Linking the evidence: intermediate outcomes in medical test assessments

LP Staub, S Dyer, SJ Lord, RJ Simes

Abstract

Objectives

To review how health technology assessments (HTA) of medical tests incorporate intermediate outcomes in conclusions about the effectiveness of tests on improving health outcomes.

Methods

Systematic review of English-language test assessments in the HTA database from January 2005 to February 2010, supplemented by a search of the websites of International Network of Agencies for Health Technology Assessment (INAHTA) members.

Results

149 HTAs from eight countries were assessed. Half evaluated tests for screening or diagnosis, a third for disease classification (including staging, prognosis, monitoring), and a fifth for multiple purposes. In 71 HTAs (48%) only diagnostic accuracy was reported, while in 17 (11%) evidence of health outcomes was reported in addition to accuracy. Intermediate outcomes, mainly the impact of test results on patient management, were considered in 61 HTAs (41%). Of these, 47 identified randomized trials or observational studies reporting intermediate outcomes. The validity of these intermediate outcomes as a surrogate for health outcomes was not consistently discussed; nor was the quality appraisal of this evidence. Clear conclusions about whether the test was effective were included in about 60% of HTAs.

Conclusions

Intermediate outcomes are frequently assessed in medical test HTAs, but interpretation of this evidence is inconsistently reported. We recommend that reviewers explain the rationale for using intermediate outcomes, identify the assumptions required to link intermediate outcomes and patient benefits and harms, and assess the quality of included studies.

Keywords: Assessment, International, Test Evaluation

Introduction

The clinical effectiveness of a new medical test is determined by the extent to which incorporating the test into clinical practice ultimately improves patient health outcomes. This depends on a series of factors. For example, the clinical effectiveness of positron emission tomography (PET) in the assessment of patients with head and neck cancer for radiotherapy depends on its accuracy to delineate the tumour, changes in the radiotherapy regimen following PET, and consequences of these changes on patient survival and quality of life (19).

Randomized controlled trials (RCTs) of tests that capture the entire clinical pathway between testing and health outcomes provide direct evidence of the clinical effectiveness of a test. Although ideal, these studies are not often done and are sometimes not feasible (4). For fast evolving technologies like medical tests, reviewers will rarely find direct trial evidence and therefore must often rely on evidence about test accuracy and other factors to draw conclusions about clinical effectiveness.

Within the test evaluation framework of Fryback and Thornbury (7), these factors can be regarded as critical steps along the clinical pathway linking the use of the test to patient health outcomes (Figure 1). Diagnostic accuracy is a measure of how well a test identifies patients with and without a disorder, commonly reported as test sensitivity and specificity (5). For the purpose of this report, we have defined the direct consequences of test results, such as changes in therapeutic decisions, that can have downstream consequences for health outcomes, as ‘intermediate’ test outcomes. Health outcomes refer to measurement of the health state of patients, which are ideally measured in treatment RCTs (17).

All these outcomes are relevant in the assessment of medical tests. Information from studies investigating test accuracy can sometimes be directly linked with health outcomes from RCTs showing that treatment for the target condition is effective to draw conclusions about the health benefits of detecting disease (15). This requires that the spectrum of disease defined by the new test is representative of cases included in the treatment RCTs.

If test accuracy and health outcomes cannot be directly linked, studies reporting intermediate outcomes — those occurring between accuracy and health outcomes — may provide additional information to strengthen conclusions about the effectiveness of a new test (Figure 1). Studies of intermediate outcomes may demonstrate that the test information has an impact on clinical decision-making, for example, by changing decisions about treatment or about the ordering of further tests. An observational study of 71 patients with head and neck cancer showed that PET changed the management plan for 32% of patients (70% when additional lesions were detected by PET, 11% when there were no additional lesions) (24). Clearly, this change in management plan does not by itself provide evidence of improved health outcomes. Hence, studies on intermediate outcomes need careful interpretation.

