UNIVERSITY OF LEEDS

This is a repository copy of *Towards a framework for outcome-based analytical performance specifications: a methodology review of indirect methods for evaluating the impact of measurement uncertainty on clinical outcomes*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/148859/

Version: Accepted Version

Article:
Smith, AF orcid.org/0000-0001-7709-1869, Shinkins, B orcid.org/0000-0001-5350-1018, Hall, PS et al. (2 more authors) (2019) Towards a framework for outcome-based analytical performance specifications: a methodology review of indirect methods for evaluating the impact of measurement uncertainty on clinical outcomes. Clinical Chemistry, 65 (11). pp. 1363-1374. ISSN 0009-9147

https://doi.org/10.1373/clinchem.2018.300954

© 2019, American Association for Clinical Chemistry. This is an author produced version of an article published in Clinical Chemistry. Uploaded in accordance with the publisher's self-archiving policy.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Towards a framework for outcome-based analytical performance specifications: a methodology review of indirect methods for evaluating the impact of measurement uncertainty on clinical outcomes

Authors: Alison F. Smith[1, 2], Bethany Shinkins[1,2], Peter S. Hall[4], Claire T. Hulme[1,2,5], Mike P. Messenger[2,3]

Affiliations:

[1]Test Evaluation Group, Academic Unit of Health Economics, University of Leeds, Leeds, UK

[2]NIHR Leeds In Vitro Diagnostic (IVD) Co-operative, Leeds, UK

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

[4] Cancer Research UK Edinburgh Centre, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, UK

[5]Health Economics Group, University of Exeter, Exeter, UK

Corresponding Author Contact Details:

Alison F. Smith

Research Fellow

Test Evaluation Group, Academic Unit of Health Economics, University of Leeds, Leeds, UK

Keywords

Measurement performance specifications; measurement uncertainty; analytical error; evidence-based laboratory medicine

Journal Categories

Evidence-Based Laboratory Medicine and Test Utilization (TUO)

Manuscript details:

Word count: 4,342

Number of tables: 3

Number of figures: 3

Supplemental material: Yes

List of Abbreviations

EFLM = European Federation of Clinical Chemistry and Laboratory Medicine

ROC = Receiver operator characteristic

AUC = Area under the curve

CV = coefficient of variation

SD = standard deviation

EQA = External Quality Assessment

QALY = quality adjusted life year

Abstract

Background: For medical tests that have a central role in clinical decision-making, current guidelines advocate outcome-based analytical performance specifications. Given that empirical (clinical-trial style) analyses are often impractical or unfeasible in this context, the ability to set such specifications is expected to rely on indirect studies to calculate the impact of test measurement uncertainty on downstream clinical, operational and economic outcomes. Currently however, a lack of awareness and guidance concerning available alternative indirect methods is limiting the production of outcome-based specifications. Our aim therefore was to review available indirect methods and present an analytical framework to inform future outcome-based performance goals.

Content: A methodology review consisting of database searches and extensive citation tracking was conducted to identify studies using indirect methods to incorporate or evaluate the impact of test measurement uncertainty on downstream outcomes (including clinical accuracy, clinical utility and/or costs). Eighty-two studies were identified, most of which evaluated the impact of imprecision and/or bias on clinical accuracy. A common analytical framework underpinning the various methods was identified, consisting of three key steps: (1) calculation of "true" test values; (2) calculation of measured test values (incorporating uncertainty); and (3) calculation of the impact of discrepancies between (1) and (2) on specified outcomes. A summary of the methods adopted is provided, and key considerations discussed.

Conclusions: Various approaches are available for conducting indirect assessments to inform outcome-based performance specifications. This study provides an overview of methods and key considerations to inform future studies and research in this area.

Introduction

Although systematic and random variation around measured test values (henceforth, measurement uncertainty) is now routinely documented within the clinical laboratory, the potential impact of this uncertainty on downstream clinical, operational and economic outcomes is rarely quantified. Meanwhile, evaluation of the impact of measurement uncertainty on clinical outcomes has become a recurring recommendation in protocols for determining analytical performance specifications. In their recently updated guidance, for example, the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) stipulate that, for medical tests that "have a central role in the decision-making of a specific disease or clinical situation and where cut-off/decision limits are established", specifications should be based on the effect of analytical performance on the clinical outcome [termed "Model 1"], as opposed to basing specifications on biological variation ["Model 2"] or state of the art measurements ["Model 3"] (1).

Two types of studies are suggested to inform specifications under Model 1: (i) direct outcome studies (i.e. analyses based solely on empirical data, such as randomised controlled trials evaluating the impact of varying analytical procedures on outcomes); or (ii) indirect outcome studies (i.e. analyses using non-empirical approaches, such as decision analytic modelling, to determine the impact of varying procedures on outcomes) (2). Since (i) is often unfeasible or impractical due to ethical, financial and time constraints associated with robust end-to-end test-outcome studies, the indirect methods of (ii) are expected to play the dominant role in this context (3).

Despite general agreement that outcome-based specifications provide the best mechanism to ensure tests best serve patients' needs, studies in this area remain uncommon. A primary reason often cited for this concerns the inherent difficulties in conducting direct outcomes studies (1, 3). It is likely, however, that a lack of awareness and specific guidance concerning alternative indirect methods that may be employed is also a key limiting factor. The aim of this study therefore was to review methodological approaches used in previous indirect assessments and outline an analytical framework to inform future outcome-based performance specifications.

Methods

A literature search was conducted in November 2017 across four databases (Ovid Medline(R), Embase, Web of Science (core collection) and Biosis Citation Index) and covering a 10 year publication period (2008 to November 2017). The search was subsequently updated in 2019 (covering the period 2008 to March 2019). The search strategy (provided in the **Supplemental Appendix**) combined key terms relating to (a) tests, (b) measurement uncertainty, and (c) simulation/ methodology. From those studies identified via the database searches, subsequent citation tracking (including extensive backwards and forwards tracking) was conducted to identify additional studies published on any date (i.e. including studies published before 2008).

Studies were included if they met the inclusion criteria shown in **Table 1**. Studies were required to include an assessment of downstream outcomes including: clinical accuracy (the ability of a test to distinguish between patients with and without a specified condition, or identify a change in condition), clinical utility (the ability of a test to impact on healthcare management decisions or patient health outcomes) and/or cost-effectiveness (the ability of a test to produce an efficient impact on health outcomes in relation to cost). Note that studies using indirect methods at any stage of the analysis were eligible for inclusion; this means, for example, that several method-comparison studies (an essentially empirical study design) were nevertheless included in cases where an indirect method was subsequently used to assess the impact of identified measurement discrepancies on outcomes.

