results in the group without disease is constant in average (location) and spread (shape)—This assumption is not fulfilled if the spectrum of non-disease changes—if, for example, the secondary care setting involves additional causes of false positives due to comorbidity not seen in primary care.

The ratio of disease to non-disease (pretest probability) is constant—If this were the case, we could use the post-test probability ("predictive" values) directly. However, this assumption is often not fulfilled—for example, the pretest probability is likely to be lowest with screening tests and greatest with tests in referred patients. This likely inconstancy is the reason for using Bayes's theorem to "adjust" the post-test probability for the pretest probability of each different setting.

All measures of test performance need the first two assumptions to be fulfilled. The importance of the last four assumptions is shown in the table, although they may not be necessary in every instance; occasionally the assumptions may not be fulfilled but, because of compensating differences, transferability is still reasonable.

Assessing transferability of discrimination and prediction

How should a study be designed to ensure that its transferability can be determined? We need first to distinguish artefactual variation from real variation in diagnostic performance. Artefactual variation arises when studies vary in the extent to which they share design features, such as whether consecutive patients were included or the reference standard and index test were read blind to each other. Once such sources of variation have been ruled out, we may explore the potential sources of true variation.[9] The issues to consider are similar to those for assessing interventions. For interventions, we consider patient, intervention, comparator, and outcome (PICO).[10 11] To ensure that readers have the necessary information to decide on the transferability of a diagnostic study to their own setting, five components need to be taken into account in design and presentation of a study.

Target condition and reference standard

The target condition and reference standard need to be carefully chosen. For example, in a study of clinically relevant tests to assess stenosis of the carotid artery, it would be sensible to dichotomise angiographic stenosis at the level of angiographic abnormality above which, on currently available evidence, the benefits of treatment outweigh harm, and to use this as the reference standard. Error in the reference standard should be minimised—for example, by better methods or multiple assessments. Any information about the accuracy of the reference standard will help interpretation.

Discriminative or predictive measures?

Assessment of the discrimination of a test requires measures such as the area under the receiver operating characteristic curve or diagnostic odds ratio. However, for estimating the probability of disease in individuals, likelihood ratios (or sensitivity and specificity) are needed, with additional information on how the tests were calibrated. Studies should include information about calibration; inclusion of selected example material, such as *x* rays of lesions, will help to clarify what thresholds have been used.

Clinical problem and population

This question defines how the initial cohort should be selected for study—for example, a new test for carotid stenosis could be considered for all patients referred to a surgical unit. However, ultrasound quantifies the extent of a stenosis reasonably accurately, so investigators may choose to restrict the study of a more expensive or invasive test to patients in whom the ultrasound result is near the decision threshold for surgery. A useful planning tool is to draw a flow diagram of how patients are selected to make up the population with the clinical problem of interest. This flow diagram shows what clinical information has been gathered, what tests have been done, and how the results of those tests determine entry into the population in which the clinical problem of interest is being studied. A good example is given in a recent paper on the assessment of imaging tests in the diagnosis of appendicitis in children.[12]

Replacement or incremental value of the test?

A key question is whether a test is being assessed as a replacement (substitution) for an existing test (because it is better or just as good and cheaper) or whether the test adds value when used in addition to specified existing tests. This decision will also be a major determinant of how the data should be analysed.[13–15]

Reasons for variability

Between test types or readers—Data should be presented on the variability between different readers or types of test and on tools to help calibration, such as standard radiographs[16 17] or laboratory quality control measures. The extent to which other factors, such as experience or training, affect reading adequacy is also helpful.

Between subgroups of the study population—Data on individuals should be available for determining the influence on test performance of the following variables: the spectrum of disease and non-disease (for example by estimating "specificity" within each category of "non-disease"); the effect of other test results, taking account of logical sequencing of tests (simplest, least invasive, cheapest are generally first); any other characteristics (for example, age and sex) that could influence test performance.

Between settings—Test performance needs to be compared in several populations or centres, as has been done for the general health questionnaire[18] and predictors of coma.[19] Variability between settings can also be explored across different studies by using meta-analytic techniques.[20 21] Studies should also explore the following sources of variability between settings that are not accounted for by the within-setting characteristics outlined in the previous section. These sources may be primary, secondary, or tertiary care; prevalence of the target condition; country or time period. Residual differences between settings should be explored to judge the extent to which there is inexplicable variability that may limit test applicability.

Summary

There is merit in studies with heterogeneous study populations. They allow exploration of the extent to which the performance of a diagnostic test depends on prespecified predictors, and how much residual variation exists. The more variation there is in study populations, the greater the potential to know how the test will perform in various settings.

We thank Petra Macaskill, Clement Loy, Andre Knottnerus, Margaret Pepe, Jonathan Craig, and Anthony Grabs for comments on the book chapter on which this paper is based.

