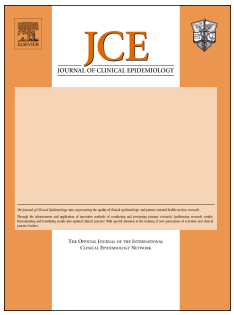


Methodological review reveals essential gaps and inconsistencies in clinical claims, effects and outcomes in HTA reviews of diagnostic tests

Jacqueline Dinnes, Clare Davenport, Isobel M. Harris, Lavinia Ferrante di Ruffano, Sue Mallett, Yemisi Takwoingi, Jonathan J. Deeks, Chris Hyde

PII: S0895-4356(25)00373-7

DOI: https://doi.org/10.1016/j.jclinepi.2025.112040

Reference: JCE 112040

To appear in: *Journal of Clinical Epidemiology*

Received Date: 6 June 2025

Revised Date: 26 September 2025

Accepted Date: 28 October 2025

Please cite this article as: Dinnes J, Davenport C, Harris IM, Ferrante di Ruffano L, Mallett S, Takwoingi Y, Deeks JJ, Hyde C, Methodological review reveals essential gaps and inconsistencies in clinical claims, effects and outcomes in HTA reviews of diagnostic tests, *Journal of Clinical Epidemiology* (2025), doi: https://doi.org/10.1016/j.jclinepi.2025.112040.

Methodological review reveals essential gaps and inconsistencies in clinical claims, effects and outcomes in HTA reviews of diagnostic tests

Authors:

Jacqueline Dinnes [a,b*] ORCID ID: 0000-0003-1343-7335; j.dinnes@bham.ac.uk

Clare Davenport [a,b] ORCID ID: 0000-0001-5659-4889; c.f.davenport@bham.ac.uk

Isobel M Harris [a] ORCID ID: 0000-0001-8125-3832; isobel.harris4@nhs.net

Lavinia Ferrante di Ruffano [a 1] ORCID ID: 0000-0002-2004-0638; lavinia.ferrante@york.ac.uk

Sue Mallett [c] ORCID ID: 0000-0002-0596-8200; sue.mallett@ucl.ac.uk

Yemisi Takwoingi [a,b] ORCID ID: 0000-0002-5828-9746; y.takwoingi@bham.ac.uk

Jonathan J Deeks [a,b 2] ORCID ID: 0000-0002-8850-1971; j.deeks@bham.ac.uk

Chris Hyde [d 2] ORCID ID: 0000-0002-7349-0616; c.j.hyde@exeter.ac.uk

a Department of Applied Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, UK

b NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK

c Centre for Medical Imaging, University College London, London, UK

d Exeter Test Group, University of Exeter Medical School, Exeter, UK

[1] present address: York Health Economics Consortium, University of York, York, UK

[2] denotes joint senior authorship

***Corresponding author:** j.dinnes@bham.ac.uk (J Dinnes)

Word count: 3843

Abstract

Background:

Essential first steps in performing a health technology assessment (HTA) for a diagnostic test include: consideration of the clinical pathway in which the test will be used, specifying the clinical claim for the test (how the test may add benefit, introduce harm or have other disadvantages beyond impact on the individual patient), and specifying the outcomes that would need to be measured to assess whether the test achieves its aims. We aimed to examine how a test evaluation framework (TEF) outlining the intended and unintended effects of tests could support the HTA process, and to identify additional ways in which tests add benefit or introduce harm.

Method:

We included 45 HTAs reporting 50 review questions. The study focused on HTA reports with a full English-language evidence review, clear methods and results sections, and evaluations of a single testing strategy or technology type. We looked for mechanisms of effect included in, and additional to, a TEF previously published by our group.

Results:

The clinical pathway and positioning of the new test were described in 98% of review questions (49/50), and illustrated in 62% (31/50). The test's clinical claims were easily identifiable in 56% (28/50). Claims, mechanisms of effect and pre-specified outcomes were frequently not coherent. For instance, at least one constituent test effect mechanism (mainly timing- and confidence-related mechanisms) could not be linked to pre-specified outcomes in 54% of reviews. Most HTAs (41, 82%) listed outcomes to be evaluated in the evidence reviews that we were unable to link to the claims for the tests (acceptability of the test, test failures, accuracy, therapeutic yield and effectiveness).

Four mechanisms of effect additional to those in the existing TEF were identified. Two concerned impact on individuals beyond the person being tested and two concerned organisational impact.

Conclusions:

Important gaps and inconsistencies in the reporting of test claims and associated outcomes in HTA reviews risks incomplete appraisal of a test's impact to patients and the healthcare system. We recommend tools are developed to support and standardise this complex process. This could be facilitated by tools in development and an expanded TEF.

Keywords: diagnostic tests; health technology assessment; value proposition; test effects; test accuracy; clinical pathway

Plain language summary

Health Technology Assessments (HTAs) are reports that assess whether medical tests, technologies, or treatments are worth using. HTA reports consider both effectiveness (how well something works) and costs. Tests can affect patients in many ways. It is important to understand the accuracy of a test as well as other impacts it may have on an individual's health and care. An important first step for an HTA of a test is to describe how the new test fits into regular medical care. A second step is to describe how the test could benefit, or harm, patients compared to regular care. HTA authors can then make sure that the result of using the test is measured using the right outcomes. We looked at 50 HTAs of medical tests that were written in English. We found that almost all reviews (98%) explained where and when the test would be used in patient care. About two-thirds of them (62%) included diagrams to show this. Only half clearly stated how the test could benefit, or harm, patients. There was often a disconnect between the expected benefits and harms from use of the test (clinical claim for the test) and the effects from the test that the authors of the review set out to assess. This could lead to the wrong conclusions about how well the test works and whether it's worth the cost. We also described broader effects of tests on healthcare systems and society. A more structured approach to help describe how a new test fits into regular medical care and to identify what a test claims to do is needed. This would help to ensure that all important outcomes are measured.

What is new

Key findings

- Application of a framework of accuracy and non-accuracy-based effects of tests to a set of diagnostic health technology assessments (HTAs) identified a broad range of test effects either in the clinical claim for the test (or the "value proposition") or as outcomes to be measured in the evidence review.
- Clinical pathways and positioning of the new test were well described, although were supported with pathway illustrations in less than two thirds of included HTAs; approximately half described the clinical claim for the test in an easily identifiable section.
- There was a frequent mismatch between test effect mechanisms identified in the clinical claim for the test and the outcomes considered in the evidence reviews. Some test effects were only identified in the test claims while others were missing from the test claims but measured as outcomes, raising concerns about potential bias in estimating the effectiveness and cost-effectiveness of diagnostic tests.

