Evaluation of diagnostic tests: from accuracy to outcome
Lijmer, J.G.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Randomised comparisons of medical tests:
Sometimes invalid, not always efficient

Patrick M.M. Bossuyt, Jeroen G. Lijmer, Ben W.J. Mol

Lancet 2000; 356:1844-1847
Abstract

Background The number of randomised diagnostic trials is slowly increasing. Such trials evaluate the effectiveness of a single diagnostic test or compare two or more tests. In this paper, we show that trials on diagnostic tests are sufficiently different from therapeutic trials to deserve special attention in designing them and additional caution in examining their results.

Randomised Comparisons of Tests Most randomised designs of tests do not offer an assessment of the test as such. Instead, they put both the test and the subsequent treatments at evaluation. Trial efficiency can be improved, by moving the point of randomisation from the decision point whether or not to test up to the point where a management decision has to be made what to do with the test results. With a single test, only patients with a test result that could lead to a difference in outcome through a change in treatment need to be included. Using the same logic, a trial that compares two tests can restrict follow-up to patients with discordant test results. Any trial should have a clear and prespecified link between test results and management decisions.

Conclusion Trialists, funding agencies, ethics committees as well as editors and readers of medical journals have to be aware that diagnostic tests warrant special attention. This can help us in moving towards a health care system where decisions on diagnostic tests are well supported by valid high quality research.
Introduction
With increasing pressure on health care budgets, decision-makers scrutinise procedures and techniques in order to maintain quality while controlling costs. Diagnostic procedures are no. Poor accuracy limits the clinical value of diagnostic test but, by themselves, high sensitivity and specificity cannot guarantee an improvement in patient outcome. Some workers have called for data on how test results affect clinical judgement, whereas others want to know how test results are used in making patient-management decisions. The most decisive evidence for judging the effectiveness of diagnostic measures should come from randomised comparisons, and the number of randomised trials with diagnostic tests is rising.

Randomized comparisons
Randomised comparisons have several advantages over other methods of comparing medical interventions. Random assignment of patients to the strategies under study should prevent any bias in the selection of patients: differences at baseline between groups of patients have to be attributed to chance. This basic principle opens up the application of experimental statistical design, such as testing for significance and calculating confidence intervals. Randomised controlled trials are also attractive from a pragmatic point of view: if randomisation coincides with a choice between two management strategies, trial design closely mimics existing clinical dilemmas. The precision of the estimates will mainly depend on the number of individuals included. One trial design is said to be more statistically efficient than another if it yields more precise estimates when applied in similarly sized groups. For example, when the two treatments under study are equally effective, designs with balanced randomisation are more efficient than trial design with alternative randomisation schemes. Similarly, a design is said to be more cost-efficient if it yields more precise estimates for the same total study cost.

Trials of a single test
Although trials are often undertaken for issues in therapy and prevention, there is no a priori reason why they should not be used to resolve difficulties in diagnosis and monitoring. Yet one should keep in mind how tests affect patient outcome. The most common way tests can affect patient outcome is when the information from these tests is used to guide decisions to start, withhold, modify, or stop treatment.

Consider a hypothetical situation in which current clinical management consists of all patients receiving the same treatment, for example pregnant women with
intrauterine growth retardation (IUGR) being managed in hospital. Women in whom doppler ultrasonography shows normal velocity profiles in the umbilical artery are thought to be at low risk for perinatal complications. For this subgroup, home care might be an alternative management option. How should we test the claim that pregnant women with IUGR benefit from the routine use of Doppler ultrasonography?

To answer this question, two strategies have to be compared. The conventional, reference strategy is admission for all: the alternative strategy includes doppler ultrasonography, with home-care for test-negative patients and admission for patients with abnormal flow in the umbilical artery. A straightforward experimental translation of this comparison is the randomised clinical trial in figure 1A. If such a trial is done, the neonatal outcome of each group can be determined. A comparison of the group results offers a measure of the effectiveness of doppler ultrasonography as an aid to management.

Unfortunately, this randomised design does not offer an assessment of the test. Instead, it evaluates both the test and the two treatments. When patient outcome in the two arms of the trial is comparable, one could conclude that a normal flow in the umbilical artery accurately selects patients for whom there is no difference in outcome between home-care and hospital management. However, when there is no difference in patient outcome in all patients satisfying the inclusion criteria for this trial, the results of the two arms will always be similar, irrespective of the accuracy of doppler ultrasonography. If all patients are worse off with home care, any strategy based on doppler ultrasonography will yield poorer results. No matter how good the test is in terms of intrinsic quality of the doppler measurements and reliability, they cannot lead to an improvement in patient outcome in these circumstances. If there is a (sub)group of patients that is better with home care, and outcome does not depend on type of management for the remaining patients, then the expected outcome in the doppler ultrasonography group will always be better than in the hospital group, regardless of the intrinsic quality or accuracy of the test. In that case, random subselction could not keep the home care strategy from being more beneficial. Consequently, it is not possible to make conclusions about the effectiveness of the test itself, as long as it remains unclear to what degree results of such a trial depend on the new treatment, accurate selection through the test, or both.

