



Use of Decision Modelling in Economic Evaluations of Diagnostic Tests: An Appraisal and Review of Health Technology Assessments in the UK

Yaling Yang¹ · Lucy Abel¹ · James Buchanan² · Thomas Fanshawe¹ · Bethany Shinkins³

Published online: 14 December 2018
© The Author(s) 2018

Abstract

Diagnostic tests play an important role in the clinical decision-making process by providing information that enables patients to be identified and stratified to the most appropriate treatment and management strategies. Decision analytic modelling facilitates the synthesis of evidence from multiple sources to evaluate the cost effectiveness of diagnostic tests. This study critically reviews the methods used to model the cost effectiveness of diagnostic tests in UK National Institute for Health Research (NIHR) Health Technology Assessment (HTA) reports. UK NIHR HTA reports published between 2009 and 2018 were screened to identify those reporting an economic evaluation of a diagnostic test using decision analytic modelling. Existing decision modelling checklists were identified in the literature and a modified checklist tailored to diagnostic economic evaluations was developed, piloted and used to assess the diagnostic models in HTA reports. Of 728 HTA reports published during the study period, 55 met the inclusion criteria. The majority of models performed well with a clearly defined decision problem and analytical perspective (89% of HTAs met the criterion). The model structure usually reflected the care pathway and progression of the health condition. However, there are areas requiring improvement. These are predominantly systematic identification of treatment effects (20% met), poor selection of comparators (50% met) and assumed independence of tests used in sequence (32% took correlation between sequential tests into consideration). The complexity and constraints of performing decision analysis of diagnostic tests on costs and health outcomes makes it particularly challenging and, as a result, quality issues remain. This review provides a comprehensive assessment of modelling in HTA reports, highlights problems and gives recommendations for future diagnostic modelling practice.

Key Points for Decision Makers

A diagnostic test-specific checklist to assess decision modelling has been developed and piloted.

The models in Health Technology Assessments tended to be of relative high quality but also suffered key problems including lacking justification of comparators, lacking model validation, insufficient efforts to examine structural uncertainty and obtain treatment effects data as well as assuming independence of tests in sequence.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s41669-018-0109-9>) contains supplementary material, which is available to authorized users.

✉ Yaling Yang
yaling.yang@phc.ox.ac.uk

Lucy Abel
lucy.abel@phc.ox.ac.uk

James Buchanan
james.buchanan@dph.ox.ac.uk

Thomas Fanshawe
thomas.fanshawe@phc.ox.ac.uk

Bethany Shinkins
b.shinkins@leeds.ac.uk

¹ Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK

² Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Richard Doll Building, Old Road Campus, Oxford OX3 7LF, UK

³ Academic Unit of Health Economics, Leeds Institute of Health Sciences, University of Leeds, Worsley Building, Clarendon Way, Leeds LS2 9LJ, UK

1 Introduction

Diagnostic tests play an important role in the clinical decision-making process by providing information that enables patients to be identified and stratified to the most appropriate treatment and management strategies. The value of diagnostic tests and its implication for technology assessment has long been recognised and discussed [1, 2]. Historically, the primary focus of test evaluation research has been the estimation of clinical accuracy. However, a recent systematic review of test–treatment randomised controlled trials (RCTs) demonstrates that improvements in test accuracy are rarely an indicator of patient health benefit [3]. Therefore, to evaluate the impact of a new diagnostic test on patient health outcomes, it must be examined as part of a broader test–treatment management strategy.

Test–treatment RCTs are uncommon due to the many challenges in conducting such trials and, of those that have been carried out, many have failed to deliver robust results [4]. Decision analytic modelling is now recognised as a practical alternative [4, 5]. It facilitates the evaluation of both the economic (i.e. cost) and clinical (i.e. health benefit) impacts of introducing a new test simultaneously. It also overcomes a number of additional methodological hurdles common to diagnostic test evaluation: the need to link evidence from a number of different sources, the lack of long-term outcome data in scenarios where only intermediate endpoints (e.g. test accuracy) have been measured or where only short-term follow up is possible and the need to compare many interventions (e.g. testing strategies/pathways), which may not be feasible/practical within a single RCT [6].

There is currently limited understanding of how decision analytical modelling has been used in economic evaluations of diagnostic tests. Existing reviews have appraised economic evaluations of diagnostic tests more generally but have not focused specifically on the modelling. Further, much of the literature assessing the economic evaluation of diagnostic tests is now outdated. In 1999, Severens and van der Wilt published a review of economic evaluations of diagnostic tests and found that many failed to adhere to guidelines for economic evaluation [7]. A further review of cost-utility studies of diagnostic tests in 2011 found that many models excluded key parameters such as test safety or the potential harms from testing [8].

Other more recent work has focused on how diagnostic evidence has been incorporated into economic models [9–11]. Doble et al. reviewed models of companion diagnostics for targeted oncology therapies, reporting that out of a total of 30 included studies, 12 studies incorporated test accuracy parameters in the model whereas the remaining 18 studies only included the cost of testing [12]. A

final study reviewed the use of evidence in decision models based on published UK National Institute of Health Research (NIHR) Health Technology Assessment (HTA) reports [13]. This study briefly assessed the models used but covered all HTAs rather than focusing on diagnostics. This work has recently been further developed into a proposed analytical framework for diagnostic and prognostic test assessments for HTAs [14].