<<Table 1>>

All screening (including initial title/abstract screening, full text screening, and citation tracking) was conducted by the primary reviewer (AS). A data extraction form was developed (including items on key study, test, and method details) and piloted on the first 10% of included studies. Subsequent full data extraction of included studies was conducted by the primary reviewer and double checked by one of four secondary reviewers (BS, MM, CH and PH). Regular meetings with all authors were conducted to review the ongoing study findings and resolve (via group consensus) any inclusion and/or extraction uncertainties.

Results

Study characteristics

A total of 82 studies were identified (see **Figure 1**). Regarding data extraction checking, 35 papers (43%) were checked by BS; 16 (20%) by CH; 16 (20%) by MM; and 15 (18%) by PH. Agreement between reviewers across extraction items was >99%.

Study characteristics are summarized in **Table 2**, and details of measurement uncertainty components and test outcomes evaluated are provided in **Table 3**. Most studies focused on evaluating tests or devices used for the purposes of monitoring, diagnosis and/or screening across four key disease areas: diabetes or glycemic control, cardiovascular diseases, cancer and metabolic or endocrine disorders. Imprecision was most commonly addressed, followed by bias and total error, and studies primarily evaluated clinical accuracy outcomes.

<<Figure 1>>

<<Table 2>>

<<Table 3>>

Aim of analyses

Most studies were conducted with the objective of either: (i) determining/ informing analytical performance specifications (4-22); (ii) exploring the impact of uncertainty allowed by current performance specifications (23-34); or (iii) evaluating the potential impact of measurement uncertainty on outcomes (without explicitly defining specifications) (35-78). A final group of studies consisted of “incidental” analyses, in which the impact of measurement uncertainty on outcomes was incorporated within the analysis but was not part of the primary study aim (79-85).

Methodology Framework

Based on the included studies, a common analytical framework underpinning the various approaches to evaluating the impact of measurement uncertainty on outcomes was identified. This framework consists of three key steps: (1) calculation of "true" test values; (2) calculation of measured test values (i.e. incorporating measurement uncertainty); and (3) calculation of the impact of discrepancies between (1) and (2) on the outcome(s) under consideration. An outline of the various methods adopted within this framework is provided below and summarized in **Figure 2**. A summary table detailing the methods used in each individual study is provided in **Supplemental Table 1**.

1. Step one: calculation of "true" test values

Calculation of "true" test values was based either on empirical data values (5, 7, 9-11, 18, 21, 26, 30-32, 34-37, 39-42, 45, 49-53, 56-58, 60, 61, 64, 66-69, 71, 74, 77, 78, 85) and/or simulated values (4-6, 8, 12-17, 19, 20, 22-25, 27-29, 33, 36, 38, 43, 44, 46-48, 54, 55, 59, 62, 63, 65, 70, 72-76, 79-84).

Studies using empirical data here included: (i) method comparison and external quality assessment (EQA) studies, which utilized indirect methods to determine the impact of discrepancies between empirical reference (i.e. "true") test measurements vs. index (i.e. uncertain) test measurements on specified outcomes (e.g. using the "error grid" approach outlined in Step 3) (35, 37, 41, 42, 51, 53, 56-58, 60, 64, 66-69, 71, 75, 78); and (ii) studies which derived uncertain measurements from "true" empirical data values using various (non-empirical) approaches outlined in Step 2 (5, 7, 9-11, 18, 21, 26, 30-32, 34, 36, 39, 40, 45, 48-50, 52, 61, 77, 85).

Studies using simulation methods here used a range of approaches – the simplest of which was to assume a fixed set of individual "true" values specified across the measurement range and simulate uncertainty around these values (see Step 2) (12, 16, 27, 33, 36, 38, 79, 83, 84). Whilst this approach does not require any simulation for the "true" measurements per se, the values here are nevertheless generated rather than using real-world data directly. An extension of this approach is to assume a uniform distribution to describe the "true" frequency distribution(s): that is, assume a constant probability of occurrence for each test value along a specified measurement range, and draw from this distribution within the simulation (14, 17, 19, 44, 55). Alternatively, the expected likelihood of test values was often modelled using Gaussian (i.e. normal) or log-Gaussian frequency distributions, specified using published or empirical data on the expected mean and variance of test values (4-6, 8, 13-15, 20, 46, 47, 59, 63, 65). Other infrequently adopted parameterizations included mixed Gaussian distributions (54, 62), multivariate Gaussian distributions (where correlations between tests are known (43)) and the exponential distribution (82). Non-parametric simulation approaches were also used, based on sampling with replacement from an empirical dataset (18, 30). Finally, several studies used simulation techniques (22, 23, 70, 74, 75), or utilized findings from previously published simulation studies (24, 25, 73, 76), but did not clearly report details regarding the calculation of "true" baseline values.

An important issue with respect to the estimation of "true" test values concerns how well the underlying data may be considered a reliable proxy for the truth. A handful of studies attempted to directly address this issue, by "stripping" known measurement uncertainty from baseline "true" test values via statistical adjustment: imprecision, for example, can be removed from the variance term of a specified Gaussian/log-Gaussian distribution using a reverse form of the "sum of squares rule"; whilst bias can be removed from the mean term (7-10, 13, 15, 31). In general, however, the likelihood that the adopted "true" test values would in fact be representative of the truth was either implicitly assumed or not discussed.

2. Step two: calculation of measured test values (incorporating measurement uncertainty)

Approaches to the calculation of measured test values predominantly fell into four broad categories: (1) empirical assessment (35, 37, 41, 42, 51, 53, 56-58, 60, 64, 66-69, 71, 74, 78), (2) graphical assessment (5, 7, 9-11, 36), (3) computer simulation (4-6, 8, 12, 14-25, 27-31, 34, 38, 39, 44, 46, 49, 50, 52, 54, 55, 59, 61-63, 65, 70, 72-77, 79-85), or regression analysis (26, 32, 43, 47).

Studies using empirical assessment here included method-comparison studies (35, 37, 41, 42, 53, 56-58, 60, 64, 66-69, 71, 75, 78) and an EQA study (51) which based "true" test values on the specified reference test and measured values on the index test measurements.