Current guidelines on conducting and reporting HTAs of medical tests do not provide explicit criteria about when to include intermediate outcomes, what assumptions are necessary when linking evidence of accuracy with intermediate outcomes and health outcomes, and how to assess the quality of primary studies that examine intermediate outcomes (1;6;18;20). Given this lack of guidance, we sought to understand how, and to what extent, different test outcomes are being incorporated into HTAs in current practice. We document what outcomes beyond test accuracy are being used in current HTAs of medical tests when direct evidence of health outcomes is lacking. This review focuses on intermediate outcomes and how this evidence is interpreted to draw conclusions about the clinical effectiveness of new tests.

Methods

Identification of HTA reports

We first searched the websites of HTA organizations that are INAHTA members to identify English-language test assessments published between January 2005 and February 2010 (search date, 12 February 2010). This pilot search confirmed the wide range of approaches in current test evaluation and helped refine the extraction of the data.

For the main review we then searched the HTA database (http://www.crd.york.ac.uk/crdweb) for test evaluations with a sensitive search strategy using the terms diagnos* OR test* AND english:la (search date, 25 March 2010). We included test HTAs with a primary focus on test accuracy, intermediate outcomes, and/or patient health outcomes. Reviews of outcomes peripheral to our study, such as patient or clinician confidence and testing or screening compliance, were not examined further. To be eligible for our review, HTAs had to be reports of human studies with a full report in English. We excluded methodological reviews, horizon scanning studies, newsletters, pure economic studies, reviews comparing different generations of the same technology, and guidelines for tests already used in clinical practice.

Assessment of HTA reports

We extracted general information about the name of the test, the proposed role of the test, the disease and patient group to be tested, and outcomes mentioned for each eligible HTA. Reports were classified according to the type of investigated test: screening (asymptomatic populations) (9); diagnosis (detecting or excluding disorders in symptomatic populations) (13); disease classification in patients with established diagnosis (including staging, prognosis, monitoring) (8;22); or combinations of these purposes. Where more than one research question, indication, or test was included in an HTA, the first indication identifying studies on intermediate outcomes was used. All included HTAs were independently reviewed by two investigators (SD, LS).

We compiled descriptive statistics of the frequencies of the types of tests, disease areas, and the types of reported outcomes in the HTAs. Where applicable, we classified the reported intermediate outcomes and summarized the kinds of primary studies on intermediate outcomes and how the quality of these studies was assessed. We also examined how this evidence was interpreted in the HTAs to support conclusions about the clinical effectiveness of the test. HTAs were classified as providing clear conclusions if they made a clear positive or negative statement about the clinical effectiveness based on the evidence presented or if they judged there was not enough evidence to support definitive conclusions. HTAs were classified as not providing a clear conclusion about clinical effectiveness if they did not provide any statement about the likely impact of the test on health outcomes or did not state that the evidence available was insufficient for these conclusions.

Results

Characteristics of identified HTA reports

We identified 318 non-duplicate records. Ninety-seven of these were excluded because the main focus was not test evaluation; 38 did not present data on accuracy, intermediate outcomes or patient health outcomes; 22 were horizon scanning reports or economic evaluations; and 12 were guidelines for tests already in use.

The included 149 HTAs were prepared by 18 agencies in eight countries. The types of tests evaluated were for screening (24%), diagnosis (25%), disease classification of established diagnosis (32%), or multiple purposes (19%). The most common disease areas were oncology (38%) and the circulatory system (17%), followed by endocrine and metabolic diseases (6%), infectious diseases (5%), and multiple disease areas (6%) (Table 1). Additional information and weblinks to all included HTAs are available in Supplementary Table 1.

Accuracy

Seventy-one of the 149 included HTAs (48%) reported solely on diagnostic accuracy. In 42 (59%) of these assessments we found a clear conclusion about the clinical effectiveness of the test. These conclusions were negative (that is, the test was not effective) in 19 assessments and positive (the test was effective) in 16, while in the remaining 7 assessments the authors argued that there was not enough evidence to support definitive conclusions about the effectiveness of the test to improve health outcomes.