In the UK, the National Institute for Health and Care Excellence (NICE) Diagnostic Assessment Programme was only established in 2010. The economic evaluation of diagnostic tests and the appropriate methods for such analyses is a relatively new area of research under active development. Specifically, model-based economic evaluations of diagnostics require careful consideration: model structure must accurately reflect the clinical care pathway, key decision points and disease progression. Parameter selection should reflect the best evidence available, and all analytical uncertainties must be fully explored. The challenges in this field have recently been highlighted by a Medical Research Centre call for methodology research to support the assessment of diagnostic technologies and the development of diagnostic service guidelines for health-care decision making such as by NICE [15]. This review provides an update on the assessment of model-based economic evaluations of diagnostics tests, which is urgently needed to inform and contribute to methodological and policy developments such as the NICE diagnostic assessment process in this area.

2 Methods

There were four key steps in this review: (i) identification of relevant UK NIHR HTA reports, (ii) development and piloting of a reporting checklist to assess model-based economic evaluations of diagnostic tests, (iii) data extraction and (iv) data analysis.

2.1 Identifying Diagnostic UK National Institute of Health Research (NIHR) Health Technology Assessments (HTAs) to be Included in the Review

The titles and abstracts of all UK NIHR HTA reports published between May 2009 and July 2018 were independently screened by two authors (TF and BS), with disagreements resolved by discussion with LA and YY as required, to identify model-based economic evaluations of diagnostic tests. Reports were excluded if the primary role of the test(s) was not diagnosis (e.g. screening, treatment selection, prognosis or monitoring). The methods used to synthesise test accuracy

data and the extent that these results informed health economic model parameters have been examined in a parallel study [16].

2.2 Developing and Piloting of a Reporting Checklist to Assess Model-Based Economic Evaluations of Diagnostic Tests

A rapid search of the literature was conducted to identify existing checklists or guidelines intended for the development or reporting of decision models in health care. Three checklists were identified: Philips et al. [17], Ramos et al. [18] and the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement [19]. Several reports related to good research practices for modelling produced jointly by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Society for Medical Decision Making (SMDM) were also identified [20–25].

A reporting checklist to assess model-based economic evaluations of diagnostic tests was developed based on these checklists and guidelines, primarily derived from the Philips et al. checklist [17]. The checklist was piloted (by LA) on five HTA reports and then modified to include key items related to diagnostic tests. Two reviewers (LA and YY) then independently reviewed and extracted data for a sub-group of 24 HTA reports using the refined checklist and the results were compared. Any differences and disagreements were identified and discussed between LA, YY and JB. This resulted in further refinement of the questions and the associated scoring system to maximise consistency.

The final checklist is presented in Table 1. It contains 39 questions and covers 11 key domains of good modelling practice. For each question, HTA reports were classified as having ‘not met’ (scored as 0), ‘partially met’ (scored as 0.5), or ‘met’ (scored as 1) the stated criterion. ‘Partially met’ was used to reflect situations in which reports implied that some work had been done to meet the criteria. For example, for the question ‘Have diagnostic accuracy data been derived from high quality data sources’, the report might indicate that this had been done for most diagnostic accuracy parameters but not all. Judgements were commonly made on the basis of self-reported evidence or efforts from the authors of the HTA reports rather than conducting independent appraisals. Not applicable (‘n/a’) was also possible for some questions. For example, if a decision tree model was developed, the question ‘Does the cycle reflect the progression and transition of health states?’ would not be applicable, since that is a Markov-specific

model characteristic. Detailed information on how the responses were defined can be found in Appendix 1 (see Electronic Supplementary Material).

2.3 Data Extraction

Background information (e.g. year of publication, condition of interest, test type and population of interest) was extracted from each UK NIHR HTA report. Test type was categorised as (i) imaging, (ii) genetic, (iii) laboratory, (iv) point of care. These categories were not necessarily mutually exclusive; for example, a genetic test could also be a laboratory test. Tests that did not fall into one of these categories were classified as ‘other’.

The four following additional details were also extracted from each report: (i) details of the model structure, (ii) whether any formal assessment tools were used to evaluate the quality of studies providing data to inform model parameters, (iii) which outcome measures were reported and (iv) the perspective adopted for the analysis.

The checklist was then completed for each report by a single reviewer (LA). The checklist was independently completed by a second reviewer (YY) for eight randomly selected HTAs. Agreement between the two reviewers was examined. Agreement was defined as absolute agreement where both reviewers gave the same responses to a question. A partial agreement was where one reviewer gave ‘met’ but the other gave ‘partially met’, or where ‘partially met’ and ‘not met’ were given. The proportions of absolute and partial agreements between reviewers for individual questions and domains are reported.

All extracted data was stored and analysed in Microsoft Excel.

2.4 Data Analysis

The characteristics of the HTA reports were summarised and the proportion of responses to the eleven assessment domains across all included HTAs was calculated. For example, the first domain ‘Decision problem and scope specified’ contains five questions. Therefore, the denominator (the total possible responses) is $N*5$ where N is the number of included HTAs. Then the sum of responses being ‘not met’, ‘partially met’ and ‘met’ for the five questions among the N HTAs were calculated as the numerator, and the proportion of each response was calculated.

Responses were broken down by checklist domain, test type and year. Responses to individual questions were also reviewed to identify areas of particular weakness.

