Leeds Institute of
Health Sciences

FACULTY OF MEDICINE AND HEALTH

UNIVERSITY OF LEEDS

WORKING PAPER NUMBER: WP17_5

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]

AFFILIATIONS:

1. Academic Unit of Health Economics, Leeds Institute of Health Sciences, University of Leeds, United Kingdom (UK).
2. National Institute of Health Research (NIHR) Diagnostic Evidence Cooperative (DEC) Leeds, UK.
3. Leeds Centre for Personalised Medicine and Health, University of Leeds, UK.
4. Edinburgh Cancer Research Centre, University of Edinburgh, Edinburgh, UK.

CORRESPONDING AUTHOR CONTACT DETAILS:

Alison F. Smith

Leeds Institute of Health Sciences

Level 11, Worsley Building

Clarendon Way

Leeds

LS2 9NL

Tel: 0131 343 30847

Email: a.f.c.smith@leeds.ac.uk

DISCLAIMER

This series enables staff and student researchers based at or affiliated with the AUHE to make recent work and work in progress available to a wider audience. The work and ideas reported here may not always represent the final position and as such may sometimes need to be treated as work in progress. The material and views expressed in the series are solely those of the authors and should not be quoted without their permission.

JEL classification: I10 (Health, General); C50 (Modeling, General)

Additional Information

Funding & Disclaimer: AS is supported by a National Institute for Health Research (NIHR) Doctoral Fellowship DRF-2016-09-084. 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.

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

Competing Interests: The authors declare that they have no competing interests.

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 final draft of the paper [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 Rodriguez Lopez (University of Leeds, UK), who provided technical advice in developing the search strategy, and Professor Christopher McCabe (University of Alberta, Canada) who provided feedback on the initial and final draft of the paper.

Keywords: Health Technology Assessment, Systematic Review, Measurement Uncertainty, Economic Evaluation, In-vitro techniques.

Abstract

Introduction: Numerous factors contribute to uncertainty in test measurement procedures, and this uncertainty can 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.

Aim: To provide an introduction to key concepts in measurement uncertainty and explore, via systematic review, current methods utilised within HTAs for the assessment of measurement uncertainty.

Methods: HTAs of in-vitro tests including a model-based economic evaluation were identified via the Centre for Reviews and Dissemination (CRD) HTA database, HTA authority websites and citation checking. Data extraction was conducted to explore the specific components of measurement uncertainty assessed and methods utilised, and results were narratively synthesised.

Results: Of 107 identified HTAs, 20 (19%) attempted to assess components of measurement uncertainty: 15 included an ‘intermediate’ assessment (e.g. literature review or laboratory survey); 4 also included components within the economic model; and 1 considered measurement uncertainty within the model only. The specific components assessed and methods adopted differed across studies. In particular, several techniques were employed within the economic models to incorporate data on test agreement, total error or biological and analytical variability.

Conclusion: Various approaches have been adopted within a minority of HTAs to attempt to capture the impact of measurement uncertainty on test outcomes; uncertainty remains around when such assessments are required and appropriate methodology for conducting analyses, particular within economic evaluations.

Background

In-vitro tests have become an indispensable feature of modern healthcare. As well as informing initial diagnoses, tests are increasingly used to monitor ongoing disease status and inform personalised treatment decisions according to predictions of treatment response and toxicity. The past decade has seen significant growth in the number of tests being developed. In the field of genetic testing, for example, over 51,000 tests are currently available for more than 10,000 conditions in the USA alone [1]. Consequently tests command an increasing proportion of the financial market: in 2016 the global market share for in-vitro diagnostics (dominated by USA and Europe) was estimated at US$ 60.22 billion, and this is expected to grow at a steady rate to $78.74 billion by 2021 [2].

The key challenge for healthcare decision makers is determining which tests should be adopted into routine clinical practice, often in the face of restricted healthcare budgets. 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, organizational and ethical issues concerning adoption [3, 4]. Since their emergence in the 1970’s, the role of HTAs has rapidly increased – there are now over 50 HTA agencies registered worldwide, affecting decision making for over 1 billion people across 33 countries [5]. 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 [6, 7].

There are many characteristics of tests which warrant a different approach to evaluation compared to their pharmaceutical counterparts. Most importantly, tests do not directly impact patient health outcomes but rather exert an indirect influence by informing treatment management decisions: the value of a test thus depends both on the ability to provide correct information on disease classifications, and the subsequent ability of that information to produce a change in health care management that leads to a meaningful impact on overall health. HTAs of tests therefore focus on the assessment of (i) *clinical accuracy* – the ability of a test to correctly identify patients with and without a given condition (usually measured according to clinical sensitivity and specificity or predictive values for diagnostic tests, and hazard/odds ratios or relative risks for prognostic tests) and (ii) *clinical utility* – the subsequent impact of a test on health outcomes.

Typically greater investment has been devoted by test manufacturers into developing evidence on clinical accuracy, with utility often estimated according to intermediate outcomes such as treatment management changes. Methodology research in this field has similarly centred on the appropriate evaluation of clinical accuracy, whether it be within the construct of primary studies [8-13] or reviews [14-19]. Particular focus has concerned the statistical estimation of accuracy within meta-analyses – a process that is complicated by the existence of correlation between test sensitivity, specificity and the cut-off threshold, as well as the common occurrence of significant heterogeneity and missing data [20-23].

Analysis of *cost-effectiveness* (the ability of a test to produce an efficient impact on health outcomes in relation to cost) also plays an increasing role within HTAs, particularly across European countries operating within publically funded and essentially ‘fixed’ healthcare budgets. In the common absence of largescale clinical trials evaluating patient outcomes, economic evaluations of tests frequently utilise the ‘linked evidence’ approach: estimating long term cost and health consequences by linking test accuracy data with treatment effectiveness and resource use data, within the construct of an economic decision model [24, 25]. Research here has here focused on the appropriate utilisation of clinical accuracy data within economic decision models, as well as identifying nuances required in the formulation of models for tests compared to pharmaceuticals [26-28].

In contrast, an area rarely explored from an HTA perspective concerns the assessment of test *measurement uncertainty*. Certainty in test measurement relates to the ability of a test to reliably and accurately measure the *measurand* (i.e. substance) of interest in the test sample. For example, when we conduct a blood glucose test, how sure can we be that the value that the test reports is representative of the true blood glucose value we wish to measure – or, in other words, how much might we expect the observed value to differ from the true value? Answering this question is not a straightforward task: there are many different components which may contribute to uncertainty in measurement, and different methods may be utilised to combine individual elements of uncertainty into an aggregated summary statistic.

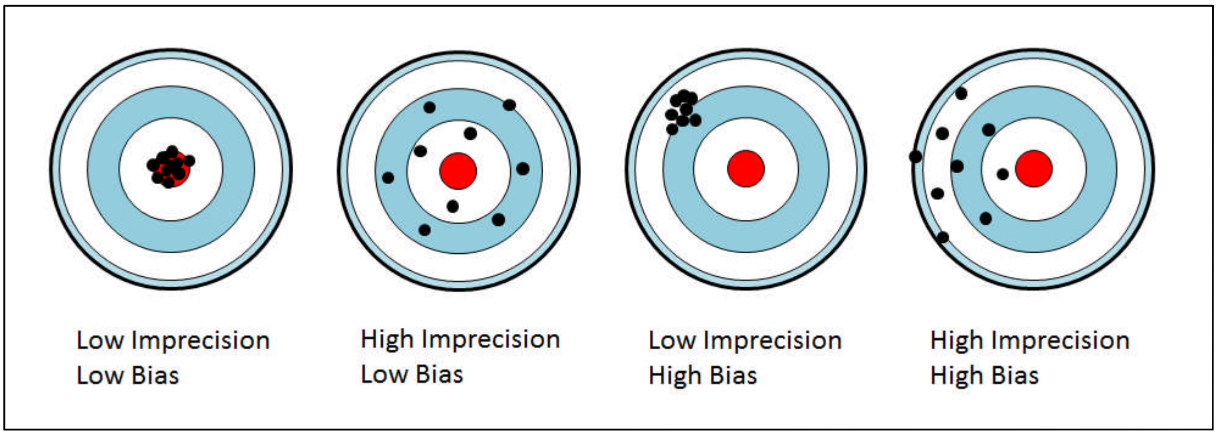
For readers unfamiliar with the field of measurement uncertainty an introduction to key concepts and relevance to the HTA context is provided in the section below. A corresponding list of relevant terminology is provided in the Appendix. It should be noted here that nomenclature in this field is notoriously inconsistent – *analytical validity, technical performance*, *measurement error* and *accuracy* are just a few terms that have been used to describe broadly similar concepts within the field of metrology (the science of measurement), and this inconsistency pervades through to the individual components of measurement uncertainty. For simplicity we use the umbrella term of ‘measurement uncertainty’ and signpost synonymous terms for individual components within the Appendix.

Introduction to Measurement Uncertainty: Key Concepts

1. *Precision and Trueness*

The central components of measurement uncertainty are *precision* (the closeness of agreement between repeated tests) and *trueness* (the closeness of agreement between observed test results and the underlying 'true' value). Precision is characterised by the absence of *imprecision* (i.e. random error) in measurement, whilst trueness is characterised by the absence of *bias* (i.e. systematic error) in measurement. These concepts are illustrated using an example of markers on a bullseye board in Figure 1. In the absence of imprecision and bias, the markers are closely aligned around the bullseye target: introducing imprecision leads to more widely scattered results, whilst introducing bias leads to a shift in the central point around which the points are clustered.

