Supplementary Material

Title: Recommendations for overcoming methodological challenges to health economic modelling that arise when comparing in vitro diagnostics with imaging tests

Authors: Karina Watts MBiolSci[1], Osvaldo Ulises Garay MSc[2], Lavinia Ferrante di Ruffano PhD[1], Phillip Kittel MScl[2], Emily Gregg PhD[1], Vincent Heller MSc[2], Deborah Watkins MSc[1], A. Joy Allen PhD[3], Bethany Shinkins DPhil[4], Sue Mallet DPhil[5], Stuart Mealing MSc[1], Hayden Holmes PGDip[1*]

[1] York Health Economics Consortium (YHEC), University of York, York, United Kingdom

[2] Roche Diagnostics International, Rotkreuz, Switzerland

[3] Roche Diagnostics Limited, West Sussex, United Kingdom

[4] Warwick Applied Health, University of Warwick, Coventry, United Kingdom

[5] Centre for Medical Imaging, University College London, London, United Kingdom

***Corresponding author:**

Hayden Holmes

York Health Economics Consortium, University of York, York, United Kingdom

Hayden.holmes@york.ac.uk

Methods review: methods

A non-systematic methods review was undertaken using a targeted search of HTA guidance and additional desk-based research. After consultation with experts, HTA guidance published by the National Institute for Health and Care Excellence (NICE) and the Canadian Agency for Drugs and Technologies in Health (CADTH) were prioritised because of their expertise in the evaluation of diagnostic tests.

The methods used for the methods review are summarised in Table S1. Following searching and screening, HTA submissions and articles were selected for extraction based on an *a priori* set of prioritisation criteria that aimed to identify the most relevant comparisons of IVDs and imaging tests.

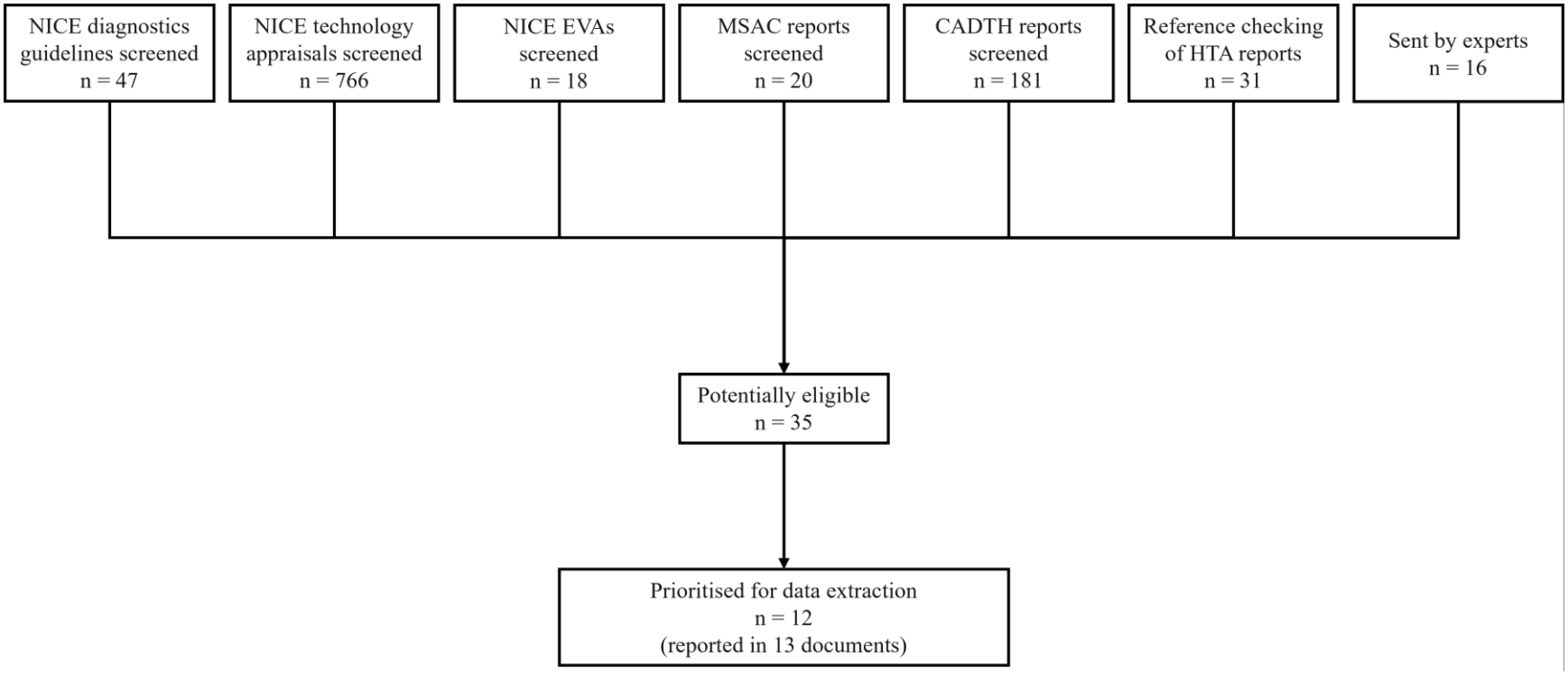
Table S1: Methods review: methods summary