An alternative method, first appearing in 1980, is based on applying hypothetical measurement uncertainty to "true" values via graphical manipulation (5, 7, 9-11, 36). This approach centers on plotting the cumulative percentage frequency of "true" values on the probit scale (x-axis) as a function of "true" values on the logarithmic scale (y-axis); assuming that the log-transformed data are Gaussian, then in the bimodal case (where healthy and diseased populations are modeled separately), cumulating the healthy (diseased) population from high (low) values results in two straight lines sloping in opposite directions for each population (i.e. forming an 'X' on the plot). The addition of negative (positive) bias is then explored by shifting the straight lines to the left (right) on the x-axis; whilst the addition of imprecision is explored by rotating each line around their mean value (i.e. broadening the 95% confidence interval of the values on the probit scale). Given a specified cut-off threshold, the proportion of false positives and negatives at a particular level of bias and imprecision can be read off directly from this plot, by observing the point at which healthy/diseased populations cross the threshold line.

In response to modern computational capabilities, the graphical method has been superseded by computer simulation approaches which can accommodate more complex specifications of the measurand distribution and measurement uncertainty. The most flexible and widely adopted approach in the identified studies was based on iterative simulation, with uncertainty added on to “true” test values according to a specified error model – a function relating measured test values to baseline “true” values plus specified components of measurement uncertainty (14, 17-19, 28-30, 34, 54, 62, 79, 82-84). This method is largely attributed to the seminal 2001 paper by Boyd and Bruns (14) – the first study of this kind to clearly specify the error model as a mathematical function (as opposed to earlier (4-6) and later (21-25, 44, 49, 52, 70, 72, 73, 76, 77, 80, 81, 85) studies limited to textual descriptions or indirect referencing). An example of a typical error model is as follows:

$$\mathbf{Test_{mesaured} = Test_{true} + [\, Test_{true} * N(0,1) * CV \,] + Bias} \qquad \mathbf{(1)}$$

where $Test_{true}$ is the “true” measurement value; $Test_{measured}$ is the observed test value measured with imprecision (coefficient of variation [CV%]) and absolute bias (Bias); and N(0,1) is a normal distribution (mean = 0, standard deviation [SD] = 1) applied with the CV% value in order to produce a spread of Gaussian-distributed results around $Test_{true}$.

The error model iterative simulation approach works as follows: (i) a random draw is taken from the distribution of “true” values to generate a value for $Test_{true}$; (ii) components of measurement uncertainty are applied to $Test_{true}$ according to the error model formula to simulate a value for $Test_{measured}$ (this may require random number draws – for example in equation (1) a random draw from N(0,1) is required for the application of imprecision); (iii) points (i) and (ii) are repeated (e.g. 10,000 times to simulate 10,000 $Test_{true}$ and $Test_{measured}$ values) for a given level of measurement uncertainty (e.g. CV% = 5% and Bias = 5%); and (iv) points (i) to (iii) are repeated for varying levels of measurement uncertainty (e.g. CV%

ranging from 0-20% and Bias ranging from +/-10% in 1% increments). This iterative process can be efficiently implemented using standard statistical software, such as Excel or R.

Rather than iteratively adding on uncertainty via error model simulation, an alternative approach is to incorporate uncertainty directly within a specified probability distribution (e.g. incorporating bias within the mean term, and imprecision within the variance term of a Gaussian or log-Gaussian distribution). This distribution can be applied iteratively around individual "true" values (12, 16, 18, 27, 30, 38, 46, 59, 61), or at a population level, by adjusting a specified "true" population distribution to include additional uncertainty (8, 15, 31, 63, 65).

The remaining studies used regression analysis (26, 32, 43, 47), other one-off methods (12, 13, 33, 40, 45, 48), or reported insufficient details regarding simulation techniques to determine the exact method employed (74, 75). Within the identified regression analyses, bias or total error was applied as a multiplicative factor to baseline measurements within a specified regression model, with the resulting impact on the regression output (e.g. likelihood ratio) explored. Details of studies using other one-off/ indeterminate methods can be found in **Supplemental Table 1**.

3. Step three: calculation of the impact on test outcomes

The final step is to assess the impact of deviations between "true" and measured values on the outcome(s) of interest.

Most studies focused on evaluating clinical accuracy (4-13, 15, 16, 20, 26-29, 31-33, 38, 39, 43, 45-52, 55, 59, 61-63, 65, 79-85). In this case the calculation is generally straightforward: the rate of change in mis-categorizations (e.g. false positive/negative diagnoses) is determined according to the change in the proportion of measured values pushed above or below the given test cut-off threshold(s) used to define disease status or inform treatment decisions, compared to the “true” value classifications. This was the typical approach taken in studies using the graphical and simulation approaches outlined in Step 2, for example.

Several studies evaluated the impact of measurement uncertainty on treatment management decisions (14, 18, 21, 30, 35, 37, 41, 42, 51, 53, 56-58, 60, 64, 66-69, 71, 74, 75, 78). Most of these were method-comparison studies which determined the impact of measurement deviations on treatment decisions using error grid analysis (35, 37, 41, 42, 53, 56-58, 60, 64, 66-69, 71, 74, 78). Two studies similarly employed the error grid approach, but used simulated (rather than empirical) reference and index test measurements (74, 75). First developed in the 1980s, the original error grid aimed to evaluate the potential impact of measurement discrepancies between self-monitoring blood glucose devices and laboratory reference measurements in terms of insulin dosing errors (35). Using a scatter plot of reference vs. index test measurements, the plot was divided into five error grid “zones” according to assumed severity of associated dosing errors (from zone A = clinically accurate results; to zone E = erroneous results leading to dangerous failure to detect and treat). More recently studies have attempted to build on this approach, for example by expanding on the small sample of experts used to define the initial error grid (37, 74, 75), accounting for temporal aspects of measurement (41), or applying the same methodology to alternative clinical settings (64).

Others have attempted to incorporate the impact of measurement uncertainty on patient health outcomes (17, 19, 22, 23, 44, 54, 70, 72). All of these studies related to evaluations of monitoring devices for glycemic control, in which health outcomes such as hypoglycemia and hyperglycemia were determined using decision analytic models based around sequential glucose measurements (incorporating measurement uncertainty via the error model simulation approach, for example). Combined with data on insulin dose administrations (resulting from measured values), and additional factors such as patient insulin sensitivity and gluconeogenesis, these models were used to track patients’ response to administered doses and resulting health outcomes.