Table 1 Checklist for model-based economic evaluations of diagnostic tests

Domain	Question	
1. Decision problem and scope specified	1. Is there a clear statement of the decision problem?	
	2. Is the perspective of the model stated clearly?	
	3. Has the target population been identified?	
	4. Are the model inputs consistent with the stated perspective?	
	5. Are the primary outcomes of the model consistent with the perspective, scope and overall objective of the model?	
2. Identification and description of comparators	6. Have all feasible and practical options been identified?	
	7. Have the comparators being evaluated been clearly described?	
	8. If comparators have been excluded from the evaluation, have these exclusions been justified?	
3. Appropriate data identification	9. Are the data identification methods transparent, systematic and appropriate given the objectives of the model?	
4. Sufficient detail for data incorporation	10. Have all data incorporated into the model been described and referenced in sufficient detail?	
	11. Where choices have been made between data sources, are these justified appropriately?	
	12. Are transition probabilities calculated appropriately?	
	13. Has discounting been conducted?	
5. Quality and incorporation of test accuracy data	14. Has the quality of the test accuracy data been assessed?	
	15. Have diagnostic accuracy data been derived from high quality data sources (hierarchy of evidence)?	
	16. Are tests in sequence treated dependently, where appropriate?	
6. Quality and incorporation of treatment data	17. Has the quality of the treatment effect data been assessed?	
	18. Have relative treatment effects been derived from high quality data sources (hierarchy of evidence)?	
7. Source and incorporation of cost data	19. Has the source of cost data been presented clearly?	
	20. Have costs been inflated to a specific year, where appropriate?	
8. Source and incorporation of utility data	21. Is the source for the utility weights referenced and justified?	
	22. Are the utilities incorporated into the model appropriately?	
9. Model structure	23. Have the reasons behind the type of decision analytic model chosen been fully described and justified?	
	24. Has a systematic review of existing economic evaluations been carried out?	
	25. Is the structure of the model consistent with a coherent theory of the health condition under evaluation?	
	26. Are the structural assumptions underpinning the model transparent and justified?	
	27. Have the methods used to extrapolate short-term results to final outcomes been documented and justified?	
	28. Has the time horizon been stated and justified?	
	29. Has cycle length of Markov models been justified?	
	10. Uncertainty	30. Has parameter uncertainty been addressed via sensitivity analysis?
		31. Has probabilistic sensitivity analysis been carried out? If not, has this omission been justified?
		32. If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?
33. If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?		
34. Have structural uncertainties been addressed via sensitivity analysis?		
35. Have alternative assumptions related to final outcomes been explored through sensitivity analysis?		
36. Has value of information analysis been done?		

Table 1 (continued)

Domain	Question
11. Validity	37. Has the face validity been reviewed by someone external to the model developers?
	38. Has the mathematical logic of the model been assessed? (e.g. using null and extreme values)
	39. Have the model and its results been compared to the findings of other models and studies, and any disagreements or inconsistencies been explained (cross-validity)?