Figure 1. Bullseye illustration of precision and bias



Precision is routinely explored by observing the level of dispersion (i.e. imprecision) in repeated test measurements, which is expressed as a coefficient of variation (CV)[1] or standard deviation (SD) [29-31]. Various levels of precision may be measured, depending on how many factors expected to affect test performance are altered during the measurement

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

procedure: the key variables of consideration – denoted here as ‘*m* factors’ – are time, operator, calibration, environment and equipment. The most elementary form of precision, *repeatability,* concerns the level of variation in results when all *m* factors are held constant (i.e. conducting repeated tests one after another in the same batch or run). Alternatively, maintaining testing within the same laboratory but altering one or more *m* factors provides a measure of *intermediate precision*. Finally the highest measure of precision, *reproducibility*, relates to observed variation when conducting repeated testing across different laboratories in which scenario all of the *m* factors would be expected to vary. The relationship between precision and measurand concentration can be illustrated using a “precision profile” plot, which commonly exhibits a U-shaped profile (i.e. high precision at low and high concentration levels, with lower precision within the reportable range of the test).

Assessment of bias relies on comparative analysis of results from the new test of interest (termed the index test) versus the ‘true’ target value. In reality this ‘true’ value is unknown and must be estimated using a specified comparator (termed the reference test). This should ideally be based on a *reference measurement procedure* – an officially validated test method which has been shown to accurately measure the measurand of interest, and can therefore be used as a reliable proxy for the ‘true’ test value; alternatively a *certified reference material* may be used, consisting of samples of known composition produced under tightly controlled manufacturing or in-house procedures to provide reliable sources of measurement [29, 32, 33]. In the absence of such procedures or materials, index test(s) are often compared to results obtained from other laboratories partaking in external quality assessment (EQA) schemes, or against established ‘gold standard’ tests used in routine clinical practice. Finally, new and existing tests are often compared against each other (i.e. without a reliably proxy for the true measurand value), in order to ascertain if substitution of one test for another would result in significant changes in practice.

Various statistical techniques may be employed to assess the level of bias in comparative measurements. The most popular and accepted method, Bland-Altman analysis, is based around a scatter plot showing the dispersion of observed differences between the index and reference test results against the average concentration of the two measurements [34, 35]. This plot allows inspection of various measures including mean difference, limits of agreement[2] and outliers, and can further help to identify whether or not any systematic error in measurement is independent or not of the concentration of the measurand (i.e. constant vs. proportional bias, respectively). Other approaches include regression analysis and measures of inter-rater agreement (e.g. Cohen's kappa statistic or the intra class correlation coefficient), although issues with the validity of these methods have been highlighted in the literature [35].

An additional component important in the analysis of trueness is that of test *selectivity*: the ability of a test to identify the target measurand of interest as opposed to any other components in the test sample. Selectivity depends on the level of *interference* and *cross-reactivity* in the test sample; that is, the existence of obstruction from substances in the test sample which either inhibit the process of binding with the target measurand (interference) or are mistaken for the target measurand leading to 'unintentional' binding (cross-reactivity). For example, heightened levels of bilirubin (a natural by-product of liver functioning) is known to produce bias in certain tests for Creatinine (used routinely to monitor kidney functioning) [36]. Typically these issues are explored at the stage of manufacturer test validation (by deliberately 'spiking' test samples with suspected interferents) and should be considered in any subsequent analyses [37].

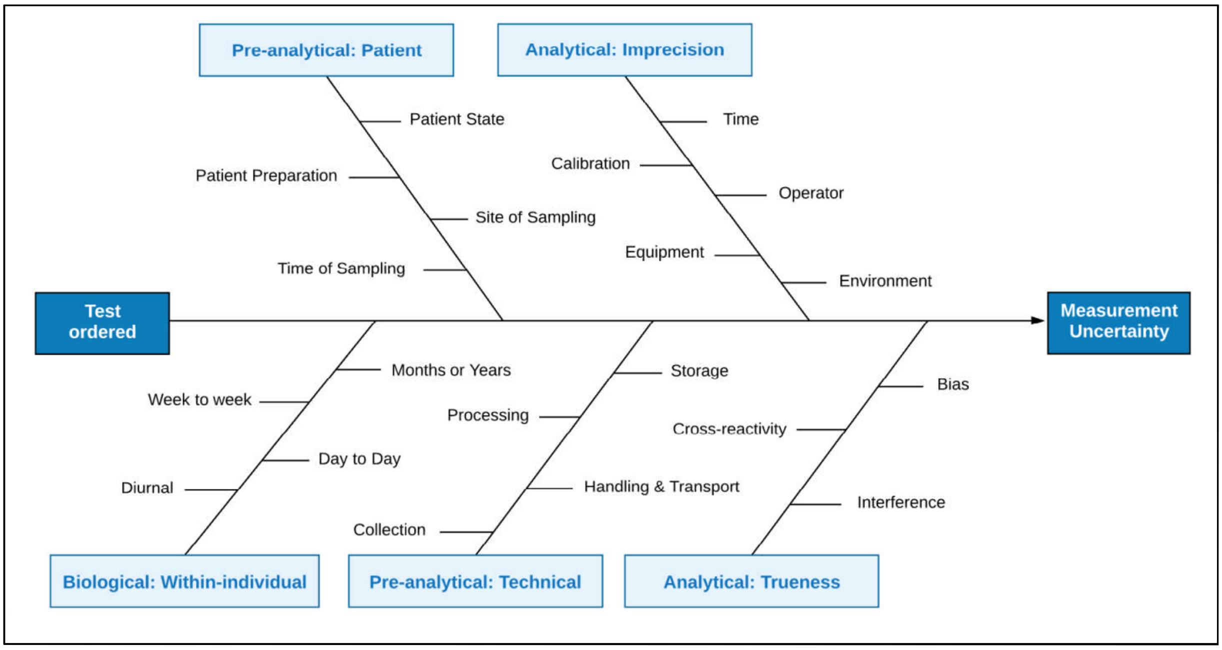
[2] For the 95% limits of agreement, for example, this is defined as the mean difference plus and minus 1.96 times the standard deviation of differences.

2. *Pre-analytical, Analytical and Biological Factors*

Both precision and trueness may be affected by numerous factors along the testing pathway, which may be categorised according to what point along the pathway they occur: the *pre-analytical phase* concerns processes occurring prior to the point of sample analysis, whilst the *analytical phase* covers processes occurring at the point of sample analysis. Pre-analytical factors therefore include patient variables such as preparation and health state, as well as technical variables such as sample collection method and test storage. Analytical factors meanwhile include affects such as the method of sample preparation and analysis, batch-to-batch variation and laboratory environment [38]. A final factor relevant to the pre-analytical phase (but typically distinguished within a third category) is that of *biological variation*: variation caused by fluctuations in the concentration of bodily fluid components within an individual over time. Such variations may occur over the course of a day (due to factors such as hydration, temperature and stress), or over weeks and months (due to factors such as menstrual cycles and seasonal impacts) [39].

For any given testing procedure, the range of elements expected to have a potential influence on measurement uncertainty can be summarized in a ‘feather diagram’. Figure 2 illustrates a generalised example, in which factors are grouped by category and can be considered to follow a (roughly) chronological order from the initial test request through to obtaining the test result. In order to determine the associated impact of such factors, ‘split-sample’ analyses (in which two equivalent portions of the same sample are analysed using alternative procedures) are routinely conducted for pre-analytical and analytical effects, whilst biological variation is typically assessed by observing the level of variation in test results within health populations over time.

Figure 2. Generalised feather diagram depicting factors which may contribute to measurement uncertainty



3. *Limits and Range*

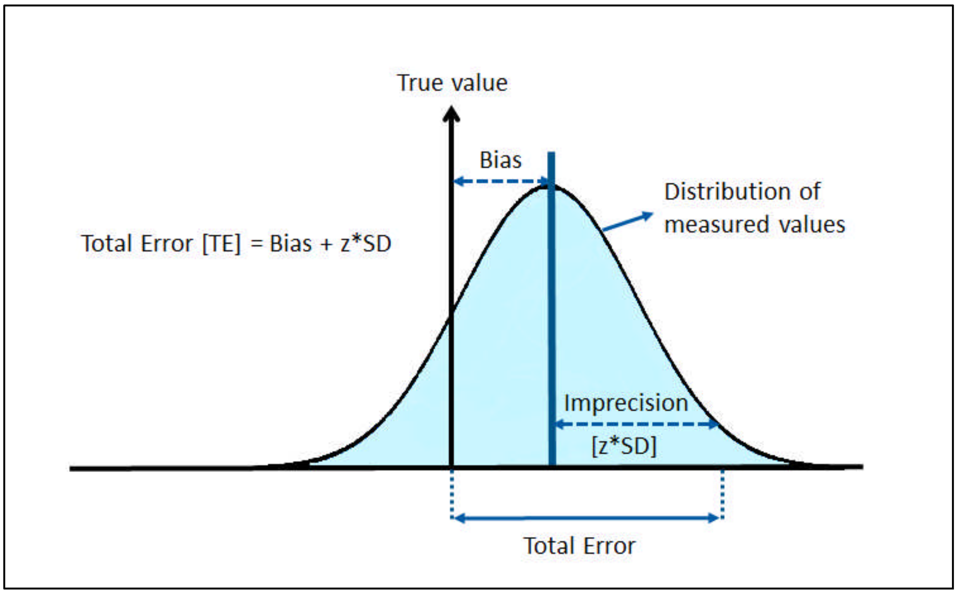
Various limits can be specified which describe the smallest concentration of a measurand that can be reliably measured for a given testing procedure. These are (i) the *limit of blank (LoB),* defined as the highest (apparent) concentration of measurand expected to be identified when processing blank samples (i.e. samples containing zero quantity of measurand); (ii) *the limit of detection (LoD),* defined as the lowest measurand concentration which the test can reliably distinguish from the LoB; and (iii) the lower *limit of quantification (LoQ),* defined as the lowest concentration of measurand which the test can detect with a specified level of precision and trueness [51]. The reportable range of a test is then typically determined by the interval from the lower LoQ and the upper LoQ (i.e. the highest concentration of measurand which can be detected with a specified level of precision and trueness). Thus the concepts of limits are important insofar as they determine the boundaries against which testing is reasonably conducted and reported.

