PharmacoEconomics (2018) 36:823–835
https://doi.org/10.1007/s40273-018-0638-1

SYSTEMATIC REVIEW

The Role of Measurement Uncertainty in Health Technology Assessments (HTAs) of In Vitro Tests

Alison F. Smith[1,2] · Mike Messenger[2,3] · Peter Hall[4] · Claire Hulme[1,2]

Published online: 3 March 2018

Abstract

Introduction Numerous factors contribute to uncertainty in test measurement procedures, and this uncertainty can have a significant impact on the downstream clinical utility and cost-effectiveness of testing strategies. Currently, however, there is no clear guidance concerning if or how such factors should be considered within Health Technology Assessments (HTAs) of tests.

Objective The aim was to provide an introduction to key concepts in measurement uncertainty for the HTA community and to explore, via systematic review, current methods utilised within HTAs.

Methods HTAs of in vitro tests including a model-based economic evaluation were identified via the Centre for Reviews and Dissemination (CRD) HTA database and key reimbursement authority websites. Data were extracted to explore the specific components of measurement uncertainty assessed and methods utilised. The findings were narratively synthesised.

Results Of 107 identified HTAs, 20 (19%) attempted to assess components of measurement uncertainty: 15 did so via some form of pre-model assessment (such as a literature review or laboratory survey); four also included components within the economic model; and one considered measurement uncertainty within the model only. One study quantified the impact of measurement uncertainty on cost-effectiveness and found that this parameter significantly changed the results, but did not impact the overall decision uncertainty.

Conclusion A minority of HTAs identified from this review used various approaches to assess and/or incorporate the impact of measurement uncertainty, indicating that these assessments are feasible. Uncertainty remains around best practice methodology for conducting such analyses; further research is required to ensure that future HTAs are fit for purpose.

This article presents independent research arising from a Doctoral Research Fellowship (Alison Smith, DRF-2016-09-084) supported by the National Institute for Health Research (NIHR). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40273-018-0638-1) contains supplementary material, which is available to authorized users.

✉ Alison F. Smith
a.f.c.smith@leeds.ac.uk

[1] Academic Unit of Health Economics, Leeds Institute of Health Sciences, University of Leeds, Leeds, UK

[2] National Institute of Health Research (NIHR) Leeds In Vitro Diagnostic Co-operative (IVDC), Leeds, UK

[3] Leeds Centre for Personalised Medicine and Health, University of Leeds, Leeds, UK

[4] Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, UK

Key Points for Decision Makers

Variation in test measurement procedures can result in systematic and/or random variation in test results (i.e. measurement uncertainty).

This uncertainty can have a significant impact on the clinical utility and cost-effectiveness of testing strategies, but is not currently routinely considered with Health Technology Assessments (HTAs).

A systematic review identified a minority of HTAs (n = 20/107; 19%) that have used various approaches to incorporate the impact of components of measurement uncertainty within a pre-model assessment (n = 19; such as a literature review or laboratory survey) and/or within the economic model (n = 5).

Uncertainty remains around best practice methodology for conducting such analyses; further research is required to ensure that future HTAs are fit for purpose.

1 Background

All measurements are subject to uncertainty, whether it be determining the distance between two objects, the level of CO_2 in the atmosphere or the pressure exerted within a mechanical system. In vitro clinical tests are no exception. The time of day a sample is taken, mode of sample transportation and time between sample collection and analysis are just a few examples of a multitude of factors that can influence the concentration of substances within a test sample, thereby altering the reported test value and introducing uncertainty.

The consequence of this uncertainty is that any observed test value may be different to the ‘true’ underlying target value one wishes to measure. This can impact the clinical accuracy of a test (the ability of a test to correctly identify patients with and without a given condition) if measured values are incorrectly observed as lying above or below the test cut-off threshold used to determine disease classifications.[1] If, as a consequence, a meaningful proportion of patients receive inappropriate healthcare interventions, patients’ health may be compromised and unnecessary costs accrued. Understanding and quantifying the magnitude of test measurement uncertainty, as well as the subsequent impact on downstream test outcomes, is therefore critical in order to ensure that testing procedures are implemented only when net health benefits are expected to be obtained.

Across the developed world, the established gold-standard tool for informing evidence-based healthcare decisions is the Health Technology Assessment (HTA): a multidisciplinary process to systematically examine the safety, efficacy and cost-effectiveness of new healthcare interventions, and identify any social, organisational and ethical issues concerning adoption [1, 2]. In response to the growing importance of in vitro tests, many HTA and reimbursement authorities now include such technologies within their remit, and some institutions—such as the National Institute for Health and Care Excellence (NICE) in the UK—have established separate programmes of assessment for tests distinct to pharmaceuticals [3, 4]. These assessments typically focus on three key domains: (1) *clinical accuracy*—the ability of a test to correctly identify patients with and without a given condition; (2) *clinical utility*—the subsequent impact of a test on health outcomes; and (3) *cost-effectiveness*—the ability of a test to produce an efficient impact on health outcomes in relation to healthcare expenditure.

The impact of measurement uncertainty within HTA assessments is, in our experience, not routinely considered. Indeed current guidance in this area is unclear: both NICE in the UK and the Canadian Agency for Drugs and Technologies in Health (CADTH)—world leaders in technology assessments—make no mention of measurement uncertainty within their current methodology guidance, for example [4, 5]. The Medical Services Advisory Committee (MSAC) in Australia is the only authority we are aware of that specifies the need to evaluate such evidence, using the associated terminology of *analytic validity* [6]. However, whilst stipulating that such data should be reviewed, MSAC offer no recommendations regarding how these data should be assessed or utilised within subsequent clinical and economic assessments.

In order to establish if and how measurement uncertainty is currently being addressed within HTAs, and in particular within economic evaluations, a systematic review of reports published by internationally recognised HTA agencies [registered with the International Network of Agencies for HTA (INAHTA)] and including an economic decision model was conducted. In addition, for readers unfamiliar with the field of measurement uncertainty, a brief introduction to key concepts in the field is first provided, focusing on the case of quantitative tests (i.e. measuring the quantity or concentration of analyte within a sample, typically assessed against a given disease cut-off

[1] A simulated example of the impact of measurement uncertainty on clinical accuracy is provided in the Electronic Supplementary Material (Sect. 1).

threshold).[2] A corresponding table of relevant terminology can be found in the Electronic Supplementary Material (Sect. 2), and further key texts are recommended for interested readers [7–10].