Patient health outcomes

In addition to accuracy, evidence of patient health outcomes was reported in 17 HTAs (11%). Common outcomes were treatment success, disease progression, and treatment complication rates. Thirteen of the 17 HTAs (76%) had clear conclusions about the clinical effectiveness of the test. These conclusions were positive in 6 HTAs and negative in one. In 6 HTAs it was concluded that evidence for final conclusions about the clinical effectiveness of the test was lacking.

Intermediate outcomes

A total of 61 HTAs (41%) identified intermediate outcomes that were deemed relevant to answer the reviewers' research question. Of these, 14 did not identify any primary studies but included a theoretical discussion of intermediate outcomes. In the remaining 47, primary studies reporting on intermediate outcomes were included. Change in patient management was reported in 33 HTAs (70%) and was by far the most common intermediate outcome (Table 2). Measures of patient management included changes in medication (dose, time to discontinuation), surgical procedures (surgery avoided, postponed, or added), radiotherapy (target field, dose), ordering of further tests, hospitalization rates, duration of treatment, and referral to specialists.

Other intermediate outcomes reported were downstream patient adherence to other interventions (e.g. motivation to cease smoking or lose weight, mammography uptake), impact of testing on subsequent visits to health services or hospital admissions, change in definitive diagnosis or reducing the number of differential diagnoses, and impact on time delays (time to diagnosis, time to transfer to operative care, length of hospital stay).

In 33 HTAs (70%), at least some of the included studies reported intermediate outcomes in sufficient detail to allow an interpretation of test consequences in the clinical pathway. For example, these studies did not simply mention that patient management was changed, but specified what changes occurred by reporting rates of patients in whom surgery was avoided or chemotherapy increased. However, only 17 HTAs included studies that compared intermediate outcomes according to test results, for instance, differences in measured time to diagnosis between test positives and negatives.

Design and quality assessment of primary studies on intermediate outcomes

Studies that reported intermediate outcomes included randomized trials of tests and observational studies. In 21 HTAs, RCTs were included that measured intermediate outcomes as the primary endpoint. In 14 of these HTAs, trials also reported health outcomes. In 12 HTAs, observational diagnostic before-after designs (10) were included to provide evidence about intermediate outcomes. These studies compared planned patient management before and after test results had been made available to clinicians. In 14 HTAs other observational studies were included, of which 5 compared the consequences of testing, such as hospital admission rates, with the rates of historic controls before the test was in use.

The quality of studies on intermediate outcomes was considered in 34 of the 47 HTAs. In 14 HTAs the authors used published quality-rating tools to assess intermediate outcomes. Some of these tools had originally been developed for diagnostic accuracy studies (e.g. QUADAS (26): 4 HTAs), some for randomized trials of clinical interventions (e.g. Jadad scale (12): 10 HTAs). In 13 HTAs the authors adapted existing tools for diagnostic accuracy studies for the appraisal of intermediate outcomes. In 7 HTAs the authors developed their own quality-assessment tools, for example checklists based on recommendations by Guyatt et al (10). The results of the quality assessment were clearly reported in 30 HTAs.

Interpretation of the evidence of intermediate outcomes

Of the 47 HTAs that identified studies of intermediate outcomes, 17 mentioned in the methods section a specific test evaluation framework or guidelines describing how evidence from different outcomes was integrated. The Fryback and Thornbury framework (7) was mentioned in 5 HTAs, while 12 Australian HTAs cited the MSAC Guidelines (18) for the assessment of diagnostic technologies. Furthermore, 9 HTAs applied an overall quality rating of the body of evidence to their review.

The relationship between intermediate and patient health outcomes was considered in 31 HTAs; however, the uncertainty around assumptions linking intermediate outcomes with health benefits was inconsistently discussed. The validity and limitations of linking patient management with health outcomes was discussed in most cases (28 HTAs). In 22 HTAs these discussions were at least partly supported with data from included studies on health outcomes, but were based on untested assumptions in the other cases.

