BMJ

Comparative accuracy: assessing new tests against existing diagnostic pathways

Patrick M Bossuyt, Les Irwig, Jonathan Craig and Paul Glasziou

BMJ 2006;332;1089-1092
doi:10.1136/bmj.332.7549.1089

Correction A correction has been published for this article. The contents of the correction have been appended to the original article in this reprint. The correction is available online at:
http://bmj.com/cgi/content/full/332/7554/1368-a

Notes

Diagnosis

Comparative accuracy: assessing new tests against existing diagnostic pathways

Patrick M Bossuyt, Les Irwig, Jonathan Craig, Paul Glasziou

Most studies of diagnostic accuracy only compare a test with the reference standard. Is this helpful?

Evaluating diagnostic accuracy is an essential step in the evaluation of medical tests.[1 2] Yet unlike randomised trials of interventions, which have a control arm, most studies of diagnostic accuracy do not compare the new test with existing tests.

We propose a modified approach to test evaluation, in which the accuracy of new tests is compared with that of existing tests or testing pathways. We argue that knowledge of other features of the new test, such as its availability and invasiveness, can help define how it is likely to be used, and we define three roles of a new test: replacement, triage, and add-on (fig 1).

Knowing the future role of new tests can help in designing studies, in making such studies more efficient, in identifying the best measure of change in accuracy, and in understanding and interpreting the results of studies.

Replacement

New tests may differ from existing ones in various ways (table 1). They may be more accurate, less invasive, easier to do, less risky, less uncomfortable for patients, quicker to yield results, technically less challenging, or more easily interpreted.

For example, biomarkers for prostate cancer have recently been proposed as a more accurate replacement for prostate specific antigen. A rapid blood test that detects individual activated effector T cells (SPOT-TB) has been introduced as a better way to diagnose tuberculosis than the tuberculin skin test. Myelography has been replaced in most centres by magnetic resonance imaging to detect spinal cord injuries, not only because it provides detailed images, but also because it is simpler, safer, and does not require exposure to radiation (table 2).

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Centre, University of Amsterdam, Amsterdam 1100 DE, Netherlands
Patrick M Bossuyt *professor of clinical epidemiology*

continued over

BMJ 2006;332:1089–92

Screening and Test Evaluation Program, School of Public Health, University of Sydney, Australia
Les Irwig *professor of epidemiology*
Screening and Test Evaluation Program, School of Public Health, University of Sydney, Department of Nephrology, Children's Hospital at Westmead, Sydney, Australia
Jonathan Craig *associate professor (clinical epidemiology)*
Department of Primary Health Care, University of Oxford, Oxford
Paul Glasziou *professor of evidence based medicine*
Correspondence to: P M Bossuyt p.m.bossuyt@amc.uva.nl

Table 1 Some features of three sets of diagnostic tests

	Replacement test (detecting herniated discs)		Triage test (detecting pulmonary embolism)		Add-on test (detecting distant metastases)	
Features	New test (magnetic resonance imaging)	Existing test (myelography)	New test (D-dimer)	Existing test (spiral computed tomography)	New test (positron emission tomography)	Existing test (computed tomography and ultrasound)
Accuracy	High	High	Low	High	High	High
Invasiveness	Non-invasive	Invasive	Non-invasive	Non-invasive	Non-invasive	Non-invasive
Waiting time	Yes	Yes	No	Yes	Yes	No
Knowledge and skills needed	Moderate	Moderate	Low	Moderate	Moderate	Moderate
Interpretable	Most tests	All tests	All tests	Most tests	Most tests	Most tests
Cost	High	High	Low	Higher	High	Medium

