Medical Tests—White Paper Series

Using the Principles of Randomized Controlled Trial Design To Guide Test Evaluation

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of the Agency for Healthcare Research and Quality (AHRQ). Therefore, no statement in this report should be construed as an official position of AHRQ or the U.S. Department of Health and Human Services.

This report has been published in edited form: Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009 Sep-Oct;29(5):E1-E12. Epub 2009 Sep 22.

Suggested citation: Lord SJ, Irwig L, Bossuyt PMM. Using the principles of randomized controlled trial design to guide test evaluation. Medical Tests—White Paper Series. Agency for Healthcare Research and Quality: Rockville: MD. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=350

Using the Principles of Randomized Controlled Trial Design To Guide Test Evaluation

Authors:

Sarah J. Lord, M.B.B.S., M.S.[a,b]
Les Irwig, M.B.B.Ch., Ph.D.[b]
Patrick M.M. Bossuyt, Ph.D.[c]

[a]National Health and Medical Research Council Clinical Trials Centre, University of Sydney, Australia

[b]Screening and Test Evaluation Program, School of Public Health, University of Sydney, Australia

[c]Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, the Netherlands

Using the Principles of Randomized Controlled Trial Design To Guide Test Evaluation

Abstract

The decision to use a new test should be based on evidence that it will improve patient outcomes or produce other benefits without adversely affecting patients. In principle, long-term randomized controlled trials (RCTs) of test-plus-treatment strategies offer ideal evidence of the benefits of introducing a new test relative to current best practice. However, long-term RCTs may not always be necessary.

We advocate using the hypothetical RCT as a conceptual framework to identify what types of comparative evidence are needed for test evaluation. Evaluation begins by stating the major claims for the new test and determining whether it will be used as a replacement, add-on, or triage test to achieve these claims. A flow diagram of this hypothetical RCT is constructed to show the essential design elements, including population, prior tests, new test and existing test strategies, and primary and secondary outcomes. Critical steps in the pathway between testing and patient outcomes, such as differences in test accuracy, changes in treatment, or avoidance of other tests, are displayed for each test strategy.

All differences between the tests at these critical steps are identified and prioritized to determine the most important questions for evaluation. Long-term RCTs will not be necessary if it is valid to use other sources of evidence to address these questions. Validity will depend on issues such as the spectrum of patients identified by the old and new test strategies.

Introduction

Tests are generally used to provide diagnostic, prognostic, or predictive information to guide treatment decisions. Test evaluation is undertaken to investigate whether this information improves patient outcomes or whether the new test produces other benefits, such as improved safety or reduced costs.

Randomized controlled trials (RCTs) that allocate patients to the new test strategy or current best practice provide ideal evidence of the net benefits or harms of introducing a new test. These RCTs should have long-term followup to capture all immediate and downstream consequences of testing, including the effects of any changes in treatment. However, long-term RCTs comparing test-plus-treatment strategies may not always be available, feasible, or even necessary. In some situations more efficient RCT designs may be possible.[1,2] In other situations, comparative evidence about the safety and accuracy of the test from observational studies may suffice because trials have already demonstrated the benefits of treatment for the cases detected.[3]

Optimizing the efficiency of an RCT, or determining what other study designs may suffice, begins by describing the pathway by which the new test is expected to improve patient outcomes. Careful scrutiny of this pathway is undertaken to identify critical steps that will determine the effectiveness of the test and the comparisons needed to investigate these steps. For instance, if a new test is intended to reduce patient morbidity by detecting additional cases of disease, the most important question is: What is the effect of treatment in the extra cases detected? An efficient RCT design would therefore focus on comparing treatment vs. no treatment in this patient subgroup only: Those who test negative on the existing test but positive on the new test.

Exactly the same principles apply when planning a systematic review of evidence for test evaluation. If the benefits of an existing test-plus-treatment strategy have already been established, the type of evidence needed depends on how it will be used to alter the existing pathway. This is determined by the proposed attributes of the new test and whether it will be positioned as a replacement for the existing test, an add-on test after the existing test, or as a triage test before the existing test.[4]

To date, guidance for systematic reviews of diagnostic tests has predominantly focused on the methods for assessing test accuracy rather than how to assess the consequences of test results and other test attributes on patient outcomes. The purpose of this paper is to: (1) describe how a hypothetical RCT offers a useful conceptual framework to identify what types of comparative evidence are needed to evaluate a new test; (2) describe how the type of evidence needed varies according to whether the new test will be used as a replacement, add-on, or triage test and the intended benefits; and (3) identify situations where RCTs assessing the entire test-plus-treatment pathway are essential for conclusions about the impact of a new test on patient outcomes or where studies assessing test accuracy, safety, and other immediate or intermediate outcomes may suffice.

Box 1 summarizes how the hypothetical RCT can be used to guide test evaluation.