3 Results

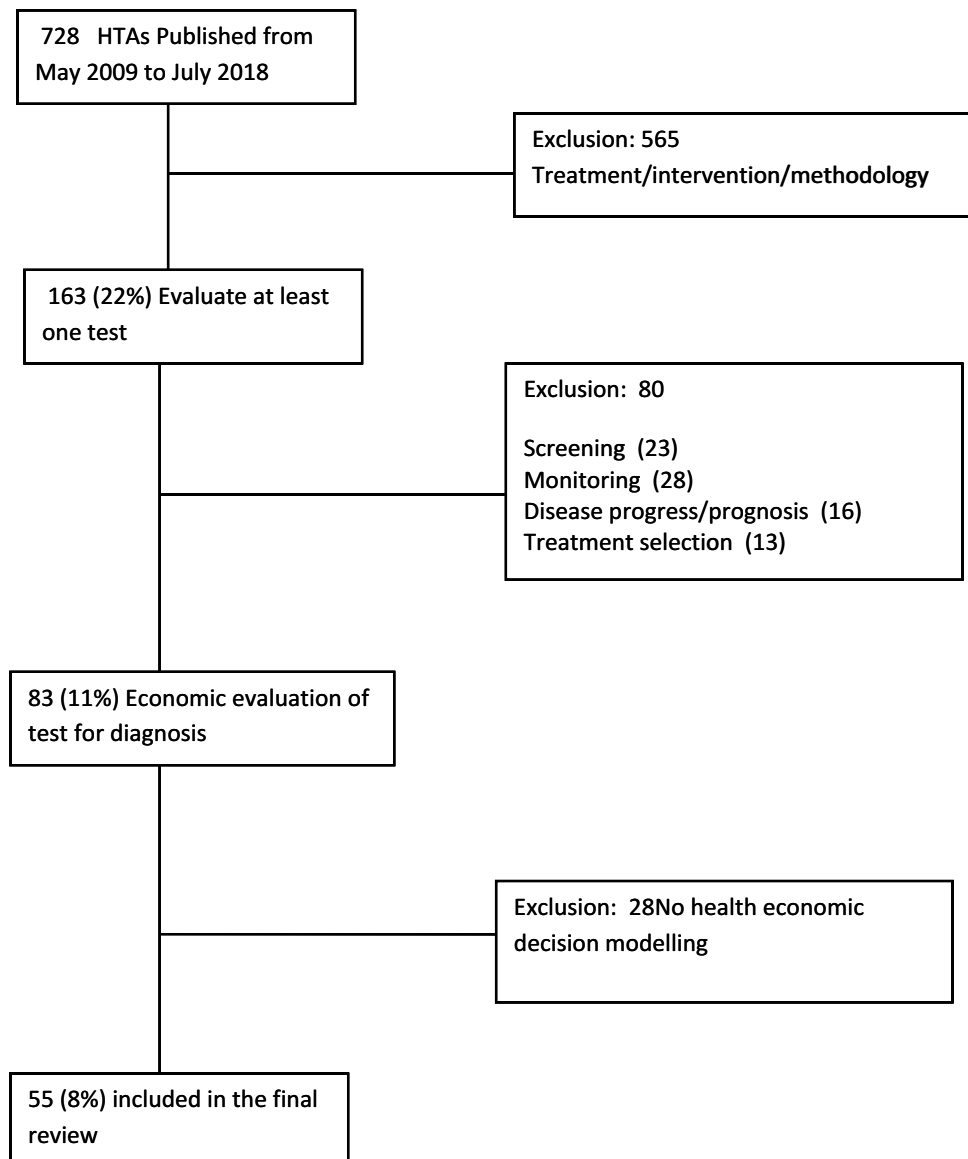
3.1 Search Results

The search results and the screening-for-inclusion process are reported in Fig. 1.

3.2 General Description of the Conditions, Tests and Models in the Included HTAs

Characteristics of the 55 included HTAs reviewed can be found in Table 2. The health conditions to be diagnosed included various chronic diseases (44%, $n = 24/55$), cancer (27%, $n = 15/55$), acute conditions (16%, $n = 9/55$) and

Fig. 1 Search results. *HTAs* Health Technology Assessments



infections (13%, $n = 7/55$). Four types of diagnostic test were identified in the reviewed HTAs: (i) imaging, (ii) genetic, (iii) laboratory, (iv) point of care. Imaging technologies were the most commonly evaluated tests (40%, $n = 22/55$). Of those that did not fall into one of the defined test types, two were clinical decision rules (i.e. where a combination of signs, symptoms and test results determine a patient's diagnosis according to a specified algorithm). In four of the models, the test under evaluation was not compared with a reference standard [26–29]. The remaining models either compared the results of the test under evaluation with another test (55%, $n = 30/55$) or to clinical criteria/follow-up (38%, $n = 21/55$).

Decision trees were the most commonly applied model structure (69%, $n = 38/55$), either alone (42%, $n = 23/55$)

or in combination with a Markov model (27%, $n = 15/55$). This approach stratifies a simulated patient cohort into different groups based on the result of the test, using disease prevalence and test accuracy to determine the proportions of individuals in each group. Markov modelling was implemented in around a third of the HTAs (36%, $n = 20/55$). This method is used to capture longer-term effects (e.g. life-time disease progression).

The vast majority of models (95%, $n = 52/55$) measured outcomes using quality-adjusted life-years. The number of cases detected or avoided were reported in the remaining three models [26, 30, 31]. All analyses adopted the perspective of the NHS or NHS and social care, which is recommended for this type of evaluation by NICE in the UK [32]. Sixteen HTAs included in the review were

Table 2 Summary description for the HTAs included in the review

	Categories	Number of studies in each category [n (%)]*($N = 55$)
Condition to be diagnosed	Cancer	15 (27)
	Chronic diseases	24 (44)
	Infections	7 (13)
	Acute conditions	9 (16)
Type of Test Evaluated	Genetic	7 (13)
	Imaging	22 (40)
	Lab-based	15 (27)
	Point of Care	6 (11)
	Others	5 (9)
Year of Publication	2009	5 (9)
	2010	2 (4)
	2011	5 (9)
	2012	3 (5)
	2013	11 (20)
	2014	6 (11)
	2015	8 (15)
	2016	7 (13)
	2017	5 (10)
	2018 (till July)	3 (5)
Was a reference standard reported?	Yes, another test	30 (55)
	Yes, clinical criteria or clinical follow up	21 (38)
	No	4 (7)
What type of modelling was implemented?	Decision tree only	23 (42)
	Markov model only	5 (9)
	Decision tree and Markov	15 (27)
	Individual patient simulation model	7 (13)
	Discrete event simulation and dynamic transmission model	1 (2)
	Decision tree and individual patient simulation model	1 (2)
	Decision tree and discrete event simulation model	3 (5)
What outcome measures were reported?	QALY	52 (95)
	Case detected or avoided	3 (5)

*May not add up to 100% due to rounding

Diagnostic Assessment Reports (DAR) and provided an evidence base for use to support development NICE diagnostic guidance [33]. Which HTA is linked to which specific NICE diagnostic guidance can be found in online supplementary material containing full data extraction at <https://figshare.com/s/643c9a69d956780eb0a8>.

3.3 Assessment of Model-Based Evaluations of Diagnostic Tests

The degree of agreement between data extracted from the random sample of eight HTAs using the modified checklist by the two reviewers is reported in Table 3. Agreement depended on the quality domains concerned and whether absolute or partial agreement was measured. Among the 11 assessment domains, two achieved $\geq 80\%$ absolute

agreement, another six achieved $> 60\%$ absolute agreement and the remaining three domains achieved around 50% absolute agreement. Six domains achieved $> 80\%$ partial agreement, four domains achieved $> 60\%$ and only one domain achieved around 50% partial agreement.