Methods review methods	**Methods detail**
Protocol	The review aims and methods were detailed in a pre-specified protocol.
Study eligibility	**Eligible studies** compared any IVD with any imaging test (including endoscopy), used in isolation or as part of a broader strategy. We excluded studies that: ● Compared an IVD with an IVD (unless an imaging test was also being evaluated). ● Evaluated an index test other than IVD and/or imaging test. ● Evaluated a comparator test other than IVD and/or imaging test. ● Evaluated tests used for purposes other than diagnosis, staging or monitoring (e.g. for guiding surgery or to support prognosis). ● Did not include de novo health economic modelling. ● Reported ongoing studies (i.e. without results). ● Were brief reports, such as horizon scanning reports.
Identifying relevant studies	A targeted search of HTA guidance published by NICE or CADTH was undertaken by a single reviewer, with a second reviewer checking 10% of irrelevant submissions. As planned in the protocol, searches of the MSAC website were performed due to the identification of fewer than 10 relevant HTAs. This was performed using targeted word terms. Reference checking of otherwise eligible HTAs which did not have a de novo model was undertaken. Relevant studies were also sought from experts in test evaluation methodology.
Study selection	Obviously irrelevant records were removed by a senior reviewer. Single reviewer selection of relevant titles, abstracts and full texts was undertaken, with first 10% at each stage screened independently and in duplicate for consistency checks. Disagreement was resolved through discussion. Two reviewers examined all eligible studies and selected 15 to be prioritised for data extraction through consensus (number reflected time and resources available). The following diagnostic comparisons were prioritised: ● 1. Diagnostic, IVD (index) replaces imaging (comparator) ● 2. Diagnostic, IVD (comparator) replaces imaging (index) ● 3. Diagnostic, IVD as add-on or triage to imaging (or vice versa) ● 4. Monitoring, IVD (index) replaces imaging (comparator)