4. Total Error (TE) and Uncertainty of Measurement (UM)

Different elements of uncertainty as illustrated in Figure 2 may be combined to estimate an aggregate value summarizing overall measurement uncertainty. Two main approaches to this end have been adopted in the literature: the *total error (TE)* approach and the *uncertainty of measurement* (U_M) approach.

The TE approach was originally promoted by Jim Westgard in the USA in the 1970's and became the dominant technique across laboratories over subsequent decades [40]. Briefly, TE is calculated as the linear sum of systematic and random error (i.e. bias and imprecision), as illustrated in Figure 3. Assuming that random error can be approximated by a normal (Gaussian) distribution, the estimate of imprecision (expressed as an SD) is multiplied by a '*z* factor' to cover a required level of confidence; in order to cover a 95% confidence interval, for example, a *z* value of 1.96 (often rounded to 2) is used. The resulting TE estimate provides an upper bound (i.e. worst case scenario) for the level of error which may occur for a given measurement.

Figure 3. Illustration of TE calculation



The uncertainty of measurement (U_M) approach emerged as a dominant method within the metrology field in the 1990's in the cornerstone 'Guide to the Expression of Uncertainty in Measurement' (GUM) document [41]. U_M is defined as "a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand". This parameter, characterised as an SD, is calculated according to the following procedure: (1) identify all elements associated with uncertainty along the testing pathway (often illustrated using a feather diagram); (2) determine the uncertainty around each of those elements; (3) combine the uncertainties using, for example, the sum of squares rule (see equation 1) or computer simulation (e.g. Monte Carlo simulation); and (4) assign a 'coverage factor' to determine an expanded region of uncertainty covering a specified probability level (see equation 2).

For example, for a given testing procedure, suppose that four elements along the testing pathway (A, B, C and D) have been identified as contributing standard uncertainties (U_a, U_b, U_c and U_d) to the measurement procedure. Assuming that the given uncertainties are independent of each other, these can then be combined using the following calculation.

$$(1) \qquad U_M = \sqrt{(U_a)^2 + (U_b)^2 + (U_c)^2 + (U_d)^2}$$

Then suppose that we want to be able to define a confidence interval around an observed test result to cover a 95% probability level. Assuming that the uncertainties are normally distributed, a coverage factor of 1.96 (or 2) is applied to give the expanded uncertainty ($U_{M'}$) as follows:

$$(2) \qquad U_{M'} = U_M * 1.96$$

The GUM originally proposed a "bottom up" procedure, in which all individual components of uncertainty in a testing pathway were required to be identified and separately measured.

More recently, owing to the recognized impracticalities of identifying and characterising all individual uncertainties, an alternative "top down" procedure has been endorsed [42]. This method recognises that high level data (e.g. from EQA schemes or validation studies) will capture multiple components of measurement uncertainty, and can therefore be used in place of having to determine the contribution of multiple individual elements.

There has been much debate in the literature regarding the differences and relative merits of using either the TE or U_M approach [43-46]. Particular issues include the exclusion of systematic error in U_M (any bias in this approach is expected to be a-priori identified and eliminated, which many have argued is not always appropriate), and conversely, the inability of TE to accommodate additional sources of uncertainty such as biological variation. Nevertheless both methods have maintained widespread adoption: U_M is most widely used across Europe (being endorsed by key bodies such as the International Organisation for Standardization [ISO]), whilst TE has maintained greatest popularity in the USA. In the context of this study, both metrics are considered to be viable measures of overall measurement uncertainty and are included in the subsequent analysis.

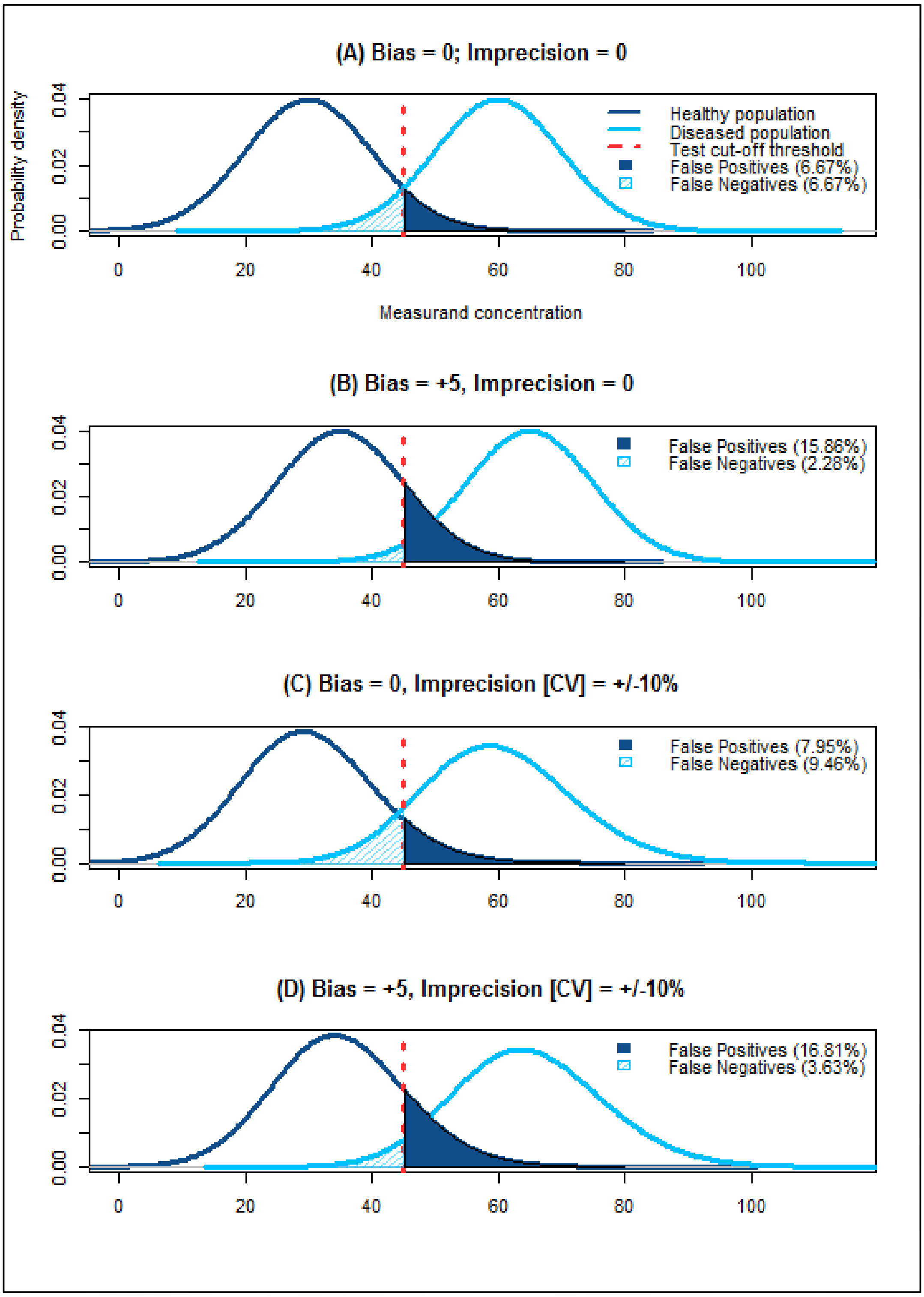
5. *HTA context*

When considering the clinical accuracy of a test, disease classification errors may occur – even in the face of perfect measurement – as a result of a natural overlap in the distribution of healthy and diseased populations in terms of the given measurand. The consequence of measurement uncertainty is that any observed test value may be different to the 'true' underlying value that one wishes to measure, which in turn may impact clinical accuracy if test values are incorrectly observed as lying above or below the test cut-off threshold used to determine disease classifications.

Consider the example in Figure 4. Part (A) illustrates the case of perfect measurement for hypothetical healthy and diseased populations, assuming both can be approximated by normal distributions with means of 30 and 60, respectively, and equivalent standard deviations of 10. In this case the test cut-off threshold has been set at 45 (with results >45 indicating disease) to deliver equal proportions of false positive and false negative results (6.67%). The subsequent figures (parts B to D)[3] illustrate the impact of imprecision and bias, assuming that the threshold remains constant. In part B, introducing a uniform positive bias of +5 leads to a rightwards shift in the central position of the distributions and results in an increase in false positives (15.86%) and decrease in false negatives (2.28%); in part C introducing concentration-dependent imprecision (CV = ±10%) leads to more widely distributed and skewed distributions, resulting in an increase in both the false positive (7.95%) and false negative (9.46%) proportions; and in part D introducing bias and imprecision simultaneously both shifts the distributions to the right and increases the dispersion, resulting in an overall increase in the false positives (16.81%) and decrease in the false negatives (3.63%).

[3] In this example, the distributions of healthy (*H*) and diseased (*D*) populations are based on 10^7 simulations drawn from distributions $H\sim N(30,10)$ and $D\sim N(60,10)$ respectively. Bias (α) is applied by adding α to each population mean i.e. $H'\sim N(30+\alpha,10)$ and $D'\sim N(60+\alpha,10)$. Imprecision [CV = ±β%] is dependent on concentration and thus applied at the individual simulation level: for the *i*th simulation from the original *H* and *D* populations (i.e. H_i and D_i), imprecision is applied as an additional random draw from the distributions $N(0,H_i{*}(\beta/100))$ and $N(0,D_i{*}(\beta/100))$ respectively.

Figure 4. Illustrative example: impact of bias and imprecision on clinical accuracy



If a significant proportion of patients receive inappropriate health care interventions as a consequence of disease classification errors, this may have an impact on the clinical utility and cost-effectiveness of the test. Furthermore, depending on the cost of the test, associated treatments and downstream outcomes, only a small proportion of test misclassifications may be required to render a test ineffective or economically non-viable. Thus, within the context of HTAs, it is important to know whether there are any issues with measurement uncertainty in order to correctly ascertain the impact of a test on health outcomes and healthcare resources.

