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.
Diagnostic testing and Prognosis: The randomised controlled trial in diagnostic research

Jeroen G. Lijmer and Patrick M.M. Bossuyt

Prepared for:

Summary

- Diagnostic test evaluations should focus on the likelihood that tests detect clinical events of interest and the effect that tests can have on these events by the way in which test results affect subsequent management decisions.
- Randomised controlled trails of diagnostic tests are feasible and several designs are possible.
- Randomised controlled trails of diagnostic tests can be made more efficient by randomising only patients with the test result of interest.
- A randomised controlled trail of diagnostic tests should incorporate a pre-specified link between test and treatment options to ascertain validity and generalizability.
- Sample-size calculations need special attention and have to include an estimation of the discordance rate.
Why bother about the prognostic impact of a diagnostic test?
For scientific purposes, it is worth knowing whether or not a result from a medical test corresponds to the truth. Can this value be trusted? Is this truly a sign of disease? These are the first questions that come to the mind in the evaluation of medical tests.

From a patient perspective, mere knowledge about the present, true state of things is in most cases not enough. In relieving health problems, information will in itself not suffice. Patients will only benefit from diagnostic test if the information generated by that test is correctly used in subsequent decisions to take action to restore or maintain a patient's health.

There are multiple ways in which medical tests can affect a patient's health. First, undergoing the test itself can have an impact. The adverse effects range from slight discomfort and temporary unpleasantness to lasting side effects or death. On the other side, undergoing an elaborate procedure can also have a non-specific positive effect on patient complaints – regardless of the information that results from it. This can be called the "placebo" effect of testing. We know very little about the magnitude and modifying factors of this context effect.

In addition to the effects from the diagnostic procedure itself, the information generated by the test also influences patients. Providing information on the likely cause of one's health problems or other aspects of health status can have both a positive and a negative effect, albeit limited. As patients, we want to be informed about the origin of our complaints, even in the absence of cure. Such information may enable us to find better ways of handling them, by developing strategies to limit their disabling impact on our daily activities.

In these cases, it is not just the present state of health that is of interest, but also the future course of disease. It then follows that the value of information from diagnostic tests lies not only in the past (where did this come from) or the present (how is it), but also in the future. Hence, the relevance of diagnostic information is closely related to prognosis: the implications for the future course of the patient's condition.

The first section of this chapter discusses the evaluation of a single test, starting from an evaluation of its prognostic value, and then moving on to the consequences for treatment. It closes with a presentation of randomised designs for evaluating test-treatment combinations. The second section contains an elaboration of the methods for comparing and evaluating multiple test strategies, also including randomised clinical trials. The chapter ends with a discussion on practical issues.
How to measure prognostic impact of a test

A recent example of the assessment of the prognostic value of a test can be found in the literature on the management of carotid disease. Several studies have examined the need to perform duplex ultrasonography in patients with a cervical bruit without further symptoms of cerebrovascular disease. To answer this question, an assessment has to be made of the value of duplex ultrasonography. Such an evaluation will often look at the amount of agreement between the index test (duplex ultrasonography) and the reference test (the best available method to reveal the true condition of the carotid arteries). In this case, the reference test will mostly likely be conventional angiography. If properly conducted, a two by two table can be constructed after the study is done and all indicators of diagnostic accuracy can be calculated. Unfortunately, many of the evaluation studies in diagnostic techniques for carotid stenosis performed so far did not meet the design requirements for an unbiased and useful evaluation.

From a patient perspective, one could successfully argue that it is not so much the correspondence with “the truth” that should be of concern, especially not in asymptomatic patients. For these patients, the true value of the information should come from the strength of the association between data on the presence and the severity of carotid stenosis and the likelihood of vascular events in the near future. The appropriate reference standard for such an evaluation will not be a diagnostic procedure. Instead, one should look for clinical information collected through a meticulous follow-up of all patients subjected to the index test.