2 An Introduction to Measurement Uncertainty

2.1 Precision and Trueness

The central components of measurement performance are *precision* [characterised by the absence of random error (i.e. *imprecision*) in measurement] and *trueness* [the absence of systematic error (i.e. *bias*) in measurement]. Increased imprecision and/or bias in measurement results in increased measurement uncertainty.

Imprecision [expressed as a coefficient of variation (CV)[3] or standard deviation (SD)] is explored by observing the degree of dispersion in repeated test measurements [11–13]. The level of imprecision measured depends on how many factors expected to affect test performance (including time, operator, calibration, environment and equipment) are altered during the measurement procedure. Holding all factors constant (i.e. within-batch testing) measures *repeatability*; altering one or more factors within the same laboratory measures *intermediate precision*; whilst conducting testing across different laboratories (in which all factors would be expected to vary) measures *reproducibility*.

Analysis of trueness meanwhile (typically assessed according to % bias, regression analysis or difference plots) relies on comparative analysis of results from the test of interest (the *index test*) versus the 'true' target value. In reality this 'true' value is unknown and must be estimated using a specified *reference test*, ideally based on officially validated test methods or samples of known composition [but often also based on consensus data from external quality assessment (EQA)[4] schemes or established 'gold standard' test results].[5] Alternatively, new tests may be compared against each other (without a reliably proxy for the truth) in order to ascertain the level of between-test discordance.

An important feature in the evaluation of trueness is test *selectivity*: the ability of a test to identify the target analyte of interest as opposed to other sample components. Selectivity depends on the level of obstruction from substances in the test sample which either inhibit the process of binding with the target analyte (i.e. *interference*) or are mistaken for the target analyte, leading to 'unintentional' binding (i.e. *cross-reactivity*).

2.2 Pre-analytical, Analytical and Biological Factors

Both precision and trueness can be affected by numerous factors along the testing pathway, including (1) *biological variation*—fluctuations in the quantity of bodily fluids within an individual over time; (2) variation in *pre-analytical factors*—processes occurring prior to the point of sample analysis; and (3) variation in *analytical factors*—processes occurring at the point of sample analysis. These can be summarised in a 'feather diagram'; the generalised example illustrated in Fig. 1 shows key factors grouped by category and following a (roughly) chronological order from the initial test request through to obtaining the final result.

2.3 Limits and Range

Various limits can be specified which determine the boundaries against which testing is reasonably conducted. These are (1) the limit of blank (LoB), defined as the highest (apparent) quantity of analyte expected to be identified when processing blank samples; (2) the limit of detection (LoD), defined as the lowest quantity reliably distinguish from the LoB; and (3) the limits of quantification (LoQ), defined as the lower and upper quantities a test can measure with a specified level of precision and trueness. Identified limits are routinely used to inform the *reportable range* of a test.

2.4 Summary Measures

Different elements of uncertainty, as illustrated in Fig. 1, may be combined to estimate a summary measure of uncertainty. Two main approaches to this end have been adopted in the literature: total error (TE) and uncertainty of measurement (U_M). Briefly, TE is calculated as the linear sum of bias and imprecision, in which imprecision is multiplied by a 'z factor' to cover a required region of confidence (e.g. at the 95% confidence level, TE = bias + 1.96 * imprecision). U_M on the other hand is a measure of dispersion (i.e. SD), calculated by combining individual

[2] Other types of tests include (a) semi-quantitative tests, based on quantitative measurements which are subsequently grouped into a number of discrete categories (e.g. high/medium/low risk) and (b) qualitative tests, which may or may not be derived from quantitative measurement and report whether or not an analyte/feature is present (i.e. positive/negative result). The metrics of measurement uncertainty adopted for (a) are much the same as for quantitative tests (with potential extra complexity occurring due to the addition of categories); whilst for (b) more simplified assessments are typically required (see [10]).

[3] CV = the ratio of the SD to the mean, multiplied by 100.

[4] Also known as proficiency testing.

[5] Note that these approaches may misrepresent bias if either the major method adopted by participating laboratories in EQA schemes or the gold-standard test are themselves inaccurate.