What this adds to what is known related to methods research within the field of clinical epidemiology

- Additional effects from tests, including broader health system and societal benefits, were observed that are not captured by the existing framework. Framework expansion is needed to capture test effects that operate above the individual level.

What is the implication, what should change now

- A structured tool to systematically identify test claims, distinguish outcome types, and better integrate clinical pathways would enhance transparency, reduce bias, and support the HTA evaluation process.

1 Background

The value of diagnostic tests ultimately lies in the degree to which their use impacts on health outcomes and health service delivery. Tests are components of a broader patient management strategy, with multiple diagnostic and therapeutic decision points, such that the ability of a test to add benefit or to cause harm goes beyond simple measures of its diagnostic accuracy. Evaluating the impact of a test requires us to consider the diagnostic pathway in which the test will be used, and to identify how the test might affect processes and decision points along that pathway in comparison to current practice. These essential steps enable evaluators to select outcomes that capture potential points of impact and assess whether introducing the test is clinically effective and safe.

We previously published a conceptual test effects framework (TEF) for assessing the value of diagnostic tests [1]. The TEF was developed by iteratively revising a preliminary theoretical model using a set of test–treatment randomised controlled trials (RCTs) to identify and explore how different test–treat strategies might impact on health outcomes. The resulting framework summarises the ways in which diagnostic tests can affect patient health outcomes, categorised under four main headings [1]: direct test effects on the patient, altering clinical decisions and actions, changing time-frames of decisions and actions, and influencing patient and clinician perceptions. The framework is intended to assist researchers designing new studies or compiling evidence generation pathways, as well as those evaluating existing evidence, for example when producing health technology assessments (HTA). The TEF may, however, be limited by its derivation exclusively from test-treatment RCTs. Such trials are methodologically challenging due to the complexity and multi-staged nature of the interventions evaluated [2, 3], are relatively scarce [4], and are mainly conducted in high income country settings. The expense and practical challenges of conducting test-treatment RCTs also means it is likely that the tests they evaluate are not representative of the full spectrum of available tests, while the outcomes they evaluate may be biased towards intermediate questions of process or short-term health impact as opposed to measures of long-term health or wider health service outcomes [2].

HTA of tests requires clear identification of the clinical claims for the test (also referred to as the 'value proposition' [5]) in order to inform which outcomes are needed to fully assess that clinical claim (i.e. to ensure that the introduction of the test leads to greater benefit than harm (or disbenefit) in comparison to the current standard of care) This is particularly critical when there is an absence of direct evidence for the health impact of the new technology and decision-makers frequently rely on linking evidence from multiple studies to evaluate the effectiveness and cost-effectiveness of tests [6, 7]. Nevertheless, a previous review of international HTA organisations' methods guidance documents identified inconsistencies in how claims for tests were identified and used to underpin subsequent HTA methods used [7].

To illustrate how the TEF could be used to support the HTA process, we aimed to identify whether the original TEF adequately captures the intended and unintended effects of diagnostic technologies typically evaluated in HTAs. We also explored whether there are additional ways in which tests add benefit, introduce harm or have other disadvantages beyond impact on the individual patient.

2 Methods

We identified completed HTAs of tests ('diagnostic HTAs') and examined the different ways in which changing a test or introducing a new test strategy were considered to impact patient health, healthcare delivery or health service organisation. Our objectives were to:

1. obtain empirical evidence of the frequency of different test effects, both intended and unintended,

2. report whether these test effects were considered in the HTAs' claims for the tests and/or the outcomes measured,
3. identify effects not included in the original TEF.

2.1 Data source

We collated publicly available documents supporting diagnostic HTAs from the websites of seven HTA organisations. These organisations were previously identified as having the most well–developed methods for conducting diagnostic HTAs [7] (Fig 1). All document types whose aim was to review the evidence for a medical test were collated, including scoping documents, protocols, systematic and non–systematic reviews, rapid reviews, economic models and guidance documents.

2.2 Selection of diagnostic HTAs

Eligible HTAs were published between 2010 and 2020 and could evaluate any medical test used to detect a condition in a suspected population (diagnosis or targeted screening) or identify the stage of known disease (staging). HTAs accompanied by full, stand-alone, English-language evidence reviews (i.e. with dedicated Methods and Results presented separately from any additional evidence retrieved for health economic model parameters) were selected for assessment (referred to throughout as 'HTAs' or 'HTA reviews').

We aimed to include between 50 and 100 review questions, which was our unit of analysis. We purposively sampled HTAs until we reached a point of saturation in terms of diversity of test technologies, types of test comparison and application to disease type and clinical setting.

2.3 Data extraction and synthesis

HTA review characteristics were extracted from the Background or Methods sections of each review.

For each evidence review, two researchers examined it to:

1. map out the patient management strategies being compared,
2. identify the clinical claim for the evaluated test strategies,
3. identify the pre-specified outcomes,
4. use the original TEF to identify all likely test effects, represented either in the claim for the test and/or by the outcomes to be measured for the evidence review,
5. identify additional ways in which the evaluated tests might impact on outcomes at the level of the individual patient, or at the level of healthcare delivery or health service organisation.

Extraction of full review characteristics was performed by one researcher and checked independently by a second. All extractions were discussed by at least three researchers, to ensure agreement.

Test effect mechanisms from the TEF were categorised thematically into five groups outlined in Box 1. Mechanisms were tabulated and summarised according to whether they were identified in the clinical claim for the test, as outcomes to be measured or both. Mechanisms were further grouped according to technology type. Any mechanisms or impacts from a test's introduction that were not represented by the TEF were noted. All data are presented descriptively.

Box 1 Thematic categorisation of test effect mechanisms

i. impact: patients impacted via diagnostic decisions, actions and health outcomes (i.e. test accuracy, diagnostic and therapeutic yield, effectiveness of test-treat strategy, adherence to treatment),

ii.	feasibility and interpretation of the test: acceptability of, and contraindications to, the test, test failures and ease of interpretation,
iii.	safety: procedural harms or benefits,
iv.	timing: changing timeframes of testing, decisions and actions (i.e. time to test delivery and test result, time to diagnosis and treatment), and
v.	confidence: influence on patient and clinician perceptions and confidence in decisions and actions, whether diagnostic or therapeutic.