Figure 1. Prognostic study
Figure 1 illustrates the general design of such a study. All patients with cervical bruits without previous cerebrovascular disease are eligible for the study. A duplex ultrasonography of the right and left common and internal arteries is performed in all patients and the percentage of stenosis is measured. Ideally, none of the patients receives treatment. Subsequently patients are followed by regular outpatient visits and telephone interviews. The following clinical indicators of poor outcome are recorded: TIA, stroke, myocardial infarction, unstable angina, vascular deaths and other deaths.

With data recorded in such a study standard diagnostic accuracy measures can be calculated to express the prognostic value of a test. Table 1, based on data published by Lewis et al. shows a positive and negative predictive value of 47% and 80%, respectively, in predicting a poor outcome for a stenosis > 80%, as detected on duplex. They also showed that the relative risk of a stenosis > 50% for a TIA or stroke was 2.3. However, insufficient data was presented to reconstruct the 2x2 table for this cut-off point.

The study in figure 1 can provide an answer to the question whether or not a test is able to discriminate between different risk categories for a specific event. Such prognostic information, although of value to patients and health care professionals, does not answer the question as to whether there is an intervention that can improve the prognosis of these patients. To respond to the latter question it is necessary to compare the prognosis for different treatment strategies.

<table>
<thead>
<tr>
<th></th>
<th>poor outcome</th>
<th>favourable outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>stenosis &gt; 80%</td>
<td>63 (47%)</td>
<td>72 (53%)</td>
</tr>
<tr>
<td>stenosis &lt; 80%</td>
<td>113 (20%)</td>
<td>451 (80%)</td>
</tr>
<tr>
<td>Total</td>
<td>176</td>
<td>523</td>
</tr>
</tbody>
</table>

**Randomised designs for a single test**

A slight modification of the design in figure 1 allows us to measure the prognostic value of a test within the context of subsequent clinical decision-making. Instead of treating all patients in an identical way one can randomly allocate patients to one of the two treatment strategies, establishing the prognostic value of the test in each arm, in a way that is similar to the previous example.

A straightforward comparison of patient outcome in the two treatment arms provides an answer as to which treatment is the most effective for all patients included in the trial. Moreover, an analysis stratified by test result offers the
possibility to compare the effectiveness of the treatment options for groups with identical test results.

This type of design and analysis can be illustrated with another example from the field of cerebrovascular disease. In the management of acute stroke patients the role of intravenous anticoagulation and duplex ultrasonography of the carotid arteries is unclear. A large trial has been performed, with as its primary objective to document the efficacy of unfractionated heparin in the treatment of acute stroke. A secondary objective was an evaluation of the role of duplex ultrasonography in selecting patients for anticoagulation.\textsuperscript{2, 3} A simplified version of the design of this trial is outlined in figure 2. Patients with evidence of an ischemic stroke, with symptoms present for more than 1 hour but less than 24 hours, were eligible for the study. A duplex ultrasonography of the right and left common and internal arteries was performed in all included patients. Subsequently, patients were randomised to treatment with an unfractionated heparin or placebo and followed for 3 months. A favourable outcome after stroke was defined as a score of I or II on the Glasgow Coma scale and a score of 12 to 20 on the modified Barthel Index.

Table 2a and 2b show the prognostic value of Duplex ultrasonography in each trial arm. An odds ratio can be calculated for each table. These odds ratios can be interpreted as measures of the natural prognostic value (table 2b) and the prognostic value with intervention (table 2a), respectively. Another presentation of the same data gives us table 2 c and 2 d, which provide us with information on the treatment effect in both test result categories.

Figure 2.
Diagnostic testing and prognosis: The RCT in diagnostic research

We will call the odds ratios of the latter two tables treatment effect in test normals and treatment effect in test abnormals. In case the test discriminates well between patients that benefit from treatment and those that do not, the treatment effect in test abnormals will differ from the treatment effect in test normals. The ratio of the odds ratios of these two tables can therefore be used as a measure of the prognostic impact of the test.