The design in Figure 1A can produce estimates of the effectiveness of the combination of test and treatment, but it is not necessarily efficient. The only patients contributing to the expected difference between the two trial arms is a subgroup of those with a normal doppler measurement. In the upper arm of figure 1A patients undergo home care. If the same patient had been randomised to the
Randomised comparisons of medical tests: Sometimes invalid, not always efficient

lower arm they would be admitted. In contrast, all patients with abnormal doppler measurements do not contribute to the expected difference between the two trial arms, since in both arms these patients are admitted.

Intuitively, one can see that trial efficiency can be improved by moving the point of randomisation from the decision point whether or not to test, to what to do with the test results. If the costs of the test are low compared with the costs of following up patients and monitoring outcome, randomising only test-negative patients (figure 1B) offers a more efficient design.

figure 1. Trial designs of a single test. IUGR = intrauterine growth retardation; R= randomisation process.; US = ultrasonography
Nienhuis and colleagues\(^2\) randomised 150 pregnant women with IUGR, with the design as shown in Figure 1A. In the doppler group, 32/74 had abnormal results and were admitted to the hospital. In the control group, doppler measurements were made but not disclosed to the clinician or midwife in attendance, and all women were admitted to hospital. 32/76 of the control group had abnormal results. Randomising only the patients who had normal doppler measurements would have reduced the number of included patients to 86, thereby reducing also the costs of follow-up and database management.

If one wanted a true evaluation of the test itself, a comparison of the doppler ultrasonography strategy with one based on a random subselection of patients could be considered. However it is unlikely that this design would be acceptable, except in areas of extreme practice variation. A more acceptable option would be to apply the test to all patients before randomisation, irrespective of the test result. If the trial is designed with enough power to allow subgroup analysis, assessing the effectiveness separately in test-positive and test-negative patients can lead to evidence-based practice recommendations. If the test results favour the new treatment in test-positive patients but not in test-negative patients, there can be a rationale for introducing the test in clinical practice. In all other cases, there is no reason for its routine use.

**Trials comparing two tests**

Concerns of validity and efficiency also applies when two tests are compared. As an example, consider scintigraphy in patients with ischemic heart disease, and whether or not to schedule patients for percutaneous transluminal coronary angioplasty. Scintigraphy can assess the functional impact of coronary lesions but, because of the costs, risks, and side-effects of percutaneous transluminal coronary angioplasty only patients with a lesion that sufficiently affects perfusion are referred for revascularisation.

A trial to test if intracoronary flow velocity measurement to evaluate the haemodynamic impact of coronary lesions can replace scintigraphy as the diagnostic tool of choice is shown in figure 2A. The design would use random allocation of consenting consecutive patients to either scintigraphy or intracoronary flow velocity measurement and compare the outcome in the respective groups. Subsequent management of patients is based on the information from the tests: test-positive patients undergo angioplasty, all others do not. A comparison of patient outcome data in both arms will produce a summary measure of the effectiveness of the test. Provided the usual criteria for trial quality are met, the resulting expression of effectiveness will be internally and externally valid. Such a trial, however, is not
Randomised comparisons of medical tests: Sometimes invalid, not always efficient

necessarily very efficient. The main reason is that tests themselves do not affect outcome, but only subsequent management decisions.

Consider once again the subgroup of patients that can contribute to a difference in average outcome between the two groups. Patients who tested positive with both scintigraphy and intracoronary flow velocity measurement would undergo angioplasty irrespective of randomisation results; their outcome would be determined by the treatment, not the test. Similarly, patients that test negative with both tests would not undergo angioplasty; their outcome would also be unaffected by the choice of the test. The only subgroup of patients that can contribute to the expected difference between the two trial arms are those that are negative on one test but positive on the other.

In the design shown in figure 2B, all patients undergo both tests and patients with concordant test results are managed accordingly. Their follow-up can be monitored, but is not essential for comparison of the tests. By contrast, patients with discordant test results are randomly allocated to undergo angioplasty or not. Follow-up information can be used to test intracoronary flow velocity measurement produces similar health outcomes. In addition to a straightforward comparison of both management options, the results can be assessed conditionally on the two tests.