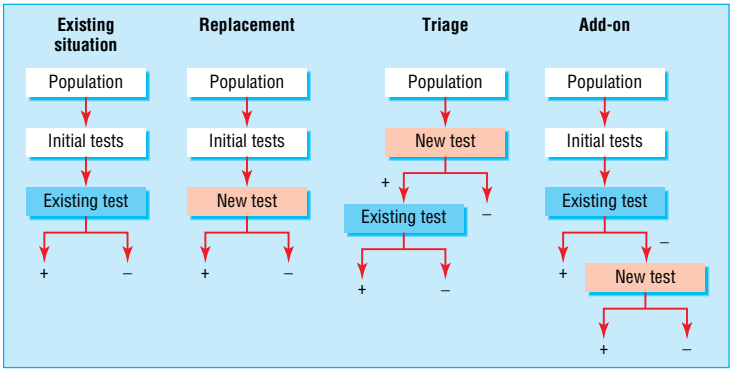
Study designs

To find out whether a new test can replace an existing one, the diagnostic accuracy of both tests has to be compared. As the sensitivity and specificity of a test can vary across subgroups, the tests must be evaluated in comparable groups or, preferably, in the same patients.[3]

Studies of comparative accuracy compare the new test with existing tests and verify test results against the same reference standard. One possibility is a paired study, in which a set of patients is tested with the existing test, the new test, and the reference standard. Another option is a randomised controlled trial, in which patients are randomly allocated to have either the existing test or the new test, after which all patients are assessed with the reference standard.

A paired study design has several advantages over a randomised trial: the patients evaluated by both tests are absolutely comparable and it may be possible to use fewer patients. Randomised trials are preferred if tests are too invasive for the old and new tests to be done in the same patients; if the tests interfere with each other, or when the study has other objectives, such as assessing adverse events, the participation of patients in testing, the actions of practitioners, or patient outcomes. Randomised controlled trials are currently being used to compare—for example—point of care cardiac markers with routine testing for the evaluation of acute coronary syndrome.

Full verification of all test results in a paired study is not always necessary to find out whether a test can act as a replacement. For example, one study compared testing for human papillomavirus DNA in self collected vaginal swabs with Papanicolaou smears to detect cervical disease and performed colposcopy (the reference standard) in all patients who tested positive on one or both of these tests.[4] For that reason, the sensitivity and specificity of the two tests could not be calculated, but the relative true and false positive rates could still be estimated, which allowed the accuracy of the two tests to be compared against the reference standard.[5–7]



Roles of tests and positions in existing diagnostic pathways

Triage

In triage, the new test is used before the existing test or testing pathway, and only patients with a particular result on the triage test continue the testing pathway (figure). Triage tests may be less accurate than existing ones and may not be meant to replace them. They have other advantages, such as simplicity or low cost.

An example of a triage instrument is the set of Ottawa ankle rules, a simple decision aid for use when ankle fractures are suspected.[8] Patients who test negative on the ankle rules (the triage test) do not need radiography (the existing test) as this makes a fracture of the malleolus or the midfoot unlikely. Another example is plasma D-dimer in the diagnosis of suspected pulmonary embolism. Patients with a low clinical probability of pulmonary embolism and a negative D-dimer result may not need computed tomography, as pulmonary embolism can be ruled out (table 2).[9]

Study designs

The triage test does not aim to improve the diagnostic accuracy of the current pathway. Rather, it reduces the use of existing tests that are more invasive, cumbersome, or expensive. Several designs can be used to compare the accuracy of the triage strategy with that of the existing test. In a fully paired study design, all patients undergo the triage test, the existing test, and the reference standard.

Designs with limited verification can be used here as well, as the primary concern is to find out whether disease will be missed with the triage test and how efficient the triage test is. One option is to use a paired design and verify the results only of patients who test negative on the triage test but positive on the existing test. This will identify patients in whom disease will be missed if the triage test is used as well as patients in whom the existing test can be avoided.

Add-on tests

Other new tests may be positioned after the existing pathway. The use of these tests may be limited to a subgroup of patients—for example, when the new test is more accurate but otherwise less attractive than existing tests (fig 1). An example is the use of positron emission tomography after ultrasound and computed tomography to stage patients with cancer. As positron emission tomography is expensive and not available in all centres, clinicians may want to restrict its use to patients in whom conventional staging did not identify distant metastases (table 1). Another example is myocardial perfusion imaging after stress (exercise) to detect coronary artery disease in patients with normal resting electrocardiograms (table 2).