Nevertheless, whilst establishing a strong evidence base regarding certainty in measurement has consistently featured as a core component of test development and adoption pathways [47-49], this has not commonly, to our best knowledge, been addressed within HTAs. Current HTA 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 [7, 50]. The Medical Services Advisory Committee (MSAC) in Australia is the only authority we are aware of that has specified the need to evaluate such evidence, using the terminology of analytic validity [51]. However, whilst stipulating that the HTA literature review should include analytical validity outcomes, MSAC offer no recommendations regarding how this data should be assessed, or if or how it should be utilised within subsequent clinical accuracy, utility and economic assessments.

In order to establish if and how measurement uncertainty is currently being addressed within test assessments, a systematic review of international HTAs of in-vitro tests was conducted. In light of the increasingly important role of health-economic evaluations within global health care decision making, and given the uncertainty around how measurement uncertainty may be included within economic evaluations specifically, the focus of this review was on HTAs that included a decision-model based economic evaluation.

Methods

The review protocol was published in advance on the PROSPERO database (CRD42017056778). 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.

The primary source for the review was the Centre for Reviews and Dissemination (CRD) Health Technology Assessment (HTA) database. A search strategy (provided in online supplementary content) combining key terms on in-vitro tests and economic decision models was developed and ran in March 2017. In addition online records of key HTA authorities expected to be the largest contributors of HTAs of tests were checked: NICE in the UK [52], the Agency for Drugs and Technologies in Health (CADTH) in Canada [53] and the Medical Services Advisory Committee (MSAC) in Australia [54]. For all included studies, citation checking was conducted to identify any further relevant HTAs.

Records were stored in and 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. For studies identified as including an assessment of measurement uncertainty, data was extracted on the specific components assessed and the methods utilised, with 10% of data extraction independently checked by the secondary reviewer and results narratively synthesised.

A broad definition of measurement uncertainty including a range of components as outlined in the introduction was adopted: in addition to those items listed in Figure 2, data on TE, U_M, LoD, LoQ, reportable range and test failure rates was also included.

Results

From 1908 citations identified from the database search, 211 were shortlisted for full-text screening. Agreement between the two independent reviewers at abstract screening was good (k=0.85). One hundred and seven studies met the final inclusion criteria: 90 were identified from the database and a further 17 from HTA authority websites and via citation checking.

A summary of all study characteristics is provided in Table 1 and Figure 6 shows the number of total HTA reports, and reports including a point-of-care test (POCT), by year of report. The majority of studies were conducted within the UK, and there has been a gradual but inconsistent rise in the number of HTAs published per year since 1999.

Figure 5. PRISMA flow diagram of search results

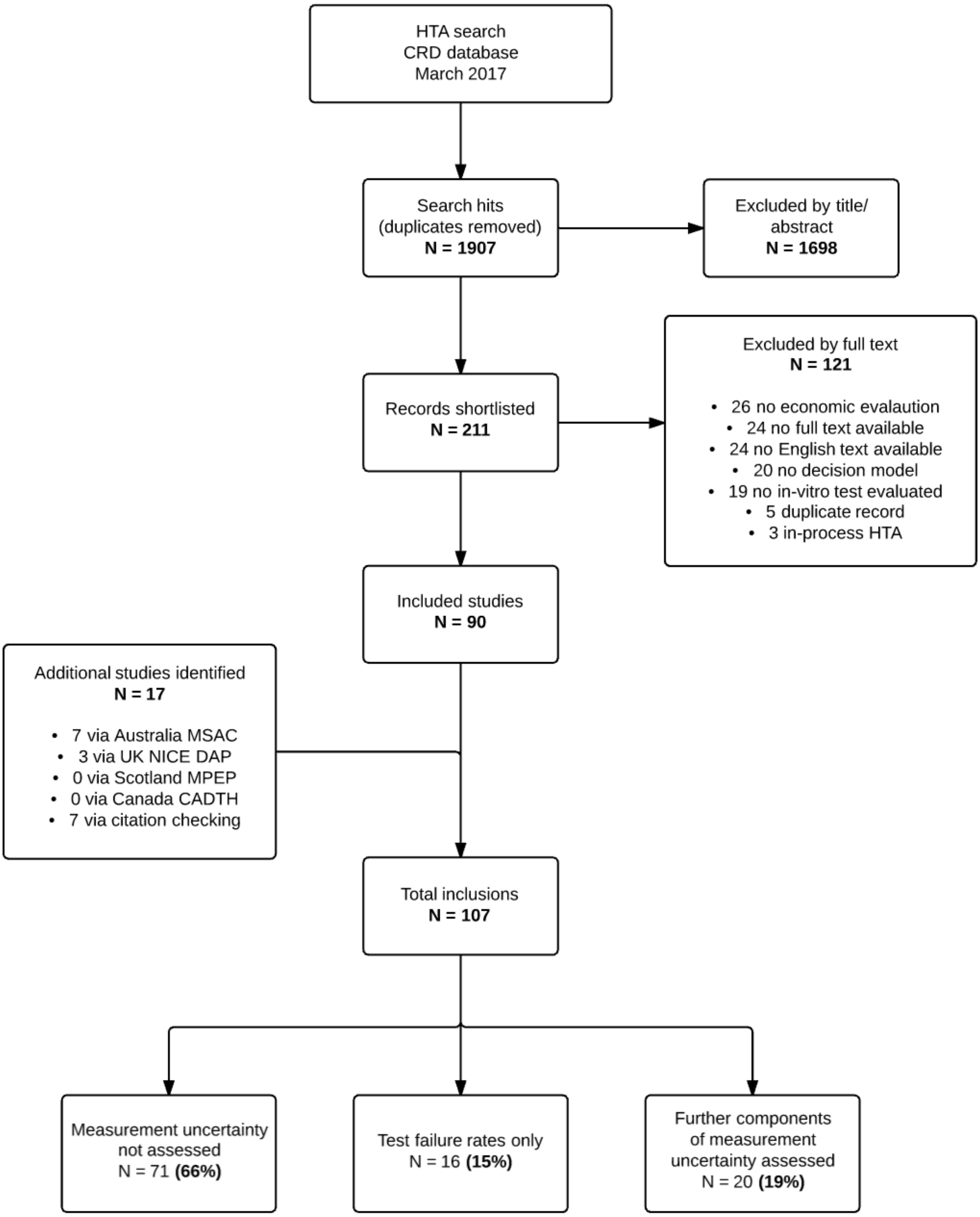
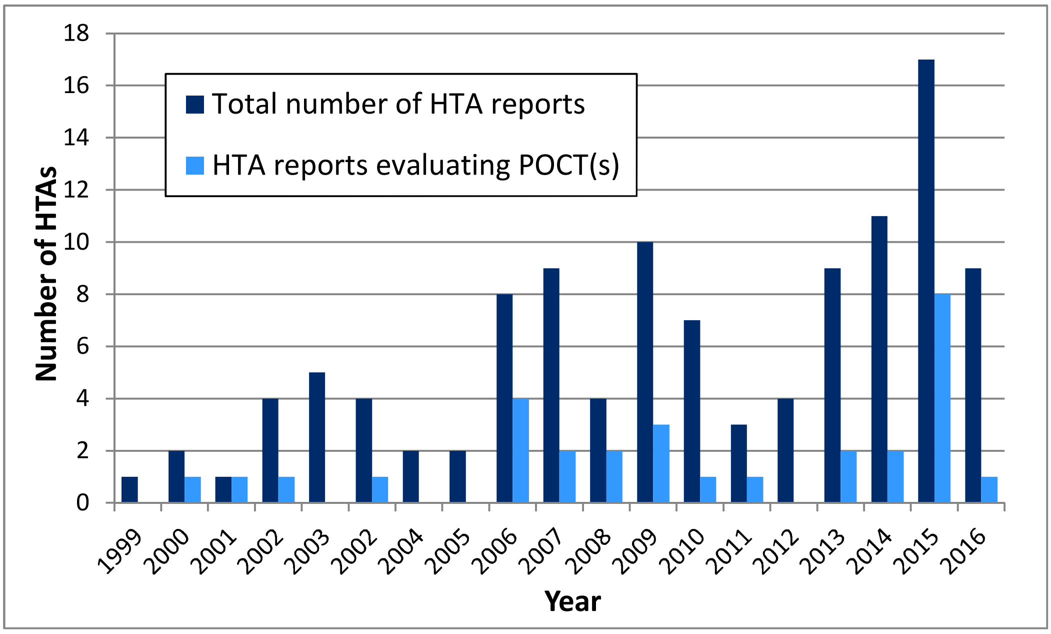


Table 1. Summary of all HTA study characteristics[4]

	Number of HTA reports (total = 107)	%
Country		
UK	66	62%
Canada	17	16%
Australia	15	14%
Belgium	3	3%
USA	3	3%
Ireland	2	2%
Italy	1	1%
Disease Area		
Cancer	36	34%
Pregnancy care & screening	14	13%
Cardiology	12	11%
Haematology	12	11%
Infections	13	12%
Diabetes	6	6%
Gastroenterology	5	5%
Other	9	8%
Type of test(s) Evaluated		
Laboratory tests only	85	79%
POCT- clinician led	18	17%
POCT- self led	5	5%
Primary Role of Test(s)		
Diagnosis	39	36%
Screening	37	35%
Prognosis	14	13%
Monitoring	9	8%
Predictive	6	6%
Other	2	2%
Type of Evaluation		
Cost-utility	53	50%
Cost-effectiveness	36	34%
Cost-utility & cost-effectiveness	17	16%
Cost-consequences	1	1%
Type of Economic Model		
Decision Tree	48	45%
Cohort Markov	22	21%
Tree + Markov	17	16%
Patient level simulation	12	11%
Infectious disease/ dynamic	6	6%
Not reported	2	2%

[4] Note: one report assessed a POCT in both a self-led and clinician led scenario and is therefore counted in both of these classifications under ‘Type of test(s) evaluated’.