Box 1. Using the hypothetical randomized controlled trial (RCT) to guide test evaluation

- State the major claims for the new test to improve patient outcomes.
- Describe whether it will be used as a replacement, add-on, or triage test to achieve these claims.
- Design a hypothetical RCT of the test-and-treatment strategy to investigate these claims. Construct a flow diagram to show its key elements, including target population, prior tests, and the pathway linking testing to patient outcomes.
- Label on the flow diagram all differences between the new and existing test strategy that can be linked to patient outcomes.
- Specify the type of comparative evidence needed to measure these differences.
- Outline what assumptions need to be made if comparative accuracy and other observational studies of elements of the flow diagram will be used instead an RCT
- Judge whether these assumptions are reasonable.

Using the RCT Analogy in Test Evaluation

Planning an evaluation of a new test is analogous to developing a protocol for an RCT for the same purpose. The first task is to clarify the claim and list the primary intended changes in patient, cost, or other health service outcomes. This is analogous to defining the primary study objective for an RCT. All other potential changes in outcomes can then be listed as secondary objectives.

Potential patient benefits of testing include reducing patient mortality or morbidity, improving health-related quality of life, or other effects, such as reduced patient discomfort, anxiety, or inconvenience. To define all potential changes in outcomes, clinical experts involved in the diagnosis and management of the target test population can advise on the most likely role of the new test relative to current practice and consider how patients will be better off and how they might be worse off. It may be helpful to refer to a checklist of possible patient outcomes to assist this process. (See the paper by Bossuyt and McCaffery among these White Papers and in *Medical Decision Making*.[5])

The incremental benefits and harms of a new test may occur along one or more pathways. One pathway involves a series of steps linking improved test accuracy with changes in patient management and the effects of these management changes on patient outcomes. Changes in patient outcomes can occur along this test-treatment pathway because of a change in treatment or further testing. In another pathway, a new test may affect patient outcomes through other attributes of the procedure itself, such as safety and acceptability, without changing patient management. Tests may also affect patient outcomes through the patient's emotional, cognitive, or behavioral response to the test result and clinical management.[5]

The next task is to determine the type of evidence needed to investigate these claims. We propose that those evaluating evidence for decisionmaking imagine the design of an RCT to measure all specified patient outcomes and construct a flow diagram to map out the key elements of this trial, including the target test population, prior tests, index test strategy, comparator strategy, and outcomes, as they would appear in an RCT protocol. This flow diagram can be used to determine the type of

comparative evidence needed to demonstrate a difference in patient outcomes between the new vs. existing test strategy. The validity of the evidence available can be appraised using this hypothetical RCT as a benchmark. Conceptually, this approach is fundamentally different from, and should precede, the use of a decision model to integrate the evidence available or the development of a clinical algorithm to guide practice, although the flow diagram can be adapted for these purposes should the test evaluation identify adequate comparative evidence.

Constructing a Test Evaluation Flow Diagram

Display Existing Test Strategy

The first step when constructing the flow diagram is to display the target population and the current best test-treatment pathway for managing these patients. This pathway represents the comparator strategy. It may involve no prior testing if the new test is intended for primary screening or for monitoring treatment, or it may be an existing test strategy with subsequent management defined by the test result.

Display New Test Strategy

The next step is to define the new test and describe the alternative test-treatment pathway proposed using this test. This involves identifying where in the existing sequence of tests the new test will be used; whether it will be used as a replacement, add-on, or triage test for existing tests; and management following positive and negative test results. This pathway should be displayed alongside the existing test pathway on the flow diagram. Prior tests that are common to each pathway can be listed with the definition of the target population. Figure 1a shows a generic test evaluation flow diagram that can be used for a replacement test. Figures 1b and 1c show how the test-treatment pathway varies if the new test will be used as an add-on or triage test.

Identify Critical Comparisons

All differences between the new and existing test strategy that can be linked to immediate or downstream consequences for the patient, both benefits and harms, can be labeled as critical comparisons on the flow diagram. These differences will determine the effectiveness of the new test and can be used to formulate the research questions for evaluation. As shown in Figure 1, differences in test safety and test accuracy always deserve consideration. If a difference in test sensitivity or specificity is identified, changes in management for the extra true or false positive or negative test results also need to be considered together with evidence about the impact of these changes in management on patient outcomes.

Differences in test attributes along other pathways will be at least as important in many cases.[5] These may include:

- Other consequences of the test procedure itself, such as improved access for patients.
- Other consequences of the test results, such as additional clinical information about prognosis without altering treatment selection.
- Other consequences of clinical decisions, such as increased adherence to treatment or adoption of healthy behaviors among patients testing positive using the new test.

Priority can be assigned to pathways and comparisons within pathways that are directly associated with higher order effects such as patient mortality and morbidity. For example, if a new test is more sensitive than the existing test, priority is assigned to the test-treatment pathway. Within this pathway, the difference in treatment effects for the extra cases detected is the most critical comparison determining changes in patient outcomes.

The key elements of the hypothetical RCT can be used to define criteria for selecting relevant evidence to address each comparison. Multiple flow diagrams may be needed if the role of the new test or the type of comparator differs for different patient groups—for example, if the test will be used as an add-on test in primary care but as a triage test in tertiary care. Three examples to show how the flow diagram can be used to map out critical comparisons, select evidence, and judge the need for a long-term RCT are discussed below and summarized in Table 1.