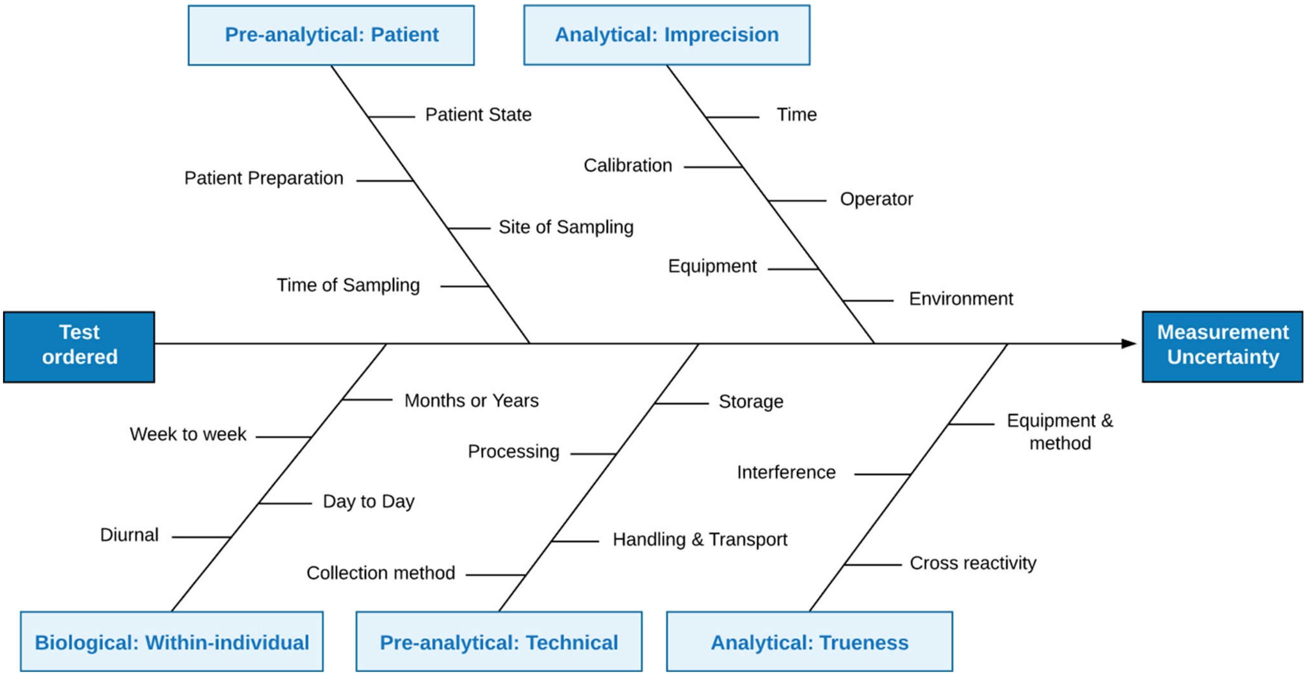

Fig. 1 Feather diagram depicting factors that may contribute to measurement uncertainty

uncertainties occurring along the testing pathway (e.g. using propagation of error rules) multiplied by a 'coverage factor' to similarly capture a specified region of confidence. Whilst TE represents an upper bound on the level of deviation from truth expected to occur in a given measurement, U_M defines a confidence interval around the observed result that is expected to contain the true value. Although there is an ongoing debate within the literature as to the relative merits of each approach [14–17], within the context of this study, both metrics are considered to be viable measures of overall measurement uncertainty.

3 Methods

The review protocol was published in advance on the PROSPERO database (CRD42017056778). The primary source was the Centre for Reviews and Dissemination (CRD)[6] HTA database; this consists of completed and ongoing HTAs from INAHTA-registered HTA authorities (49 at the time of conducting the review)[7] in addition to 20 other CRD-recognised HTA organisations, and includes reports from national reimbursement authorities (e.g. NICE) as well as publically funded research councils [e.g. the UK National Institute for Health Research (NIHR)]. As such it is a principle resource for HTAs expected to directly influence national healthcare decisions.

A search strategy (see the Electronic Supplementary Material, section 3) combining key terms on in vitro tests and economic decision models was developed and run in March 2017. All HTAs including a model-based economic evaluation and evaluating an in vitro test (including diagnostic, screening, prognostic, predictive and monitoring tests) across any disease area, human population or setting and reported since 1999 with a full HTA report available in English were included.

Records were managed using Endnote V 7.2 (Thompson Reuters). All titles and abstracts were screened by a primary reviewer, and 10% were independently screened by a secondary reviewer. Full papers were subsequently screened by the primary reviewer only; any uncertainties regarding inclusions were checked with the secondary reviewer. For studies identified as including an assessment of measurement uncertainty, data were extracted on the specific components assessed and the methods utilised, with 10% of data extraction independently checked by the secondary reviewer. A broad definition of measurement uncertainty was adopted, including all components listed in Fig. 1, as well as data on TE, U_M, limits (LoB, LoD and LoQ), reportable range and test failure rates. Results were narratively synthesised.

[6] Whilst the maintenance of other CRD databases (DARE and NHS EED) ceased in 2015, the HTA database continued to be maintained at the time of conducting this review and into 2018.

[7] As per the INAHTA membership eligibility criteria, these are non-profit organisations assessing healthcare technologies, relating to a regional or national government, funded at least 50% by public sources and providing free access to reports on request (see http://www.inahta.org/).

In addition to the HTA database, online records of key reimbursement authorities expected to be the largest contributors of relevant HTAs (NICE, CADTH and MSAC) were cross-checked by the primary reviewer [18–20]. Citation checking of included studies was also conducted to identify any further relevant HTAs.

4 Results

A total of 107 studies were included (see Fig. 2), and agreement between reviewers at abstract screening was good (k = 0.85).[8] A summary of study characteristics is provided in the Electronic Supplementary Material (section 4). The majority of studies were conducted within the UK (62%), followed by Canada (16%) and Australia (14%), with a gradual rise in the frequency of annual HTA publications since 1999.

Of the 107 identified HTAs, 71 (66%) did not evaluate measurement uncertainty. Sixteen (15%) incorporated data on test failure rates only (e.g. test failures included as an item within a literature review and/or as a parameter within the economic model) and were therefore of limited interest. Twenty studies (19%) considered further components of measurement uncertainty (see Table 1) [21–42]. The majority of these were published from 2009 onwards, and evaluated one or a small number of measurement uncertainty components (including imprecision, trueness, biological variability and pre-analytical or analytical effects) within some form of assessment *prior* to the economic model, such as a literature review or laboratory survey. These evaluations are henceforth denoted ‘pre-model assessments’. Five studies incorporated measurement uncertainty within the economic model itself: four in addition to a pre-model assessment [22, 30, 31, 39]; one within the model only [21]. These studies used a range of techniques—including individual patient simulation and Monte Carlo simulation—to incorporate data on test agreement [39], biological and analytical variability [21, 30, 31] or TE [22] (see Table 2).

5 Discussion

5.1 Review Findings

Despite limited guidance in this area, assessment of test measurement uncertainty has been attempted in a minority of HTAs (n = 20; 19%) indicating that such analyses are feasible.

The majority of studies (n = 19) included measurement uncertainty within some form of pre-model assessment, such as a literature review or laboratory survey. Indeed the frequency of these assessments appears to have been increasing in recent years; this may reflect the fact that more HTAs of tests are being conducted in general, a growing awareness of the importance of measurement uncertainty, and/or increasing availability of relevant data upon which to base such evaluations. On the whole, however, these studies were considered to be partial assessments: most considered one or a limited set of measurement uncertainty components and none formally assessed (i.e. beyond a general discussion) the potential quantitative impact of measurement uncertainty on clinical accuracy or utility.