Figure 6. Number of HTA reports by year of publication



Assessments of measurement uncertainty

Of the 107 identified HTAs, 36 studies (34%) included some form of assessment of measurement uncertainty. Sixteen of those (15%) were limited to an assessment of test failure rates only; these were considered to be of limited interest and are therefore not discussed further. Twenty studies (19%) considered further components of measurement uncertainty: 15 (14%) included some form of 'intermediate' assessment (consisting of a systematic [n=11] or non-systematic [n=2] literature review, laboratory survey and systematic review [n=2], database analysis [n=2], or clinical study [n=2]); 4 (4%) also included components within the economic model in addition to an intermediate assessment (consisting of database analysis [n=2], clinical study [n=1] or systematic review [n=1]); and 1 (1%) considered measurement uncertainty within the economic model only. A summary of these studies is provided in Table 2.

Table 2. Summary of HTA reports (n=19) including components of measurement uncertainty in an intermediate assessment and/or in the economic decision model

Study	Test characteristics			Intermediate Assessments		Economic Decision Model Assessments	
	POCT?	Disease area	Primary role of test	Method	Components of measurement uncertainty included	Method	Components included
Auguste *et al.* 2016 (UK) [55]	-	Infection (TB)	Diagnosis	Systematic review	Trueness (Kappa statistic, discordance); test failures	-	-
Freeman *et al*. 2016 (UK) [56]	-	Gastro.	Monitoring	Systematic review	Trueness (Bland-Altman analysis, Cohen's Kappa); test failures	-	-
Hay *et al.* 2016 (UK) [57]	POCT: clinician-led	Other (urology)	Diagnosis	Clinical study	Trueness (kappa statistic); test failures	-	-
Stein *et al.* 2016 (UK) [58]	-	Cancer	Prognosis	Pathology study	Trueness (kappa statistic, discordance)	Test agreement data informed uncertainty around model parameters	Test agreement
Freeman *et al*. 2015 (UK) [59]	-	Cancer	Monitoring	Systematic review	Trueness (Bland-Altman analysis, deming regression); test failures	-	-
Harnan *et al.* 2015 (UK) [60]	POCT: self-led	Other (asthma)	i) Diagnosis ii) Monitoring	Systematic review	Trueness (Bland-Altman analysis, correlation coefficients); test failures	-	-
Kessels *et al*. 2015 (AUS) [61]	-	Pregnancy care & screening	Diagnosis	Systematic review	Selectivity; test failures	-	-
MSAC 2015 (AUS) [62, 63]	-	Cancer	Prognosis	Literature review	Selectivity	-	-
Nicholson et al. 2015 (UK) [64]	-	Cancer	Diagnosis	Systematic review	Precision (intermediate and reproducibility); trueness (recovery); LoB, LoD, LoQ; interference; linearity; range; pre-analytical effects; stability; test failures	-	-
Perera *et al.* 2015 (UK) [65]	-	Cardiology	Monitoring	Analysis of IPD	Biological and analytical variation	Regression of IPD + model simulations	Biological and analytical variation
Sharma *et al.* 2015 (UK) [66]	POCT: self-led	Haematology	Monitoring	Literature review	Precision (reproducibility); trueness (r correlation coefficient)	-	-
Farmer *et al.* 2014 (UK) [67]	-	Diabetes	Screening	Analysis of IPD	Biological and analytical variation	Regression of IPD + model simulations	Biological and analytical variation

Westwood *et al.* 2014 (UK) [68]	-	Cancer	Predictive	Systematic review + survey	Proportion of tumour cells needed; LOD; test failures	-	-
Westwood *et al.* 2014 (UK) [69]	-	Cancer	Predictive	Systematic review + survey	Proportion of tumour cells needed; test failures	-	-
Ward *et al.* 2013 (UK) [70]	-	Cancer	Prognosis	Systematic review	Precision (intermediate and reproducibility); trueness (concordance).	-	-
M.A.S 2010 (CA) [71]	-	Cancer	Prognosis	Systematic review	Precision (intermediate and reproducibility); test failures	-	-
Pearson *et al.* 2010 (UK) [72, 73]	POCT: clinician-led	Gastro.	Diagnosis	Systematic review	Biological variability; distribution in faeces; faecal matrix; interference; stability; patient compliance; normal range	-	-
Gailly *et al*. 2009 (BEL) [74]	POCT: self-led	Haematology	Monitoring	Systematic review	Precision (repeatability and intermediate); test failures	-	-
MSAC 2001 (AUS) [75]	POCT: clinician-led	Cardiology	Prognosis	Systematic review	Trueness (% bias); precision (repeatability and reproducibility); TE; analytical effects (site, operator and sample type)	Monte Carlo simulation	TE (bias + 1.96*CV)
Marks *et al.* 2000 (UK) [76]	-	Cardiology	Screening	-	-	Rate of false positives set equal to coefficient of biological and analytical variability	Biological and analytical variation

M.A.S = Medical Advisory Secretariat; UK = United Kingdom; AUS = Australia; CA = Canada; BEL = Belgium; POCT = point of care test; Gastro. = Gastroenterology; IPD = individual patient data; LOB = limit of blank; LOD = limit of detection; LOQ = limits of quantification; TE = total error; CV= coefficient of variation.

Of 19 HTAs including an intermediate assessment, 18 were reported from 2009 onwards. The type and number of components of measurement uncertainty evaluated varied across studies and included imprecision, trueness, test agreement, biological variability and pre-analytical or analytical effects. Of the five studies that included components of measurement uncertainty within the economic decision model, two were reported over 15 years ago and utilised decision tree models [75, 76], whilst three were published more recently and utilised either a cohort Markov model [58] or individual patient simulation models [65, 67]. These models incorporated data on test agreement [58], biological and analytical variability [65, 67, 76] or TE [75] and used a range of alternative methods, outlined below.

Stein et al. 2016

In the most recent study, Stein and colleagues conducted an evaluation of genomic tests for distinguishing between patients with high and low risk of breast cancer recurrence [58]. The primary analysis utilised clinical outcomes data from a feasibility trial in which patients were randomised to receive either standard care (chemotherapy for all) or test-directed treatment. Further tests were evaluated using data from a supplementary pathology study; the level of between-test discordance was used to determine the level of uncertainty around test-dependent cancer recurrence rates in the model.

Farmer et al. 2014 & Perera et al. 2015

The next two studies both utilized individual patient data containing repeated test measurements over time to model serial testing scenarios whilst accounting for variation in patient baseline test values, rates of disease progression, and biological variation and analytical variability [65, 67]. In the study by Farmer and colleagues, for example, the authors estimated the accuracy of repeated albumin-creatinine ratio (ACR) testing for diabetes screening using alternative intervals by constructing a longitudinal hierarchical linear model for log ACR, consisting of individual intercepts (baseline ACR), gradients (change in log ACR per year of age) and biological variation [67]. The rate of true and false diagnoses over time was determined by simulating true and observed ACR values based on this model. A similar approach was adopted by Perera and colleagues for the assessment of repeated measures of cholesterol for the monitoring of patients with or at risk of cardiovascular disease [65].

MSAC 2001

Of the two older studies, MSAC utilised Monte Carlo simulations to estimate the impact of TE on clinical accuracy for a cholesterol POCT [75]. Each patient in the simulation was assigned a true cholesterol level according to the distribution observed in the Australian population; two test results were then generated based on the reported level of TE (95% CI = +/- 8%) derived from a preceding literature review, with a diagnosis of elevated cholesterol based on the average of the two results against a given cut-off threshold. Clinical accuracy in the decision tree model was then based on a weighted average of the probabilities of misclassification errors across the cholesterol range assessed. The impact of TE was explored via sensitivity analyses, which indicated that TE would not alter the overall decision uncertainty (since all results remained above the specified cost-effectiveness threshold of AUS$100,000), but did have a significant impact on the base case Incremental Cost-Effectiveness Ratios (ICERs) (resulting in a 24% drop from $133,934 to $101,419 per life year gained, when reducing TE from 8% to 0%).

Marks et al. 2000

Marks and colleagues assessed a screening test for familial hypercholesterolemia [76]. They reported that they set the proportion of false positive test results in the model (6.5%) equal to a “coefficient of biological and analytical variability” identified from a cited paper [77]. The authors provided no further details on this analysis.

Discussion

Review findings

This review provides the first definitive evidence that, despite limited guidance in this area, assessment of test measurement uncertainty has been attempted in a minority of HTAs, and that within those studies wide variation exists in terms of the types of components assessed and methodology adopted. There appears to be no trend in terms of where the relevant HTAs were conducted (i.e. country), the disease area or test characteristics. However, such assessments appear to be increasing in frequency in recent years: this may reflect either the fact that more HTAs of tests are being conducted in general, a growing awareness of the importance of measurement uncertainty, and/or increasingly availability of relevant data upon which to base such evaluations.

Most of the HTAs that considered measurement uncertainty did so via some form of intermediate assessment, such as a literature review or laboratory survey. On the whole these were considered to be partial assessments: most considered one or a limited set of measurement uncertainty components and for the majority of assessments based on a literature review, the primary aim was to identify evidence on clinical accuracy or utility. A further 5 studies were identified which considered components of measurement uncertainty within the economic model. Of those, the most recent study by Stein and colleagues (2016) was not really an attempt to account for measurement uncertainty, but rather the authors here utilised data on test discordance to allow evaluation of additional tests in the model [58]. Meanwhile the oldest study by Marks and colleagues (2000) is perhaps most interesting as an example of what not to do [76]. Here the authors set the proportion of false positive results equal to a given level of biological and analytical variability (i.e. imprecision). This approach fails to account for the dependence of test misclassifications on the position of values relative to the test cut-off threshold, as well as measurement uncertainty. These dependencies mean that a 6.5% coefficient of variability (CV) will not necessarily lead to a 6.5% false positive rate: in the hypothetical example discussed in the introduction (Figure 4), for example, applying a 6.5% CV leads to a false positive rate of 7.24%.