Study designs

Add-on tests can increase the sensitivity of the existing pathway, possibly at the expense of specificity.[10] Alternatively, add-on tests may be used to limit the number of false positives after the existing pathway. For example, the specificity of two screening questions for depression used by general practitioners is improved by asking whether help is needed, but sensitivity is not affected.[11]

More efficient methods other than fully paired or randomised designs with complete verification can be used to evaluate the effect of the add-on test on diagnostic accuracy. In the first example, the difference in accuracy between the existing staging strategy and the additional use of positron emission tomography will depend exclusively on the patients who are positive on positron emission tomography (the add-on test). A study could therefore be limited to patients who were negative after conventional staging (the existing test) with verification by the reference standard of only those who test positive on positron emission tomography. This limited design allows us to calculate the number of extra true positives and false positives from using the add-on test.

Discussion

Several authors have proposed a multiphase model to evaluate medical tests, with an initial phase of laboratory testing and a final phase of randomised trials to compare outcome between groups of patients assessed with new tests or existing tests.[12–15] An intermediate phase is multivariable modelling to measure whether a text provides more information than is already available to the doctor.[16] We propose a model based on comparative accuracy, which compares new and existing testing pathways, and takes into account how the test is likely to be used.

A series of questions should be considered when a new test is evaluated:

- What is the existing diagnostic pathway for the identification of the target condition?
- How does the new test compare with the existing test, in accuracy and in other features?
- What is the proposed role of the new test in the existing pathway: replacement, triage, or add-on?
- Given the proposed role, what is the best measure of test performance, and how can that measure be obtained efficiently?

Table 2 Examples of proposed replacement, triage, and add-on diagnostic tests

Target condition	New test	Existing test or pathway
Replacement		
Intracerebral haemorrhage	Magnetic resonance imaging	Computed tomography
Prostate cancer	Autoantibody signatures	Prostate specific antigen
Breast cancer	Digital mammography	Plain film mammography
Iron deficiency anaemia in infants	Reticulocyte haemoglobin content	Haemoglobin
Colorectal cancer and polyps	Faecal DNA	Faecal occult blood testing
Colorectal cancer and polyps	Computed tomography colonography	Double contrast barium enema
Spinal cord compression	Magnetic resonance imaging	X ray myelography
Micrometastases in sentinel lymph nodes	Supervised automated microscopy	Routine pathology.
Childhood tuberculosis	T cell based rapid blood test	Tuberculin skin test
Acute coronary syndrome	Cardiac troponin	Serial CK
Triage		
Pulmonary embolism	D-Dimer	Computed tomography
Ankle fracture	Ottawa ankle rules	X ray
Down's syndrome	Triple test and nuchal translucency on ultrasound	Sampling of chorionic villus
Heart failure	B-type natriuretic peptide	Echocardiogram
Breast cancer with axillary lymph node metastases	Sentinel node biopsy	Axillary clearance
Cervical cancer	Human papillomavirus DNA	Colposcopy
Add-on		
Depression	"Would you like help" question	Two screening questions
Small cell lung cancer	Positron emission tomography	Conventional staging
Breast cancer with axillary lymph node metastasis	Radiocolloid mapping	Lumpectomy with sentinel node biopsy
Parkinson's disease	Neuroimaging with 123I and single photon emission computed tomography	Clinical evaluation
Acute ischaemic stroke	Computed tomography angiography	Non-contrast head computed tomography
Coronary artery disease	Myocardial perfusion scan	Electrocardiogram

Not all of these new tests will have the intended role in practice.

To determine whether a new test can serve as a replacement, triage instrument, or add-on test, we need more than a simple estimate of its sensitivity and specificity. The accuracy of the new testing strategy, as well as other relevant features, should be compared with that of the existing diagnostic pathway. We have to determine how accuracy is changed by the addition of the new test. These changes are dependent on the proposed role of the new test.