3 Results

Of the 1837 documents identified, 105 reported eligible test evaluations that were accompanied by full English-language evidence reviews (Figure 1). The majority of eligible HTAs (102/105) were published by three organisations (AHRQ, MSAC and NICE), two of which (MSAC and NICE) provided more clearly focused review questions (evaluating a single testing strategy or technology type, usually for a single target condition) and were prioritised for extraction (n=59). Forty-five of the 59 HTA reviews, reporting 50 separate review questions were included up to the point of saturation; 18 (40%) were conducted for MSAC (reporting 21 review questions) and 27 (60%) for NICE (reporting 29 review questions). A list of all included HTA reviews is provided in Supplementary Appendix 1. All results are described with a denominator of 50 review questions.

Figure 1: PRISMA style figure documenting HTAs per organisation and process of inclusion

Footnote to Figure 1

AHRQ *- Agency for Healthcare Research and Quality;* ***CADTH*** *- Canadian Agency for Drugs and Technologies in Health;* ***ER*** *– evidence review;* ***HTA*** *– health technology assessment;* ***IQWiG*** *- Institute for Quality and Efficiency in Health Care;* ***MSAC*** *- Medical Services Advisory Committee;* ***NICE*** *- National Institute for Health and Care Excellence;* ***SBU*** *- Swedish Agency for Health Technology Assessment and Assessment of Social Services;* ***ZIN*** *- National Health Care Institute.*

3.1 Topic areas, test roles and test comparisons

As summarised in Table 1, included reviews covered a representative range of topics and test roles, with possible under-representation of triage comparisons (8/50, 4%). Comparisons were generally of single tests (32, 64%), rather than multiple. For three quarters of review questions (38, 76%) the index technology was compared with the same type of technology. For example, for 18 of the 24 (75%) evaluations of in-vitro devices (IVDs) the comparator test was also an IVD. Similarly, 14 of the 18 imaging review questions compared index and comparator tests in the same imaging category (8/10 radiological imaging evaluations, 3/3 endoscopic test evaluations and 3/5 optical imaging assessments).

Table 1: Summary of topic areas and test comparisons

Characteristic		Subgroup							Identified n (%)
Topic area		Cancer							13 (26)
		Cardiovascular							6 (12)
		Gastrointestinal							7 (14)
		Infection / infectious diseases							6 (12)
		Genetic mutations							5 (10)
		Obstetrics or Gynaecology							4 (8)
		Other*							9 (18)
Setting		Primary only							5 (10)
		Secondary only							34 (68)
		Secondary or tertiary							5 (10)
		Multiple (including primary or community settings)							6 (12)
Test comparison		Single index versus single comparator							32 (64)
		(Multiple) index vs (Multiple) comparator							18 (36)
Role of test		Add-on							14 (28)
		Replacement							17 (34)
		Triage							4 (8)
		More than one possible role							15 (30)
Comparison of testing strategies		**Comparator strategies**							
Index tests ↓		**Clinical**	**IVD**	**Radiological imaging**	**Endoscopy**	**Optical**	**Physiological**	**Clinical + IVD + other test type**	**Total n (%)**
IVD		1	18					5	**24 (48)**
Imaging	**Radiological**		1	8	1				**10 (20)**
	Endoscopic				3				**3 (6)**
	Optical		1		1	3			**5 (10)**
Physiological			1				5	1	**7 (14)**
Clinical + IVD ± imaging								1	**1 (2)**
Total n (%)		**1 (2)**	**21 (42)**	**8 (16)**	**5 (10)**	**3 (6)**	**5 (10)**	**7 (14)**	

IVD – in vitro diagnostic; **±** - with or without

*Other topics include allergy (1), diabetes (2), haematology (1), hepatology (1), renal (2), respiratory (1), sleep apnoea (1).

Table 2: Clinical pathway description and claim for the test

	Category	Total (N=50) n (%)
Reporting of clinical pathway	Diagram provided	31 (62)
	Described using text only	18 (36)
	Not documented	1 (2)
Reporting of claim for the test	Clearly labelled, dedicated section	28 (56)
	Description of technology / Index test section	14 (28)
	Multiple sections or not well reported	8 (16)
Coherence of claim for test	All mechanisms in the claim for the test were linked to pre-specified outcomes*	23 (46)
	Mechanisms identified in the claim for the test were not linked to pre-specified outcomes	27 (54)
Outcomes	All pre-specified outcomes were identifiable in the claim for the test	19 (18)
	Additional outcomes were pre-specified that were not included in the claim for the test	41 (82)

*mechanisms here are those from the original test evaluation framework

3.2 Clinical pathway and clinical claims for the tests

The clinical pathway, and the proposed position of the new test within it, was described in 98% of review questions (49/50), and illustrated figuratively in 62% (31/50) (Table 2). Around half (28, 56%) stipulated the clinical claim for the test in a clearly labelled dedicated section, while 28% (14/50) included the test claim as part of the description of the technology. For eight (16%) review questions, the test claim was less easily identified and was reported in multiple sections of the background.

3.3 Coherence of test claims and outcomes measured by the evidence reviews

In approximately half of the review questions (23, 46%), all test effect mechanisms that were identified in the claims for the test could be linked to measurable outcomes specified in the evidence review (Table 2). Mechanisms identified in the clinical claim for the test that were not linked to outcomes assessed by the HTA reviews (27, 54%) included: acceptability of the test, confidence-related mechanisms and timing-related effects.

The majority of HTAs (41, 82%) also listed outcomes to be assessed by evidence reviews that were not represented in the claim for the test. Examples included: acceptability of the test, test failures, accuracy, therapeutic yield and effectiveness.

3.4 Frequency of test effect mechanisms grouped according to themes

Figure 2 shows the frequency of test effect mechanisms organised by theme and according to whether and how they were captured in the HTA reviews. The data underlying Figure 2 are reported in Supplementary Appendix 2.

Figure 2: Distribution of mechanisms identified in clinical claims for the test, as measurable outcomes, or both

Footnote for Figure 2:

Dx – diagnosis; **Rx** - treatment
"n" represents the number reporting at least one mechanism from each group, e.g "CONFIDENCE (n=14)" indicates that at least one confidence-related mechanism was identified for 14 of the 50 review questions

Mechanisms directly related to the 'impact' theme were the most commonly considered across review questions either in the test claim or as outcomes to be measured. Accuracy (49/50), therapeutic yield (treatment choices) (42/50) and effectiveness of a test-treat strategy (47/50) mechanisms were particularly well captured. In contrast, adherence to treatment as a result of using a test was identified in only six evidence review questions. Notably, accuracy and effectiveness mechanisms were often captured only as outcomes to be measured by the evidence reviews (29 of 49 instances for accuracy and 19 of 47 instances for effectiveness), and were not identified *a priori* as a claim of the test being evaluated.