Methods review methods	Methods detail
	● 5. Monitoring, IVD (comparator) replaces imaging (index) ● 6. Monitoring, IVD as add-on or triage to imaging (or vice versa) ● 7. Diagnostic imaging vs imaging Although our focus was IVD vs imaging comparisons, we extended to comparisons between imaging tests due to not identifying a sufficient number of otherwise eligible HTAs. This decision was taken at protocol stage. We also prioritised to try to capture different types of IVD and imaging tests, conducted in a range of health conditions.
Data extraction	A data extraction sheet was created in Microsoft Excel and piloted on two HTAs. A single reviewer conducted the data extraction, and another reviewer checked all extractions for three HTAs. Disagreements were resolved through discussion. Elements extracted: ● Study details (bibliographic details). ● Setting (clinical area, population, healthcare setting). ● Index test (technology, value proposition of the technology, indicated use, target condition, shift in pathway position). ● Comparator test (technology, description of test). ● Modelling methods (type of economic evaluation, model approach used, sensitivity analysis, PICOTs, reference standard used, key model assumptions, key sources of uncertainty). ● Challenges or limitations of the model. ● Potential solutions to challenges/limitations. ● Model conclusions. ● HTA decision.
Synthesis and analysis	Extracted data were synthesised using categorisation where relevant; for example, the diagnostic comparisons, type of limitations reported, and which aspects of the comparison (or value proposition) these limitations were related to. Each included study was discussed with the project team to identify additional limitations and solutions that may not have been reported within the study.

Abbreviations: CADTH, Canadian Agency for Drugs and Technologies in Health; HTA, Health Technology Assessment; IVD, in vitro diagnostic; MSAC, Medical Services Advisory Committee; NICE, National Institute for Health and Care Excellence; PICOT, Patient/Population, Intervention, Comparison, Outcome, and Time.

Figure S1: Flow diagram showing study selection



Abbreviations: CADTH, Canadian Agency for Drugs and Technologies in Health; EVA, early value assessment; HTA, Health Technology Assessment; MSAC, Medical Services Advisory Committee; NICE, National Institute for Health and Care Excellence.

Methods review: results

The methods review identified 35 potentially eligible studies, of which 12 studies (reported in 13 documents) were prioritised for data extraction: 6 HTAs (published 2007 to 2024), 6 journal articles (published 2015 to 2022), and 1 National Institute for Health and Care Research report (published 2007). No relevant reports were identified from the CADTH website. The characteristics of these extracted studies are reported in Table S2.

Table S2: Study characteristics of extracted studies (n=12)

Characteristics	**n(%)**
Source	
NICE DG search	3(25)
MSAC website search	1(8)
Journal article identified through reference checking	4(33)
Journal article known to the authors	4(33)
Role of test (priority 1 to 7*)	
Diagnostic, IVD (index) replaces imaging (comparator) (priority 1)	1(8)
Diagnostic, imaging (index) replaces IVD (comparator) (priority 2)	1(8)
IVD as add-on (n=1) or triage (n=6) to imaging (priority 3)	7(58)
New imaging technology as an add-on to existing imaging (n=2) or to replace existing imaging test (n=1) (priority 7)	3(25)
Purpose of testing	
Diagnosis	9(75)
Monitoring	0(0)
Screening	3(25)
Setting	
Primary	2(17)
Secondary	8(67)
Both	1(8)
Not reported	1(8)
Target condition	
Suspected heart failure	1(8)
Traumatic brain injury	1(8)
Alzheimer's disease	2(17)
Pulmonary arterial hypertension	1(8)
Lung cancer	1(8)
Inflammatory bowel disease	1(8)
Suspected brain tumour	1(8)
Colorectal cancer	1(8)
Ischaemic stroke	1(8)
Hepatocellular carcinoma	2(17)
Type of analysis used	
Cost-utility analysis	3(25)
Cost-effectiveness analysis	7(58)
Exploratory economic model	1(8)
Cost-comparison analysis (with a cost-effectiveness analysis)	1(8)

Characteristics	n(%)
Model used	
Decision tree	6(50)
Markov model	1(8)
Decision tree and Markov model	2(17)
Decision tree and state transition model	1(8)
Cost calculator	1(8)
Not reported (assumed to be decision tree and Markov model)	1(8)

Abbreviations: DG, diagnostic guidance; IVD, in vitro diagnostics; MSAC, Medical Services Advisory Committee; NICE, National Institute for Health and Care Excellence.