Nine final studies included an assessment of costs or cost-effectiveness (7, 8, 11, 24, 25, 40, 73, 76, 77). Four were based on a simple assignment of expected costs of misdiagnoses to rates of false positive/negative results (7, 8, 11), or expected costs of adverse events applied to simulated health outcomes data (77). One study included a more comprehensive costing analysis, in which the potential financial implications of calibration bias in serum calcium testing was explored (40). The remaining four studies all utilized the previous work of Breton and Kovatchev (2010), in which the impact of reduced glucose meter imprecision on glycemic events was simulated using a published simulation platform (23). Two studies constructed simple cost-consequence decision models, combining the Breton and Kovatchev (2010) findings with data on patient population numbers, glucose meter costs, and the rate of myocardial infarctions resulting from glycemic outcomes, to estimate annual cost savings associated with improved meter precision (73, 76). Two more recent studies conducted full cost-effectiveness analyses, using cohort Markov (i.e. state-transition) models to link the data on improved glycemic control and reduced glycemic event rates, with data on diabetes complication rates, patient health-related quality of life and health service costs (24, 25). Using these models the authors were able to estimate the incremental cost per additional quality adjusted life year (QALY) associated with reduced device error.

<<Figure 2>>

Discussion

Review findings

Based on our methodology review findings, a three-step analytical framework underpinning the various approaches to determining the impact of measurement uncertainty on outcomes was identified (see **Figure 2**). Key points for consideration within this framework are discussed below.

With regards to Step 1 (calculation of “true” test values), the primary advantage of using either empirical data or informed parametric distributions is that, by accounting for the expected frequency of values, population-level conclusions (such as analytical performance specifications) may be derived. In contrast, the primary drawback of the fixed-values approach, and by extension the uniform distribution approach (assuming this is not a realistic parameterization), is that population-level conclusions cannot be derived. Nevertheless, such approaches may be useful for exploring the impact of measurement uncertainty in specific scenarios – for example, to explore the impact of uncertainty on test values close to the test cut-off threshold.

A question that must be considered when using either empirical or parametric distributions, is how well the underlying data may be considered to represent the truth. If values used to inform the “true” distributions are themselves subject to measurement uncertainty (even if this uncertainty is expected to be small), then all subsequent analyses may be affected by this confounding factor and care should be taken when asserting absolute maximum bounds for imprecision and bias. A handful of studies did attempt to address this issue using statistical adjustment methods however this approach depends on having reliable information on the expected measurement uncertainty contained in the baseline “true” measurement values and can only be used when modelling test values as parametric distributions (7-10, 13, 15, 31).

A second consideration in the adoption of parametric distributions concerns the appropriateness of the assumed parametric form. Whilst a minority of studies provided some form of justification for the parametric choice (e.g. using the Kolmogorov–Smirnov test for normality), a common implicit assumption was that data would be likely to be Gaussian or log-Gaussian distributed. The validity of this assumption is not always clear, however.

Within Step 2 (calculation of measured test values) computer simulation methods offer the most flexible approach for exploring alternative specifications and levels of measurement uncertainty. In the context of setting performance goals, studies based on method-comparison analyses are of limited use given the fact that alternative levels of measurement uncertainty cannot be efficiently explored, and analyses using the graphical method suffer from the issue that non-Gaussian parameterisations or non-constant/ non-linear specifications of bias or imprecision cannot be accommodated. The error model approach is particularly useful in this respect. While the example formula provided in Equation (1) specifies one CV% element representing total imprecision, additional elements of imprecision (e.g. pre-analytical, analytical and biological) may be separately specified. Alternative characterisations of imprecision may also be defined: for example, using (i) a fixed SD, (ii) different SD/CV values for different sections of the measurement range, or (iii) imprecision defined as a linear/ non-linear function of $Test_{true}$. Similarly bias may also be characterised in alternative ways.

With regards to Step 3 (calculation of the impact on outcomes), a further advantage of the simulation approach is that, by sampling over a range of bias and imprecision values, the joint impact of these components on outcomes can be clearly explored. In particular, several studies used contour plots to present their findings (14-19, 21, 30, 34, 62): an example, provided in **Figure 3**, represents a hypothetical case in which bias and imprecision have been applied (according to equation (1)) to normally distributed healthy [N(30,5)] and diseased

[N(60,10)] populations. The plotted lines indicate at which values of imprecision and bias a given value of clinical sensitivity/specificity is maintained. For example in this case, at imprecision=0, increasing positive bias decreases clinical specificity and increases clinical sensitivity, whilst negative bias has the opposite effect. Based on this plot, we expand on the typical contour plot to show how maximum allowable bounds for imprecision and bias can be identified according to specified minimum requirements for clinical accuracy. Suppose, for example, that we require sensitivity to remain above 90% and specificity to remain above 80% in order to maintain expected health utility gains. The region of acceptable analytical bias and imprecision values for this specification of clinical accuracy is illustrated by the shaded region of the contour plot – from this we can see that, if bias is zero we can tolerate up to 20% imprecision, whilst if imprecision is zero we can tolerate -8 to +6 units of absolute bias. Plots such as this one offer an effective means of highlighting acceptable bounds for measurement uncertainty.

<<Figure 3>>

Whilst most studies focused on the intermediate outcome of clinical accuracy, ideally technologies should be evaluated in terms of their influence on “end-point” outcomes i.e. health outcomes (clinical utility), operational and/or cost-effectiveness outcomes. Several of the identified studies utilized analytic decision modeling techniques to determine the impact of measurement uncertainty on health outcomes: while these all related to the context of glycemic control devices, decision models can feasibly be used to explore any clinical pathway of interest, subject to data availability. Within the field of health technology assessment, for example, decision models are routinely employed to evaluate the expected clinical utility and cost-effectiveness of novel tests, by linking data on disease prevalence and test clinical accuracy (e.g. the proportion of correct and incorrect diagnoses), with downstream data on the expected change in patient management, patient compliance to treatment and treatment effectiveness (often referred to as the “linked-evidence approach”) (86-88). Although this approach is more resource- and data-intensive, and care must be taken to ensure that the model structure appropriately reflects key aspects of the clinical pathway, it nevertheless has the advantage of explicitly capturing the impact of additional parameters (e.g. treatment effectiveness) on end-point outcomes (which may not always produce expected or intuitive results) and uncertainty around the exact values of these parameters can be quantitatively characterised in the model framework (89). We identified two recent studies which utilized health-economic models to estimate the cost-effectiveness of improved analytical performance (24, 25). These studies explored a limited set of fixed imprecision levels relating to pre-existing performance specifications: future studies could extend this methodology to explore a broader range of measurement uncertainty values (e.g. by linking error-model simulations with the downstream health-economic modelling) and derive de novo performance specification based on maintaining or optimizing cost-utility and cost-effectiveness outcomes.