The two most computationally advanced studies were those that evaluated repeated testing scenarios using patient level simulation models [65, 67]. By utilising individual patient data on repeated measures over time, these authors were able to incorporate the effect of biological variation as well as potential pre-analytical and analytical variation (assuming sampling and laboratory conditions altered over time). Whilst this approach is interesting from a statistical and modelling perspective, it is also the most demanding in terms of data requirements and computational resources; such an approach would therefore likely be challenging within typical HTA studies due to restricted timelines, resources, and unavailability of individual patient data.

The final study utilised data on TE to assign 95% confidence intervals around test values and simulate the impact on clinical accuracy [75]. This approach correctly attempts to account for the relationship between true values, measurement uncertainty, and the cut-off threshold; however there are potential issues with use of TE within this framework. TE linearly combines both systematic and random components of error: assuming bias acts in a fixed direction, using TE to assign both sides of a confidence interval will overestimate uncertainty; furthermore, bias and imprecision may at times work in opposite directions and could therefore partially mitigate or cancel each other out. A better approach may be to simulate the impact of systematic bias and random uncertainties as separate components – this would further enable sensitivity analyses to be conducted to explore the potential of resolving systematic errors (i.e. by resetting bias to zero), or moving the cut-off threshold to mitigate the impact of such errors.

Proposed priorities for future research

Based on the findings from this review, three key questions for consideration in future research are apparent: (1) when should assessment of measurement uncertainty be required; (2) what components of measurement uncertainty should be assessed; and (3) what type of analysis should be utilised.

1. When should assessment of measurement uncertainty be required?

This review confirms that measurement uncertainty is currently considered within a minority of relevant HTAs. Two key justifications may be driving the common decision to forgo such analyses. First, the assumption may be that issues in measurement performance will be dealt with at the stage of test regulation (e.g. the Food and Drug Administration [FDA] or Clinical Laboratory Improvement Amendments [CLIA] in the USA, or Conformité Europeenne [CE] marking across Europe). This is contrary to the fact that regulatory authorities are primarily committed to establishing that manufacturers provide *true* statements regarding the analytical performance of a given test, rather than ensuring that the stated level of measurement uncertainty is *acceptable* in any way [40]. In addition, whilst newly reformed EU regulation (set to come into full force in 2022) demands greater evidence on the clinical accuracy of new tests, there remains no requirement to assess the impact of stated uncertainties on downstream clinical utility or cost-effectiveness which, as we have noted, may be significant even in the face of small or "acceptable" measurement uncertainty values by regulation standards. For these reasons it is inappropriate to assume that regulation assures that measurement uncertainty will not have a meaningful impact on the clinical utility or cost-effectiveness of a test.

The second potential justification lies in the assumption that any effects of measurement uncertainty will be captured within clinical accuracy studies, where available. The validity of this assumption will depend on the specific characteristics of any given set of clinical accuracy studies in relation to key drivers of measurement uncertainty. For example, where clinical accuracy estimates are based on studies conducted across multiple centres and avoiding any known interferents and confounding pre-analytical/ analytical processes, then it may be reasonable to assume that the synthesised accuracy data would likely capture the impact of expected uncertainties and provide reliable estimates. Meanwhile, studies conducted within a single centre; which adopt tightly controlled patient sampling and laboratory protocols unlikely to be reproducible in routine testing; or which fail to appropriately account for known interferents and influential pre-analytical or analytical factors, will likely fail to capture relevant uncertainties and potentially produce inapplicable or biased results.

As it stands, there is no validated checklist or mechanism by which to determine whether or not clinical accuracy data is likely to be "suitable" with regard to capturing measurement uncertainty. Determining a feasible and reliable process to inform this decision is therefore a key priority for future research. Whilst any such process would need to be feasible within the context of restricted HTA timelines and resources, it is expected that a basic understanding of the types of factors likely to influence or bias test measurement would be required (such that measurement uncertainty would need to be at least initially considered within any HTA in order to ensure reliable results). Feasibility of such analyses would be greatly improved if manufacturers, regulators, laboratories and clinical researchers made the relevant data on measurement uncertainty more readily and easily available for HTA reviewers, such as through active participation in open databases. Likewise, analyses would benefit from greater transparency and clarity of reporting concerning measurement procedures within clinical accuracy studies also, in order to ensure that relevant aspects of measurement processes can be ascertained.

2. *What components of measurement uncertainty should be assessed?*

Previous HTAs including assessments of measurement uncertainty have tended to consider a limited set of components. It seems reasonable to propose that if measurement uncertainty is to be robustly assessed, then the core measures of imprecision and bias, as well as any potential influences of biological, pre-analytical and analytical effects, should be considered. It may be that evaluations would require particular emphasis depending on the specific properties of the test under assessment: imprecision, for example, may be of particular concern for any new POCT yet to be validated outside of the laboratory. Nevertheless there would need to be reasonable justification for excluding any of the aforementioned components of measurement uncertainty from the analysis.

A separate question concerns whether or not it is appropriate to combine individual measures of measurement uncertainty within a summary statistic, and, if so, which metric should be adopted. As previously discussed there is continuing debate in the literature as to the relative merits of TE vs. U_M, and we have highlighted potential problems with the combination of systematic and random errors into a single estimate. It may be that direct use of such summary statistics is not the best approach anyway, since this precludes the possibility of exploring the impact of specific components of measurement uncertainty on the cost-effectiveness results. Future research in this area could explore the impact of adopting alternative approaches to combining elements of measurement uncertainty on the corresponding clinical utility and cost-effectiveness results.

3. *What type of analysis should be utilised?*

To date the majority of HTA analyses of measurement uncertainty have relied upon some form of systematic literature review, considered to be the most appropriate means of reviewing published evidence for a given topic. An ongoing issue with the conduct of reviews in this field is the appropriate means of assessing the quality of identified evidence. Whilst guidelines exist for the appropriate conduct and reporting of the relevant primary studies (e.g. BRISQ [78], STROBE-ME, RIPOSTE [79]), there is no widely adopted quality assessment framework specifically for studies reporting measurement uncertainty. We are aware of one relevant framework in development: the Quality Assessment of Measurement Procedures (QAMPs) checklist [80, 81]. Once finalised this tool should help enable more robust assessments of the quality of measurement procedures reported in clinical studies or trials. In addition to the central literature review, grey literature in the form of test inserts, manufacturer data and EQA schemes – not historically widely published – is likely to be a crucial source of data in this field. In situations where paucity of data is a particular issue, or where publication bias is suspected, survey data may also provide an additional useful means of collecting unpublished laboratory data and of supplementing evidence identified from the review.

A small number of HTAs have so far attempted to account for measurement uncertainty within the economic model. Within the 5 studies identified in this review, various techniques were employed to incorporate data on test agreement, TE and biological and analytical variability. In addition to examining the feasibility of alternative simulation methods, future studies in this area could also explore: (i) the impact of using alternative types of summary data on measurement uncertainty, e.g. U_M vs. TE vs. modelling individual components; (ii) methods for combining data on measurement uncertainty with clinical accuracy data; (iii) methods to account for dependent and/or non-normal components of measurement uncertainty; and (iv) methods to capture for the impact of concentration level on measurement uncertainty (e.g. accounting for non-linear effects). There is also currently an underutilised opportunity to include components of measurement uncertainty within model sensitivity analyses, as well as value of information and implementation analyses which are becoming increasingly popular

[82]. Such evaluations could help to identify key components of uncertainty in measurement procedures, and highlight areas for future research where necessary.

A further opportunity yet to be explored within HTAs, is that of utilising the economic decision model to inform test pre-analytical and analytical performance goals. According to the 2015 Milan consensus [83], performance criteria for tests (i.e. acceptable levels of bias or imprecision in measurement) should be determined according to the following hierarchy of models: (1a) direct outcome studies (e.g. clinical trial data); (1b) indirect outcome studies (e.g. simulation or decision analysis); (2) based on components of biological variation (e.g. by minimising the ratio of 'analytical noise' to the biological signal); or (3) based on state-of-the-art comparisons (e.g. relating results to the highest level of performance technically achievable or achieved by other laboratories). In practice performance criteria are often set according to analyses falling under (2) or (3) in this hierarchy, whilst the proposed use of decision modelling would instead constitute an example of a (1b) study on this hierarchy. This type of analysis at the stage of an HTA evaluation could enable existing test performance goals to be adjusted and optimised in light of new information, and ensure that test processes are optimised based on expected patient outcomes. The practicalities and feasibilities of such an approach within the HTA context is another area worthy of future investigation.

A final point to consider is that, whilst the focus of this analysis has been on in-vitro test evaluations, many of these issues are likely to be applicable to imaging and in-vivo technologies. In particular, all of the issues highlighted here will be relevant to pharmacological studies which utilise tests as surrogate outcome measures.

Strengths and Limitations

This review considers papers published in English language since 1999. It is possible that relevant analyses may have been conducted prior to this date or in other languages that the current search will have missed. Nevertheless, this is the first comprehensive and 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 in this area.

Conclusions

Various approaches have been adopted within a minority of HTAs to attempt to capture the impact of measurement uncertainty on test outcomes; uncertainty remains around when such assessments are required, and appropriate methodology for conducting analyses. Further research is proposed to resolve these questions and ensure that future HTAs are fit for purpose.

References

1. National Center for Biotechnology Information [NCBI]. Genetic Testing Registry. 2017 [06/07/2017]. Available from: https://www.ncbi.nlm.nih.gov/gtr/.

2. Marketsandmarkets.com. In Vitro Diagnostics/IVD Market by Product (Instruments, Reagents, Software), Technology (Immunoassay, Clinical Chemistry, Molecular Diagnostics, Hematology), Application (Diabetes, Oncology, Cardiology, Nephrology, Infectious Diseases) - Forecast to 20212016 06/07/2017 2016]. Available from: http://www.marketsandmarkets.com/Market-Reports.