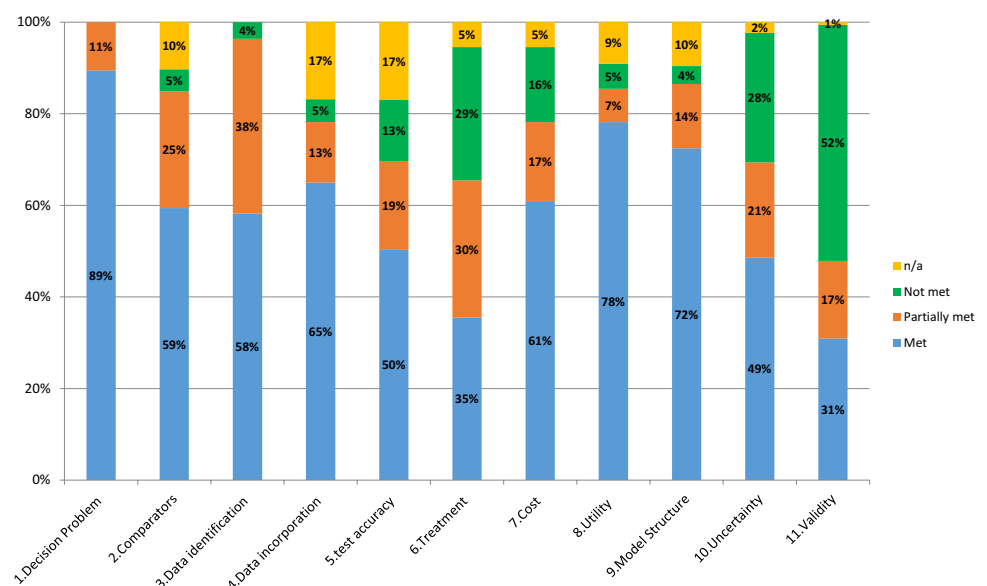
The results from the assessment of the decision analytic models are summarised in Fig. 2. Specification of the decision problem and scope was generally clearly reported (89% of reports met the criteria, 11% partially met). Evaluation of the validity of the models was generally poor, with over half failing to meet or even partially meet the criteria. It should be noted that reviewer agreement for this domain was low, which could highlight some general uncertainty in what would be required and reported to evaluate the validity of a model. The quality of the treatment data incorporated into

Table 3 Percentage of absolute and partial agreement of data extracted by two reviewers

Quality domain	Absolute agreement (%)	Partial agreement (%)
1. Decision problem and scope specified	90***	95***
2. Identification and description of comparators	46*	67**
3. Appropriate data identification	63**	81***
4. Sufficient detail for data incorporation	69**	81***
5. Quality and incorporation of test accuracy data	54*	81***
6. Quality and incorporation of treatment data	63**	75**
7. Source and incorporation of cost data	63**	84***
8. Source and incorporation of utility data	81***	69**
9. Model Structure	66**	79**
10. Uncertainty	77**	88***
11. Validity	46*	54*

* $\leq 60\%$, ** 60–80%, *** \geq over 80%

Fig. 2 Overall performance of models in the 55 included HTAs (2009–2018). HTAs Health Technology Assessments



the models was also of particular concern, with only 35% of reports meeting the required criteria.

For the remaining domains, the proportion of responses that fully met the criteria ranged from 49 to 78%. Most of the remaining reports partially met the criteria, although there were a few domains where a notable proportion failed to either meet or partially meet the criteria (uncertainty 28%, source and incorporation of cost data 16%, quality and incorporation of test accuracy 13%).

A breakdown of the assessment results by type of diagnostic test and publication date is reported in Appendix 2 (see Electronic Supplementary Material). In general, the assessment results were similar across the different types of tests. Across the seven year groups analysed (2009/10, 2011/12, 2013, 2014, 2015, 2016, 2017/18), there is no obvious trend observed in modelling assessment and there are variations across assessment domains.

3.4 Issues of Particular Concern in the Assessment

In Table 4, on the basis of reviewing the performance of domains and questions within each domain, we report the results for those questions that scored poorly and which are regarded as of particular importance (full data extraction results can be found as online supplementary material in <https://figshare.com/s/643c9a69d956780eb0a8>).

Where comparators were excluded from their model, only half of HTAs clearly justified reasons for exclusion. The quality of treatment effect data included in the models was not systematically evaluated; only 20% fully met this criterion and just over half (51%) failed to meet it at all. When the issue is relevant, most HTAs (68%) assumed independence between the tests; that is, the result of one test does not alter the diagnostic accuracy of the second so the combined sensitivity and specificity of sequential tests were obtained by simply multiplying the probabilities of each.

Only just over half (53%) of HTAs reported clearly how transition probabilities were calculated. Only 26% clearly described and justified the choice of distributions used for each parameter and 64% partially did this. Sensitivity analyses to explore uncertainty in the model structure were rarely reported with none of the HTAs fully meeting this criterion, 78% not meeting and 22% partially meeting. Similarly, face validity and cross-validity were rarely properly done.

4 Discussion

This study assesses model-based economic evaluations of diagnostic tests featuring in UK NIHR HTAs published between 2009 and 2018. To achieve this, we developed an assessment checklist tailored specifically to account for the nuances of evaluating diagnostic tests using decision analytic modelling. Most items included in the checklist are applicable to critical appraisal of decision models in general, but some items capture specific key aspects of modelling diagnostics, such as the linkage between test accuracy data to patient management decisions and the efficacy of any treatments given.

In general, the reviewed models scored moderately well on the assessment, with around half of the included HTAs fully or partially meeting most of the criteria. All analyses adopted the perspective of the NHS or NHS and social care, which is recommended for this type of evaluation by NICE in the UK [32]. Our assessment did, however, highlight some areas for possible improvement.