Strengths and limitations:

In light of the sustained international focus on outcome-based analytical performance specifications, it is expected that the indirect approaches outlined in this study will become increasingly important. The analytical framework presented in this study provides a useful starting point to inform future studies in this area, by clearly outlining available methods in sufficient detail to enable practical implementation, and highlighting possible advantages and limitations to consider under each approach. Whereas previous studies have provided commentaries and general reviews of various approaches to setting analytical performance specifications (3, 90, 91), this is the first methodology review to focus specifically on indirect methods for setting outcome-based performance specifications.

As a methodology review, the aim of this study was not to systematically identify all evidence, but rather to ensure that key examples of relevant methods were identified. While we attempted to make the database search as sensitive as possible, due to the vast volume of literature in this area we necessarily had to focus the search strategy by: (i) concentrating on terms related to in-vitro biomarkers, (ii) including a filter for simulation and methodology terms, and (iii) restricting the initial database search period to 10 years. Extensive citation tracking was additionally conducted, extending into preceding years, in order to ensure that seminal papers informing modern practices would be identified in addition to current state-of-the-art methodology. Although we believe that this two-stage strategy will have captured key methodologies, not all relevant material relating to each method will have been identified and we cannot therefore draw definitive conclusions regarding the frequency that each method has been used. Nevertheless, we believe our findings provide a valuable overview of indirect study methods and an informative starting point for future studies in this area.

Acknowledgements

The authors would like to thank the following individuals for their feedback on the project plan and/or manuscript: Christopher Hyde (Exeter, UK), Christopher Bojke (Leeds, UK), Rebecca Kift (Leeds, UK), Joy Allen (Newcastle, UK), Jon Deeks (Birmingham, UK), James Turvill (York, UK), Natalie King (Leeds, UK) and the anonymous reviewers.

Funding

Alison Smith is supported by the NIHR Doctoral Research Fellowship programme (DRF-2016-09-084). Dr Bethany Shinkins and Dr Mike Messenger are also supported by the NIHR Leeds In Vitro Diagnostics Co-operative. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Ceriotti F, Fernandez-Calle P, Klee GG, Nordin G, Sandberg S, Streichert T, et al. Criteria for assigning laboratory measurands to models for analytical performance specifications defined in the 1st EFLM Strategic Conference. Clin Chem Lab Med 2017;55:189-94.
2. 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. Clin Chem Lab Med 2015;53:833-5.
3. Horvath AR, Bossuyt PM, Sandberg S, St John A, Monaghan PJ, Verhagen-Kamerbeek WD, et al. Setting analytical performance specifications based on outcome studies–is it possible? Clin Chem Lab Med 2015;53:841-8.
4. Groth T, Hakman M, Hällgren R, Roxin L-E, Venge P. 4.5. Diagnosis, size estimation and prediction of acute myocardial infarction from S-myoglobin observations. A system analysis to assess the influence of various sources of variability. Scand J Clin Lab Invest 1980;40:Suppl:S111-24.
5. Hørder M, Petersen PH, Groth T, Gerhardt W. 4.3. Influence of analytical quality on the diagnostic power of a single S-CK B test in patients with suspected acute myocardial infarction. Scand J Clin Lab Invest 1980;40:Suppl:S95-100.
6. Jacobson G, Groth T, Verdier C-HD. 4.1. Pancreatic iso-amylase in serum as a diagnostic test in different clinical situations. A simulation study. Scand J Clin Lab Invest 1980;40:Suppl:S77-84.
7. Petersen P, Rosleff F, Rasmussen J, Hobolth N. 4.2. Studies on the required analytical quality of TSH measurements in screening for congenital hypothyroidism. Scand J Clin Lab Invest 1980;40:Suppl:S85-93.
8. Groth T, Ljunghall S, De Verdier C-H. Optimal screening for patients with hyperparathyroidism with use of serum calcium observations. A decision-theoretical analysis. Scand J Clin Lab Invest 1983;43:699-707.
9. Nørregaard-Hansen K, Petersen PH, Hangaard J, Simonsen E, Rasmussen O, Horder M. Early observations of S-myoglobin in the diagnosis of acute myocardial infarction. The influence of discrimination limit, analytical quality, patient's sex and prevalence of disease. Scand J Clin Lab Invest 1986;46:561-9.
10. Wiggers P, Dalhøj J, Petersen PH, Blaabjerg O, Hørder M. Screening for haemochromatosis: Influence of analytical imprecision, diagnostic limit and prevalence on test validity. Scand J Clin Lab Invest 1991;51:143-8.
11. Arends J, Petersen PH, Nørgaard-Pedersen B. 6.1. 2.3 Prenatal screening for neural tube defects, quality specification for maternal serum alphafetoprotein analysis. Ups J Med Sci 1993;98:339-47.
12. Kjeldsen J, Lassen JF, Petersen PH, Brandslund I. Biological variation of International Normalized Ratio for prothrombin times, and consequences in monitoring oral anticoagulant therapy: computer simulation of serial measurements with goal-setting for analytical quality. Clin Chem 1997;43:2175-82.
13. von Eyben FE, Petersen PH, Blaabjerg O, Madsen EL. Analytical quality specifications for serum lactate dehydrogenase isoenzyme 1 based on clinical goals. Clin Chem Lab Med 1999;37:553-61.
14. Boyd JC, Bruns DE. Quality specifications for glucose meters: assessment by simulation modeling of errors in insulin dose. Clin Chem 2001;47:209-14.

