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

Updated information and services can be found at:
http://bmj.com/cgi/content/full/332/7549/1089

These include:

References
This article cites 15 articles, 11 of which can be accessed free at:
http://bmj.com/cgi/content/full/332/7549/1089#BIBL
3 online articles that cite this article can be accessed at:
http://bmj.com/cgi/content/full/332/7549/1089#otherarticles

Rapid responses
One rapid response has been posted to this article, which you can access for free at:
http://bmj.com/cgi/content/full/332/7549/1089#responses
You can respond to this article at:
http://bmj.com/cgi/eletter-submit/332/7549/1089

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top left of the article

Topic collections
Articles on similar topics can be found in the following collections

Diagnostics tests (394 articles)

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

To order reprints follow the “Request Permissions” link in the navigation box

To subscribe to BMJ go to:
http://resources.bmj.com/bmj/subscribers
integrity of medical research must be of top priority to protect study participants and future patients. This principle outweighs concerns over confidentiality, provided that safeguards are established to minimise threats to the competitive interests of investigators and sponsors.

We thank Tony Hope for helpful discussions and Evelyne Decullier for comments and help in obtaining legal references from France.

Contributors and sources: All authors have experience in method-ological research and clinical trials, and several authors have compared cohorts of trial protocols with publications (AWC, DG, FC, DGA). Some authors also have expertise in biostatistics research/training (RG, JAS, FC) and health law (JAS); RU has conducted research into the ethical aspects of when consent can be waived. This article arose from discussions regarding various experiences with accessing protocols: AW, DG and DGA contributed to the conception, background research, and drafting of the article. RU, JAS, DG, and FC contributed to the background research and drafting of the article. AWC is the guarantor.

Competing interests: None declared.

4 Chan AW, Kekelj-Rokic K, Schmidt I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. CMAJ 2004;171:355-40.
8 Decullier E, Leduc-Jérôme F. Fate of biomedical research protocols and publication bias in France: retrospective cohort study. BMJ 2005;331:19.
(Accepted 11 March 2006)
Roles of tests and positions in existing diagnostic pathways

Table 1 Some features of three sets of diagnostic tests

<table>
<thead>
<tr>
<th>Features</th>
<th>Replacement test (detecting herniated discs)</th>
<th>Triage test (detecting pulmonary embolism)</th>
<th>Add-on test (detecting distant metastases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New test</td>
<td>Existing test</td>
<td>New test (o-dimer)</td>
<td>New test (position emission tomography)</td>
</tr>
<tr>
<td>(magnetic resonance imaging)</td>
<td>(myelography)</td>
<td></td>
<td>(computed tomography)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Invasiveness</td>
<td>Non-invasive</td>
<td>Non-invasive</td>
<td>Non-invasive</td>
</tr>
<tr>
<td>Waiting time</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Knowledge and skills needed</td>
<td>Moderate</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Interpretable</td>
<td>Most tests</td>
<td>All tests</td>
<td>Most tests</td>
</tr>
<tr>
<td>Cost</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Knowledge and skills needed</td>
<td>Moderate</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Interpretable</td>
<td>Most tests</td>
<td>Most tests</td>
<td>Most tests</td>
</tr>
<tr>
<td>Cost</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

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.

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. 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.

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. 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 o-dimer in the diagnosis of suspected pulmonary embolism. Patients with a low clinical probability of pulmonary embolism and a negative o-dimer result may not need computed tomography, as pulmonary embolism can be ruled out (table 2).

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. 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.

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 (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 (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. An intermediate phase is multivariable modelling to measure whether a test provides more information than is already available to the doctor. 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 | Serum CK |

Triage

Pulmonary embolism  | d-Dimer | Computed tomography |
| Ankle fracture  | Ottawa ankle rules | X-ray |
| Down's syndrome | Triple test and nuchal translucency on ultrasonound | 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 papilloma virus DNA | Cytoscopy |
| 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 | Radiocolloidy 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. Purposed add-on tests should be assessed after the existing diagnostic
A complaint that changed my practice

The family asked to meet me. Their daughter had recovered from meningococcal septicemia, and they wanted to know why I hadn’t diagnosed it when they saw me that morning six weeks ago at the GP surgery. A few hours after I had treated her for an upper respiratory tract infection, her parents noticed a rash on her legs and took her straight to the accident and emergency department, where the seriousness of her condition was recognised.

The letter of complaint arrived a few weeks after she was discharged: How had I missed the diagnosis? And how was it that the emergency doctor who had seen their daughter at home a few hours before me had also dismissed her illness?

My stomach wrenched with anger and frustration. Can’t they see? That’s the whole point: two doctors a few hours apart both made the same clinical judgment that this was a viral illness. There was nothing that morning to indicate meningitis or septicemia. To the family, the fact that two doctors had failed them compounded their criticism of the quality of care they received: to me, that double failure showed the difficult reality of their daughter dying. They lost the doctor; they could have lost their daughter. I felt the parents’ fear, and I understood their terror. They had trusted him to keep her safe. They needed a doctor to walk with them, support them, and give meaning to their fears. The child got worse and nearly died, they lost the doctor; they could have lost their daughter.

The complaint wasn’t about diagnostic skills or statistical significance, but on how the accuracy of the existing testing pathway was 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.


(Accepted 11 March 2006)
Cardiopulmonary bypass is suggested to be a contributing factor to neurocognitive dysfunction after conventional coronary artery bypass grafting surgery.

Concerns were raised about graft patency after off-pump coronary artery bypass grafting surgery, which is at least as safe as conventional coronary artery bypass grafting surgery for clinical outcome.

What this study adds

Rates of graft patency are similar between coronary artery bypass grafting surgery using cardiopulmonary bypass and off-pump coronary artery bypass surgery.

Patients in the off-pump bypass group showed better preservation of neurocognitive function at six weeks and six months and better clinical outcomes required for this trial aimed to test equivalence rather than superiority. Despite the inherent weakness of this approach, we tried to minimise its effect.22

We thank Derek Robinson (Sussex University) for help with statistical planning.

Contributors: See bmj.com.

Competing interests: None declared.

Ethical approval: This study was approved by the Royal Brompton and Harefield research ethics committee.


Corrections and clarifications

Comparative accuracy: assessing new tests against existing diagnostic pathways

In this Analysis and Comment article by Patrick M Bossuyt and colleagues (BMJ 2006;332:1089-92) 6 May, the authors made a small error in table 2 that does not affect the ideas they discussed. Among the proposed add-on diagnostic tests, they listed neuroimaging with 123I for Parkinson's disease. The enigmatic and potentially confusing 123I should have been (123I)β-carboxymethoxy-3-(4-i odophenyl)-tropane. Cancer charity is to "borrow" candidate drugs shelved by companies

In this news article by Susan Mayer (BMJ 2006;332:1112 (13 May), doi:10.1136/bmj.332.7550.1112) we made an error in a quote attributed to Harpal Kumar when stating that Cancer Research UK is government funded. In fact, the charity gets very little government funding and is almost entirely funded by public donations. Bump starting the heart

An editorial oversight led to this story on the BMJ family highlights page (BMJ 2006;332:1145, 13 May) missing its citation: it came from Heart 2006;52:460.