A small minority of studies (n = 5) utilised data on test measurement uncertainty within the economic model. Of those, the most recent (Stein et al. 2016) was not a direct attempt to account for measurement uncertainty, but rather the authors here utilised between-test discordance data as a means of evaluating additional tests in the model [39]. Meanwhile the oldest study (Marks et al. 2000) is most interesting as an example of what *not* to do [21]. Here the authors simply set the proportion of false positive results equal to a given level of biological and analytical variability (i.e. imprecision), which fails to account for the dependence of test misclassifications on the position of values relative to the test cut-off threshold. In contrast, the approach taken by MSAC correctly accounted for this dependency, by first assigning ‘true’ test values, simulating the addition of measurement uncertainty to generate observed values (in this case, using TE to define a confidence interval around the true value),[9] and then comparing these results against the given cut-off threshold to determine the proportion of misdiagnoses [22]. Similarly the more recent studies by Farmer et al. (2014) and Perera et al. (2015) simulated the addition of uncertainty on top of ‘true’ baseline values; in this case also accounting for the impact of uncertainty in the rate of baseline health and disease progression within repeated testing scenarios using regression analysis of longitudinal individual patient data [30, 31]. A key drawback with this approach, however, concerns the data and computational resources required, which would likely pose challenges within typical HTA timelines.

Only the MSAC study explicitly explored the impact of variation in measurement uncertainty on cost-effectiveness [22]. Here the authors found that, whilst variation in TE

[8] Note all discrepancies were a result of the primary reviewer being more inclusive than the secondary reviewer.

[9] A key question for this study, however, concerns the validity of using TE (which combines both random and systematic error) to assign both sides of a confidence interval; assuming bias acts in a fixed direction, for example, this approach will overestimate uncertainty.

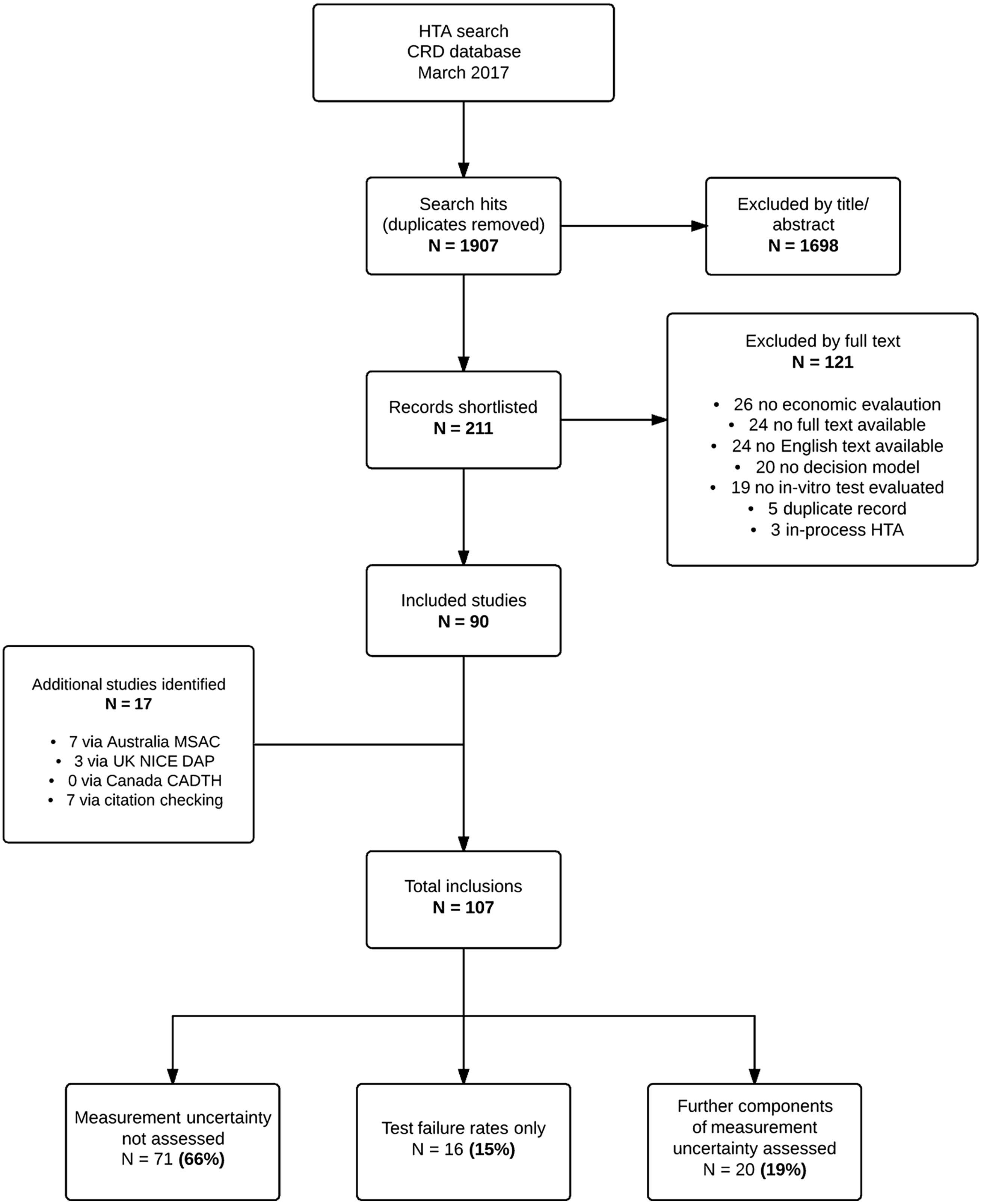

Fig. 2 PRISMA flow diagram of search results. *CADTH* Canadian Agency for Drugs and Technologies in Health, *CRD* Centre for Reviews and Dissemination, *DAP* Diagnostics Assessment Programme, *HTA* Health Technology Assessment, *MSAC* Medical Services Advisory Committee, *NICE* National Institute for Health and Care Excellence

Table 1 Summary of HTA reports (*n* = 20) including components of measurement uncertainty in a pre-model assessment and/or the economic decision model