*Note: No studies were identified for priority 4 (monitoring, IVD [index] replaces imaging test [comparator]), 5 (monitoring, IVD [comparator] replaces imaging [index], or 6 (monitoring, IVD as an add-on or triage to imaging or vice versa).

Each included study and its extraction was discussed at roundtable amongst authors (LFR, DW, KW, EG, HH) to ensure all challenges had been identified. Overall, 85 challenges were extracted from the 13 included studies. These were grouped into using two levels of categories (Table S3).

Table S3: Categories used to group limitations identified in the literature

Data availability / Applicability	**Test uptake**
	Versions of tests
	Variation in current practice (imaging)
	Procedural harm
	Diagnostic accuracy
	Intermediate results
	Clinical decision-making / confidence
	Health outcomes
	Mismatching PICO
	Poor quality data
	Multiple
Interoperator variability	Accuracy
	Decision-making
	Test fails
Differential access	Access to tests
	Accuracy
Differential impact	Adherence

	Procedural harm
	Test fails
	Accuracy
	Time to diagnosis
	Decision-making and health outcomes
Test fails	Varying image quality (patient factors)
	Varying image quality (technical factors)
	Repeated testing
Test uptake	Acceptability
	Contraindications

Abbreviations: PICO, Patient/Population, Intervention, Comparison, Outcome.

Preliminary challenges (pre-interview)

During the workshop, the results of the methods review were presented and a selection of case studies identified in the review were discussed in detail. The research team used these discussions to develop a list of preliminary challenges (Table S4).

Table S4: Preliminary challenges identified from methods review and prioritised following workshop

Challenges
1. Variability in imaging pathways The existing imaging pathway may be highly variable in terms of the tests conducted and their sequencing, which creates uncertainty in the nature of the comparator.
2. Variability in position of new test When considering adding an IVD to an existing imaging pathway (or vice versa), the most effective role of the new test (add-on, replacement or triage) may also be uncertain.
3. IVD changes the care setting Introducing an IVD can alter the setting of care from a centralised, specialist centre to a more local setting. In these comparisons, patient burden to attend testing (such as travelling time and cost, taking time off work) might be reduced. This is not often considered in models and decision making.
4. Incidental findings Incidental findings have unclear clinical significance. While they could be beneficial, they may also result in additional investigations that carry risks of procedural harm and overtreatment. This issue is much more common for imaging tests, and introduces more complex modelling including for unrelated conditions.
5. Threshold for referring to imaging changes while waiting for IVD triage result In acute or chronic settings with long healthcare waiting times where an IVD is being introduced to triage for an existing imaging test, the patient may deteriorate, or other test results may become available that could change the threshold for referring the patient for the imaging test. The resulting variability in the diagnostic pathway is challenging to incorporate into the model.
6. Contraindications It may not be possible to conduct one of the tests being compared in all indicated patients due to contraindications. This tends to affect imaging tests more frequently and can make the structure of the model more complex.
7. Patient preference It may not be possible to conduct either the imaging test or IVD in all patients due to patient refusal or preference.
8. Variation in test failure and interpretable outputs The risk of the test failing to produce an output that can be interpreted is likely to vary between an IVD (which is highly standardised and thus more likely to produce a result) and an imaging test (which can fail to provide an output that can be interpreted due to patient characteristics, such as obesity or presence of dense tissue, and / or technical factors, such as poor image quality). This can result in missing diagnostic information and additional diagnostic testing, which may impact on costs and health-related quality of life.
9. Lack of data to parameterise the model Evidence from existing primary studies may not be available for key model inputs (including accuracy or decision-making), for the comparator imaging test.
10. Uncertain test accuracy data Test accuracy data for the comparator imaging test are not robust, not applicable to the PICO / decision problem, or do not reflect clinical practice; or the data for either test is conflicting or has high uncertainty. This creates uncertainty.
11. Different versions of imaging tests in use