Competing interests: None declared.

1 Begg C. Biases in the assessment of diagnostic tests. *Stat Med* 1987;6: 411-23.
2 Moons K, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* 1997;8:12-7
3 Detrano R, Janosi A, Lyons KP, Marcondes G, Abbassi N, Froehlicher VF. Factors affecting sensitivity and specificity of a diagnostic test: the exercise thallium scintigram. *Am J Med* 1988;84:699-710.
4 Hlatky M, Pryor DB, Harrrell FE, Califf RM, Mark DB, Rosati RA, et al. Factors affecting sensitivity and specificity of exercise electrocardiography. *Am J Med* 1984;77:64-71.
5 Rozanski ADG, Berman D, Forrester JS, Morris D, Swan HJC. The declining specificity of exercise radionuclide ventriculography. *N Engl J Med* 1983;309:518-22.
6 Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;17:135-40.
7 Molarius ASJ, Sans S, Tuomilehto J, Kuulasmaa K. Varying sensitivity of waist action levels to identify subjects with overweight or obesity in 19 populations of the WHO MONICA Project. *J Clin Epidemiol* 1999;52:1213-24.
8 Starmans R, Muris JWM, Fijten JH, Schouten HJ, Pop P, Knotterus JA. The diagnostic value of scoring models for organic and non-organic gastrointestinal disease, including the irritable-bowel syndrome. *Med Decis Making* 1994;14:208-16.
9 Gatsonis C, McNeil BJ. Collaborative evaluations of diagnostic tests: experience of the Radiology Diagnostic Oncology Group. *Radiology* 1990;175:571-5.
10 Sackett D, Straus S, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practise and teach EBM*. Edinburgh: Churchill Livingstone, 2000.
11 Richardson W, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. *ACP J Club* 1995; 123:A12-3.
12 Garcia Pena BM, Mandl KD, Kraus SJ, Fischer AC, Fleisher GR, Lund DP, et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. *JAMA* 1999;282:1041-6.
13 Marshall R. The predictive value of simple rules for combining two diagnostic tests. *Biometrics* 1989;45:1213-22.
14 Biggerstaff, B. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Stat Med* 2000;19:649-63.
15 Chock C, Irwig L, Berry G, Glasziou P. Comparing dichotomous screening tests when individuals negative on both tests are not verified. *J Clin Epidemiol* 1997;50:1211-7.
16 Beam C, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Arch Intern Med* 1997;156:209-13.
17 D'Orsi C, Swets J. Variability in the interpretation of mammograms. *N Engl J Med* 1995;332:1172.
18 Furukawa TGD. Cultural invariance of likelihood ratios for the general health questionnaire. *Lancet* 1999;353:10.
19 Zandbergen EGJ, de Haan RJ, Stouenbeek CP, Koelman CP, Hijdra A. Systematic review of early predictors of poor outcome in anoxic-ischaemic coma. *Lancet* 1998;352:1808-12
20 Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120:667-76.
21 Rutter C, Gatsonis C. Regression methods for meta-analysis of diagnostic test data. *Acad Radiol* 1995;2(suppl 1):S48-56.

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

A tale of two blisters

In 1924, when I was 4 years old, there was an outbreak of smallpox in Sunderland, my home town. Everyone was getting vaccinated with three inoculations of cowpox vaccine scratched into the skin over the deltoid area on the left arm. After a few days three blisters would appear, while the surrounding area became red and inflamed. To avoid pain, those who had been vaccinated wore red ribbons on their arms to prevent other people from bumping into them. As I had been vaccinated in infancy, I did not require revaccination, but my mother knew that I needed a red ribbon like everyone else so, as a wise parent, she put one on my arm.

In September 1945 I was a surgeon lieutenant in the Royal Navy in Ceylon when I was sent to Sumatra with a naval landing party to Belawan Deli, the port for Medan, the principal city in the northern part of the island. We were supposed to be getting the Japanese out, but in fact we were getting the Dutch back in, being protected from the Indonesians by armed Japanese sentries. One day when I was walking down the village street I saw a woman coming towards me wearing a red sarong and a red turban. Her face and arms were covered with blisters. I immediately decided that it must be smallpox, but then I quickly thought it over and decided that if that was the diagnosis then she should be much too ill to be walking around. I was so struck by her appearance that I ran around the block, so that I could walk past her again and get another look. I still could not see whether there was umbilication of the blisters, which I had been taught was a key diagnostic distinction between chickenpox and smallpox, and I could not make a diagnosis. When I got back to naval headquarters I asked a Dutch medical officer, and he told me that it was yaws—a tropical spirochaetal disease that is a probable variant of syphilis.

Blisters are seldom what they seem. Skimmed milk masquerades as cream.

M G Jacoby *Patchogue, New York, USA*