Identifying the Type of Comparative Evidence Needed for Test Evaluation

1. The Replacement Test

A new replacement test may be introduced to improve patient outcomes by improving treatment selection along the test-treatment pathway if it is more sensitive and/or specific than the existing test or by providing benefits along other pathways due to other attributes.

Critical comparisons and evidence

If the new test is intended to be more sensitive than the old test, the critical comparisons are:

- First, does treatment of the extra true positive cases detected improve patient outcomes?
- Second, what is the difference in sensitivity and specificity of the new test for detecting extra cases of disease?

Trials may have already demonstrated the efficacy of treatment among cases detected by the existing test to help address the first question. However this evidence may not apply to the extra cases detected by the new test if they represent a different spectrum of disease, as discussed below for add-on tests.

If the major intended benefit of the new test is improved specificity, the critical comparisons are:

- First, what are the benefits of fewer false positive findings?
- Second, what is the difference in sensitivity and specificity of the new test?

If the new test is more specific but less sensitive than the old test, careful assessment of the trade-off between these benefits vs. the harms of additional false negatives will be needed.

Alternatively, if the new test is intended to provide other attributes such as improved safety, the critical comparisons are:

- First, is the new test at least as sensitive and specific as the existing test so patient management is not compromised?
- Second, what is the difference in adverse event rates or other relevant outcomes?

Example: liquid-based cytology (LBC) as a replacement for the Pap test in cervical screening

Using the flow diagram in Figure 1a, consider the critical comparisons and types of evidence needed to evaluate LBC as a replacement for the conventional Pap test for cervical cancer screening (Table 1). We already have evidence that Pap test screening programs reduce the incidence and mortality of cervical cancer.[6] For the purpose of this simplified example, we will assume the major claims for LBC are reduced cervical cancer incidence due to improved test sensitivity and reduced unsatisfactory slide rates.

Test-treatment pathway

The first priority is to compare the sensitivity and the specificity of LBC vs. the Pap test. Two meta-analyses have reported LBC has similar sensitivity and specificity to the Pap test for the detection of high-grade cervical abnormalities.[7,8] Based on this evidence, we conclude that LBC will not directly lead to a change in management. A long-term RCT of the entire new test-plus-treatment strategy is not needed because the efficacy of the existing test-plus-treatment strategy has already been established and the new test will not alter this pathway.

The exception is if the two tests have similar sensitivity and specificity but do not have perfect agreement and detect patients in a different spectrum of disease. For example, two cytology tests being compared might have the same sensitivity when using the total number of true low-grade and high-grade lesions detected, but one test might detect a higher proportion of high-grade cytological abnormalities and a lower proportion of low-grade cytological abnormalities than the other test.

Other pathways

The next step is to assess differences in test safety and other attributes of the test that may be linked to patient outcomes along other pathways. LBC is a safe procedure and is associated with the same level of patient discomfort as the Pap test, so no differences in immediate patient outcomes are anticipated as a result of the test procedure. A comparison of unsatisfactory slide rates using LBC vs. Pap tests takes next priority. A reduced unsatisfactory slide rate for LBC may provide immediate patient benefits by reducing patient recall rates, associated inconvenience, and costs, as well as providing potential downstream benefits, such as improved adherence to screening protocols.

Short-term RCTs would provide ideal evidence about a difference in unsatisfactory slide rates between the two test strategies. At least one such trial has been performed.[9] If followup were taken to the next screening round, such RCTs would also provide valuable evidence about a difference in patient adherence.

Long-term RCTs are not required if assumptions linking these intermediate outcomes to long-term patient benefits, including reduced cancer incidence and mortality, appear to be reasonable.

2. The Add-On Test

Add-on tests are generally introduced to improve patient outcomes through improved treatment selection by increasing the sensitivity or specificity of a testing strategy. They may be used on all patients, in which case the test-treatment pathway resembles that shown in Figure 1a for replacement tests, or reserved for a subset of patients—for example, the addition of a more sensitive test for those testing negative

on the existing test, as shown in Figure 1b and discussed in the breast magnetic resonance imaging (MRI) example below.

Critical comparisons

The critical comparisons for add-on tests are:

- First, to establish the treatment effects of detecting extra cases of disease if adding the new test increases sensitivity, or the benefits of avoiding further tests or treatment if adding the new test increases specificity.
- Second, to assess the incremental sensitivity and specificity of adding the new test to estimate the additional proportion of patients tested who will benefit.

Traditional cross-sectional paired accuracy studies to compare the sensitivity and specificity of the new vs. existing test strategy in all patients may not be required. As highlighted in Figure 1b, the only difference in management between the new and existing test strategies occurs among patients testing positive using the new test following a negative result on the existing test. Thus verification of test results by the reference standard for this subpopulation with discordant test results is sufficient. All other patients would receive the same treatment in each arm of the hypothetical RCT and do not contribute beyond chance to any difference in treatment outcomes between the tests.