The study in figure 2 provides information on the treatment effect in all test result categories. In practice, it will not always be necessary or ethical to randomise all patients, as uncertainty may exist only for patients with a specific - say abnormal - test result. This will be the case when there is information available that the prognosis for normal test results is good and that patients with such results need no intervention. A logical translation of such a question into a study design would be to randomise only patients with abnormal test results between the different treatment options.

Consider the first example of duplex ultrasonography in patients with cervical bruises. Such a trial could provide evidence that the natural history of patients with a stenosis of less than 50% have a good prognosis.

**Table 2.**

<table>
<thead>
<tr>
<th></th>
<th>unfractionated heparin</th>
<th>placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>poor outcome</td>
<td>favourable outcome</td>
</tr>
<tr>
<td>stenosis &gt; 50%</td>
<td>38 (32%)</td>
<td>82 (68%)</td>
</tr>
<tr>
<td>stenosis &lt;50%</td>
<td>121 (23%)</td>
<td>400 (77%)</td>
</tr>
<tr>
<td>Total</td>
<td>159</td>
<td>482</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stenosis larger than 50% or occlusion</th>
<th>poor outcome</th>
<th>favourable outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>unfraction. heparin</td>
<td>38 (32%)</td>
<td>82 (68%)</td>
</tr>
<tr>
<td>placebo</td>
<td>51 (47%)</td>
<td>58 (53%)</td>
</tr>
<tr>
<td>Total</td>
<td>89</td>
<td>140</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stenosis smaller than 50%</th>
<th>poor outcome</th>
<th>favourable outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>unfraction. heparin</td>
<td>121 (23%)</td>
<td>400 (77%)</td>
</tr>
<tr>
<td>placebo</td>
<td>116 (22%)</td>
<td>409 (78%)</td>
</tr>
<tr>
<td>Total</td>
<td>237</td>
<td>809</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>comparison of strategies</th>
<th>poor outcome</th>
<th>favourable outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplex US</td>
<td>154 (24%)</td>
<td>491 (76%)</td>
</tr>
<tr>
<td>no Duplex US</td>
<td>167 (26%)</td>
<td>467 (74%)</td>
</tr>
<tr>
<td>Total</td>
<td>321</td>
<td>958</td>
</tr>
</tbody>
</table>

Duplex US in table e: Decision whether or not to give UFH is based on duplex ultrasonography.

The odds ratios and their 95% CI of table a to e are: 1.5 (0.99-2.4), 3.1 (2.0-4.8), 0.53 (0.31-0.90), 1.1 (0.80-1.4) and 0.88 (0.68-1.1). The relative odds ratio of a/b or c/d is 0.48.
Chapter 11

The trial outlined in figure 3 can subsequently answer the question if therapy can improve the prognosis of patients with a stenosis of 50% or more. As in the first example, all patients with cervical bruits without previous cerebrovascular disease are eligible for the study. A duplex ultrasonography of the right and left common and internal arteries is performed in all patients to measure the percentage of stenosis. Subsequently, if the stenosis is 50% or more patients are randomly assigned to receive either aspirin 325 mg a day or placebo. The clinical endpoints, TIA, stroke, myocardial infarction, unstable angina, vascular deaths and other deaths are recorded during follow-up.

Coté and colleagues performed such a trial in 1995. They randomised 372 neurologically asymptomatic patients with a carotid stenosis of 50% or more between aspirin and placebo. By comparing the outcomes in both treatment arms the effectiveness of treating patients with a stenosis of 50% or more with aspirin was evaluated (treatment effect in test abnormals). In 50 of the 188 patients receiving aspirin and 54 of the 184 patients receiving placebo a clinical event was measured during follow-up, yielding an adjusted hazard ratio (aspirin versus placebo) of 0.99 (95% CI, 0.67 to 1.46). The authors concluded that aspirin did not have a significant long-term protective effect in asymptomatic patients with high-grade stenosis (more then 50%).