Using the evidence of intermediate outcomes, 27 of 47 (57%) HTAs drew clear conclusions about the clinical effectiveness of the investigated technology. These conclusions were positive in 15 and negative in 7. A lack of evidence to make conclusions was concluded in 5.

Discussion

We have reviewed how the international HTA community deals with the challenges of evaluating medical tests, with particular focus on the common situation where no direct evidence exists that a test improves health outcomes. Half of 149 HTAs reported evidence about the consequences of testing beyond accuracy, with 41% considering intermediate outcomes. Overall only about 60% of 149 HTAs drew clear conclusions about the clinical effectiveness of the test based on the evidence available. Here we will discuss the use of evidence of the impact of test results on patient management, the most frequently used intermediate outcome, and make recommendations about the interpretation of this evidence in HTAs of tests.

The use of intermediate outcomes is well established in test evaluation frameworks. Fryback and Thornbury’s six-tiered model (7) is arguably the most prominent of these frameworks, and similar schemes have been proposed (14). They share the basic principle of a hierarchy of types of outcome, starting with technical efficacy at the lowest level and then progressing sequentially to diagnostic accuracy, diagnostic thinking, therapeutic impact, patient health outcomes, and societal aspects. In this hierarchy, therapeutic impact provides higher level evidence of test effectiveness than accuracy. When a test has been shown to be accurate and its purpose is to improve treatment selection, change in patient management is a necessary condition for the test to improve health outcomes. It is, however, not a sufficient condition, because the test result is often only one of several factors influencing patient management, and a change of management does not necessarily lead to improved outcomes. Hence, intermediate outcomes may help answer some questions about the consequences of testing but leave reviewers with open issues about how to judge whether this evidence is an adequate surrogate for patient health outcomes.

To make valid judgments when evaluating change in patient management, we propose a structured approach that starts with making a claim about what change in patient management will occur as a consequence of the test results and how this is expected to lead to improved health outcomes. The type of management change specified and assumptions required to infer impact on health outcomes will then inform the formulation of research questions for the test HTA (Box 1). This approach is similar to the methodology of realist synthesis developed for complex policy interventions (21). Indeed, change in patient management may provide important evidence for realist reviews of tests.

The first consideration is whether evidence of test impact on change in patient management is needed for drawing conclusions about the clinical effectiveness of a test. When direct evidence of test impact on health outcomes is not available, the value of measuring patient management depends on the role the test has in the clinical pathway (3). If a new test is proposed to replace a more expensive or invasive existing test without changing practice, accuracy may suffice to recommend the new test. For example, evidence of improved or at least similar sensitivity of new fecal DNA analyses compared with the common fecal occult blood tests in colorectal cancer screening may be enough to recommend the new method, provided it is reasonable to assume that a positive test result from the new test will have the same consequences on patient management as a positive test from the old test (23).

When the consequences of test results are not well established, evidence about patient management will be relevant for assessment. In these situations, the second step for reviewers is to specify what management changes are anticipated and the assumptions required to link the management changes to change in health outcomes (Box 1). These assumptions are critical to interpretation of the evidence and ideally should be tested. We found that the key assumptions were identified in most HTAs we reviewed but not all. Evidence from published studies was often used to support these assumptions. Expert opinion is required to infer whether evidence of effective treatment from these studies can be applied to the new setting which includes the test in review. In the assessment of PET for head and neck cancer, a panel of oncologists and radio-oncologists judged that increased radiotherapy due to PET-detected additional lymph node metastases is likely to improve health outcomes based on existing evidence of the effectiveness of radiotherapy on cervical lymph node metastases (19). Such a judgment needs to weigh up the likelihood and extent of the benefits of changed management against its potential harms. However, in many of the reviewed assessments the statements of assumptions could not easily be located; they were often somewhat hidden in the discussion. We suggest giving this important issue a more prominent place in a dedicated paragraph of test HTAs.