3. Banta D, Jonsson E. History of HTA: introduction. International journal of technology assessment in health care. 2009;25(S1):1-6.

4. World Health Organisation (WHO). Health technology assessment. 2017 [08/08/2017]. Available from: http://www.who.int/medical_devices/assessment/en/

5. The International Network of Agencies for Health Technology Assessment (INAHTA). Welcome to INAHTA 2017 [07/08/2017]. Available from: http://www.inahta.org/.

6. Newland A. NICE diagnostics assessment programme. Annals of The Royal College of Surgeons of England. 2011;93(5):412.

7. National Institute of Health and Clinical Excellence (NICE). Diagnostics Assessment Programme manual 2011 07/08/2017. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-diagnostics-guidance.

8. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Radiology. 2015;277(3):826-32.

9. Bossuyt PM, Reitsma JB, E Bruns D, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clinical chemistry and laboratory medicine. 2003;41(1):68-73.

10. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC medical research methodology. 2003;3(1):25.

11. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.

12. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. REporting recommendations for tumour MARKer prognostic studies (REMARK). European journal of cancer. 2005;41(12):1690-6.

13. Reitsma J, Rutjes A, Whiting P, Vlassov V, Leeflang M, Deeks J. Chapter 9: Assessing methodological quality. Cochrane handbook for systematic reviews of diagnostic test accuracy version. 2009;1(0):1-27.

14. De Vet H, Eisinga A, Riphagen I, Aertgeertes B, Pewsner D. Chapter 7: Searching for studies. Cochrane handbook for systematic reviews of diagnostic test accuracy version 0.4. The Cochrane Collaboration. 2008.

15. Jordan JL, Hayden JA, Irvin E, Parker R, Smith A, van der Windt DA. Protocol: a systematic review of studies developing and/or evaluating search strategies to identify prognosis studies. Systematic reviews. 2017;6(1):88.

16. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Annals of internal medicine. 2008;149(12):889-97.

17. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Medical Research Methodology. 2002;2(1):9.

18. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. Journal of clinical epidemiology. 2006;59(3):234-40.

19. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). Journal of clinical epidemiology. 2010;63(8):854-61.

20. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health technology assessment (Winchester, England). 2005;9(12):1-113, iii.

21. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of clinical epidemiology. 2005;58(10):982-90.

22. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2006;8(2):239-51.

23. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. Journal of clinical epidemiology. 2008;61(11):1095-103.

24. Merlin T, Lehman S, Hiller JE, Ryan P. The “linked evidence approach” to assess medical tests: a critical analysis. International journal of technology assessment in health care. 2013;29(3):343-50.

25. di Ruffano LF, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. Journal of clinical epidemiology. 2012;65(3):282-7.

26. 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 medical research methodology. 2017;17(1):56.

27. 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 in Health. 2010;13(8):952-7.

28. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Medical Decision Making. 2008;28(5):650-67.

29. International Organization for Standardization (ISO). Accuracy (trueness and precision) of measurement methods and results [ISO 5725:1-6]. Geneva, Switzerland: 1994.

30. Clinical and Laboratory Standards Institute (CLSI). Evaluation of Precision of Quantitative Measurement Procedures; Approved Guideline - Third Edition. CLSI document EP05-A3. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

31. Westgard J. Method validation: the replication experiment. Basic Method Validation 3rd ed Madison, WI: Westgard QC, Inc. 2008:114-22.

32. Westgard J. Method validation: the comparison of methods experiment. Basic Method Validation 3rd ed Madison, WI: Westgard QC, Inc. 2008:124-36.

33. NCCLS. Method Comparison and Bias Estimation Using Patient Samples; Approved Guideline- Second Addition. NCCLS document EP9-A2 [ISBN 1-56238-472-4]. NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898, USA 2002.

34. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet. 1986;327(8476):307-10.

35. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PloS one. 2012;7(5):e37908.

36. Greenberg N, Roberts WL, Bachmann LM, Wright EC, Dalton RN, Zakowski JJ, et al. Specificity characteristics of 7 commercial creatinine measurement procedures by enzymatic and Jaffe method principles. Clinical chemistry. 2012;58(2):391-401.

37. Clinical and Laboratory Standards Institute (clsi). Interference Testing in Clinical Chemistry; Approved Guideline- Second Edition. . CLSI document EP7-A2 [ISBN 1-56238-584-4]. Clinical and Laboratory Standards Institute, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA 2005.

38. Plebani M. Exploring the iceberg of errors in laboratory medicine. Clinica chimica acta. 2009;404(1):16-23.

39. Fraser CG. Biological variation: from principles to practice: Amer. Assoc. for Clinical Chemistry; 2001.

40. Westgard JO, Barry PL, Quam EF, Ehrmeyer SS. Basic method validation: training in analytical quality management for healthcare laboratories: Westgard Quality Corporation; 1999.

41. BIpm I, IFcc I, IUpAc I. OIML, Guide to the Expression of Uncertainty in Measurement (GUM). International Organization for Standardization, Genève. 1995:11.

42. BIPM I, IFCC I, IUPAC I, ISO O. Evaluation of measurement data—guide for the expression of uncertainty in measurement. JCGM 100: 2008. Citado en las. 2008:167.

43. Panteghini M, Sandberg S. Total error vs. measurement uncertainty: the match continues. Clinical Chemistry and Laboratory Medicine (CCLM). 2016;54(2):195-6.

44. Oosterhuis WP, Theodorsson E. Total error vs. measurement uncertainty: revolution or evolution? Clinical Chemistry and Laboratory Medicine (CCLM). 2016;54(2):235-9.

45. Rozet E, Rudaz S, Marini R, Ziemons E, Boulanger B, Hubert P. Models to estimate overall analytical measurements uncertainty: Assumptions, comparisons and applications. Analytica chimica acta. 2011;702(2):160-71.

46. 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.

47. Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clinica chimica acta. 2014;427:49-57.

48. Hornberger J, Doberne J, Chien R. Laboratory-developed test—SynFRAME: an approach for assessing laboratory-developed tests synthesized from prior appraisal frameworks. Genetic testing and molecular biomarkers. 2012;16(6):605-14.

49. Sanderson S, Zimmern R, Kroese M, Higgins J, Patch C, Emery J. How can the evaluation of genetic tests be enhanced? Lessons learned from the ACCE framework and evaluating genetic tests in the United Kingdom. Genetics in Medicine. 2005;7(7):495-500.

50. Drugs CAf, Health Ti. Guidelines for the economic evaluation of health technologies: Canada. Guidelines for the economic evaluation of health technologies: Canada: CADTH; 2006.

51. Medical Service Advisory Committee. Technical Guidelines for preparing assessment reports for the Medical Services Advisory Committee – Service Type: Investigative (Version 3.0). 2017.

52. National Institute of Health and Care Excellence (NICE). Guidance and advice list 2017 [March 2017]. Available from: https://www.nice.org.uk/guidance/published?type=dg.

53. Canadian Agency for Drugs and Technologies in Health (CADTH). 2017 [March 2017]. Available from: https://www.cadth.ca/.

54. Medical Services Advisory Committee (MSAC). 2017 [March 2017]. Available from: http://www.msac.gov.au/.

55. 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 Technology Assessment. 2016;20(38).

56. 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 Technology Assessment. 2016;20(83):1-288.

57. 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 Technology Assessment. 2016;20(51).

58. 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 Technology Assessment. 2016;20(10).

59. 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 Technology Assessment. 2015;19(91).

60. 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).

61. 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.

62. 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.

63. 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.

64. 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.

65. 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 cost-effectiveness modelling. Health Technol Assess. 2015;19(100).

66. 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 Technology Assessment. 2015;19(48).

67. 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 Technology Assessment. 2014;18(14).

68. 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).

69. 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).

70. 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 Technology Assessment 2013;17(44).

71. Medical Advisory Secretariat. KRAS testing for anti-EGFR therapy in advanced colorectal cancer : an evidence-based and economic analysis. Ont Health Technol Assess Ser [Internet]. 2010;10(25):1-49.

72. 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.

73. 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.

74. Gailly J, Gerkens S, Bruel A, Devriese S, Obyn C, Cleemput I. Use of point-of-care devices in patients with oral anticoagulation : a Health Technology Assessment. Brussels: Belgian Health Care Knowledge Centre (KCE). Belgian Health Care Knowledge Centre (KCE), 2009 Contract No.: KCE Reports vol 117C. D/2009/10.273/49.

75. Medical Service Advisory Committee. Evaluation of Near Patient Cholesterol Testing Using the Cholestech LDX [MSAC Assessment Report 1026]. 2001. Available from: http://www.msac.gov.au.

76. 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).

77. Neil H. Problems in measurement: cholesterol. et al, Prevention of cardiovascular disease: an evidence-based approach, Oxford University Press, Oxford. 1996:253-7.

78. Moore HM, Kelly AB, Jewell SD, McShane LM, Clark DP, Greenspan R, et al. Biospecimen reporting for improved study quality (BRISQ). Journal of proteome research. 2011;10(8):3429-38.

79. Masca NG, Hensor EM, Cornelius VR, Buffa FM, Marriott HM, Eales JM, et al. Science forum: RIPOSTE: a framework for improving the design and analysis of laboratory-based research. Elife. 2015;4:e05519.

80. Messenger M, Cairns D, Smith A, Hutchinson M, Wright J, Hall P, et al. A framework for the quality assessment of measurement proceedures using in vitro diagnostic medical devices (IVDs). Diagnostic and Prognostic Research. 2016;1(Suppl 1)(P19).

81. Hall P, Mitchel L, Smith AF, Cairns D, Messenger MP, Wright J, et al. The future for diagnostic tests of acute kidney injury in critical care: evidence synthesis, care pathway analysis and research prioritisation. Health Technol Assess. In press.

82. Fenwick E, Claxton K, Sculpher M. The value of implementation and the value of information: combined and uneven development. Medical Decision Making. 2008;28(1):21-32.