Figure 3. Randomizing abnormal test results

![Diagram showing randomizing abnormal test results. Patients with cervical bruits undergo duplex US screening. If the result is abnormal, patients are randomly assigned to receive aspirin 325 mg/d or placebo. Clinical endpoints are recorded during follow-up, with outcomes measured in 2 years.]
The trial in figure 3 can also provide information on the accuracy of Duplex US in predicting the outcomes of interest (natural prognostic value). This can be done by comparing the outcome in patients in the placebo arm, who all had an abnormal test result, with the outcome in patients with a normal test result. A prerequisite for this comparison is that patient management in both of these arms is similar. Table 3a and b show the crude results and possible comparisons. Note that to calculate the diagnostic accuracy of Duplex US it is necessary to correct for the sampling rate of patients with a high-grade stenosis.\(^5\)

**Alternative randomised designs**

An alternative to the design in figure 3 would be to move the point of randomisation back in time, to the point where the test results are not yet known. This comes down to the randomisation of all patients to either disclosure or non-disclosure of the results of the test.

The latter design was used to evaluate Doppler ultrasonography of the umbilical artery in the management of women with intra uterine growth retardation (IUGR).\(^6\) A total of 150 pregnant women with IUGR underwent Doppler ultrasonography and were subsequently randomised to disclosure or non-disclosure of the test results (figure 4a). In the group in which the results of the test were revealed, women were hospitalised in case of abnormal flow and discharged with outpatient management in case of normal flow. In the non-disclosure group all patients received the conventional strategy for women with IUGR of hospitalisation, regardless of their test results. The trial compared perinatal outcome, neurological development and postnatal growth between the two strategies. The trial design, depicted in figure 4, allows us to determine the natural prognostic value and the treatment effect in test abnormals. Unfortunately the authors did not report sufficient data to reconstruct the necessary 2x2 tables.

**Table 3.**

<table>
<thead>
<tr>
<th>a</th>
<th>natural prognostic value</th>
<th>b</th>
<th>treatment effect in case of ≥ 50% stenosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>poor outcome</td>
<td>favourable outcome</td>
<td>poor outcome</td>
</tr>
<tr>
<td>stenosis &gt; 50%*</td>
<td>130 (71%)</td>
<td>54 (29%)</td>
<td>184</td>
</tr>
<tr>
<td>stenosis &lt;50%</td>
<td>255 (78%)</td>
<td>72 (22%)</td>
<td>327</td>
</tr>
<tr>
<td>Total</td>
<td>385</td>
<td>126</td>
<td>511</td>
</tr>
</tbody>
</table>

\*random sample of patients with a stenosis ≥ 50%
Chapter 11

One could move the point of randomisation further back in time, to the decision whether or not to perform the test. A translation of this comparison for the Doppler US in IUGR is the RCT in Figure 4b. Women in whom IUGR has been diagnosed are randomly allocated to two strategies.

**Figure 4.** Alternative randomized designs
The first strategy consists of applying the test of interest, Doppler ultrasonography, in all women with IUGR. In case of abnormal flow a patient is hospitalised. In case of normal flow a patient is discharged with outpatient management. In the second strategy all women with IUGR are hospitalised. Subsequently neonatal outcome is observed in each trial arm. A comparison of the outcomes in the two arms offers a measure of the effectiveness of using Doppler ultrasonography in making decisions on hospitalisation. Such a design evaluates both the test and the treatment effect; it is however not possible to distinguish the treatment effect from the prognostic value of the test.

Similar outcomes in both arms will be observed if there is no difference in outcome with either home-care or clinical management in all patients satisfying the inclusion criteria for this trial. Differences in outcome are not necessarily attributed to the test. In case of a wrong choice of treatment, the outcome of the Doppler Ultrasonography arm can turn out to be inferior to the conventional strategy, no matter how good or reliable the test actually is. This same line of reasoning can also be applied in case of a superior outcome in the Doppler arm. If there is a (sub) group of patients that is better of with home care, then the expected outcome in the Doppler ultrasonography group will always be superior, regardless of the intrinsic quality or accuracy of the test.