Whether our suggestions for the improvement of the efficacy and validity of trials on diagnostic tests can be applied in practice will depend on the context of the trial. Doing both tests, and subsequently randomising patients with discordant test results, implies that the two test results will be known to the patient and the clinician at the moment of randomisation, and would only be ethically acceptable if there is reasonable doubt on how to link subsequent management to the results of these tests. Another difficulty arises when the expected benefit from a new test is to result from earlier application of adequate therapy. Comparison of rapid, versus overnight testing, for the identification of various infections, for example,\(^\text{10}\) cannot wait until the results of both tests are available before randomisation, as the advantage lies in the early guidance of antibiotic therapy. Nor is it feasible to obtain both test results in one patient in case the tests are likely to affect each others performance. For example a lymph node suspected of metastasis cannot be biopsied twice. In these examples the effectiveness of two different diagnostic approaches can be compared with a classic design.\(^\text{11}\)
Figure 2. Trial designs to compare two tests. IHD = ischemic heart disease; PTCA = percutaneous transluminal coronary angioplasty; R = randomisation process; positive scintigraphy = reversible perfusion defect; positive intracoronary flow velocity = insufficient reserve.
Other threats to validity

There are additional difficulties for those who want to translate trial results to clinical practice.

In figures 1 and 2 a clear and prespecified link between test results and management decisions is shown. Test-positive patients received one treatment, test-negative another. In some trials, such a link is absent. A recent trial\(^{12}\) reported comparable pregnancy rates in couples randomised to a subfertility work-up with a post-coital test, or to the work-up without post-coital test, and concluded that routine use of the post-coital test would increase treatment without a substantial improvement in pregnancy rates. However, clinicians participating in the trial did not receive prespecified instructions on how to manage couples according to the test results. Consequently, we do not know whether the comparable pregnancy rates resulted from poor performance of the post-coital test, management decisions guided by those tests, or low effectiveness of subsequent treatments.

If the subgroup that received the test had done better, on average, then such a trial without a prespecified link between test and treatment can only indicate that it is possible, in principle, to improve outcomes with the test. Trials without a clearly prespecified protocol are similar to single patient case histories. A positive outcome can be encouraging, buy workers might prefer more guidance in translating trial results to their own clinical practice. Unfortunately, the individual reader will never know how to replicate the findings in his own clinical setting in the absence of the prespecified link between test and treatment.

Running trials without a protocol for translating the test results to clinical management decisions is like putting pharmaceuticals to trial without prespecifying the preferred dosage, optimum route of administration, the need for monitoring, or the way to deal with side-effects. Designing such a drug trial would be hardly acceptable these days. Why then should we not apply the same stringent criteria to trials of test-treatment combinations?

Another threat to validity is the effect of clinicians' knowledge of the randomisation outcome on clinical management. In a randomised trial on magnetic-resonance pelvimetry in breech presentation at term,\(^{13}\) pelvimetry results of one group were reported to the obstetricians, who used them as a basis for elective caesarean section. In the control group, pelvimetry results were not disclosed until 8 weeks post partum. Although there were precisely defined minimum acceptable pelvimetry criteria for vaginal delivery of a term breech, the obstetricians were aware that magnetic-resonance pelvimetry either had shown a normal pelvis, or that the result of magnetic-resonance pelvimetry was not disclosed. It is likely that
obstetricians were more inclined to vaginal delivery when reassured by magnetic- 
resonance pelvimetry. Indeed, the duration of the first stage of labour in women 
that eventually underwent emergency sectio caesarean was almost 3 h longer in the 
magnetic- resonance pelvimetry group than in the control group.\textsuperscript{14}

Assisting clinical-decision making is not the only way that medical tests can affect 
patient’s health. Undergoing the test itself can have an impact. On the negative side, 
these effects range from slight discomfort and temporary unpleasantness, to lasting 
side effects or death.\textsuperscript{15} On the positive side, undergoing an elaborate procedure can 
also have a positive nonspecific effect on patient complaints irrespective of the 
information that results from it.\textsuperscript{16,17} Lawson and colleagues reported that 20/135 
people tested for Huntington’s disease had an adverse event in the first year after 
testing,\textsuperscript{18} and in prenatal screening for Down symptom, parents were stressed after 
false-negative test results.\textsuperscript{19}

Designing trails to cover the impact of test on health requires creativity and care. 
Standard recipes will not suffice. Awareness that diagnostic tests warrant special 
attention can help in moving towards a health care system where decisions on 
diagnostic tests are well supported by valid high quality research.

References
1. Guyatt GH, Tugwell PX, Feeny DH, Drummond MF, Haynes RB. The role of before-after 
   studies of therapeutic impact in the evaluation of diagnostic technologies. Journal of 
3. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of 
4. Pocock SJ. Clinical trials: A practical approach. 1st ed. Chichester: John Wiley & Sons, 
   1983.
   methodological quality associated with estimates of treatment effects in controlled trials. 
7. Walter SD. Determination of significant relative risks and optimal sampling procedures in 
   prospective and retrospective comparative studies of various sizes. Am J Epidemiol 
8. Bossuyt PM, Lijmer JG. Traditional health outcomes in the evaluation of diagnostic tests. 