Example: MRI as an add-on test for staging early breast cancer

Breast MRI is proposed to detect additional tumor foci in women with a diagnosis of early breast cancer on mammography and ultrasonography planned for breast-conserving surgery (BCS) (Table 1). RCTs have demonstrated that BCS plus adjuvant radiotherapy is a safe and effective alternative to mastectomy with a similar risk of disease recurrence in women with stage I-II disease.[10] If mammography detects multicentric or multifocal disease, mastectomy is recommended based on evidence that this population is at higher risk of local recurrence following BCS plus radiotherapy than women with unifocal disease.[11,12] The major intended benefit of MRI is to improve overall and/or recurrence-free survival by detecting extra cases with multicentric or multifocal disease who will benefit from conversion from BCS to mastectomy. This treatment comparison, therefore, takes priority for the evaluation. The other critical comparisons are the magnitude of the increase in sensitivity; the extent to which true positive test findings lead to a change in management and, therefore, patient outcomes; and the consequences of false positive findings. These issues are discussed separately below.

Test-treatment pathway

Treatment effects. Randomized comparisons are needed to assess the efficacy of converting from BCS to mastectomy in women with additional tumor foci detected by MRI that are mammography occult (Pathway A*, Figure 1b). Should the evaluation proceed without this RCT evidence? This depends on judgments about the plausibility of assumptions that the treatment effects observed in women with mammogram-detected disease will equally apply to this new subpopulation, and the potential consequences should these assumptions later be proven incorrect.

Table 2 uses a hypothetical example to describe how estimates of treatment effects for the extra cases detected by a new test may vary according to different assumptions about patient prognosis and treatment response. In theory, the absolute benefits of a new treatment depend on three factors: the patient risk of future disease

events without this treatment (prognosis), the relative effectiveness of the new treatment, and the risks of treatment. The absolute risk reduction equals the patient baseline risk times the relative risk reduction minus the risks of treatment.

These concepts can be applied to the breast MRI example. If it is reasonable to assume mastectomy provides the same relative effects for patients detected by either test, the number of tumor recurrences avoided at 10 years per 1,000 extra cases detected will vary proportional to the prognosis of these cases when treated by standard BCS plus adjuvant radiotherapy alone. On the one hand, if the extra cases show a prognosis similar to cases detected by mammography and ultrasonography, the same absolute treatment benefits can be expected (Scenario 1, Table 2). One the other hand, if the extra cases show a similar prognosis to mammography and ultrasonography negative "noncases," the addition of MRI will not be warranted based on existing RCT evidence that BCS is adequate for this low-risk group (Scenario 2, Table 2).

Alternatively, patient prognosis and treatment response for the extra MRI-detected cases may lie somewhere between these extremes (Scenarios 3 and 4, Table 2). Observational studies can sometimes offer useful evidence about differences in prognosis between patient groups. However prognostic studies are unlikely to be feasible in this example, where the goal would be to compare long-term outcomes for the extra cases of multicentric or multifocal disease detected by MRI with cases detected by mammography alone when both groups are managed with BCS plus adjuvant radiotherapy. Even so, regardless of whether prognostic information is available, RCTs would still be needed to test the assumption that mastectomy provides the same relative effects for patients detected by either test. This would involve randomizing women with multicentric or multifocal disease to mastectomy or BCS plus adjuvant radiotherapy and comparing the treatment effects between subgroups of women defined by mammography or MRI alone using a statistical test for interaction.

Test accuracy. If comparative accuracy studies with verification of all test results are not available, cross-sectional studies that verify MRI-positive, mammography-negative patients will suffice to estimate and compare the rate of extra true positive and false positive findings. It is not essential to verify concordant negative test results or mammography and ultrasonography test-positive results because this information will not affect treatment decisions or patient outcomes.

Change in management. Evidence about the impact of the new test on changes in management is needed if there is uncertainty about whether all additional patients testing positive using the new test will receive the same treatment as cases detected by the existing test, or whether all patients missed by the existing test will not otherwise receive this treatment. Such uncertainty may arise if further testing occurs before treatment or if other clinical factors or patient preference influence treatment decisions. Even taking into account the use of needle biopsy to detect false positive MRI results, some women with a positive MRI finding of multiple tumor foci may still opt for BCS and adjuvant radiotherapy, or some women with a negative mammogram may still proceed to mastectomy due to clinical findings indicating more widespread disease at surgery. Therefore, it will not be possible to infer the additional rate of conversion to mastectomy following MRI based on the rate of extra true positive MRI-detected cases reported by accuracy studies alone. Ideally, short-term RCTs would be available to quantify the difference in mastectomy conversion rates

for true positive cases with and without MRI. Other sources of evidence would be accuracy studies with a period of followup or before-and-after studies reporting the proportion of women with a true-positive finding who go on to convert from BCS to mastectomy.[13]

Consequences of false positive findings. Finally, what are the consequences of false positive findings? The immediate consequences, such as unnecessary needle biopsies, surgery, or other interventions and costs, can be determined using data from consecutive series of tested patients that include a period of clinical followup. For example, a systematic review of breast MRI included data from accuracy studies that reported on the management of positive MRI findings.[14] This review found that although many false positive findings are investigated by needle biopsy, conversion from wide local excision to more extensive surgery due to false positive findings occurred in around 5 percent of women.