If assumptions that changes in patient management are likely to improve outcomes appear to be reasonable, the third step is a review of the evidence for changed management (Box 1). Included studies need to report their results with a minimum standard of detail in order to be interpretable. Simply reporting a rate of ‘overall change’ is not informative. Information about the direction and extent of changed treatment after a positive and negative test result is needed to estimate the impact on health outcomes. The assumptions used for these conclusions should be explicitly stated as discussed above. Disappointingly, in only about a third of reviewed HTAs were the included primary studies sufficiently reported to allow an interpretation of changed patient management stratified by test result. Interpretation also requires information about test accuracy to determine what proportion of patients receives a change in management based on a correct diagnosis and what proportion has management changed due to a false positive or false negative test result.

In the fourth step, the quality appraisal of this evidence, reviewers have to judge whether the included studies are able to demonstrate a true change in patient management (Box 1). The different study designs are prone to varying types of bias (25). If these studies do not measure actual management in patients randomly allocated to different test strategies, the outcome is often a hypothetical assessment of planned management in a patient cohort, so it remains unclear to what extent the measured changes in planned management reflect actual clinical practice. These limitations always need consideration. We also found inconsistent use of different appraisal tools. For a systematic review evaluating the added value of structural neuro-imaging with computed tomography or magnetic resonance imaging compared with current practice in the assessment of psychotic patients (2), the authors adapted an appraisal tool commonly used for accuracy studies (QUADAS) to assess the included diagnostic before-after studies. Their subsequent publication of this method (16) is an important step towards a more consistent appraisal of these studies. However, the sources of bias relevant to accuracy studies, particularly in the verification of test results with the reference standard, do not apply to assessing the impact of test information on downstream health outcomes. More important are the types of bias encountered in intervention studies, such as differences in patient characteristics between tested groups, differences in the measurement of outcomes, or differences in the reporting of outcomes (11). In addition, appraisal should include assessing the validity of the study authors' assumptions for inferring that management is a good proxy for outcomes.

Finally, the conclusions of test HTAs should have a clear statement as to whether the use of the test is recommended (Box 1). They should also explain whether the test is accurate, changes patient management and improves health outcomes; and reviewers should specify on what basis the recommendation about the use of the test was drawn.

This review has some limitations. Because of financial and time restraints we included only English-language assessments. We believe that our sample is representative of HTAs in the current published English literature, but the extent to which the results can be applied to other

HTA settings is debatable. However, the primary aim of this review was to document the range of approaches to test evaluation used by different agencies. We believe that the HTAs used here are appropriate to document this issue. Some of the information extracted for this review was subjective, such as whether conclusions about the effectiveness of tests on improving health outcomes were clearly stated. Although two investigators (SD, LS) independently rated the included assessments and agreed on a consensus rating in cases of initial disagreement, these judgements cannot be fully objective. Finally, in undertaking this review, we have presented a framework for test evaluation that has been used by the Australian MSAC. We are aware that different agencies may hold slightly different views; we anticipate this review will stimulate discussion about the use of intermediate outcomes in medical test assessments. In particular, we have identified the need for further research in the HTAi community to establish criteria for assessing the quality of primary studies and judging the validity of assumptions when using patient management as a surrogate for health outcomes. We hope that the recommendations in our Box can be a departure point for these discussions.

In conclusion, we have demonstrated that intermediate outcomes are frequently used in medical test HTAs, but interpretation of this evidence is inconsistently reported. We recommend that reviewers routinely explain the rationale for using intermediate outcomes to investigate a claim about impact on health outcomes, identify the assumptions required to link intermediate outcomes and patient benefits and harms, and assess the quality of included studies.

Figure legends

Figure 1

Clinical pathway and determinants of the clinical effectiveness of a medical test: accuracy, intermediate outcomes (e.g. patient management) and health outcomes

Box 1

Incorporating evidence of test impact on patient management in HTAs of medical tests

References

(1) AHRQ. *Methods Guide for Medical Test Reviews (Draft)*. Rockville, MD: Agency for Healthcare Research and Quality; 2010.