1. Poor reporting and/or justification of comparators

In economic evaluations, it is important to include all possible comparators or clearly justify their exclusion because the evaluation is, by nature, a comparison in which the selection of comparators may change the con-

Table 4 Checklist questions with poor performances n (%)

Questions	Met	Partially met	Not met	N (Excluding n/a)
If comparators have been excluded from the evaluation, have these exclusions been justified?	19 (50)	15 (39)	4 (11)	38
Has the quality of the treatment effect data been assessed?	10 (20)	15 (29)	26 (51)	51
Are transition probabilities calculated appropriately?	19 (66)	6 (21)	4 (14)	29
Are tests in sequence treated dependently, where appropriate?	11 (32)	8 (24)	15 (44)	34
If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified?	12 (26)	30 (64)	5 (11)	47
Have structural uncertainties been addressed via sensitivity analysis?	0 (0)	12 (22)	43 (78)	55
Has the face validity been reviewed by someone external to the model developers?	16 (29)	15 (27)	24 (44)	55
Have the model and its results been compared to the findings of other models and studies, and any disagreements or inconsistencies been explained (cross-validity)?	27 (50)	8 (15)	19 (35)	54

The percentages might not add up to 100% due to rounding. Percentages were calculated by n/N (i.e. n/a are excluded)

clusions of the analysis. We found that reporting of this aspect was poor and the selection of comparators was rarely justified. Where authors did justify the exclusion of comparators, the reasons given were very vague (e.g. ‘not applicable to the NHS’, ‘lack of evidence’, ‘beyond the scope’), making it impossible to judge whether these justifications were valid. Authors are recommended to provide justification for their choice of comparators, ideally including a horizon scan of the diagnostic landscape with clinical experts’ opinions.

2. Poor reporting of transition probability calculations

Transition probabilities are key model parameters. In decision tree models of diagnostic pathways, the prevalence of the target condition and the accuracy of the test are often used to inform transition probabilities. However, in Markov models, calculation of transition probabilities between health states is typically more complex, in many cases relying on survival or clinical event data. In this context, the reporting of transition probability calculations was poor, making it difficult to judge the validity of these parameters. This also prevents other researchers from rebuilding the model in future evaluations.

3. Treatment effect data—identification, quality assessment and synthesis (or lack of)

Treatments are an important component of the test–treatment pathway, linking the effects of diagnostics to patient health outcomes. However, in this review, we noted a stark difference in the rigour in which treatment effect data was identified, quality assessed and synthesised compared with diagnostic accuracy data. Treatment effect data was rarely identified via systematic review and the quality of the evidence was rarely assessed. Furthermore, there was typically insufficient evidence identified to facilitate meta-analysis.

4. Lack of validation

The face validity and cross-validity of the models reviewed were rarely assessed. The lack of cross-validation may be due to a lack of other published models to compare against but, if this is the case, then it needs to be clearly reported. Examination of the face validity of the model is crucial, particularly in this context where the evidence underpinning the model is likely to come from a wide range of sources (e.g. test accuracy studies, treatment effect studies, epidemiology studies) and therefore the linkage of this data could easily produce clinically spurious results. Presenting the model and its output to a team of relevant clinicians and experts in the field is vital to ensuring it reflects real-world clinical practice as far as possible.

5. Assuming independence of tests in sequence

Diagnostic tests are often used sequentially in clinical practice and it is important to acknowledge and account

for dependencies when this occurs. If a test is introduced as a means of identifying an appropriate group of individuals (i.e. a triage test) for a confirmatory test, this will directly impact on the prevalence and spectrum of disease among those referred for the confirmatory test. If the two tests measure a similar physiological concept (e.g. inflammation), then it is likely that there will be a correlation in the misdiagnoses that each test makes. This issue was largely ignored in the reviewed HTAs; the majority (68%) assumed the test results were independent, potentially leading to biased outcomes. This issue could be accounted for by using either diagnostic accuracy studies from the tests used in sequence [34] or by attempting to model the correlation between tests in the sensitivity analysis [35, 36].

6. Structural uncertainty

Although sensitivity analysis (including probabilistic sensitivity analysis) was undertaken routinely, structural uncertainties were rarely examined in the reviewed HTAs. This is particularly pertinent to the evaluation of a novel diagnostic, as its position in and impact on the real-world care pathway and clinical practice (and hence the model structure) has not yet been demonstrated or confirmed. Model structure should ensure some flexibility to allow exploring of various scenarios within sensitivities to reflect both the ‘ideal’ scenario and the ‘messier’ reality of clinical practice.

4.1 Limitations of the Study

This review focused only on HTAs published between 2009 and 2018 in the UK, and these publications represent just a subset of all model-based economic evaluations of diagnostic tests. HTAs tend to allow more space for reporting than normal journal articles, undergo an extensive peer review process and are generally deemed of good quality. The areas for improvement outlined above are in most cases due to the inherent difficulties in performing model-based economic evaluations of diagnostic tests, rather than a reflection of the quality of the HTA reports themselves. Therefore, we believe that the issues that have been identified in this paper are likely to exist in similar studies and may be more prevalent.

Tests play different roles in patient care and the focus of this review was purely on diagnostic tests. We chose to exclude tests used for other purposes as the evidence requirements and data linkage for these models is likely to differ. Reviewing the model-based evaluations of tests used for other purposes (e.g. screening, prognosis, monitoring or treatment selection) would be an interesting addition to this work.