These examples demonstrate that it is not possible to make conclusions on the prognostic impact of the test itself using the design 4b, as long as it remains unclear to what degree results of such a trial depend on the new treatment, on accurate selection through the test, or both.

**How to compare test strategies**

In many clinical situations there are multiple tests available to examine the presence of the target condition. When one wants to compare two competing tests the first three designs introduced earlier for the evaluation of a single test have to be adapted slightly.

To compare the prognostic value of two tests, a straightforward translation of figure 1 is to perform both tests in all patients and monitor the outcome of interest during a follow-up period. Such a design is outlined in figure 5a. The data of such a study can be used to calculate and compare the prognostic value of each test, using conventional measures of diagnostic accuracy. One can also analyse the data by stratifying the results according to the possible test combinations. With two dichotomous tests this will result in a four by two table (table 4). Note that each possible combination of results on test A and test B is treated as a separate test result category, analogously to a single test with four possible result categories.
Subsequently, the predictive value or the likelihood ratio of each result category can be calculated as a measure of prognostic value.

To examine both tests in the context of subsequent clinical decision-making, it is possible to randomise all patients between two treatment strategies, similar to the design in figure 2, regardless of their test results. Figure 5b shows an example of such a design: both tests are performed and all patients are randomly allocated to one of the two treatment options. This design allows one to explore the prognostic value of both tests in each treatment arm. In addition the data of such a trial can be used to find the most effective treatment for all patients included in the trial. If statistical power allows it, subgroup analysis of the treatment effect in the four possible test result categories offers the possibility to identify the most effective treatment option for patients in the respective categories.

Although the previous design allows for many different explorations, only some are relevant from a clinical perspective. When two tests are compared, one of them is often already used in clinical practice and decisions on subsequent management are made based on this test. Let us assume that, in clinical practice, test positive patients are treated and test negative patients are not. If future decisions are to be made under the guidance of the new test, patients who test positive on the new test will be treated and those who test negative will not. This means that the only patients that will be managed differently are the ones who test positive on the existing test but negative on the new one, and those who test negative on the existing test but positive on the new one.

As patients with concordant test results (++) or (--) will receive the same management, it is unnecessary and in some circumstances even unethical to examine the treatment effect in these two subgroups. If a new test (B) is then examined with as goal to substitute the old, possible more invasive and/or costly, test (A), the design in figure 5c, randomising only the discordant test results, is more efficient. Subsequently the treatment effect and the predictive values of the discordant result categories (A+B- and A-B+) can be examined (see tables 5a-d).

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>+</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.
Figure 5. Designs to compare diagnostic strategies
Table 5.

<table>
<thead>
<tr>
<th>Treatment effect A+B-</th>
<th>Treatment effect A-B+</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+       B-       poor</td>
<td>A-       B+       poor</td>
</tr>
<tr>
<td>A+ B-     good</td>
<td>A- B+     good</td>
</tr>
<tr>
<td>Treatment I</td>
<td>Treatment I</td>
</tr>
<tr>
<td>Treatment II</td>
<td>Treatment II</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>treatment</th>
<th>A B</th>
<th>poor</th>
<th>favourable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+ _</td>
<td></td>
<td>_</td>
</tr>
<tr>
<td></td>
<td>- +</td>
<td></td>
<td>_</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>strategy based on A</th>
<th>strategy based on B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B</td>
<td>A B</td>
</tr>
<tr>
<td>poor</td>
<td>poor</td>
</tr>
<tr>
<td>outcome</td>
<td>outcome</td>
</tr>
<tr>
<td>A B</td>
<td>A B</td>
</tr>
<tr>
<td>+ -</td>
<td>+ -</td>
</tr>
<tr>
<td>- +</td>
<td>- +</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
</tr>
</tbody>
</table>

By transposing these tables it is possible to examine the effect of a clinical pathway based on test A or test B for patients with discordant test results (tables 5e to 5f). The difference in poor outcome rate between these two tables is, after correcting for the frequency of discordant results, equal to absolute risk difference of a clinical pathway based on test A compared to a pathway based on test B. To calculate the relative risk or the total risk of each strategy separately it is necessary to have information on the clinical event rate in each concordant group.