Mechanisms within the 'feasibility of conducting or interpreting the test' theme were identified in a total of 37/50 (74%) review questions. Two mechanisms within this theme (acceptability of the test and test failure rates) were responsible for most of these occurrences (Figure 2). The test failure mechanism primarily appeared only in the outcomes to be assessed by the HTA reviews (21 of the 24 instances where it was identified), and was infrequently identified in the clinical claim for the test (3/24). The acceptability mechanism was more often identified in the claim for the test (10 of the 21 instances) but appeared only as an outcome to be measured in 11 reviews (52% of the 21 instances).

Clinical contraindications and ease of interpretation were each considered in fewer than five evidence review questions.

Procedural harms were identified as a mechanism in 28 reviews. This mechanism was frequently picked up as part of the claim for the test (68%, 19/28), and for 16 of those 19 instances was also considered as an outcome to be measured.

'Timing' was considered by 70% (35/50) of review questions, with individual timing mechanisms identified in between eight (time to test delivery) and 23 (speed of diagnosis and time to treatment) review questions (Figure 2). Where timing mechanisms were identified as part of the clinical claims, they were also included as outcome measures on at least half of occasions.

Mechanisms related to confidence in the diagnostic decision or treatment choices were the least frequently considered by review questions (individual mechanisms identified in between two and nine reviews). 'Confidence' was usually considered as part of the claim of the test (one to six reviews) as opposed to in both the claim and the outcome (one to two reviews per mechanism) (Figure 2).

3.5 Additional effects of test introduction not originally covered by the TEF

We identified four test effects not covered by the TEF, illustrated with detailed examples in Table 3. Two effects describe impact on individuals beyond the person being tested: scenarios where tests allow a broader population to be tested; and where tests provide benefit to the wider population. Examples include increased access to testing for underserved populations who do not typically routinely access health services (e.g. rapid self-tests for HIV [8]), or where faster or more accurate diagnosis of infection has potential to reduce the risk of onward transmission of infection beyond the patient (e.g. rapid tests for TB [9]) or to confer societal value in terms of antimicrobial stewardship and antimicrobial prescribing decisions [10, 11].

Impact from changes in healthcare organisation and delivery include efficiency gains, either within a testing pathway or at the wider health service organisational level. Examples include a single test replacing several standard of care tests for diagnosis of one or more target conditions (e.g. [12] [9] [13]), or single tests allowing simultaneous diagnosis and staging of a condition (e.g. [14]) (Table 3).

Efficiency gains at the wider health service organisation level can be made by introducing a test that (i) reduces hospital admissions (e.g. [15]) or referrals (e.g. [16]) or (ii) allows more efficient use of healthcare facilities (e.g. [17]). A third route for efficiency gains occurs with the introduction of more specialist tests that may have additional training requirements but reduce the use of health service resources (e.g. [18]).

Table 3: Additional effects from tests that were not included in the original test evaluation framework (TEF)

Additional effects identified (n HTAs; n by technology type)	**Explanatory text**	**Examples from set of HTAs**
i. Enable a broader or different population to be tested (n=5; IVD 3; imaging 1, other 1)	• the test is more acceptable so more people consent to or attend for testing • the test has fewer contraindications than currently available tests • the test can be used in more settings (e.g. more transportable (does not require specialist equipment (is easier to use (requires less training • the test is used in the same setting but at a different point in the pathway	• Point-of-care antigen/antibody test vs Western blot for HIV benefits high-risk or hard-to-reach populations resistant to conventional (non point-of-care) testing [8] • New generation CT for cardiac imaging [18] broadens population (obese (high / irregular heart beat (high levels coronary calcium (previous stent or bypass) that can be imaged outside of specialist centres (e.g. those contraindicated for 64-slice CT
ii. Benefit to wider population (beyond the individual tested (n=8; IVD 8)	• antimicrobial stewardship • minimising onward transmission of disease.	• Gastro pathogen panel for gastroenteritis [17] • Point of care tests for HIV [8] or for tuberculosis [9] • Rapid tests for sepsis [10] or streptococcus A [11]
iii. Whole pathway (efficiency) of testing or treatment strategies (n=16; IVD 7, imaging 4, other 5)	• test replaces several current care tests for diagnosis of a single target condition • test replaces several current care tests for diagnosis of multiple target conditions at the same time (e.g. testing for multiple causative agents (multi-cancer early detection (MCED) tests (or whole genome sequencing for rare diseases • test replaces a pathway of tests for diagnosis and staging and/or treatment planning for a single target condition	• A single whole body 68Ga-DOTA-peptide PET-CT scan to replace >1 111Inoctreotide SPECT/CT for gastroenteropancreatic neuroendocrine tumours allows detection of diffuse disease and reduces the amount of repeated testing needed over 2 days [12] • HbA1c test to replace the random blood glucose or fasting blood glucose test and the oral glucose tolerance test [19] • Gastro pathogen panel for gastroenteritis allows detection of multiple infections at same time [17] • Contrast–enhanced US using SonoVue® allows characterisation of focal liver lesions and detection of liver metastases (allowing some CT and MRI examinations (for definitive staging) to be avoided [14]

iv. Health service effects (n=16; IVD 7, imaging 2, other 7)	• the test requires fewer visits or healthcare interactions to obtain a diagnosis or make a treatment decision • the test avoids admissions or referrals for further investigations (e.g. less expensive imaging (or fewer biopsies) • the test had additional training requirements for healthcare professionals	• High-sensitivity troponin assays for ruling out AMI avoids hospital admission or could allow earlier discharge [15] • Hand-held nitric oxide measurement for treatable (eosinophilic asthma) could reduce referrals to secondary care [16] • Gastro pathogen panel for gastroenteritis [17]: "shorter turnaround times of the tests may improve efficient use of isolation bays and allow people to be treated on open bays when infectious pathogens are not present" • Virtual chromoendoscopy added to conventional endoscopy for colorectal polyps [18] increases training costs but with benefit of fewer polyp resections with possible associated reduction in complications (with concomitant reduction in histopathology use

CT – computed tomography; HbA1c - glycated haemoglobin; HIV – human immunodeficiency virus; IVD – in vitro diagnostic; MRI – magnetic resonance imaging; PET – positron emission tomography; SPECT - single photon emission computed tomography

[a] 'current care test' may include test of treatment

[b] additional example not from our included set of HTAs

4 Discussion

We provide empirical evidence for the types and frequencies of test effect mechanisms and the linkage between claims made for a test and outcomes measured across different test technologies in HTA evidence reviews. We have illustrated that the full breadth of test effect mechanisms identified from a set of test-treat RCTs, including accuracy and non-accuracy effects, are also observed in diagnostic HTAs. We have also highlighted additional features of test effects beyond the individual patient that may be important to incorporate into an updated version of our TEF.