Study	Test characteristics			Pre-model assessments		Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Method	Components of measurement uncertainty assessed	
Marks et al. 2000 (UK) [21]	–	Cardiology	Screening	–	–	Yes
MSAC 2001 (AUS) [22]	POCT: clinician-led	Cardiology	Prognosis	Systematic review	Trueness (% bias); precision (repeatability and reproducibility); TE; analytical effects (site, operator and sample type)	Yes
Gailly et al. 2009 (BEL) [23]	POCT: self-led	Haematology	Monitoring	Systematic review	Precision (repeatability and intermediate); test failures	–
Pearson et al. 2010 (UK) [24, 25]	POCT: clinician-led	Gastro	Diagnosis	Systematic review	Biological variability; distribution in faeces; faecal matrix; interference; stability; patient compliance; normal range	–
MAS 2010 (CA) [26]	–	Cancer	Prognosis	Systematic review	Precision (intermediate and reproducibility); test failures	–
Ward et al. 2013 (UK) [27]	–	Cancer	Prognosis	Systematic review	Precision (intermediate and reproducibility); trueness (concordance)	–
Westwood et al. 2014 (UK) [28]	–	Cancer	Predictive	Systematic review + survey	Proportion of tumour cells needed; test failures	–
Westwood et al. 2014 (UK) [29]	–	Cancer	Predictive	Systematic review + survey	Proportion of tumour cells needed; LoD; test failures	–
Farmer et al. 2014 (UK) [30]	–	Diabetes	Screening	Analysis of IPD	Biological and analytical variation	Yes
Perera et al. 2015 (UK) [31]	–	Cardiology	Monitoring	Analysis of IPD	Biological and analytical variation	Yes
Sharma et al. 2015 (UK) [32]	POCT: self-led	Haematology	Monitoring	Literature review	Precision (reproducibility); trueness (*r* correlation coefficient)	–
Nicholson et al. 2015 (UK) [33]	–	Cancer	Diagnosis	Systematic review	Precision (intermediate and reproducibility); trueness (recovery); LoB, LoD, LoQ; interference; linearity; range; pre-analytical effects; stability; test failures	–
MSAC 2015 (AUS) [34, 35]	–	Cancer	Prognosis	Literature review	Selectivity	–
Kessels et al. 2015 (AUS) [36]	–	Pregnancy care and screening	Diagnosis	Systematic review	Selectivity; test failures	–

Table 1 continued

Study	Test characteristics			Pre-model assessments		Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Method	Components of measurement uncertainty assessed	
Harnan et al. 2015 (UK) [37]	POCT: self-led	Other (asthma)	(1) Diagnosis, (2) monitoring	Systematic review	Trueness (Bland-Altman analysis, correlation coefficients); test failures	–
Freeman et al. 2015 (UK) [38]	–	Cancer	Monitoring	Systematic review	Trueness (Bland-Altman analysis, Deming regression); test failures	–
Stein et al. 2016 (UK) [39]	–	Cancer	Prognosis	Pathology study	Trueness (Kappa statistic, discordance)	Yes
Hay et al. 2016 (UK) [40]	POCT: clinician-led	Other (urology)	Diagnosis	Clinical study	Trueness (Kappa statistic); test failures	–
Freeman et al. 2016 (UK) [41]	–	Gastro	Monitoring	Systematic review	Trueness (Bland-Altman analysis, Cohen's Kappa); test failures	–
Auguste et al. 2016 (UK) [42]	–	Infection (TB)	Diagnosis	Systematic review	Trueness (Kappa statistic, discordance); test failures	–

Further details of modelling studies provided in Table 2

AUS Australia, *BEL* Belgium, *CA* Canada, *Gastro* gastroenterology, *HTA* Health Technology Assessment, *LoB* limit of blank, *LoD* limit of detection, *LoQ* limits of quantification, *MAS* Medical Advisory Secretariat, *MSAC* Medical Services Advisory Committee, *POCT* point of care test, *TB* Tuberculosis, *TE* total error, *UK* United Kingdom

was not expected to alter the overall decision uncertainty (since all results remained above the specified 100,000 Australian dollars (AUS$) cost-effectiveness threshold), it was expected to have a significant impact on the base case results (resulting in a 24% drop from AUS$133,934 to AUS$101,419 per life year gained when reducing TE from 8% to 0%). This example clearly illustrates the potential impact of measurement uncertainty on cost-effectiveness, which could feasibly be of significant importance in scenarios exhibiting baseline results closer to the cost-effectiveness threshold.

5.2 Future Research

Whilst this review has identified previous HTA assessments of measurement uncertainty, outstanding uncertainties and issues require consideration before general guidance in this area can be feasibly implemented. For pre-model assessments, future studies would benefit from (currently lacking) guidance on best practice methods to conduct, synthesise and report literature reviews of measurement uncertainty, as well as appropriate methodology for utilising data from alternative resources (e.g. laboratory surveys, EQA reports and pathology studies). For economic evaluations, future case studies could explore particular considerations of interest including the following: the relative importance of various components of measurement uncertainty for different kinds of tests (e.g. diagnostic vs monitoring; laboratory vs point of care test; quantitative vs qualitative etc.); the use of alternative summary measures versus individual components of measurement uncertainty; and the feasibility of different approaches. In addition, outside the scope of HTAs, we are aware of several studies that have utilised Monte Carlo simulation methods to explore the impact of measurement uncertainty on clinical accuracy as a means of identifying test analytical performance goals (i.e. maximum allowable imprecision and/or bias in order to maintain clinical

Table 2 Details of HTA reports (n = 5) including components of measurement uncertainty within the economic model