An alternative design, using the random disclosure principle, is outlined in figure 6a. Both tests A and B are performed in all patients. Subsequently patients are randomised between a clinical pathway based on test A without disclosing the results of test B or a pathway based on test B with non-disclosure of the results of test A. The same measures and tables can be obtained from such a design as discussed for the design in figure 5.

In some situations one might want to let the point of randomisation coincide with that of the clinical decision to choose either test A or test B and act on the respective results. A recent trial used this design to study two different diagnostic approaches for the management of outpatients with dysphagia.8 Patients with dysphagia are at risk for aspiration pneumonia.
Figure 6. Alternative designs to compare diagnostic strategies

A

- Patients
- A
- B
- R

- Disclose A
- Treatment I
  - Outcome
- Disclose B
  - Treatment II
  - Outcome
- Treatment II
  - Outcome

B

- Patients with dysphagia
- R

- FEEST
- PEG
  - Outcome
  - Diet
  - Outcome
- MBS
- PEG
  - Outcome
  - Diet
  - Outcome
Modified barium swallowing swallow test (MBS) and flexible endoscopic valuation of swallowing with sensory testing (FEESST) are supposed to distinguish patients who can benefit from behavioural and dietary management from those who will need a percutaneous endoscopic gastrostomy (PEG) tube.

For the discussion we consider a simplified design as outlined in figure 6b. Outpatients presenting with dysphagia were randomly allocated to either a strategy using MBS or a strategy using FEESST to guide subsequent management. During one year of follow-up the occurrence of pneumonia was recorded in both trial arms. There were 6 cases of pneumonia in the 50 (12%) patients allocated to the FEEST strategy and 14 in the 76 (18%) patients allocated to the MBS strategy. The absolute risk difference was not significantly different from zero (risk difference 6%; 95% CI – 6% to 19%). As no patient received both tests, it is not possible to distinguish the treatment effect from the prognostic value of the tests, nor is it possible to compare the outcome in the subgroups with discordant test results.

Often a new test is introduced to complement rather than to replace existing tests. One example is where the new test is to be added to the diagnostic pathway before an older test as a triage instrument. Patients with a particular test results (say, negative) on the new test will not be subjected to the existing test. Alternatively, the new test is added after the existing test, making further refinement possible in diagnosis or treatment decisions. We refer to these two options as pre-addition and post-addition.

If a test is added at the end of a diagnostic work-up to further classify disease (post-addition) all the designs, presented in figures 2 to 4 for the single test evaluation, can be used to evaluate this new classification. For example, to evaluate the prognostic impact of a genetic test for the classification of women with breast cancer in two different subgroups, one could use a design similar to the one in figure 3. Women suspected of breast cancer are evaluated with the conventional diagnostic work-up. Subsequently only women with breast cancer are eligible for the trial. In all these women genetic tests are performed. Depending on tests results they are subsequently randomised between two types of treatment.

In case the goal of a new test is to limit the amount of people undergoing the classic diagnostic work-up (triage or pre-addition), designs in figure 5b-c and 6a can be used to evaluate the prognostic impact of such a strategy. Using the principle that only patients with test results that will actually account for the difference are randomised, one could also adapt the design of figure 5b, randomising only patients with the pair of discordant test results that will be treated differently if the new strategy is adopted. Another option is drawn in figure 7a.
Figure 7. Designs to evaluate pre-addition

A

patients

Treatments

Treatment I

Treatment II

outcome

outcome

outcome

outcome

B

patients with dyspepsia

endoscopy

proton pump inhibitor

eradication therapy

outcome

outcome

outcome

HP Serology

endoscopy

eradication therapy

proton pump inhibitor

outcome

outcome

outcome

outcome
As the difference between the two strategies comes from the group of patients who are not selected for the classic diagnostic work-up, one can randomise only these patients to either the classic work-up and treatment or management based on the results of the new test.