Incremental improvements in imaging tests over time may result in multiple versions of a single test in use at different centres and at different points in time. This results in heterogeneity in both costs and performance values which are challenging to incorporate into the model. This is particularly relevant to imaging tests using AI assistance. Incremental improvements in accuracy may also have negative consequences such as an increased rate of incidental findings (see Limitation 3). This is a less frequent issue for IVDs.
12. Inter-operator variability of imaging tests Due to their subjective nature, the accuracy of imaging tests varies according to the clinician's expertise and ability. By comparison, IVDs use highly standardised methods with objective thresholds for interpretation. Incorporating this variability into the model structure is challenging.
13. Lack of evidence on clinical decision making If evidence of how the test alters clinical decision making is not available, the model will not be able to directly link the test's accuracy through to treatment effectiveness. The issue is more acute for imaging tests where clinical interpretation is less standardised and more susceptible to variation in interpretation of findings and decision making.
14. Availability of diagnostic decision-making data when IVD introduced as add-on test When an IVD is proposed as an add-on to an existing imaging test, diagnostic decisions resulting from information available from both IVD and imaging are likely to differ from decisions made using the imaging test alone. Changes will depend on how the results of the new test will be combined with the existing test to produce a diagnostic decision.
15. Additional information provided by imaging tests The imaging test may provide information in addition to the presence / absence of disease, such as information on severity / stages, location of lesions, or number of lesions, which is not provided by the IVD. This is typically disease-specific and is more important if this information alters the treatment course.
16. Variation in turnaround time for test results The IVD and imaging test may vary in time to produce a test result. These differences are challenging to include in the model structure. This limitation considers the period up to diagnosis.
17. Variation in timing of care The IVD and imaging test may vary in the time to diagnosis, causing differential timing of treatment. This may be difficult to capture in a model. For example, a point-of-care IVD might enable a diagnosis within minutes to hours while an imaging test could take days to weeks to produce a diagnosis due to the need to wait for imaging centre availability and clinical interpretation.
18. Capacity constraints Imaging services are under increasing demand and are very likely to be prone to capacity constraints. Though this is an issue of importance to policymakers, it is challenging to capture these constraints in economic modelling.
19. IVD triage delays treatment It can be challenging to identify important trade-offs in triage comparisons. When introducing a triage IVD to identify people who will benefit from earlier treatment, false negative cases would receive delayed treatment.
20. Incorporating capital investment costs Often a new test will require substantial capital investment, such as new digital readers or pathology services. This creates a complex decision problem as it is likely new tests that can use the existing infrastructure will be considered more cost-effective than those that require new (and costly) infrastructure.
21. Additional resources are required to optimise performance of a new test When a new test is deployed, technicians and clinicians are likely to require additional resources to optimise their performance. While this applies to both IVDs and imaging tests, the requirements for imaging tests are often greater.

Abbreviations: IVD, in vitro diagnostics; PICO, Patient/Population, Intervention, Comparison, and Outcome.

Interviews

Seven semi-structured interviews were conducted with international experts in health economic modelling and HTA of diagnostic tests. The characteristics of experts are reported in Table S5.

Table S5: Interview expert characteristics

Expert code	Gender	Role	Country	Years of experience	Field of training
E1	Male	Health Economist	UK	Over 30 years	Economics, health economics
E2	Male	Professor of Health Economics	Sweden	30 years	Doctor of Medicine, health economics
E3	Male	Professor of HTA	UK	Over 25 years	Operational research
E4	Male	Health Economist	UK	Over 30 years	Health economics
E5	Male	Professor (Cancer Prevention Program) and Physician	US	Over 30 years	Economics and general science, Doctor of Medicine, Health economics,
E6	Male	Senior Lecturer in Health Economics	UK	13 years	Engineering, mathematics, health economics
E7	Female	Health Economist (Researcher)	Australia	20 years	Health Economics and Policy, Economics, Pharmacy

Abbreviations: UK, United Kingdom; US, United States.