15. Petersen PH, Brandslund I, Jørgensen L, Stahl M, Olivarius NDF, Borch-Johnsen K. Evaluation of systematic and random factors in measurements of fasting plasma glucose as the basis for analytical quality specifications in the diagnosis of diabetes. 3. Impact of the new WHO and ADA recommendations on diagnosis of diabetes mellitus. Scand J Clin Lab Invest 2001;61:191-204.
16. Petersen PH, Jørgensen LG, Brandslund I, De Fine Olivarius N, Stahl M. Consequences of bias and imprecision in measurements of glucose and HbA1c for the diagnosis and prognosis of diabetes mellitus. Scand J Clin Lab Invest 2005;65:Suppl:S51-60.
17. 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.
18. Karon BS, Boyd JC, Klee GG. Glucose meter performance criteria for tight glycemic control estimated by simulation modeling. Clin Chem 2010;56:1091-7.
19. Boyd JC, Bruns DE. Effects of measurement frequency on analytical quality required for glucose measurements in intensive care units: assessments by simulation models. Clin Chem 2014;60:644-50.
20. Petersen PH, Klee GG. Influence of analytical bias and imprecision on the number of false positive results using guideline-driven medical decision limits. Clin Chim Acta 2014;430:1-8.
21. Van Herpe T, De Moor B, Van den Berghe G, Mesotten D. Modeling of effect of glucose sensor errors on insulin dosage and glucose bolus computed by LOGIC-Insulin. Clin Chem 2014;60:1510-8.
22. 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;60:1500-9.
23. Breton MD, Kovatchev BP. Impact of blood glucose self-monitoring errors on glucose variability, risk for hypoglycemia, and average glucose control in type 1 diabetes: an in silico study. J Diabetes Sci Technol 2010;4:562-70.
24. McQueen RB, Breton MD, Craig J, Holmes H, Whittington MD, Ott MA, Campbell JD. Economic value of improved accuracy for self-monitoring of blood glucose devices for type 1 and type 2 diabetes in England. J Diabetes Sci Technol 2018;12:992-1001.
25. McQueen RB, Breton MD, Ott M, Koa H, Beamer B, Campbell JD. Economic value of improved accuracy for self-monitoring of blood glucose devices for type 1 diabetes in Canada. J Diabetes Sci Technol 2016;10:366-77.
26. Turner MJ, Baker AB, Kam PC. Effects of systematic errors in blood pressure measurements on the diagnosis of hypertension. Blood Press Monit 2004;9:249-53.
27. Jorgensen LG, Petersen PH, Brandslund I. The impact of variability in the risk of disease exemplified by diagnosing diabetes mellitus based on ADA and WHO criteria as gold standard. International Journal of Risk Assessment and Management 2005;5:358-73.
28. Turner MJ, Irwig L, Bune AJ, Kam PC, Baker AB. Lack of sphygmomanometer calibration causes over- and under-detection of hypertension: a computer simulation study. J Hypertens 2006;24:1931-8.
29. Turner MJ, van Schalkwyk JM, Irwig L. Lax sphygmomanometer standard causes overdetection and underdetection of hypertension: a computer simulation study. Blood Press Monit 2008;13:91-9.
30. Karon BS, Boyd JC, Klee GG. Empiric validation of simulation models for estimating glucose meter performance criteria for moderate levels of glycemic control. Diabetes Technol Ther 2013;15:996-1003.

31. Kuster N, Cristol JP, Cavalier E, Bargnoux AS, Halimi JM, Froissart M, et al. Enzymatic creatinine assays allow estimation of glomerular filtration rate in stages 1 and 2 chronic kidney disease using CKD-EPI equation. Clin Chim Acta 2014;428:89-95.
32. Åsberg A, Odsæter IH, Carlsen SM, Mikkelsen G. Using the likelihood ratio to evaluate allowable total error–an example with glycated hemoglobin (HbA1c). Clin Chem Lab Med 2015;53:1459-64.
33. Kroll MH, Garber CC, Bi C, Suffin SC. Assessing the impact of analytical error on perceived disease severity. Arch Pathol Lab Med 2015;139:1295-301.
34. Lyon ME, Sinha R, Lyon OA, Lyon AW. Application of a simulation model to estimate treatment error and clinical risk derived from point-of-care International Normalized Ratio device analytic performance. J Appl Lab Med 2017;2:25-32.
35. Clarke WL, Cox D, Gonder-Frederick LA, Carter W, Pohl SL. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. Diabetes care 1987;10:622-8.
36. Petersen PH, de Verdier C-H, Groth T, Fraser CG, Blaabjerg O, Hørder M. The influence of analytical bias on diagnostic misclassifications. Clin Chim Acta 1997;260:189-206.
37. Parkes JL, Slatin SL, Pardo S, Ginsberg BH. A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. Diabetes care 2000;23:1143-8.
38. Sölétormos G, Hyltoft Petersen P, Dombernowsky P. Progression criteria for cancer antigen 15.3 and carcinoembryonic antigen in metastatic breast cancer compared by computer simulation of marker data. Clin Chem 2000;46:939-49.
39. Rouse A, Marshall T. The extent and implications of sphygmomanometer calibration error in primary care. J Hum Hypertens 2001;15:587.
40. Gallaher MP, Mobley LR, Klee GG, Schryver P. The impact of calibration error in medical decision making. Washington: National Institute of Standards and Technology 2004.
41. Kovatchev BP, Gonder-Frederick LA, Cox DJ, Clarke WL. Evaluating the accuracy of continuous glucose-monitoring sensors: continuous glucose–error grid analysis illustrated by TheraSense Freestyle Navigator data. Diabetes Care 2004;27:1922-8.
42. Baum JM, Monhaut NM, Parker DR, Price CP. Improving the quality of self-monitoring blood glucose measurement: a study in reducing calibration errors. Diabetes Technol Ther 2006;8:347-57.
43. Nix B, Wright D, Baker A. The impact of bias in MoM values on patient risk and screening performance for Down syndrome. Prenat Diagn 2007;27:840-5.
44. Raine III C, Pardo S, Parkes J. Predicted blood glucose from insulin administration based on values from miscoded glucose meters. J Diabetes Sci Technol 2008;2:557-62.
45. Elloumi F, Hu Z, Li Y, Parker JS, Gulley ML, Amos KD, Troester MA. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med Genomics 2011;4:54.
46. Schlauch RS, Carney E. Are false-positive rates leading to an overestimation of noise-induced hearing loss? J Speech Lang Hear Res 2011;54:679-92.
47. Wright D, Abele H, Baker A, Kagan KO. Impact of bias in serum free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-A multiples of the median levels on first-trimester screening for trisomy 21. Ultrasound Obstet Gynecol 2011;38:309-13.
48. Drion I, Cobbaert C, Groenier KH, Weykamp C, Bilo HJ, Wetzels JF, Kleefstra N. Clinical evaluation of analytical variations in serum creatinine measurements: why laboratories should abandon Jaffe techniques. BMC nephrology 2012;13:133.