83. Sandberg S, Fraser CG, Horvath AR, Jansen R, Jones G, Oosterhuis W, et al. Defining analytical performance specifications: consensus statement from the 1st Strategic Conference of the European Federation of Clinical Chemistry and Laboratory Medicine. Clinical Chemistry and Laboratory Medicine (CCLM). 2015;53(6):833-5.

Appendix

Table A1. Key terminology relating to measurement uncertainty[5]

	Synonyms	General definition	Definition from CLSI Harmonized Terminology Database (using standard, internationally preferred terms where provided)
Measurand	Analyte	Substance intended to be measured by a given test.	Quantity intended to be measured.
Precision	-	The closeness of agreement between repeated tests.	Closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions.
Imprecision	-	Random error in measurement.	The random dispersion of a set of replicate measurements and/or values expressed quantitatively by a statistic, such as standard deviation or coefficient of variation.
Trueness	Accuracy	The closeness of agreement between observed test results and the underlying 'true' value.	Closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value.
Bias	-	Systematic error in measurement.	Estimate of a systematic measurement error.
Repeatability	Within-run precision; Intra-assay precision; Intra-operator precision	Level of imprecision observed when conducting repeated testing one after another (in the same batch or run) on the same day, by the same operator, using the same method and equipment and in the same laboratory.	Measurement precision under a set of repeatability conditions of measurement. Repeatability condition of measurement = condition of measurement, out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time.
Intermediate precision	Within-laboratory precision; inter-operator precision	Level of imprecision observed when conducting repeated testing within the same laboratory but altering one or more of the following factors: time, operator, calibration, environment and equipment.	Measurement precision under a set of intermediate precision conditions of measurement. Intermediate precision conditions of measurement = condition of measurement, out of a set of conditions that includes the same measurement procedure, same location, and replicate measurements on the same or similar objects over an extended period of time, but may include other conditions involving changes.
Reproducibility	Between-laboratory precision	Level of imprecision observed when conducting repeated testing across different laboratories, in which the following factors would be expected to vary: time, operator, calibration, environment and equipment.	Measurement precision under reproducibility conditions of measurement. Reproducibility conditions of measurement = condition of measurement, out of a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects.

[5] This table provides two definitions for each term: first, the general definitions used in this paper, which are intended as far as possible to be accessible to a lay audience; second, formal definitions from the Clinical and Laboratory Standards Institute (CLSI) Harmonized Terminology Database (https://clsi.org/resources/harmonized-terminology-database/), using internationally preferred terms where these are provided.

Reference Measurement Procedure	-	An officially validated test method which has been shown to accurately measure the measurand of interest.	Measurement procedure accepted as providing measurement results fit for their intended use in assessing measurement trueness of measured quantity values obtained from other measurement procedures for quantities of the same kind, in calibration, or in characterizing reference materials.
Certified reference material	-	Samples of known composition produced under tightly controlled manufacturing or in-house procedures to provide reliable sources of measurement.	Reference material [defined as: material, sufficiently homogeneous and stable with reference to specified properties, which has been established to be fit for its intended use in measurement or in examination of nominal properties] accompanied by documentation issued by an authoritative body and providing one or more specified property values with associated uncertainties and traceabilities, using valid procedures.
Bland-Altman Analysis	Difference plot	A scatter plot of the difference between the new and reference test results vs. the mean of the paired results, allowing estimation of mean difference, limits of agreement, outliers and constant and proportional bias.	[Defined as "difference plot"] A plot of the difference between a measured value and a reference concentration plotted on the y-axis vs the reference concentration on the x-axis.
Selectivity	Analytical specificity	The ability of a test to measure the target measurand of interest as opposed to any other components in the test sample.	Property of a measuring system, used with a specified measurement procedure, whereby it provides measured quantity values for one or more measurands such that the values of each measurand are independent of other measurands or other quantities in the phenomenon, body, or substance being investigated
Interference	-	The existence of obstruction from substances in the test sample which either inhibit the process of binding with the target measurand.	Artifactual increase or decrease in apparent concentration or intensity of an analyte (measurand) due to the presence of a substance that reacts nonspecifically with either the detecting reagent or the signal itself.
Cross-reactivity	-	The existence of obstruction from substances in the test sample which are mistaken for the target measurand leading to ‘unintentional’ binding.	The ability of a drug, metabolite, a structurally similar compound other than the primary measurand, or even an unrelated compound to affect the assay
Pre-analytical phase	-	All processes in a testing procedure occurring prior to the point of sample analysis.	[Defined as "preexamination processes"] Processes starting, in chronological order, from the request for examination and including the examination requisition, preparation of the patient, collection of the primary sample, and transportation to or within the laboratory, and ending when the analytical examination procedure begins
Analytical phase	-	All processes in a testing procedure occurring at the point of sample analysis.	[Defined as "examination procedure"] Set of operations, described specifically, used in the performance of examinations according to a given method
Biological variation	-	Variation in the concentration of bodily fluid components either within an individual over time, or between-individuals.	Consists of within-subject (CVI, intra-individual) and between-subject (CVG, inter-individual, group) variation.
Limit of Blank (LoB)	-	The highest (apparent) concentration of measurand expected to be identified when processing blank samples (i.e. samples containing zero quantity of measurand).	The highest measurement result that is likely to be observed (with a stated probability [alpha]) for a blank sample.

Limit of Detection (LoD)	**-**	The lowest measurand concentration which the test can reliably distinguish from the LoB.	Measured quantity value, obtained by a given measurement procedure, for which the probability of falsely claiming the absence of a component in a material is β, given a probability α of falsely claiming its presence.
Limits of Quantification (LoQ)	**-**	The lower and upper concentration of measurand in a sample that a test can detect with a specified level of imprecision and trueness.	The existence of obstruction from substances in the test sample which are mistaken for the target measurand leading to ‘unintentional’ binding.
Total error (TE)	**-**	An upper limit on the expected error within a given measurement, calculated as a linear sum of random error (imprecision) and systematic error (bias).	The combined impact of any set of defined precision and bias errors that can affect the accuracy of an analytical result.
Uncertainty of Measurement (U_M)	**-**	A parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand.	Non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used.

Supplementary Material

Cochrane CRD database search strategy

Date Run: 01/03/17 11:10:24
Description:

ID	Search	[Hits]
#1	MeSH descriptor: [Diagnosis] explode all trees	[301036]
#2	MeSH descriptor: [Reagent Kits, Diagnostic] explode all trees	[351]
#3	MeSH descriptor: [Investigative Techniques] explode all trees	[440907]
#4	MeSH descriptor: [Precision Medicine] explode all trees	[251]
#5	MeSH descriptor: [Biomarkers] explode all trees	[18996]
#6	#1 or #2 or #3 or #4 or #5	[491884]
#7	"in vitro*" or test* or assay* or microarray* or "micro array*" or urinalys?s or ELISA* or diagnos* or biomarker* or marker* or signature* or investigat* (Word variations have been searched)	[450916]
#8	monitor* or screen* or prognos* or predict* or diagnos* or stratif* or detect* (Word variations have been searched)	[297700]
#9	(analytic* near/2 valid*) or sensitiv* or specific* or (positiv* near/2 predict*) or (negativ* near/2 predict*) or "true positive*" or "true negative*" or "false positive*" or "false negative*" or ((pre-test* or pretest*) near/2 probability) or ("post test*" near/2 probability) or "likelihood ratio*" (Word variations have been searched)	[136399]
#10	#7 or #8 or #9	[560305]
#11	#6 or #10	[735569]
#12	MeSH descriptor: [Economics] this term only	[63]
#13	MeSH descriptor: [Economics, Nursing] this term only	[19]
#14	MeSH descriptor: [Economics, Pharmaceutical] this term only	[244]
#15	MeSH descriptor: [Economics, Hospital] explode all trees	[1774]
#16	MeSH descriptor: [Economics, Medical] explode all trees	[105]
#17	MeSH descriptor: [Economics, Dental] explode all trees	[10]
#18	MeSH descriptor: [Costs and Cost Analysis] explode all trees	[25219]
#19	MeSH descriptor: [Fees and Charges] explode all trees	[506]
#20	MeSH descriptor: [Budgets] explode all trees	[72]
#21	MeSH descriptor: [Value of Life] explode all trees	[146]
#22	MeSH descriptor: [Quality-Adjusted Life Years] explode all trees	[4194]
#23	MeSH descriptor: [Quality of Life] explode all trees	[19488]
#24	MeSH descriptor: [Models, Economic] explode all trees	[2012]
#25	MeSH descriptor: [Markov Chains] explode all trees	[2161]
#26	cost* or pharmacoeconomic* or pharmaco-economic* or economic* or price* or pricing* or budget* or eq5d* or eq-5d* or euroquol* or euroqol* or euroqual* or euro-quol* or euro-qol* or euro-qual* or finance* or financial* or fee or fees or "economic model*" or markov* or "quality adjusted life" or qaly* or qald* or qale* or qtime* or "disability adjusted life" or daly* or SF6D or "sf 6d" or "short form 6d" or shortform6d or "health* year* equivalent*" or hye or hyes or "health utilit*" or hui or hui1 or hui2 or hui3 or disutil* or "standard gamble*" or "time trade off" or time tradeoff or tto or (value near/2 money) or (value near/2 monetary) or hql or hqol or "h qol" or hrqol or "hr qol" or pqol or qls (Word variations have been searched)	[94942]
#27	Cost* near/2 (effective* or utilit* or benefit* or minimi* or evaluat* or analy* or study or studies or consequenc* or compar* or efficienc*) (Word variations have been searched)	[43044]
#28	#12 or #13 or #14 or #15 or #16 or #17 or #18 or #19 or #20 or #21 or #22 or #23 or #24 or #25 or #26 or #27	[108553]
#29	#11 and #28	[90863]
#30	#29 in Technology Assessments	[2036]
#31	#30 Publication Year from 1999 to 2017	[1908]