Many studies to evaluate the pre-addition of a test have randomised all patients between the two different diagnostic work-ups.\textsuperscript{9, 10} One such study evaluated helicobacter pylori serology as a way to reduce the number of patients subjected to endoscopy. Lassen et al performed the trial outlined in figure 7b. Patients presenting in primary care with dyspepsia were randomly assigned to either H Pylori and eradication therapy or prompt endoscopy. In case of a negative Helicobacter Pylori test patients were still subjected to endoscopy. During a one-year follow-up the symptoms were recorded on a Likert scale.

**Choice of design**

Each of the designs discussed in figures 1 to 7 has its own advantages and disadvantages. Depending on the clinical problem one wants to answer, the type of information needed, and the costs of tests or follow-up, one design can be preferred over another.

The outlined in figures 2 to 4 can be used to evaluate a strategy with a new test compared to a classic strategy without such a test. In case of post addition, the classic strategy will consist of the classic diagnostic work-up and treatment. In case of a substitution problem any of the trial designs outlined in figure 5b to 6b can provide an answer. The designs outlined in figures 5b, 6a, 7a and 7b can provide an answer in case of a pre-addition problem.

Table 6 gives an overview of the information that can be deducted from the different designs.

**Table 6.**

<table>
<thead>
<tr>
<th></th>
<th>1, 5a</th>
<th>2</th>
<th>3, 4a</th>
<th>5b</th>
<th>5c, 6a</th>
<th>4b, 6b, 7a, 7b</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural prognostic value</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prognostic value under intervention</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>treatment effect test abnormalities</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>treatment effect test normals</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>treatment effect discordant tests</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strategy effect</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
Diagnostic testing and prognosis: The RCT in diagnostic research

The designs in figures 2 and 5b, testing all patients and randomising all between two treatment strategies, provide the most information. In addition to data on the effects of the two evaluated strategies, they provide information on the treatment effect and prognostic value of all possible test result categories. Yet these designs are not always ethical, as there is often evidence on one treatment being better for some of the test result categories. In that case a better alternative are the designs outlined in figures 3, 5c and 7a in which only the group of patients are randomised for which there is uncertainty in the subsequent management. The designs in figure 4b, 6b and 7b have frequently been used in the medical literature, probably due to their pragmatic attractiveness. In these designs the point of randomisation coincides with the decision to perform either test A or test B. From a cost-perspective these designs can be more economical than the other designs, in case of an expensive test, as on average less patients receive tests, as compared to the other designs. In case follow-up is expensive designs randomising only patients with the test category of interest (figures 3, 5c and 7) are more efficient, as less patients will be needed to achieve the same amount of statistical precision. However the latter designs are not feasible in case tests are compared that influence each other's performance. For example, it is not possible to compare two surgical diagnostic procedures, mediastinoscopy and anterior mediastinomy, for the detection of mediastinal lymphomas by performing them both in all patients as suspected lymphnodes are removed.

Practical issues
We have discussed the pros and cons of different designs to evaluate the prognostic impact of a single test or to compare different test strategies. In the design of a trial there are several other issues that should be considered in advance. In all of the examples we have presented here there was a pre-specified link between test results and management decisions. Test positive patients were to receive one treatment, test negative another. If such a link is absent, and physicians are free to select therapy for each test result, it will remain unclear to what extent poor results of the trial reflect deficiencies of the test itself, ineffective treatment options or, alternatively, incorrect management decisions. Detailed information on the treatment protocol is also necessary for others to implement the possible findings of the study. A clear specification of the treatment options and their relation with the different test results is an absolute necessity for any diagnostic study.