It may not always be easy to determine the existing pathway. In some cases, the prevailing diagnostic strategy may be found in practice guidelines. If a series of tests is in use, with no consensus on the optimal sequence, researchers must decide on the most appropriate comparator. This is similar to the problem of which comparator to use when intervention trials are designed against a background of substantial variation in practice.

As our understanding grows, or when circumstances change, the role of a test may change. The cost of positron emission tomography currently limits its use as an add-on test in most centres, whereas some centres have introduced this test or combined computed tomography and positron emission tomography at the beginning of the testing pathway.

Determining the likely role of a new test can also aid the critical appraisal of published study reports—for example, in judging whether the test has been evaluated in the right group of patients. Triage tests should be evaluated at the beginning of the diagnostic pathway, not in patients who tested negative with the existing tests. Purported add-on tests should be assessed after the existing diagnostic

Summary points

Studies of comparative accuracy evaluate how new tests compare with existing ones

New tests can have three main roles—replacement, triage, or add-on

Features of a new diagnostic test can help define its role

Knowing the likely role of new diagnostic tests can help in designing studies to evaluate the accuracy of tests and understand study results

pathway. Finding out whether a test can serve its role is not exclusively based on its sensitivity and specificity, but on how the accuracy of the existing testing pathway is changed by the replacement, triage, or add-on test.

In general, methods to evaluate tests have lagged behind techniques to evaluate other healthcare interventions, such as drugs. We hope that defining roles for new and existing tests, relative to existing diagnostic pathways, and using them to design and report research can contribute to evidence based health care.

Contributors and sources: PB, LI, JC, and PG designed and contributed to many studies that evaluated medical and screening tests. This paper arose from a series of discussions about ways to improve diagnostic accuracy studies. PB and LI drafted the first version of the article, which was improved by contributions from PG and JC. All authors approved the final version. PB is guarantor.

Competing interests: None declared.

1 Knottnerus JA, van Weel C, Muris JW. Evaluation of diagnostic procedures. *BMJ* 2002;324:477-80.
2 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Standards for Reporting of Diagnostic Accuracy Group. Standards for reporting of diagnostic accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* 2003;138:40-4.
3 Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002;324:669-71.
4 Wright TC Jr, Denny L, Kuhn L, Pollack A, Lorincz A. HPV DNA testing of self-collected vaginal samples compared with cytologic screening to detect cervical cancer. *JAMA* 2000;283:81-6.
5 Alonzo TA, Pepe MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Stat Med* 2002;21:835-52.
6 Pepe MS. *The statistical evaluation of medical tests for classification and prediction.* Oxford: Oxford University Press, 2003.
7 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.
8 Bachmann LM, Kolb E, Koller MT, Steurer J, ter Riet G. Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. *BMJ* 2003;326:417-20.
9 Van Belle A, Buller HR, Huisman MV, Huisman PM, Kaasjager K, Kamphuisen PW, for the Christopher Study. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. *JAMA* 2006;295:172-9.
10 Macaskill P, Walter SD, Irwig L, Franco EL. Assessing the gain in diagnostic performance when combining two diagnostic tests. *Stat Med* 2002;21:2527-46.
11 Arroll B, Goodyear-Smith F, Kerse N, Fishman T, Gunn J. Effect of the addition of a "help" question to two screening questions on specificity for diagnosis of depression in general practice: diagnostic validity study. *BMJ* 2005;331:884.
12 Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *CMAJ* 1986;134:587-94.
13 Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991;11:88-94.
14 Bruns DE. Laboratory-related outcomes in healthcare. *Clin Chem* 2001;47:1547-52.
15 Bossuyt PM, Lijmer JG, Mol BWJ. Randomised comparisons of medical tests: sometimes valid, not always efficient. *Lancet* 2000;356:1844-7.
16 Moons KG, Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem* 2004;50:473-6.

(Accepted 11 March 2006)