RCTs comparing the new vs. existing test strategy would provide the most valid evidence to assess differences in rates of patient anxiety and other adverse events due to these unnecessary interventions. If such trials are not available, conclusions about the benefits of adding MRI have to depend on judgments weighing the rates and consequences of extra true positives against the rates and consequences of extra false positive findings (Table 1).

3. The Triage Test

Triage tests are generally introduced to increase the safety or efficiency of a testing strategy—for example, through the avoidance of more invasive, time-consuming, or costly tests.[4] They present different comparisons from replacement tests because only a proportion of all patients tested avoid the existing test—those testing negative on the triage test, as shown in Figure 1c.

Critical comparisons

Triage tests often present trade-offs between the benefits of safer or earlier exclusion of patients without the target condition and the harms of false negatives. The critical comparisons are commonly: Is the new test at least as sensitive and specific as the existing test? What is the difference in adverse event rates or other test attributes other than accuracy?

Example: D-dimer as a triage test in suspected deep venous thrombosis

Consider the evaluation of rapid point-of-care D-dimer as a triage test prior to ultrasound in patients with suspected deep venous thrombosis (DVT) clinically assessed as low risk (Table 1). The potential benefits of this strategy include improved patient access and convenience with reduced time and costs due to the avoidance of ultrasound in patients with a negative D-dimer test.[15,16] The potential harms include increased patient morbidity due to missed diagnoses if D-dimer is less sensitive than ultrasound; increased inconvenience, time to definitive treatment, and costs for patients with a positive D-dimer who need to proceed to ultrasound; and/or increased patient anxiety and costs for false positive D-dimer findings.

Test-treatment pathway

The optimal comparative accuracy study would verify all test results with the reference standard. However, the only difference in the use of ultrasound between the test strategies occurs among patients testing negative using D-dimer (Pathway B*,

Figure 1c). Thus, studies that only compare negative D-dimer results with ultrasound can provide all the required information about clinically meaningful differences in test accuracy for the detection of DVT.

The other key critical comparisons—the effects of delayed treatment for patients with false negative results—can also be identified for consideration using the flow diagram. The consequences of delayed treatment of DVT are potentially serious and ideally measured by RCTs.

Other pathways

It is possible that the availability of D-dimer lowers the threshold for testing in patients with suspected DVT. This could be explored in short-term RCTs, which would also be ideal to compare patient convenience and other attributes.

When do we need long-term RCTs to assess trade-offs between the benefits and harms of a new test? Even if feasible, the hypothetical RCT would be unnecessary if there are good comparative studies assessing all critical comparisons, provided assumptions linking this evidence with changes in patient outcomes were judged to be reasonable. If so, decision modeling could be undertaken to integrate these data and quantify differences in patient outcomes.[17]

Discussion

We propose that considering the hypothetical RCT provides a sound conceptual framework for selecting and interpreting evidence to compare patient outcomes using a new test with current best practice. Using the RCT analogy draws the focus of test evaluation on the most critical comparisons driving the intended changes in patient outcomes. The development of a flow diagram to illustrate the proposed role of the test is helpful to identify the best measure of comparative accuracy, characterize other critical comparisons and decide whether and what sort of further research is required.

The GRADE (Grading of Recommendations, Assessment, Development and Evaluation) Working Group has recently emphasized the need to assess the consequences of test results on patient outcomes when making recommendations about the quality of evidence for a new test.[18] However, there is little guidance available about what type of evidence is needed to assess these outcomes and how to assess other attributes of the test that may have an impact on patient outcomes.

The USPSTF (U.S. Preventive Services Task Force) analytic framework for screening tests provides valuable guidance for mapping out a causal pathway linking testing with patient outcomes and identifying critical steps along the pathway, referred to as "linkages," for investigation, such as test safety, accuracy, and treatment effectiveness.[19] Differences between the new test and existing test strategies at these critical linkages will drive differences in patient outcomes. We provided examples to illustrate that these differences and the types of evidence needed vary according to whether the new test will be used as a replacement, add-on, or triage test and its intended benefits.

RCTs are needed to assess the efficacy of an existing treatment for the extra cases detected by a new more sensitive test, just as they are needed to assess the efficacy of new treatments. The same applies in reverse if the new test reclassifies some patients with a positive finding on the existing test as "disease free" or "low risk" but existing treatment trials have included this patient group. This could happen with a new triage test intended to guide the avoidance of other tests or treatment. An example is the MINDACT (Microarray In Node negative Disease may Avoid

ChemoTherapy) trial, which is designed to assess the value of a prognostic gene signature test to identify low-risk women with early breast cancer who can safely avoid chemotherapy.[20]

In situations in which comparative evidence of the diagnostic accuracy of a test is not available—for example, where the perfect reference standard does not exist—the flow diagram can be used to identify the most efficient randomized comparisons as an alternative to an RCT of the entire test-plus-treatment strategy. This approach can also be used if the test is intended to guide treatment by providing information to classify patient prognosis or to predict treatment response. In these situations, RCTs that allow a comparison of treatment effects between patients with different test results will also provide optimal evidence. For example, the benefit of testing for estrogen receptor status in women with breast cancer is supported by RCTs of tamoxifen demonstrating an interaction between a patient's estrogen receptor status and response to tamoxifen.