The checklist was developed to assess diagnostic models reported in HTAs. Although the checklist may not be able to provide a comprehensive quality appraisal of

models used in HTAs, it gives some indication of whether or to what extent important principles of analysis have been met or missed. Our checklist focuses on analytical principles and data identification, which supplements the recently published AGREEDT (an acronym of ‘‘AliGnment in the Reporting of Economic Evaluations of Diagnostic Tests and biomarkers AliGnment in the Reporting of Economic Evaluations of Diagnostic Tests and biomarkers’’) checklist with a focus on report and consideration during the early stages of modelling diagnostic tests [37]. During the development, the checklist was revised based on piloting in a subgroup of 24 HTAs. The checklist was also partially validated by comparing agreements of data extraction between the two reviewers for a random sample of eight HTAs. Subjective judgements are inevitably involved in this type of assessment study and therefore absolute agreements between reviewers are sometimes difficult to achieve, as shown in the current study. This suggests that further validation is needed for the checklist, but our work provides the basis for the development of a fully validated diagnostic-relevant checklist, which is much needed.

5 Conclusions

Given the challenges in conducting clinical studies that fully capture the impact of diagnostic testing on patient outcomes, decision analytic modelling plays a critical role in diagnostic test evaluation. In this study, we have comprehensively assessed model-based economic evaluations of diagnostic tests within recent UK HTA reports. Our review recognised the complexity and difficulty in modelling the cost effectiveness of diagnostic tests.

It also highlights that tools and guidelines are needed to evaluate and promote improved model development, the systematic review of identification of evidence and validation for the evaluation of diagnostics in the future. We call for methods to efficiently examine structural uncertainty in modelling. We suggest methodological development to address the low-quality evidence or lack of evidence on diagnostic test accuracy, especially to quantify the impact of potential interdependency of multiple test combinations in parallel or sequentially within the diagnostic pathway.

Data Availability Statement Full data extraction results can be found as online supplementary material in <https://figshare.com/s/643c9a69d956780eb0a8>.

Acknowledgements This work was undertaken at the Nuffield Department of Primary Care Health Sciences, University of Oxford. It has been presented (oral presentation) at the Methods for evaluating

Medical Tests and Biomarkers (MEMTAB) 2016 Symposium at the University of Birmingham, UK, July 19–20, 2016. It has also been presented (poster) at the Summer 2016 Health Economists’ Study Group meeting in Gran Canarias, Spain, June 21–23, 2016. We are thankful for the helpful comments from audiences of both conferences. This study was part-funded by the National Institute for Health Research (NIHR) School for Primary Care Research (SPCR) (Project No. 269). During the study, Dr Yang was funded by the NIHR Oxford Biomedical Research Centre (BRC) and NIHR Oxford Community Health Care MedTech and In Vitro Diagnostics Cooperative (MIC). Ms Abel is jointly funded by the NIHR Research Capacity Fund (RCF) and the NIHR Oxford MIC. Dr Fanshawe is partly supported by the NIHR Oxford MIC. Dr Shinkins is supported by the NIHR Leeds IVD MIC. The authors’ independence was ensured in designing, interpreting the data, writing and publishing the manuscript. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care in the UK.

Author Contributions YY led the review. She developed the idea and contributed to study design, checklist development, data extraction, data analysis and interpretation of the results and drafted the manuscript. LA extracted all data from the included HTAs and contributed to study design and data analysis. Both LA and JB contributed to checklist development. TF and BS searched and screened eligible HTAs and YY and LA were involved in the final stage of HTA identification according to defined inclusion and exclusion criteria. All authors contributed to interpretation of the results and writing up of the manuscript. All authors read and approved the final manuscript.

Compliance with Ethical Standards

Conflict of interest No conflicts of interest are reported by Yaling Yang (YY), Lucy Abel (LA), James Buchannan (JB), Thomas Fanshawe (TF) and Bethany Shinkins (BS) as authors of this manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. di Ruffano LF, 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. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. *Med Decis Making*. 1988;8(4):279–89.
3. Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, Ioannidis JPA. Diagnostic tests often fail to lead to changes in patient outcomes. *J Clin Epidemiol*. 2014;67(6):612–21.
4. di Ruffano LF, Deeks JJ. Test-treatment RCTs are sheep in wolves’ clothing (Letter commenting on: *J Clin Epidemiol*. 2014; 67: 612–21). *J Clin Epidemiol*. 2016;2016(69):266–7.
5. Pletcher MJ, Pignone M. Evaluating the clinical utility of a biomarker: a review of methods for estimating health impact. *Circulation*. 2011;123(10):1116–24.
6. Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, et al. Modelling in economic evaluation: an unavoidable fact of life. *Health Econ*. 1997;6(3):217–27.