Identifying the mechanisms by which a new test might have impact, and consequently which outcomes to measure, relies on explicit consideration of clinical pathways. We identified the need for clearer reporting of clinical pathways and clinical claims for tests; a minority (38%) of included review questions did not provide an illustrative diagram and only 56% had a dedicated, clearly labelled test claims section. We identified a tendency for HTAs to report outcomes that did not appear to be explicitly related to the claims for the tests being evaluated. Although the rationale for including outcomes that are essential parameters for a decision model can often be inferred (e.g. accuracy and effectiveness were almost universally included), this is not necessarily the case for other outcomes (e.g. those related to acceptability of a test or to procedural harms). A more explicit differentiation between structural outcomes and those related to specific test claims would clarify, for example, whether accuracy was included as a key parameter for decision modelling or because it might be either compromised or enhanced by the introduction of the test being evaluated. We did not find any such reflections in the HTAs we examined.

A more pressing concern was the frequent observation that test effect mechanisms identified in the HTAs' claims for the test could not be linked to pre-specified outcomes to be considered by the evidence reviews (occurring for 54% of review questions). Mechanisms that were not linked to outcomes for evidence reviews included those that might be considered more difficult to measure within a typical test accuracy or effectiveness study, such as the acceptability of the test, timing- or confidence-related mechanisms. The failure to translate claims into outcomes could mean that important potential benefits (or disbenefits) are not represented in the assessment of overall effectiveness and cost-effectiveness. Concerning outcomes not suggested by claims there is a risk that important proximal effects are obscured by absence of effects on longer term outcomes where an effect is only distantly plausible or too small to measure.

We observed that index technologies were almost always directly compared to the same type of technology, e.g. a new IVD compared to an existing IVD-based strategy. Over-simplification of the comparison between new and existing diagnostic strategies could occur either because of difficulties in delineating the current diagnostic standard of care contributing to comparisons that are not necessarily reflective of clinical practice [20], or because difficulties in undertaking evaluations of tests that disrupt standard of care clinical pathways (due to the introduction of a different testing modality) has led to a paucity of evidence to support a more clinically relevant comparison. Recent guidance for identifying and operationalizing the diagnostic standard of care comparator in economic models [21] may help counteract this tendency.

4.1 Strengths and limitations

We used systematic and structured methods to identify, assess and extract data from eligible HTA reviews for this methodological analysis, documented in a pre-specified protocol. All evidence reviews were assessed by two researchers and roundtable discussion of all extractions by at least three researchers ensured consistency. Despite using a structured framework to aid the systematic identification of test effects in HTA evidence reviews, the process was challenging. A lack of

consistency in reporting of the clinical claims for tests, and in the selection of outcomes to assess those claims, may have resulted in us missing relevant information. The results may not be representative of HTAs from other agencies with broader review questions, or to agencies producing non-English language reports, or, potentially, to HTAs published within the last five years. We did not identify any evaluations of AI-based tools, for example, and were unable to compare mechanisms by technology type due to small numbers of HTAs for some technology types. We did note that timing-related mechanisms, for example, were more frequently identified for IVDs, primarily due to the replacement of an existing laboratory-based test with a point-of-care device.

The protocol for this project was not registered and we have not used an open access registry for the data collected. We also could not identify a relevant reporting guideline, but have reported our Methods as transparently as possible.

The scope of our research did not include assessing the types of study used to evaluate the reported outcomes. We note that reports typically focus search efforts on identifying studies evaluating diagnostic test accuracy and RCTs of test-treat strategies or treatment effectiveness. Any limitations on study designs in our sample may have impacted the range of mechanisms observed, however such designs can be used or adapted to capture the mechanisms of interest. We were also unable to evaluate the potential impact from an HTA methods manual, which provides instructions on explicit identification and prioritisation of which outcomes to evaluate (published in 2021) [22].

4.2 Recommendations for future research

More widespread use of visual representation of pathways and use of a causal pathway approach [22], is likely to help elicit the purported added benefit from introducing a new test into a testing pathway [7]. The universal lack of comment on the process by which the claims of tests were elucidated was notable. Identifying both pathway and claim are fundamental first steps in the HTA process, therefore there may be a need for a tool to support a standardised approach for this process. Such a tool should aim to reduce the mismatch between the mechanisms of a test's clinical claim and the outcomes measured. This is particularly true for more complex questions that include multiple index or comparator tests, potentially with different associated benefits.

The tool should also help to distinguish between outcomes which flow from the clinical claim for the test, those needed to parameterise a decision model, and those which are added for "safety" or exclusion of possible harm or disbenefit. For example, the claim for a point-of-care test may centre on reducing time to diagnosis, however if there are concerns that accuracy could be compromised in comparison to current care tests, accuracy will be evaluated as an aspect of possible harm. For mechanisms that were less commonly identified in the claims for the tests, such as diagnostic confidence, it is important to understand whether these are genuinely less frequent components of test claims, are simply overlooked, or are excluded because of anticipated difficulties in their measurement. Our identification of mechanisms not represented by outcomes is concerning as this represents a potential bias resulting in inaccurate estimation of the impact of new tests.

The challenges of eliciting and combining clinical claims across several tests being compared in an HTA should be acknowledged, and points to greater complexity than might initially be anticipated [7]. Care pathway analysis [23] or process mapping [24] with input from a range of stakeholders [25], could be a useful approach to articulate both the pathway and clinical claims for a new test strategy. Recent work to develop a step-by-step guide to designing a test-management pathway for guideline developers [26], alongside the TEF [1] could provide a useful starting point. Such a tool would also help improve the transparency, standardisation and reporting of this critical HTA step, and could be incorporated into available methods guidance from HTA agencies who we would involve in the

development and piloting. The use of a formal tool could also encourage researchers to return to the claims made for the test in the light of the evidence identified, to identify gaps in the evidence base and formulate clear and useful recommendations for further research.