Study	Model details			Assessment of measurement uncertainty				
	Tests evaluated	Type of model	Base case results	Component(s) included	Source of evidence	Value(s) used	Method of incorporation	Impact on cost-effectiveness results
Marks et al. 2000 (UK) [21]	Screening test for hypercholesterolaemia (universal, opportunistic and case finding strategies)	Decision Tree	Cost per LYG: £14,842–£78,060 (universal); £21,106–£70,009 (opportunistic); £3300–£4914 (case finding)	Biological and analytical variation	Individual cited paper (no formal review)	Base case: coefficient of biological and analytical variation = 6.5%	Rate of false positives in the model set equal to the reported coefficient of biological and analytical variability	Not assessed
MSAC 2001 (AUS) [22]	Cholesterol screening POCT for coronary heart disease (vs standard lab test)	Decision Tree	Incremental cost per LYG: AUS$133,934	TE (% bias + 1.96 * %CV)	Systematic review. Calculation used average of reported CVs and total % biases	Base case: TE = 8%. Sensitivity analysis: TE = 0%, 4%, 11%	10,000 Monte Carlo simulations: (1) patients assigned a 'true' cholesterol level; (2) two observed results generated based on CI of ±8%; (3) diagnosis based on average of the two results against threshold of 6.5 mmoL/L; (4) probability of misclassifications based on weighted average across cholesterol range (2.5–9.4 mmol/L)	Incremental cost (AUS$) per LYG: $101,419 (TE = 0%); $115,615 (TE = 4%); $133,934 (TE = 8%); $151,378 (TE = 11%)

Table 2 continued

Study	Model details			Assessment of measurement uncertainty				
	Tests evaluated	Type of model	Base case results	Component(s) included	Source of evidence	Value(s) used	Method of incorporation	Impact on cost-effectiveness results
Farmer et al. 2014 (UK) [30]	Screening test (ACR) for kidney disease in diabetes patients (1-, 2-, 3-, 4- and 5-yearly intervals)	Individual patient simulation	Incremental cost per QALY (2 vs 1 year): £9601 (type 1 diabetes; SD = 34,112); £606 (type 2 diabetes; SD = 1782)	Biological and analytical variation	Retrospective analysis of longitudinal IPD databases	Estimated SD of within-measurement variability: type 1 diabetes = 0.79 (95% CI 0.73–0.86); type 2 = 0.85 (0.74–1.00). Both >100% CV	A longitudinal hierarchical model for log(ACR) was obtained from the IPD. Individual simulations as follows: (1) a representative population (*n* = 75,000) was generated; (2) baseline log(ACR) and progression rates simulated and used to calculate annual true log(ACR) post-diagnosis; (3) observed ACR derived by adding biological and analytical variation; (4) clinical accuracy assessed against gender-specific cut-off thresholds	Not assessed
Perera et al. 2015 (UK) [31]	Lipid monitoring tests for patients at risk or with cardiovascular disease	Individual patient simulation	Annual monitoring dominated all other strategies	Biological and analytical variation	Retrospective analysis of longitudinal IPD databases	Estimated SD of within-measurement variability across tests: 0.12–0.35 (male population); 0.14–0.37 (females)	Same method as above [longitudinal regression of IPD + individual simulations to model impact of progression and biological and analytical variation over time].	Not assessed
Stein et al. 2016 (UK) [39]	ODX (+ additional tests) to guide use of adjuvant chemotherapy in breast cancer patients (vs chemotherapy for all)	Decision tree + cohort Markov model	Net health benefit (QALYs) for tests vs chemotherapy for all: 6.99 QALYs (ODX); 7.16–7.20 (alternative tests)	Test discordance	*De novo* clinical pathology study	Kappa statistics for tests vs ODX: 0.40–0.53. Agreement with ODX ranged from all tests agreeing in 39% of cases to no test agreeing in 4% of cases	Predictive effect of ODX for recurrence-free survival in the model was derived from a historic ODX clinical trial. For the alternative tests, extra uncertainty was introduced in the model according to the degree of discordance between tests vs. ODX	Not assessed

Net health benefit (QALYs) = incremental QALYs − (incremental costs/cost-effectiveness threshold)

ACR albumin-to-creatinine ratio, *AUS* Australia, *CI* confidence interval, *CV* coefficient of variation, *HTA* Health Technology Assessment, *IPD* individual patient data, *LYG* life year gained, *MSAC* Medical Services Advisory Committee, *ODX* oncotype DX, *POCT* point of care test, *QALY* quality-adjusted life year, *SD* standard deviation, *TE* total error, *UK* United Kingdom accuracy) [43–46]; extending HTA evaluations to include similar assessments (which could feasibly be based on cost-effectiveness outputs in addition to clinical accuracy) is another potential avenue for exploration in future studies, which could further extend the clinical impact of HTAs.

5.3 Strengths and Limitations

This review focused on reports from INAHTA-registered and CRD-recognised HTA authorities, which are expected to reflect best practice methodologies and directly influence healthcare reimbursement and adoption decisions. Taking a broader perspective and considering all kinds of evidence which may inform healthcare decision making (e.g. standalone cost-effective assessments) would likely yield additional findings of interest; as may expanding the search to before 1999 (although the majority of relevant studies identified were from 2009 onwards) and non-English languages. Nevertheless, this is the first systematic review of its kind, which highlights both advances and issues in current approaches to HTAs and can help to inform the direction of future research and guidance in this area. Furthermore, whilst the focus of this study was on in vitro tests, many of the issues here highlighted will be of relevance to pharmacological studies utilising tests as surrogate outcome measures, as well as evaluations of imaging and in vivo technologies.

6 Conclusions

Various approaches have been adopted within a minority of HTAs to assess test measurement uncertainty. Further research is required to identify best practice methodology for conducting such analyses and to ensure that future HTAs are fit for purpose.

Data Availability Statement The dataset generated during the current study is available in the Research Data Leeds Repository (https://doi.org/10.5518/324).

Acknowledgements The authors would like to thank members of the doctoral Scientific Advisory Board (SAB) for this project, who provided feedback on the initial project plan and drafts of the paper [Professor Christopher McCabe (Institute of Health Economics, Alberta, Canada), Bethany Shinkins (University of Leeds, UK), Christopher Bojke (University of Leeds, UK), Rebecca Kift (Leeds Teaching and Hospitals Trust, UK), Joy Allen (University of Newcastle, UK), Christopher Hyde (University of Exeter, UK), Jon Deeks (University of Birmingham, UK), James Turvill (York Teachings Hospital NHS Foundation Trust, UK) and Cathie Sturgeon (UK NEQAS)], and patient and public involvement (PPI) members, who provided feedback on the initial project proposal. We would also like to thank Rocio Lopez (University of Leeds, UK), who provided technical advice in developing the search strategy.