Post-interview scoring

Table S6: Post-interview scoring of revised general recommendations with lower levels of agreement

Challenges	Recommendations	Expert 1 score	Expert 2 score	Expert 3 score	Expert 4 score	Expert 5 score	Expert 6 score	Expert 7 score
2. Variability in position of new test	2. When the role of the new IVD test is uncertain, investigating the most cost-effective sequence of tests is recommended.	2	4	5	4	5	5	5
3. Lack of data to parameterise the model	3. Expert opinion is recommended to identify whether alternative imaging tests are clinically and sufficiently similar to the comparator to provide proxy values. Elicitation is recommended to identify how the tests are similar, as well as acceptable ranges of key parameters (such as sensitivity, specificity and decision making).	4	3	4	4	4	3	4
4. Uncertain test accuracy data	4b. If evidence is not available in the relevant patient population, seeking clinical advice on the most generalisable study populations is recommended to inform the most likely direction of results.	5	3	5	4	5	5	5
	4c. When evidence is poor, conflicting or uncertain, varying accuracy parameters using sensitivity analysis is recommended to estimate the impact of changes to sensitivity/specificity and to identify the values at which the results will change direction (i.e. threshold analyses). Probabilistic approaches can be used when there is confidence in the ranges and distributions of accuracy parameters.	2	4	5	5	5	2	3
	4d. Clinical validation of uncertain accuracy limits or thresholds is recommended through clinical opinion or elicitation to advise whether ranges are realistic. Validation against other data that are not used in the model may also be useful, for example existing clinical studies or real-world data.	4	3	4	3	5	5	4
5. Lack of evidence on clinical decision making	5a. The preferred model inputs are reliable end-to-end or clinical utility studies that measure clinical decision making, are applicable to the decision problem, and evaluate the imaging test in its intended place in the pathway. Adjusting the model structure/inputs to reflect this evidence is recommended. It is not appropriate to use stand-alone efficacy data from either test when a diagnostic decision is made using combined data.	2	4	3	4	5	5	5
	5b. In the absence of direct evidence, follow the recommendation published by Shinkins et al. (2024) for identifying changes in recommended and actual management: seek clinical opinion to understand the impact to decision making, avoid making blanket assumptions about recommended patient management unless truly reflective of the clinical scenario. Adjusting the model structure/inputs to reflect this evidence is recommended.	3	4	4	4	4	5	5
	5c. Performing multivariate sensitivity analysis assuming varying patterns of clinical decision making is recommended (e.g. different patterns of referral behaviour with increasing levels of resource use). If relevant, varying assumptions on clinician compliance to management pathways is also	2	4	4	4	4	4	5

Challenges	Recommendations	Expert 1 score	Expert 2 score	Expert 3 score	Expert 4 score	Expert 5 score	Expert 6 score	Expert 7 score
	recommended to indicate the direction of results and whether this changes based on different decision-making assumptions.							
6. Additional resources are required to optimise performance of a new test	6. When optimisation costs could significantly impact accuracy or capacity, incorporating the key additional resources required to use the new test should be considered.	3	4	5	4	5	3	4
7. Incorporating capital investment costs	7a. Capital investment in changes to diagnostic infrastructure should be considered for both the proposed and existing tests.	5	2	3	4	5	4	5
	7b. Capital investment costs should always be considered but should only be included in the model if appropriate. Several analytic methods can be used to incorporate capital investment costs. It is recommended to select the method which is most applicable. A qualitative discussion of capital investment for use by multiple departments or multiple indications could also be provided. For example, a digital reader could be used for all pathology services not just one test.	3	2	5	4	4	4	4
8. Prioritising the value proposition	8. It is not always possible to include every consequence of introducing the test in the model structure because this could make the model too complex. Identifying and considering the most important consequences of introducing the new test is recommended, to help prioritise those that should be included in the model structure.	3	3	5	4	5	4	5

Abbreviations: IVD, in vitro diagnostics.