A new test can change patient outcomes by more than one mechanism, producing different effects, in different directions, at different time points in clinical care. All potential consequences of the new test strategy deserve attention when considering the overall effects of testing. The RCT analogy can guide prioritization according to the potential for higher order effects. For example, full-body computed tomography (CT) might lead to reduced morbidity and mortality due to earlier detection of treatable disease in some cases, but these benefits come at a price: the risk of radiation-induced cancer, adverse events of further tests, treatment of spurious findings, and patient anxiety. An evaluation of full-body CT should give priority to identifying evidence about rates of test effects that are associated with changes in patient risk of mortality or serious morbidity. This process can be even more challenging when comparing two tests with different attributes, different adverse event rates, and different sensitivity and specificity.

The construction of a flow diagram based on the hypothetical RCT should not be confused with the process used to construct a decision-analytic model. Decision models represent a later step in the evaluation process, where the intended outcomes, causal pathway, and best available evidence have already been defined, although the development of the flow diagram should inform the development of a subsequent decision-analytic model.

Linking evidence from different studies conducted in different populations can never provide evidence about the impact of a new test on patient outcomes of the same strength and quality as an RCT, which captures the entire causal pathway, including the unexpected and unknown pathways. Evaluators must interpret linked evidence with caution. Identifiable uncertainties about effect estimates for critical comparisons and assumptions about linkages in the pathway can be explored using decision modeling,[17] but modeling itself may also be prone to oversimplification and potential bias.

Finally, evaluating tests is not making a choice between using evidence from accuracy studies or developing RCTs. It should always involve scrutiny of the clinical situation, identification of the health claims of the new tests, and a clear definition of the evidence needed to support these claims or prove them false. Regardless of the feasibility of a long-term RCT, both practical and ethical, to quantify all the benefits and harms of testing, the principles of RCT design should guide the identification and interpretation of relevant comparative evidence.