Finally, there is scope for the original TEF [1] to be adapted as a tool to facilitate more effective identification of test effect mechanisms. Any further adaption of the TEF (e.g. development of a shiny app or other web-based tool) should accommodate additional test effects that have a wider impact beyond the individual being tested. Test effects that operate above the individual level are not uncommon for HTA reviews that are commissioned at a regional or national level, which naturally take a wider health service perspective compared to RCTs whose main purpose is usually to answer a clinically defined question. A future adaption of the TEF could provide a useful means for HTA agencies to directly address the identification of test effect mechanisms and outcomes. A further individual level 'impact' mechanism that also deserves future consideration is 'the avoidance of treatment in those who would be harmed by it', such as a genetic test to identify a variant that causes deafness in babies if they are treated with a particular type of antibiotic [27]. No examples of this mechanism were identified in our sample of HTAs, however this is an active area of research in the field of pharmacogenetics that is not adequately covered by the current TEF.

5 Conclusions

We identified important gaps and inconsistencies in the reporting of test claims and associated outcomes in HTA reviews. Since this step frames the HTA and drives both evidence review and health economic modelling, we recommend tools are developed to support this often complex process. Under-identification of all relevant mechanisms may lead to incomplete appraisal of a test's impact to patients and the healthcare system. Foremost is a tool to provide a consistent and standardised approach for identifying all mechanisms that constitute a test's claim, so that these can translate directly to outcomes. This tool should be built on an existing framework, such as the TEF, for which we have identified additional mechanisms.

Additional information

Declaration of competing interests

The authors declare none. Lavinia Ferrante di Ruffano has received direct support by commercial companies for methods work in related topics.

Funding statement

This work was funded by the Medical Research Council (grant number MR/T025328/1). JD, YT and JJD are funded by the National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre (BRC).

Author contributions

LFR, JJD, CH, SM, CD and YT conceived and designed the study and acquired study funding. JD coordinated the review, abstracted data, interpreted results, and wrote the original draft manuscript. CD abstracted data, interpreted results and reviewed and edited the manuscript. IMH screened citations, abstracted data, interpreted results, and reviewed and edited the manuscript. LFR coordinated the review, screened citations, abstracted data, interpreted results, and reviewed

and edited the manuscript. SM, YT, JJD and CH interpreted results and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments
We would like to thank the MRC TEST Advisory Group for discussion of the methods approach: Pauline Beattie, Patrick Bossuyt, Katherine Payne, Samuel Schumacher, Bethany Shinkins and Thomas Walker

Data sharing statement
No new data have been created in the preparation of this article and therefore there is nothing available for access and further sharing. All queries should be submitted to the corresponding author.

Ethics statement
This report concerns secondary research, for which ethics approval is not required.

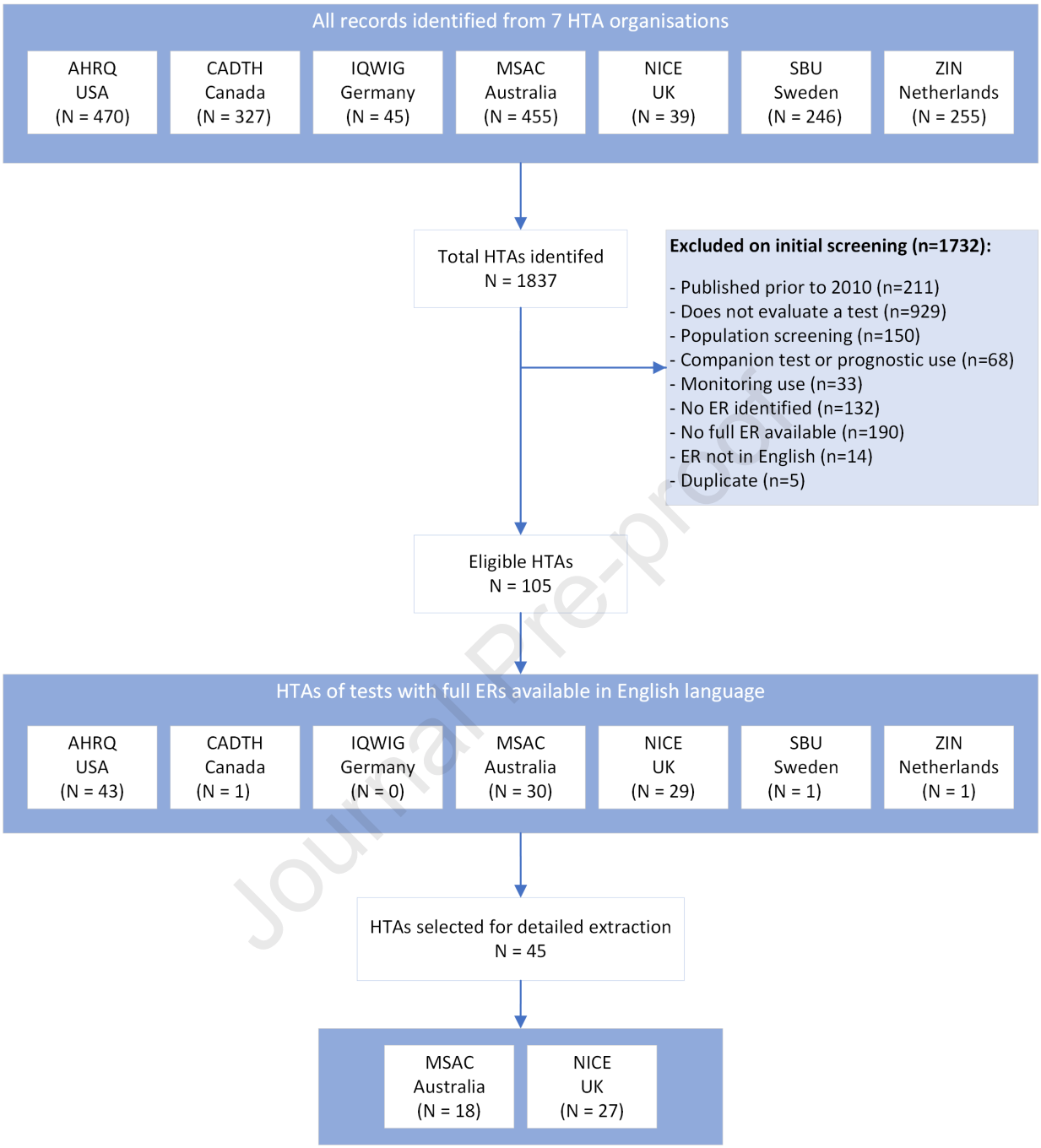
Information governance statement
This project did not involve the handling of personal information.