49. Jin Y, Bies R, Gastonguay MR, Stockbridge N, Gobburu J, Madabushi R. Misclassification and discordance of measured blood pressure from patient's true blood pressure in current clinical practice: a clinical trial simulation case study. J Pharmacokinet Pharmacodyn 2012;39:283-94.

50. Sarno MJ, Davis CS. Robustness of ProsVue linear slope for prognostic identification of patients at reduced risk for prostate cancer recurrence: simulation studies on effects of analytical imprecision and sampling time variation. Clin Biochem 2012;45:1479-84.

51. 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;233:83-90.

52. Thomas F, Signal M, Harris DL, Weston PJ, Harding JE, Shaw GM, et al. Continuous glucose monitoring in newborn infants: how do errors in calibration measurements affect detected hypoglycemia? J Diabetes Sci Technol 2014;8:543-50.

53. De Block CE, Gios J, Verheyen N, Manuel-y-Keenoy B, Rogiers P, Jorens PG, et al. Randomized evaluation of glycemic control in the medical intensive care unit using real-time continuous glucose monitoring (REGIMEN Trial). Diabetes Technol Ther 2015;17:889-98.

54. Krinsley JS, Bruns DE, Boyd JC. The impact of measurement frequency on the domains of glycemic control in the critically ill-a monte carlo simulation. J Diabetes Sci Technol 2015;9:237-45.

55. Bietenbeck A. Combining medical measurements from diverse sources: experiences from clinical chemistry. Stud Health Technol Inform 2016;228:58-62.

56. Shinotsuka CR, Brasseur A, Fagnoul D, So T, Vincent J-L, Preiser J-C. Manual versus Automated moNitoring Accuracy of GlucosE II (MANAGE II). Crit Care 2016;20:380.

57. Sutheran HL, Reynolds T. Technical and clinical accuracy of three blood glucose meters: clinical impact assessment using error grid analysis and insulin sliding scales. J Clin Pathol 2016;69:899-905.

58. Baumstark A, Jendrike N, Pleus S, Haug C, Freckmann G. Evaluation of accuracy of six blood glucose monitoring systems and modeling of possibly related insulin dosing errors. Diabetes Technol Ther 2017;19:580-8.

59. Bhatt IS, Guthrie On. Analysis of audiometric notch as a noise-induced hearing loss phenotype in US youth: data from the National Health And Nutrition Examination Survey, 2005–2010. Int J Audiol 2017;56:392-9.

60. Bochicchio GV, Nasraway S, Moore L, Furnary A, Nohra E, Bochicchio K. Results of a multicenter prospective pivotal trial of the first inline continuous glucose monitor in critically ill patients. J Trauma Acute Care Surg 2017;82:1049-54.

61. Chai JH, Ma S, Heng D, Yoong J, Lim WY, Toh SA, Loh TP. Impact of analytical and biological variations on classification of diabetes using fasting plasma glucose, oral glucose tolerance test and HbA1c. Sci Rep 2017;7:7.

62. 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;63:585-92.

63. Chung RK, Wood AM, Sweeting MJ. Biases incurred from nonrandom repeat testing of haemoglobin levels in blood donors: selective testing and its implications. Biom J 2019;61:454-66.

64. Saugel B, Grothe O, Nicklas JY. Error grid analysis for arterial pressure method comparison studies. Anesth Analg 2018;126:1177-85.

65. Rodrigues Filho BA, Farias RF, dos Anjos W. Evaluating the impact of measurement uncertainty in blood pressure measurement on hypertension diagnosis. Blood Press Monit 2018;23:141-7.
66. Piona C, Dovc K, Mutlu GY, Grad K, Gregorc P, Battelino T, Bratina N. Non-adjunctive flash glucose monitoring system use during summer-camp in children with type 1 diabetes: the free-summer study. Pediatr Diabetes 2018;19:1285-93.
67. Hansen EA, Klee P, Dirlewanger M, Bouthors T, Elowe-Gruau E, Stoppa-Vaucher S, et al. Accuracy, satisfaction and usability of a flash glucose monitoring system among children and adolescents with type 1 diabetes attending a summer camp. Pediatr Diabetes 2018;19:1276-84.
68. Freckmann G, Link M, Pleus S, Westhoff A, Kamecke U, Haug C. Measurement performance of two continuous tissue glucose monitoring systems intended for replacement of blood glucose monitoring. Diabetes Technol Ther 2018;20:541-9.
69. Hughes J, Welsh JB, Bhavaraju NC, Vanslyke SJ, Balo AK. Stability, accuracy, and risk assessment of a novel subcutaneous glucose sensor. Diabetes Technol Ther 2017;19:S21-4.
70. Breton MD, Hinzmann R, Campos-Nanez E, Riddle S, Schoemaker M, Schmelzeisen-Redeker G. Analysis of the accuracy and performance of a continuous glucose monitoring sensor prototype: an in-silico study using the UVA/PADOVA type 1 diabetes simulator. J Diabetes Sci Technol 2017;11:545-52.
71. Aberer F, Hajnsek M, Rumpler M, Zenz S, Baumann PM, Elsayed H, et al. Evaluation of subcutaneous glucose monitoring systems under routine environmental conditions in patients with type 1 diabetes. Diabetes, Obesity and Metabolism 2017;19:1051-5.
72. Kovatchev BP, Patek SD, Ortiz EA, Breton MD. Assessing sensor accuracy for non-adjunct use of continuous glucose monitoring. Diabetes Technol Ther 2015;17:177-86.
73. Schnell O, Erbach M. Impact of a reduced error range of SMBG in insulin-treated patients in Germany. J Diabetes Sci Technol 2014;8:479-82.
74. Kovatchev BP, Wakeman CA, Breton MD, Kost GJ, Louie RF, Tran NK, Klonoff DC. Computing the surveillance error grid analysis: procedure and examples. J Diabetes Sci Technol 2014;8:673-84.
75. Klonoff DC, Lias C, Vigersky R, Clarke W, Parkes JL, Sacks DB, et al. The surveillance error grid. J Diabetes Sci Technol 2014;8:658-72.
76. Schnell O, Erbach M, Wintergerst E. Higher accuracy of self-monitoring of blood glucose in insulin-treated patients in Germany: clinical and economical aspects. J Diabetes Sci Technol 2013;7:904-12.
77. Budiman ES, Samant N, Resch A. Clinical implications and economic impact of accuracy differences among commercially available blood glucose monitoring systems. J Diabetes Sci Technol 2013;7:365-80.
78. McGarraugh GV, Clarke WL, Kovatchev BP. Comparison of the clinical information provided by the FreeStyle Navigator continuous interstitial glucose monitor versus traditional blood glucose readings. Diabetes Technol Ther 2010;12:365-71.
79. Petersen PH, Soletormos G, Pedersen MF, Lund F. Interpretation of increments in serial tumour biomarker concentrations depends on the distance of the baseline concentration from the cut-off. Clin Chem Lab Med 2011;49:303-10.
80. Hu Y, Ahmed HU, Carter T, Arumainayagam N, Lecornet E, Barzell W, et al. A biopsy simulation study to assess the accuracy of several transrectal ultrasonography (TRUS)-biopsy strategies compared with template prostate mapping biopsies in patients who have undergone radical prostatectomy. BJU Int 2012;110:812-20.