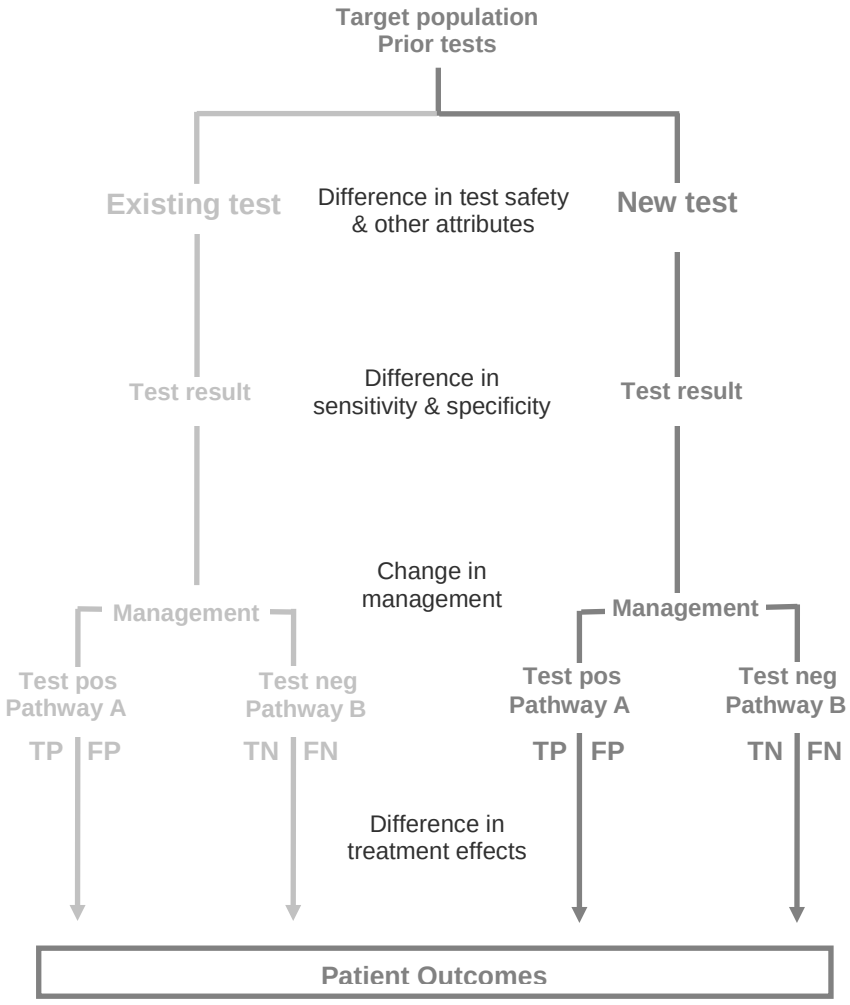
References

1. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000 Nov 25;356(9244):1844-1847.
2. Lijmer J, Bossuyt PMM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol 2009 Apr;62(4):364-373.
3. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850-855.
4. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006 May;332(7549):1089-1092.
5. Bossuyt PMM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009 Sept-Oct;29(5):E30-E38.
6. Peto J, Gilham C, Fletcher O, et al. The cervical cancer epidemic that screening has prevented in the UK. Lancet 2004 Jul;364(9430):249-256.
7. Arbyn M, Bergeron C. Liquid compared with conventional cervical cytology: a systematic review and meta-analysis. Obstet Gynecol 2008;111(1):167-177.
8. Davey E, Barratt A, Irwig L, et al. Effect of study design and quality on unsatisfactory rates, cytology classifications, and accuracy in liquid-based versus conventional cervical cytology: A systematic review. Lancet 2006;367(9505):122-132.
9. Ronco G, Cuzick J, Pierotti P, et al. Accuracy of liquid based versus conventional cytology: overall results of new technologies for cervical cancer screening: randomised controlled trial. BMJ 2007;335:28.
10. Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and surgery on early breast cancer: an overview of the randomised trials. N Engl J Med 1995;333(22):1444-1455.
11. Leopold KA, Recht A, Schnitt SJ, et al. Results of conservative surgery and radiation therapy for multiple synchronous cancers of one breast. Int J Radiat Oncol Biol Phys1989 Jan;16(1):11-16.
12. Morrow M, Strom EA, Bassett LW, et al. Standard for breast conservation therapy in the management of invasive breast carcinoma. CA Cancer J Clin 2002 Sep-Oct;52(5):277-300.
13. Guyatt GH, Tugwell PX, Feeny DH, et al. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. J Chron Dis 1986;39(4):295-304.
14. Houssami N, Ciatto S, Macaskill P, et al. Accuracy and surgical impact of magnetic resonance imaging in breast cancer staging: systematic review and meta-analysis in detection of multifocal and multicentric cancer. J Clin Oncol 2008 Jul 1;26(19):3248-3258.
15. Goodacre S, Sampson F, Stevenson M, et al. Measurement of the clinical and cost-effectiveness of non-invasive diagnostic testing strategies for deep vein thrombosis. Health Technol Assess 2006;10(15).
16. Bates SM, Kearon C, Crowther M, et al. A diagnostic strategy involving a quantitative latex D-dimer assay reliably excludes deep venous thrombosis. Ann Intern Med 2003 May 20;138(10):787-794.
17. Trikalinos TA, Siebert TA, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Mak 2009 Sep-Oct;29(5):E22-E29.
18. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106-1110.
19. Harris RP, Helfand M, Woolf SH, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J PrevMed 2001 Apr;20(3:Suppl):Suppl-35.

20. Bogaerts J, Cardoso F, Buyse M, et al. Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. Nat Clin Pract Oncol 2006 Oct;3(10):540-551.