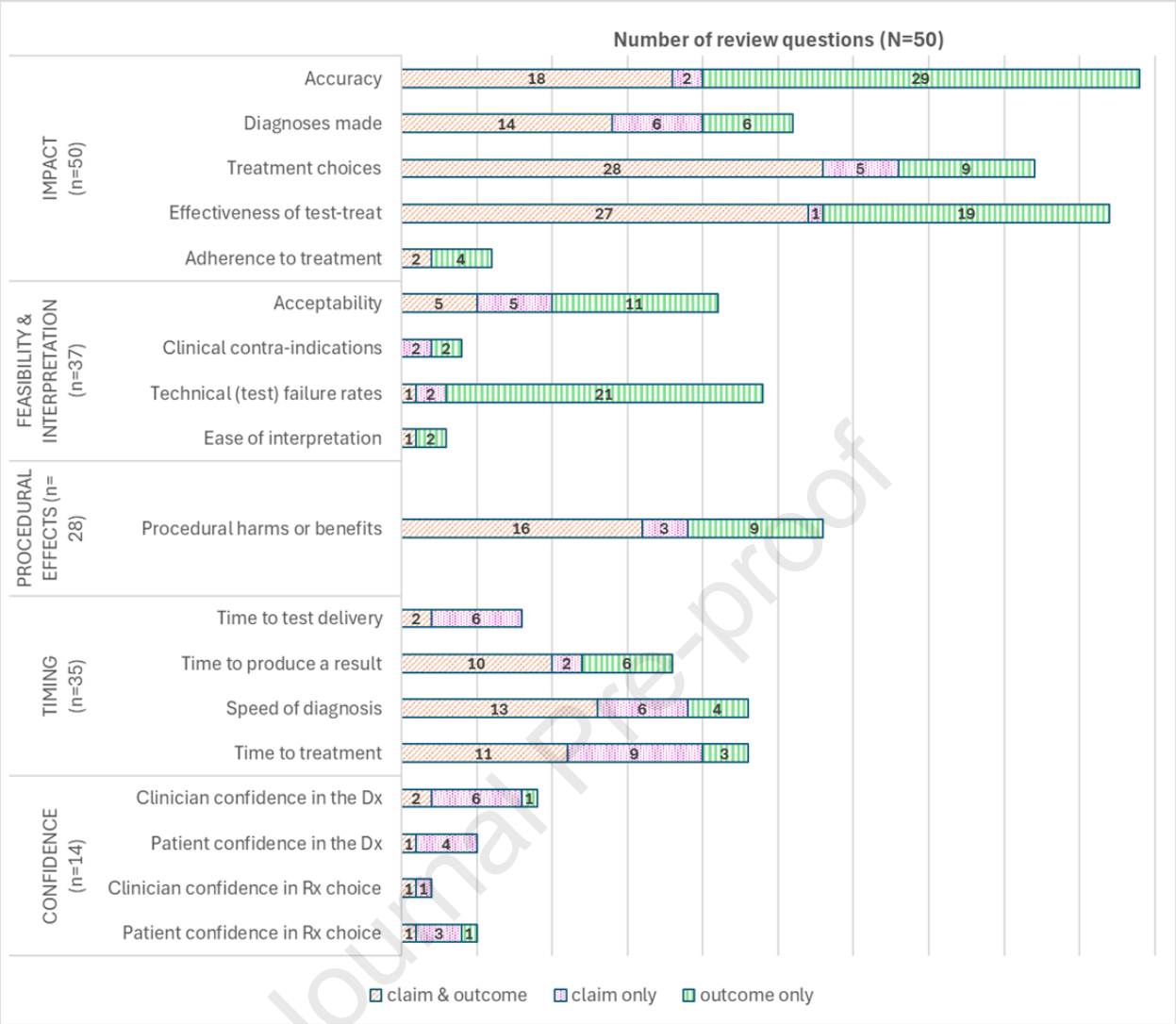
Disclaimer
This publication presents independent research commissioned by the Medical Research Council (MRC) and delivered though the National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre (BRC). The views and opinions expressed in this publication are those of the authors and not necessarily those of the MRC, the NIHR, or the Department of Health and Social Care.

References:

[1] Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ. 2012;344:e686.
[2] Ferrante di Ruffano L, Dinnes J, Sitch AJ, Hyde C, Deeks JJ. Test-treatment RCTs are susceptible to bias: a review of the methodological quality of randomized trials that evaluate diagnostic tests. BMC Med Res Methodol. 2017;17:35.
[3] Ferrante di Ruffano L, Dinnes J, Taylor-Phillips S, Davenport C, Hyde C, Deeks JJ. Research waste in diagnostic trials: a methods review evaluating the reporting of test-treatment interventions. BMC Med Res Methodol. 2017;17:32.
[4] Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. Journal of Clinical Epidemiology. 2012;65:282-7.
[5] Price CP, St John A. Anatomy of a value proposition for laboratory medicine. Clin Chim Acta. 2014;436:104-11.
[6] Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of Meta-analysis and Economic Decision Modeling for Evaluating Diagnostic Tests. Medical Decision Making. 2008;28:650-67.
[7] Ferrante di Ruffano L, Harris IM, Zhelev Z, Davenport C, Mallett S, Peters J, et al. Health technology assessment of diagnostic tests: a state of the art review of methods guidance from international organizations. International Journal of Technology Assessment in Health Care. 2023;39:e14.
[8] Ghijben P, Zavarsek S, Yong K, Ip F. Rapid point-of-care combined Antigen/Antibody HIV test to aid in the diagnosis of HIV infection, MSAC Application 1391, Assessment Report. Commonwealth of Australia, Canberra, ACT: Medical Services Advisory Committee (MSAC); 2015.
[9] Morona JK, Vogan A, Kessels S, Gum D, Milverton J, Parsons J, et al. An assessment of nucleic acid amplification testing for active mycobacterial infection. MSAC application no. 1234, Assessment Report. Commonwealth of Australia, Canberra, ACT: Medical Services Advisory Committee (MSAC); 2014.
[10] Stevenson M, Pandor A, Martyn-St James M, Rafia R, Uttley L, Stevens J, et al. Sepsis: the LightCycler SeptiFast Test MGRADE®, SepsiTest™ and IRIDICA BAC BSI assay for rapidly identifying bloodstream bacteria and fungi – a systematic review and economic evaluation. England, United Kingdom: NIHR Health Technology Assessment programme; 2016.
[11] Fraser H, Gallacher D, Achana F, Court R, Taylor-Phillips S, Nduka S, et al. Rapid Tests for Group A Streptococcal infections in people with sore throat. Diagnostic Assessment Report commissioned by the NIHR HTA Programme on behalf of the National Institute for Health and Care Excellence. England, United Kingdom: Warwick Evidence, University of Warwick; 2020.
[12] Morona JK, Mittal R. Substitution of 68Ga-DOTA-peptide PET/CT scanning in lieu of Octreotide for patients undergoing somatostatin receptor diagnostic imaging under MBS item 61369. Mini-assessment Report. Commonwealth of Australia, Canberra, ACT: Medical Service Advisory Committee (MSAC); 2017.
[13] Westwood M, Ramaekers B, Lang S, Armstrong N, Noake C, de Kock S, et al. ImmunoCAP® ISAC and Microtest for multiplex allergen testing in people with difficult to manage allergic disease: a systematic review and cost-effectiveness analysis. A Diagnostic Assessment Report. . England, United Kingdom: Kleijnen Systematic Reviews Ltd; 2015.
[14] Westwood ME, Joore MA, Grutters JPC, Redekop WK, Armstrong N, Lee K, et al. Contrast enhanced ultrasound of the liver using SonoVue®(sulphur hexafluoride microbubbles): a Diagnostic Assessment Report. . England, United Kingdom: Kleijnen Systematic Reviews Ltd.; 2012.
[15] Westwood M, van Asselt T, Ramaekers B, Whiting P, Thokala P, Joore M, et al. High-sensitivity troponin assays for the early rule-out or diagnosis of acute myocardial infarction in people with acute chest pain: a systematic review and cost-effectiveness analysis. A Diagnostic Assessment Report. England, United Kingdom: Kleijnen Systematic Reviews Ltd; 2014.