Author contributions All authors contributed to the development of the protocol for this project. AS was responsible for conducting the searches, screening, data extraction and analysis, and took leading responsibility for writing the paper. CH was partly responsible for screening and data extraction (secondary reviewer). CH, PH and MM all provided supervision throughout the project and contributed to the paper by providing comments and feedback on successive manuscript drafts.

Funding AS is supported by an NIHR Doctoral Fellowship, DRF-2016-09-084. In addition, AS and MM are supported by the NIHR infrastructure at Leeds.

Compliance with Ethical Standards

Conflict of interest AS, MM, PH and CH declare that they have no competing interests.

Ethical approval As this is a review of secondary research in the public domain ethical approval was not required.

References

1. Banta D, Jonsson E. History of HTA: introduction. Int J Technol Assess Health Care. 2009;25(S1):1–6.
2. World Health Organisation (WHO). Health technology assessment. 2017. http://www.who.int/medical_devices/assessment/en/. Accessed 08 Aug 2017.
3. Newland A. NICE diagnostics assessment programme. Ann R Coll Surg Engl. 2011;93(5):412.
4. National Institute of Health and Clinical Excellence (NICE). Diagnostics Assessment Programme manual. 2011. https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-diagnostics-guidance. Accessed 8 July 2017.
5. Canadian Agency for Drugs and Technologies in Health (CADTH). 2017. Guidelines for the economic evaluation of health technologies: Canada. 4th ed. Ottowa: CADTH.
6. Medical Service Advisory Committee (MSAC). 2017. Technical guidelines for preparing assessment reports for the Medical Services Advisory Committee—Service Type: Investigative (version 3.0).
7. Ellison SL, Rosslein M, Williams A. Quantifying uncertainty in analytical measurement. Quantifying uncertainty in analytical measurement: Eurachem; 2000.
8. Westgard JO, Barry PL, Quam EF, Ehrmeyer SS. Basic method validation: training in analytical quality management for healthcare laboratories: Westgard Quality Corporation; 1999.
9. Price CP, Bossuyt PM, Bruns DE. Tietz fundamentals of clinical chemistry. Medicine. 1976;3:1.
10. Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, et al. A standardized framework for the validation and verification of clinical molecular genetic tests. Eur J Hum Genet. 2010;18(12):1276–88.