(2) Albon E, Tsourapas A, Frew E, et al. Structural neuroimaging in psychosis: a systematic review and economic evaluation. Health Technol Assess 2008 May;12:iii-163.

(3) Bossuyt PM, Irwig L, Craig J, Glasziou P. Diagnosis - Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-1092.

(4) Bossuyt PM, Lijmer JG, Mol BW, et al. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-1847.

(5) Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-44.

(6) EUnetHTA. *HTA Core Model for Diagnostic Technologies v 1.0r*. European Network for Health Technology Assessment; 2008.

(7) Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decis Making 1991;11:88-94.

(8) Glasziou PP, Aronson JK. An introduction to monitoring therapeutic interventions in clinical practice. In: Glasziou PP, Irwig L, Aronson JK, editors. Evidence-based medical monitoring. Malden, MA: BMJ Books; 2008.

(9) Grootendorst DC, Jager KJ, Zoccali C, Dekker FW. Screening: why, when, and how. Kidney International 2009;76:694-699.

(10) Guyatt GH, Tugwell PX, Feeny DH, et al. The role of before after studies of therapeutic impact in the evaluation of diagnostic technologies. J Chronic Dis 1986;39:295-304.

(11) Higgins JPT, Altman DG. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. *Cochrane handbook for systematic reviews of interventions. Version 5.0.1 (updated September 2008)*. Cochrane Collaboration; 2008.

(12) Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1-12.

(13) Knottnerus JA, van Weel C, Muris JWM. Evidence base of clinical diagnosis - Evaluation of diagnostic procedures. BMJ 2002;324:477-480.

(14) Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13-21.

(15) 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.

(16) Meads CA, Davenport CF. Quality assessment of diagnostic before-after studies: development of methodology in the context of a systematic review. BMC Med Res Methodol 2009;9:3.

(17) Moher D, Schulz KF, Altman DG. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001;134:657-662.

(18) MSAC. *Guidelines for the assessment of diagnostic technologies*. Canberra: Medical Services Advisory Committee (MSAC), Commonwealth of Australia; 2005.

(19) MSAC. *Positron emission tomography for head and neck cancer*. Canberra: Medical Services Advisory Committee (MSAC), Commonwealth of Australia; 2008.

(20) NICE. *Diagnostics Assessment Programme; Interim Methods Statement (March 2010)*. London: National Institute for Health and Clinical Excellence; 2010.

(21) Pawson RF, Greenhalgh TF, Harvey GF, Walshe K. Realist review - a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy 2005;10:21-34.

(22) Pepe MS. Introduction. In: Pepe MS, editor. The statistical evaluation of medical tests for classification and prediction.Oxford: Oxford University Press; 2003.

(23) Piper MA, Aronson N, Ziegler KM, et al. *Special report: fecal DNA analysis for colon cancer screening. Assessment program 21(6)*. Blue Cross Blue Shield Association; 2006.

(24) Scott AM, Gunawardana DH, Bartholomeusz D, et al. PET changes management and improves prognostic stratification in patients with head and neck cancer: results of a multicenter prospective study. J Nucl Med 2008;49(10):1593-1600.

(25) Staub LP, Lord SJ, Simes RJ, et al. Using patient management as a proxy for patient outcomes in test evaluation. Methods for Evaluating Medical Tests and Biomarkers 2nd Symposium; 1-2 July 2010; Birmingham: 41.

(26) Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.

Table 1. Characteristics of 149 English-language HTAs of medical tests from 18 agencies in 8 countries, published Jan 2005-Feb 2010 (details provided in Supplementary Table)