Figure 1. Test evaluation flow diagrams

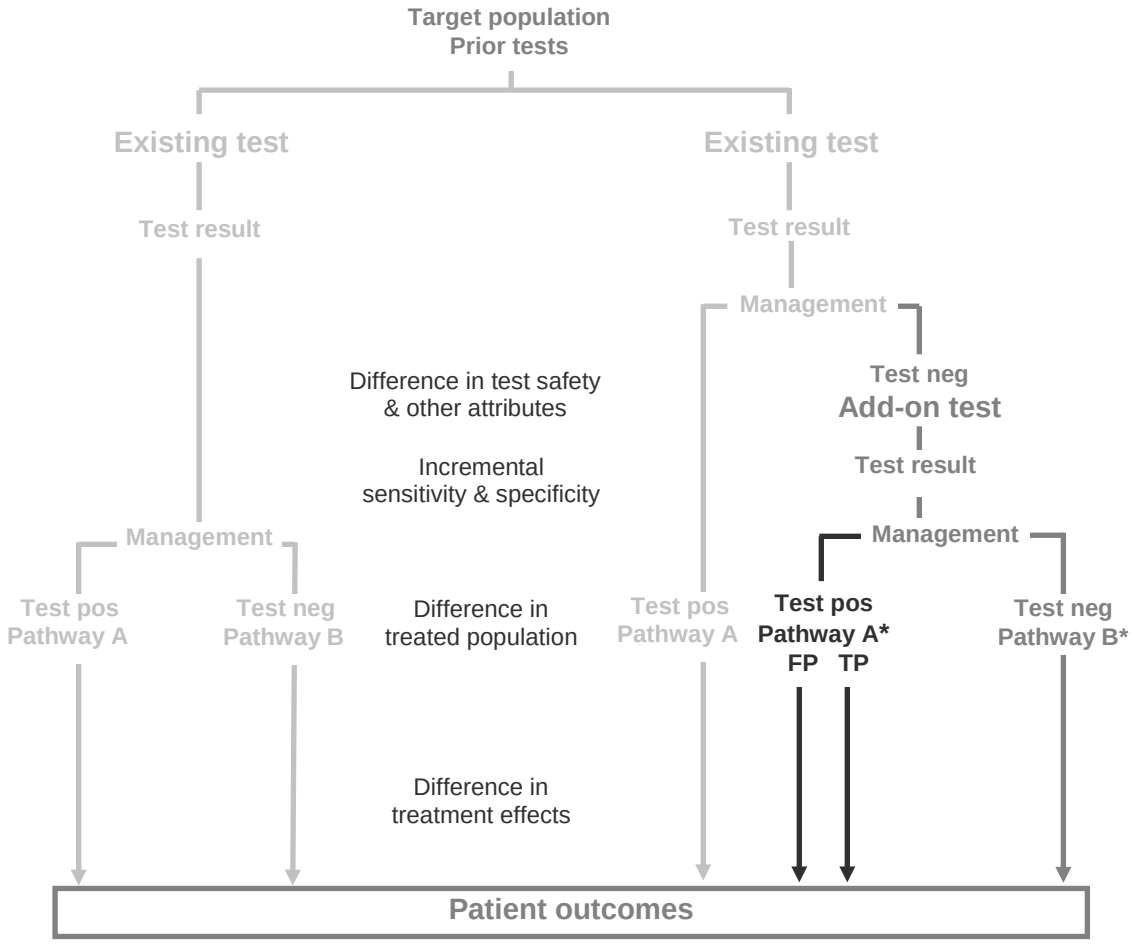
a. The replacement test



TP= true positive, FP=false positive, TN=true negative, FN=false negative

b. The add-on test

Difference in test-treatment pathway using add-on test shown in black

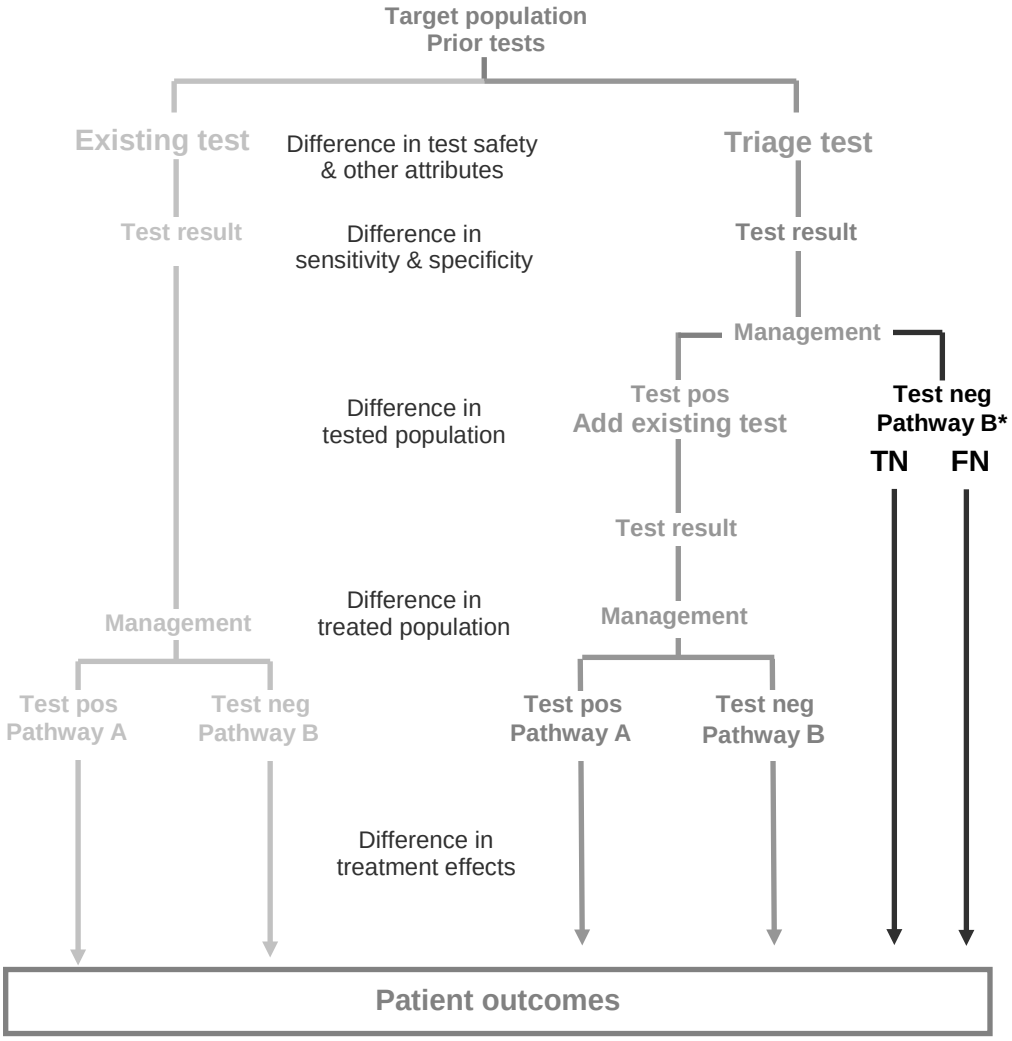


TP= true positive, FP=false positive

Pathway A* includes patients testing positive on the add-on test but negative on the existing test who would not have been assigned to treatment A using the existing test strategy.

c. The triage test

Difference in test-treatment pathway using triage test shown in black



TN = true negative, FN = false negative

Pathway B* may include patients testing negative on the triage test but positive on the existing test who would have received treatment A using the existing test strategy.