11. International Organization for Standardization (ISO). Accuracy (trueness and precision) of measurement methods and results [ISO 5725:1-6]. Geneva, Switzerland; 1994.
12. Clinical and Laboratory Standards Institute (CLSI). Evaluation of precision of quantitative measurement procedures; approved guideline—third edition. CLSI document EP05-A3. Wayne: Clinical and Laboratory Standards Institute; 2014.
13. Westgard J. Method validation: the replication experiment. Basic method validation, 3rd ed Madison: Westgard QC, Inc.; 2008:114–22.
14. Panteghini M, Sandberg S. Total error vs. measurement uncertainty: the match continues. Clin Chem Lab Med (CCLM). 2016;54(2):195–6.
15. Oosterhuis WP, Theodorsson E. Total error vs. measurement uncertainty: revolution or evolution? Clini Chem Lab Med (CCLM). 2016;54(2):235–9.
16. Rozet E, Rudaz S, Marini R, Ziemons E, Boulanger B, Hubert P. Models to estimate overall analytical measurements uncertainty: assumptions, comparisons and applications. Anal Chim Acta. 2011;702(2):160–71.
17. Kallner A. Is the combination of trueness and precision in one expression meaningful? On the use of total error and uncertainty in clinical chemistry. Clinical Chemistry and Laboratory Medicine (CCLM). 2016;54(8):1291–7.
18. National Institute of Health and Care Excellence (NICE). Guidance and advice list. 2017. https://www.nice.org.uk/guidance/published?type=dg. Accessed Mar 2017.
19. Canadian Agency for Drugs and Technologies in Health (CADTH). 2017. https://www.cadth.ca/. Accessed Mar 2017.
20. Medical Services Advisory Committee (MSAC). 2017. http://www.msac.gov.au/. Accessed Mar 2017.
21. Marks D, Wonderling D, Thorogood M, Lambert H, Humphries SE, Neil HAW. Screening for hypercholesterolaemia versus case finding for familial hypercholesterolaemia: a systematic review and cost-effectiveness analysis. Health Technol Assess. 2000;4(29):1–123. https://doi.org/10.3310/hta4290.
22. Medical Service Advisory Committee. Evaluation of near patient cholesterol testing using the Cholestech LDX [MSAC Assessment Report 1026]. 2001. http://www.msac.gov.au. Accessed Apr 2017.
23. Gailly J, Gerkens S, Van Den Bruel A, Devriese S, Obyn C, Cleemput I. Use of point-of care devices in patients with oral anticoagulation: a Health technology Assessment. Health Technology Assessment (HTA). Brussels: Belgian Health Care Knowledge Centre (KCE). 2009. KCE Reports vol 117C. D/2009/10.273/49. https://kce.fgov.be/sites/default/files/atoms/files/d20091027349.pdf. Accessed Apr 2017.
24. Pearson S, Whitehead S, Hutton J. Evidence review: value of calprotectin in screening out irritable bowel syndrome. London: Centre for Evidence-based Purchasing (CEP), 2010 Contract No.: CEP09026.
25. Whitehead SJ, Hutton J. Economic report: value of calprotectin in screening out irritable bowel syndrome. London: Centre for Evidence-based Purchasing (CEP), 2010 Contract No.: CEP09041.
26. Medical Advisory Secretariat. Gene expression profiling for guiding adjuvant chemotherapy decisions in women with early breast cancer: an evidence-based and economic analysis. Ont Health Technol Assess Ser. 2010;10(23):1–57.
27. Ward S, Scope A, Rafia R, Pandor A, Harnan S, Evans P, et al. Gene expression profiling and expanded immunohistochemistry tests to guide the use of adjuvant chemotherapy in breast cancer management: a systematic review and cost-effectiveness analysis. Health Technol Assess. 2013;17(44):1–302. https://doi.org/10.3310/hta17440.
28. Westwood M, Joore M, Whiting P, Asselt T, Ramaekers B, Armstrong N, et al. Epidermal growth factor receptor tyrosine kinase (EGFR-TK) mutation testing in adults with locally advanced or metastatic non-small cell lung cancer: a systematic review and cost-effectiveness analysis. Health Technol Assess. 2014;18(32):1–166. https://doi.org/10.3310/hta18320.
29. Westwood M, Asselt T, Ramaekers B, Whiting P, Joore M, Armstrong N, et al. KRAS mutation testing of tumours in adults with metastatic colorectal cancer: a systematic review and cost-effectiveness analysis. Health Technol Assess. 2014;18(62):1–132. https://doi.org/10.3310/hta18620.
30. Farmer AJ, Stevens R, Hirst J, Lung T, Oke J, Clarke P, et al. Optimal strategies for identifying kidney disease in diabetes: properties of screening tests, progression of renal dysfunction and impact of treatment-systematic review and modelling of progression and cost-effectiveness. Health Technol Assess. 2014;18(14):1–127. https://doi.org/10.3310/hta18140.
31. Perera R, McFadden E, McLellan J, Lung T, Clarke P, Pérez T, et al. Optimal strategies for monitoring lipid levels in patients at risk or with cardiovascular disease: a systematic review with statistical and costeffectiveness modelling. Health Technol Assess. 2015;19(100):1–442. https://doi.org/10.3310/hta191000.
32. Sharma P, Scotland G, Cruickshank M, Tassie E, Fraser C, Burton C, et al. The clinical effectiveness and cost-effectiveness of point-of-care tests (CoaguChek system, INRatio2 PT/INR monitor and ProTime Microcoagulation system) for the self-monitoring of the coagulation status of people receiving long-term vitamin K antagonist therapy, compared with standard UK practice: systematic review and economic evaluation. Health Technol Assess. 2015;19(48):1–172. https://doi.org/10.3310/hta19480.
33. Nicholson A, Mahon J, Boland A, Beale S, Dwan K, Fleeman N, et al. The clinical effectiveness and cost-effectiveness of the PROGENSA® prostate cancer antigen 3 assay and the Prostate Health Index in the diagnosis of prostate cancer: a systematic review and economic evaluation. Health Technol Assessment. 2015;19(87):1–192.
34. Medical Service Advisory Committee. Clinical utility card for heritable mutations which increase risk in breast and/or ovarian cancer. Commonwealth of Australia: Medical Services Advisory Committe (MSAC); 2015.
35. Medical Service Advisory Committee. Economic evaluation of BRCA mutations testing of affected individuals and cascade testing. Commonwealth of Australia: Medical Service Advisory Committe (MSAC); 2015.
36. Kessels SJM, Morona JK, Mittal R, Vogan A, Newton S, Schubert C, et al. Testing for hereditary mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Commonwealth of Australia, Canberra, ACT: 2015 Assessment Report 1216.
37. Harnan SE, Tappenden P, Essat M, Gomersall T, Minton J, Wong R, et al. Measurement of exhaled nitric oxide concentration in asthma: a systematic review and economic evaluation of NIOX MINO, NIOX VERO and Nobreath. Health Technol Assess. 2015;19(82):1–330. https://doi.org/10.3310/hta19820.
38. Freeman K, Connock M, Cummins E, Gurung T, Taylor-Phillips S, Court R, et al. Fluorouracil plasma monitoring: systematic review and economic evaluation of the My5-FU assay for guiding dose adjustment in patients receiving fluorouracil chemotherapy by continuous infusion. Health Technol Assess. 2015;19(91):1–321. https://doi.org/10.3310/hta19910.
39. Stein RC, Dunn JA, Bartlett JMS, Campbell AF, Marshall A, Hall P, et al. OPTIMA prelim: a randomised feasibility study of personalised care in the treatment of women with early breast cancer. Health Technol Assess. 2016;20(10):1–201. https://doi.org/10.3310/hta20100.
40. Hay AD, Birnie K, Busby J, Delaney B, Downing H, Dudley J, et al. The Diagnosis of Urinary Tract infection in Young children

(DUTY): a diagnostic prospective observational study to derive and validate a clinical algorithm for the diagnosis of urinary tract infection in children presenting to primary care with an acute illness. Health Technol Assess. 2016;20(51):1–294. https://doi.org/10.3310/hta20510.

41. Freeman K, Connock M, Auguste P, Taylor-Phillips S, Mistry H, Shyangdan D, et al. Clinical effectiveness and cost-effectiveness of use of therapeutic monitoring of tumour necrosis factor alpha (TNF-α) inhibitors [LISA-TRACKER® enzyme-linked immunosorbent assay (ELISA) kits, TNF-α-Blocker ELISA kits and Promonitor® ELISA kits] versus standard care in patients with Crohn's disease: systematic reviews and economic modelling. Health Technol Assess. 2016;20(83):1–288.
42. Auguste P, Tsertsvadze A, Pink J, Court 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. https://doi.org/10.3310/hta20380.
43. Lyon AW, Kavsak PA, Lyon OA, Worster A, Lyon ME. Simulation models of misclassification error for single thresholds of high-sensitivity cardiac troponin I due to assay bias and imprecision. Clin Chem. 2017;63(2):585–92.
44. Wilinska ME, Hovorka R. Glucose control in the intensive care unit by use of continuous glucose monitoring: what level of measurement error is acceptable? Clin Chem. 2014;60(12):1500–9.
45. Langlois MR, Descamps OS, van der Laarse A, Weykamp C, Baum H, Pulkki K, et al. Clinical impact of direct HDLc and LDLc method bias in hypertriglyceridemia. A simulation study of the EAS-EFLM Collaborative Project Group. Atherosclerosis. 2014;233(1):83–90.
46. Boyd JC, Bruns DE. Monte Carlo simulation in establishing analytical quality requirements for clinical laboratory tests: meeting clinical needs. Methods Enzymol. 2009;467:411–33.