7. Severens JL, van der Wilt G-J. Economic evaluation of diagnostic tests. *Int J Technol Assess Health Care*. 1999;15(3):480–96.
8. Fang C, Otero HJ, Greenberg D, Neumann PJ. Cost-utility analyses of diagnostic laboratory tests: a systematic review. *Value Health*. 2011;14(8):1010–8.
9. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. *Med Decis Mak*. 2008;28(5):650–67. <https://doi.org/10.1177/0272989x08324036>.
10. Peters JL, Cooper C, Buchanan J. Evidence used in model-based economic evaluations for evaluating pharmacogenetic and pharmacogenomic tests: a systematic review protocol. *BMJ Open*. 2015;5(11):e00846. <https://doi.org/10.1136/bmjopen-2015-008465>.
11. Novielli N, Cooper NJ, Abrams KR, Sutton AJ. How is evidence on test performance synthesized for economic decision models of diagnostic tests? A systematic appraisal of health technology assessments in the UK since 1997. *Value Health*. 2010;13(8):952–7.
12. Doble B, Tan M, Harris A, Lorgelly P. Modeling companion diagnostics in economic evaluations of targeted oncology therapies: systematic review and methodological checklist. *Expert Rev Mol Diagn*. 2015;15(2):235–54.
13. Cooper N, Coyle D, Abrams K, Mugford M, Sutton A. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *J Health Serv Res Policy*. 2005;10(4):245–50. <https://doi.org/10.1258/135581905774414187>.
14. Soares MO, Walker S, Palmer SJ, Sculpher MJ. Establishing the value of diagnostic and prognostic tests in Health Technology Assessment. *Med Decis Mak*. 2018;38(4):495–508.
15. MRC. Methods research to support the assessment of diagnostic health technologies and the development of guidelines for diagnostic services. 2018. <https://mrc.ukri.org/funding/how-we-fund-research/opportunities/methods-research-to-support-the-assessment-of-diagnostic-health-technologies/>. Accessed 18 Oct 2018.
16. Shinkins B, Yang Y, Abel L, Fanshawe TR. Evidence synthesis to inform model-based cost-effectiveness evaluations of diagnostic tests: a methodological review of health technology assessments. *BMC Med Res Methodol*. 2017;17(1):56.
17. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment. *PharmacoEconomics*. 2006;24(4):355–71.
18. Ramos MCP, Barton P, Jowett S, Sutton AJ. A systematic review of research guidelines in decision-analytic modeling. *Value Health*. 2015;18(4):512–29.
19. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. *BMC Med*. 2013;11(1):1.
20. Siebert U, Alagoz O, Bayoumi AM, Jahn B, Owens DK, Cohen DJ, et al. State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3. *Value Health*. 2012;15(6):812–20.
21. Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. *Med Decis Mak*. 2011;31(4):675–92. <https://doi.org/10.1177/0272989x11409240>.
22. Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. *Med Decis Mak*. 2012;32(5):722–32.
23. Caro JJ, Briggs AH, Siebert U, Kuntz KM, Force I-SMGRPT. Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Med Decis Mak*. 2012;32(5):667–77.
24. Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Med Decis Mak*. 2012;32(5):733–43.
25. Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M, et al. Conceptualizing a model: a report of the ISPOR-SMDM modeling good research practices task force-2. *Med Decis Mak*. 2012;32(5):678–89.
26. Bryant J, Cooper K, Picot J, Clegg A, Roderick P, Rosenberg W, et al. Diagnostic strategies using DNA testing for hereditary haemochromatosis in at-risk populations: a systematic review and economic evaluation. *Health Technol Assess*. 2009;13:148. <https://doi.org/10.3310/hta13230>.
27. Burch J, Hinde S, Palmer S, Beyer F, Minton J, Marson A, et al. The clinical effectiveness and cost-effectiveness of technologies used to visualise the seizure focus in people with refractory epilepsy being considered for surgery: a systematic review and decision-analytical model. *Health Technol Assess*. 2012;16:163. <https://doi.org/10.3310/hta16340>.
28. Auguste P, Tsertsvadze A, Pink J, Courtney R, Seedat F, Gurung T, et al. Accurate diagnosis of latent tuberculosis in children, people who are immunocompromised or at risk from immunosuppression and recent arrivals from countries with a high incidence of tuberculosis: systematic review and economic evaluation. *Health Technol Assess*. 2016;20(38):1–678.
29. 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. *Health Technol Assess (Winchester, England)*. 2017;21(23):1.
30. Hislop J, Quayyum Z, Flett G, Boachie C, Fraser C, Mowatt G. Systematic review of the clinical effectiveness and cost-effectiveness of rapid point-of-care tests for the detection of genital chlamydia infection in women and men. *Health Technol Assess*. 2010;14:125. <https://doi.org/10.3310/hta14290>.
31. Meads C, Sutton A, Małysiak S, Kowalska M, Zapalska A, Rogozinska E, et al. Sentinel lymph node status in vulval cancer: systematic reviews of test accuracy and decision-analytic model-based economic evaluation. *Health Technol Assess*. 2013. <https://doi.org/10.3310/hta17600>.
32. NICE. Guide to the methods of technology appraisal 2013. London: National Institute for Health and Care Excellence; 2013.
33. NICE. NICE diagnostic guidance. 2018. <https://www.nice.org.uk/guidance/published?type=dg>. Accessed 20 Oct 2018.
34. Westwood M. 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. *Health Technol Assess*. 2015. <https://doi.org/10.3310/hta19440>.
35. Severens JL, Sonke G, Laheij RJF, Verbeek ALM, De Vries Robbé PF. Efficient diagnostic test sequence: applications of the probability-modifying plot. *J Clin Epidemiol*. 2001;54:1228–37. [https://doi.org/10.1016/S0895-4356\(01\)00391-2](https://doi.org/10.1016/S0895-4356(01)00391-2).
36. Mowatt G. The diagnostic accuracy and cost-effectiveness of magnetic resonance spectroscopy and enhanced magnetic resonance imaging techniques in aiding the localisation of prostate abnormalities for biopsy: a systematic review and economic evaluation. *Health Technol Assess*. 2013. <https://doi.org/10.3310/hta17200>.
37. Kip MM, Ijzerman MJ, Henriksson M, Merlin T, Weinstein MC, Phelps CE, et al. Toward alignment in the reporting of economic evaluations of diagnostic tests and biomarkers: the AGREEDT checklist. *Med Decis Mak*. 2018;38(7):778–88.