Characteristic	n	%
Disease area (ICD 10)		
Infectious diseases (I)	7	5
Neoplasms (II)	57	38
Blood (III)	6	4
Endocrine, metabolic (IV)	9	6
Mental, behavioural (V)	8	5
Nervous system (VI)	3	2
Ear (VIII)	4	3
Circulatory system (IX)	25	17
Respiratory system (X)	3	2
Digestive system (XI)	3	2
Musculoskeletal system (XIII)	4	3
Genitourinary system (XIV)	4	3
Pregnancy, childbirth (XV)	3	2
Multiple	9	6
Other	4	3
Test type		
Screening	36	24
Diagnosis	37	25
Classification of established diagnosis*	48	32
Multiple types	28	19
Outcomes reported		
Accuracy only	71	48
Accuracy + patient health outcomes	17	11
Accuracy + intermediate outcomes	36	24
Accuracy + intermediate outcomes + patient health outcomes	25	17

*includes staging, prognosis, monitoring

Table 2. Types of intermediate outcomes reported in primary studies included in 47 HTAs of medical tests

Type of intermediate outcome	Screening	Diagnosis	Classification	Multiple	Total
Patient management	7	8	13	5	33
Time to treatment or procedure	3	3	1	1	8
Downstream patient compliance	3	1	4	0	8
Health visits, hospital admission rates	1	3	2	0	6
Change in diagnosis	0	2	3	1	6
Length of hospital stay	0	1	3	0	4
Number of potential diagnoses	0	1	0	1	2
Other	0	1	1	1	3
Total	14	20	27	9	70

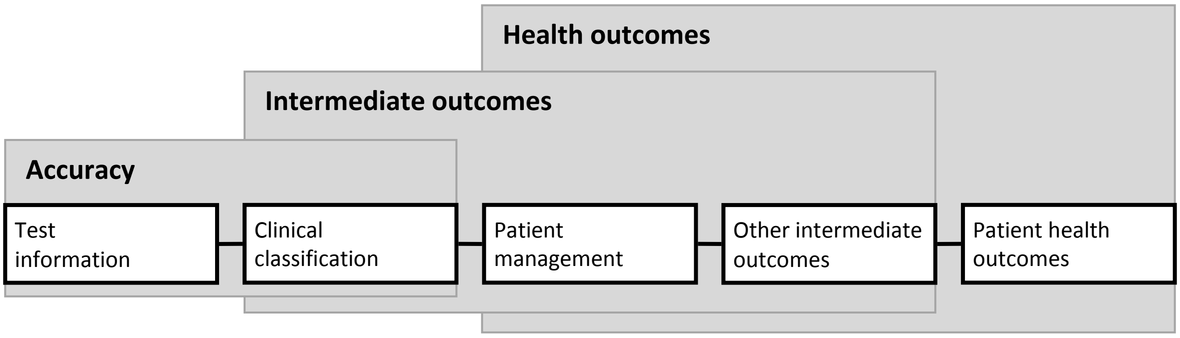
Health outcomes
Intermediate outcomes
Accuracy
Test information
Clinical classification
Patient management
Other intermediate outcomes
Patient health outcomes

Figure 1

Box 1

1. Identifying whether the consequences of test results on patient management need to be reviewed
   - Specify the consequences of test results for patient management
   - Determine whether these consequences are well defined in existing test protocols or whether a review of the evidence is needed
2. Specifying consequences of patient management for health outcomes
   - Specify test-related changes in patient management that are expected to have consequences for health outcomes
   - List key assumptions required to infer these changes in patient management will improve health outcomes (e.g. reduced harms through avoidance of invasive further tests, improved treatment selection)
   - Discuss the strength of these assumptions and the evidence they are based on
3. Reviewing patient management studies
   - Include studies that report patient management in sufficient detail: type and extent of management changes, contingent on test results
   - Use evidence of test accuracy to report whether the changes are likely to be based on correct positive or negative test results
4. Assessing the quality of included studies
   - Discuss the potential sources of bias of management studies, which include:
     - Reporting of planned (hypothetical) management versus actual management
     - Differences in patient characteristics between tested groups (selection bias)
     - Differences in the measurement of outcomes (detection bias)
     - Differences in the reporting of outcomes (reporting bias)
5. Drawing conclusions
   - Indicate whether the test is accurate, changes patient management and improves health outcomes
   - Indicate whether the test is recommended and state what evidence this conclusion is based on