Note: Yellow colour denotes lower levels of agreement, defined as a median score of 4. Recommendations that all experts agreed or strongly agreed with have been removed from this list but are included in Table 3 in the manuscript.

Table S7: Post-interview scoring of revised recommendations relevant to specific aspects of the value proposition with lower levels of agreement

Challenges	Recommendations	Expert 1 score	Expert 2 score	Expert 3 score	Expert 4 score	Expert 5 score	Expert 6 score	Expert 7 score
10. Incidental findings	10a. When considered clinically relevant and there is a policy to act on them, incidental findings and their consequences should be made explicit and considered for incorporation in the conceptual model. When incorporation is not considered appropriate, the rationale should be clearly stated.	4	3	4	3	5	5	5
11. Direct test harms	11. When relevant, considering changes in complication rates, uptake or other important consequences of differences in procedural harm is recommended.	3	3	5	5	5	4	5
12. Different versions of imaging tests in use	12b. When the evidence base is uncertain, Recommendation 4b applies.	3	3	4	4	4	5	4
	12c. If the test is not cost effective, exploring what is needed to make the test cost effective in the model is recommended. Target product profiles could be used for this purpose.	3	4	3	4	5	3	3
13. Variation in test failure and interpretable outputs	13b. Some tests require more tests per patient than other tests. This could be due to setting or test reliability, e.g. home sampling return rates versus hospital testing. Varying the number of tests per patient through scenario or sensitivity analyses, using PSA where there is good evidence of variability, is recommended.	3	4	3	4	5	5	4
14. Additional information provided by imaging tests	14. The impact of additional diagnostic information should be evaluated in terms of its potential impact to decision making and patient outcomes.	4	4	4	3	5	5	5
15. Inter-operator variability of imaging tests	15b. Time or experience-based resource use and complication rates should also be considered and varied using scenario or sensitivity analysis for each of the relevant inputs.	4	3	4	4	5	4	4
16. Variation in lead time for test results	16. It is important to consider whether differences in the lead time for test results create system benefits (such as clinician time saved) or resource use benefits (such as cost per image read). These could be included as cost savings or efficiency gains. It is also important to consider any benefits from the patient's perspective (such as comfortability during the test) and clinician's perspective (such as time saved).	4	4	2	4	3	1	4
17. Variation in timing of care	17. Clinical opinion should be sought to understand whether a material difference in time to diagnosis could delay treatment that might impact adversely on disease progression or treatment outcomes. If so, these risks should be incorporated into the model and evidenced appropriately. An appropriate time horizon should be used to estimate the difference in long-term outcomes.	5	3	5	4	5	5	4
18. Capacity constraints	18a. If capacity constraints are important to the decision problem, they should be incorporated into the conceptual model. When a new test has the potential to	4	4	4	4	5	4	3

Challenges	Recommendations	Expert 1 score	Expert 2 score	Expert 3 score	Expert 4 score	Expert 5 score	Expert 6 score	Expert 7 score
	be capacity releasing, the incremental change in resource should be considered in the results (i.e. number of radiology appointments avoided).							
	18b. Refer to 16 as a potential pragmatic approach. However, a separate capacity constraint model could be considered alongside the cost-effectiveness model.	3	4	3	4	4	3	3
19. IVD triage delays treatment	19a. It is important to understand whether IVD false negative cases have a different likelihood of progression compared with imaging false negatives. Using evidence or expert elicitation to determine this is recommended.	3	3	3	3	3	4	4
	19b. If there is a different likelihood of progression, incorporating this higher risk in the model and using an appropriate time horizon to estimate the differences in long-term outcomes is recommended.	4	3	3	1	5	4	5

Abbreviations: IVD, in vitro diagnostics; PSA, probabilistic sensitivity analysis.

Note: Yellow denotes lower levels of agreement, defined as a median score of 4. Orange colour denotes exclusion, defined as a median score of less than 4. Recommendations that all experts agreed or strongly agreed with have been removed from this list but are included in Table 4 in the manuscript.