As for each randomised controlled trial, methods to preserve allocation concealment and blinding deserve special attention. It has been shown empirically that inadequate concealment of allocation as well as inadequate blinding can lead
to exaggerated estimates of a strategy's effectiveness.\textsuperscript{13} One way to guard adequate allocation concealment is a central randomisation procedure. In some situations the use of sealed opaque envelopes with monitoring of the concealment process may be more feasible.\textsuperscript{14} Blinding of the outcome measurement for the randomisation outcome is of greater importance for some outcomes than for others, but can be implemented with the same methods as developed for therapeutic trials. Blinding of the physician or patient to the allocation is more difficult. In case two different strategies are randomised (figure 6b) one can imagine that the knowledge of the type of test influences subsequent management decisions of a physician, despite a prespecified link. For example an obstetrician might be more reassured with the results of a magnetic-resonance pelvimetry in breech presentation at term compared to manual pelvimetry, which will influence subsequent decisions to perform an emergency sectio.\textsuperscript{15} One could choose a design that randomises test results to overcome this problem. Alternatively, one could try to mask the physician by only presenting standardised test results without any referral to the type of test.

The a priori calculation of the necessary sample size for a randomised diagnostic study is not straightforward. When discussing figure 5c, we showed that the expected difference in outcome between the two test strategies results from the expected difference in the category with discordant test results only. In trials in which patients are randomised to one of two test strategies (figure 6b) a large group of participants will also not contribute to the final difference. Let us explain this with another randomised diagnostic trial from the literature in which ultrasonography was compared with clinical assessment for the diagnosis of appendicitis.\textsuperscript{16} The authors report a power of 80\% to detect a reduction in the non-therapeutic operation rate from 11\% to 2\%, by randomising 302 patients. What are the nominator and denominator of these estimated rates?

Figure 8 shows the two trial arms. A large group of patients with abnormal results in the ultrasound group, indicating operation, would also have been detected at clinical examination. The same argument stands for a subgroup of patients with a normal ultrasound. The sum of these two groups forms the total with concordant test results. As patients with concordant test results will receive the same management, their event rates will be identical except for chance differences. The rate of 11\% results from $(X+N+O)/151$. The rate of 2\% results from $(Y+N+O)/151$. The rate difference, 9\%, solely results from the events in the discordant group. By assuming a concordance rate of ultrasonography with clinical assessment of 80\% one can calculate the postulated rate difference in this discordant group: 9\%/20\% is 45\%. This could result from a rate of non-therapeutic operations of 55\% in patients with a positive clinical assessment and otherwise negative ultrasound, and a rate of
10% in patients with a positive ultrasound and otherwise negative clinical examination. (This implies that the event rate is 0% in the concordant group, which is not very likely as the authors already discuss in their introduction that 15-30% of all operations are non-therapeutic.) With some extra calculations we can show that the difference assumed by the authors implies a discordance rate of at least 80%. It would be very strange to expect such a high discordance rate in advance. This example shows that it is important to incorporate the discordance rate in sample size calculations of randomised trials of diagnostic tests.

**Figure 8.**

![Figure 8](image)

**Conclusions**

In this chapter we discussed the evaluation of the prognostic impact of tests. From a patient perspective one could argue that it is not so much the correspondence with "the truth" that should be the focus of a diagnostic test evaluation but the likelihood that such a test detects events of clinical interest, and the possibilities that exist to let test results guide subsequent clinical decision making to reduce the likelihood of these events occurring. The latter can be evaluated by evaluating a test-treatment combination in a clinical trial, for which several possible designs were discussed. The examples of published randomised diagnostic trials in this chapter show that it is feasible to perform such a thorough evaluation of a diagnostic test. Several
additional examples can be found in the literature, such as trials of mediastinoscopy, cardiotocography and MRI, and of a number of screening tests. These date even back to 1975.

In most of these trials the point of randomisation coincided with the clinical decision whether or not to perform the tests. This makes it impossible to differentiate between the treatment effect and the prognostic value of the test. Power analyses of any diagnostic trial should incorporate an estimation of the discordance rate, as differences in outcome can only be expected for patients that have discordant test results. In this chapter we have shown that a design incorporating randomisation of discordant test results is more efficient, provides more information and is less prone to bias. Most important, all of these designs require a pre-specified test-treatment link. This to allow for application of study results in other settings and to guard the internal validity of the study.

References