[16] Harnan SE, Tappenden P, Essat M, Gomersall T, Minton J, Wong R, et al. Measurement of exhaled nitric oxide concentration in asthma; NIOX MINO and NObreath. Diagnostic Assessment Report. England, United Kingdom: ScHARR, University of Sheffield; 2015.
[17] Freeman K, Mistry H, Tsertsvadze A, Royle P, McCarthy N, Taylor-Phillips S, et al. Multiplex tests to identify gastrointestinal bacteria, viruses and parasites in people with suspected infectious gastroenteritis: a systematic review and economic analysis. Diagnostic Assessment Report. England, United Kingdom: Warwick Evidence, University of Warwick; 2017.
[18] Picot J, Rose M, Cooper K, Pickett K, Lord J, Harris P, et al. Virtual chromoendoscopy for the real-time assessment of colorectal polyps in vivo: a systematic review and economic evaluation. England, United Kingdom: Southampton Health Technology Assessments Centre (SHTAC); 2016.
[19] Parsons J, Vogan A, Morona J, Schubert C, Merlin T. HbA1c test for the diagnosis of diabetes mellitus. In: HbA1c testing in the diagnosis of diabetes mellitus. MSAC Application 1267 AR, editor. Commonwealth of Australia, Canberra, ACT: Medical Services Advisory Committee (MSAC); 2014.
[20] Gopalakrishna G, Leeflang MMG, Davenport C, Sanabria AJ, Alonso-Coello P, McCaffery K, et al. Barriers to making recommendations about medical tests: a qualitative study of European guideline developers. BMJ Open. 2016;6:e010549.
[21] Graziadio S, Gregg E, Allen AJ, Neveux P, Monz BU, Davenport C, et al. Is the Comparator in Your Diagnostic Cost-Effectiveness Model "Standard of Care"? Recommendations from Literature Reviews and Expert Interviews on How to Identify and Operationalize It. Value Health. 2024;27:585-97.
[22] Medical Services Advisory Committee. Guidelines for preparing assessments for the Medical Services Advisory Committee. [Online]. Version 1.0. 2021.
[23] Graziadio S, Winter A, Lendrem BC, Suklan J, Jones WS, Urwin SG, et al. How to Ease the Pain of Taking a Diagnostic Point of Care Test to the Market: A Framework for Evidence Development. Micromachines. 2020;11:291.
[24] St John A, O'Kane M, Christenson R, Julicher P, Oellerich M, Price CP. Implementation of medical tests in a Value-Based healthcare environment: A framework for delivering value. Clin Chim Acta. 2021;521:90-6.
[25] Korte BJ, Rompalo A, Manabe YC, Gaydos CA. Overcoming Challenges with the Adoption of Point-of-Care Testing: From Technology Push and Clinical Needs to Value Propositions. Point of Care. 2020;19:77-83.
[26] Tuut MK, Gopalakrishna G, Leeflang MM, Bossuyt PM, van der Weijden T, Burgers JS, et al. Co-creation of a step-by-step guide for specifying the test-management pathway to formulate focused guideline questions about healthcare related tests. BMC Medical Research Methodology. 2024;24:241.
[27] Genedrive MT-RNR1 ID Kit for detecting a genetic variant to guide antibiotic use and prevent hearing loss in babies: early value assessment. Health technology evaluation; HTE6: National Institute for Health and Care Excellence; 2023.





Highlights

- Evaluated test effects in a set of diagnostic health technology assessments
- A range of accuracy and non-accuracy effects were identified.
- A mismatch between clinical claims for tests and outcomes measured was observed.
- Broader health system and societal benefits were also identified.
- Tools needed to systematically define clinical pathways, test claims, and outcomes

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jonathan Deeks reports financial support was provided by UK Research and Innovation Medical Research Council. Lavinia Ferrante di Ruffano reports a relationship with York Health Economics Consortium Ltd that includes: employment. Jacqueline Dinnes reports a relationship with NIHR Birmingham Biomedical Research Centre that includes: funding grants. Jonathan Deeks reports a relationship with NIHR Birmingham Biomedical Research Centre that includes: funding grants. Yemisi Takwoingi reports a relationship with NIHR Birmingham Biomedical Research Centre that includes: funding grants. Yemisi Takwoingi reports a relationship with UK Research and Innovation Medical Research Council that includes: funding grants. Christopher Hyde reports a relationship with UK Research and Innovation Medical Research Council that includes: funding grants. Jonathan Deeks reports a relationship with UK Research and Innovation Medical Research Council that includes: funding grants. Clare Davenport reports a relationship with UK Research and Innovation Medical Research Council that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.