81. Lecornet E, Ahmed HU, Hu Y, Moore CM, Nevoux P, Barratt D, et al. The accuracy of different biopsy strategies for the detection of clinically important prostate cancer: a computer simulation. J Urol 2012;188:974-80.
82. McCloskey LJ, Bordash FR, Ubben KJ, Landmark JD, Stickle DF. Decreasing the cutoff for Elevated Blood Lead (EBL) can decrease the screening sensitivity for EBL. Am J Clin Pathol 2013;139:360-7.
83. Lund F, Petersen PH, Pedersen MF, Abu Hassan SO, Soletormos G. Criteria to interpret cancer biomarker increments crossing the recommended cut-off compared in a simulation model focusing on false positive signals and tumour detection time. Clin Chim Acta 2014;431:192-7.
84. Abu Hassan SO, Petersen PH, Lund F, Nielsen DL, Tuxen MK, Sölétormos G. Monitoring performance of progression assessment criteria for cancer antigen 125 among patients with ovarian cancer compared by computer simulation. Biomark Med 2015;9:911-22.
85. Lin J, Fernandez H, Shashaty MG, Negoianu D, Testani JM, Berns JS, et al. False-positive rate of AKI using consensus creatinine-based criteria. Clin J Am Soc Nephrol 2015;10:1723-31.
86. Merlin T, Lehman S, Hiller JE, Ryan P. The “linked evidence approach” to assess medical tests: a critical analysis. Int J Technol Assess Health Care 2013;29:343-50.
87. Schaafsma JD, van der Graaf Y, Rinkel GJ, Buskens E. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. J Clin Epidemiol 2009;62:1248-52.
88. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making 2009;29:E22-E9.
89. Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. Med Decis Making 2011;31:675-92.
90. Klee GG. Establishment of outcome-related analytic performance goals. Clin Chem 2010;56:714-22.
91. Panteghini M, Ceriotti F, Jones G, Oosterhuis W, Plebani M, Sandberg S. Strategies to define performance specifications in laboratory medicine: 3 years on from the Milan Strategic Conference. Clin Chem Lab Med 2017;55:1849-56.

Tables

Table 1. Review inclusion criteria

Population	Any human population with any indication
Intervention	In-vitro test (excluding imaging) or any kind of medical device used for the purpose of screening, diagnosis, prognosis, monitoring or predicting treatment response
Comparator	Any
Outcomes	(a) Clinical accuracy e.g. - Diagnostic sensitivity and/or specificity - Positive/negative predictive values - ROC curve/ AUC analysis - Relative risks - Likelihood ratios (b) Clinical utility - Impact on treatment management decisions - Impact on patient health outcomes (c) Costs (d) Cost-effectiveness
Method	Analysis includes indirect methods (i.e. excluding purely empirical analyses) to incorporate or assess the impact of one or more components of measurement uncertainty (below) on one or more outcomes (above): - Bias (e.g. calibration or method bias) - Imprecision (e.g. repeatability, within-laboratory or between-laboratory imprecision) - Pre-analytical or analytical effects - Summary metrics (e.g. total error [TE] or uncertainty of measurement [U_M])
Study type	Full paper relating to an original study
Language	Full text in English
Year of publication	Database search: January 2008 – March 2019 Citation tracking: any data

ROC = Receiver operator characteristic; AUC = Area under the curve

Table 2. Study characteristics

	N	%
Year of publication		
Pre-2008 (identified via citation tracking alone)	25	30%
2008 – 2009	3	4%
2010 – 2011	7	9%
2012 – 2013	9	11%
2014 – 2015	18	22%
2016 – 2017	13	16%
2018-2019	7	9%
Clinical area[a]		
Diabetes & glycemic control	43	52%
Cardiovascular diseases	17	21%
Cancer	10	12%
Metabolic & endocrine disorders	8	10%
Kidney disorders	3	4%
Prenatal screening	3	4%
Noise induced hearing loss	2	2%
Role of test[a]		
Monitoring	44	54%
Diagnosis	24	29%
Screening	11	13%
Prognosis	7	9%

[a]Several studies included a test or tests used in multiple clinical areas or roles (hence total percentages under these categories sum to >100%).

Table 3. Components of measurement uncertainty included and test outcomes assessed

	N	%
Component(s) of measurement uncertainty included[a]		
Imprecision:		
Analytical	31	38%
Pre-analytical / combined pre-analytical and analytical	8	10%
Non-specific	11	13%
Total	**50**	**61%**
Bias:		
Analytical	18	22%
Calibration bias	9	11%
Non-specific	9	11%
Pre-analytical / combined pre-analytical and analytical	2	2%
Between-method bias	1	1%
Total	**39**	**48%**
Total error:		
Method-comparison study	18	22%
EQA study	2	2%
Other	6	7%
Total	**26**	**32%**
Biological variation included?		
Yes - included as a separate element	13	16%
Yes - combined with imprecision	5	6%
Total	**18**	**22%**
Primary test outcome assessed[a]		
Clinical accuracy	45	55%
Clinical utility:		
Impact on treatment management	23	28%
Impact on health outcomes	13	16%
Costs	7	9%
Cost-effectiveness	2	2%

[a] Several studies included multiple components of measurement uncertainty or assessed multiple test outcomes (hence total percentages under these categories sum to >100%).

Figure captions

Figure 1. PRISMA flow diagram of included studies

Figure 2. Summary box outlining the three-step analytical framework, primary methods identified for each step in the framework, and key questions for consideration in future analyses

Figure 3. Example contour plot based on simulations using the error model approach (adding increasing magnitudes of bias and imprecision onto assumed "true" measurand values). The contour lines indicate what level of clinical accuracy is achieved across the range of bias and imprecision inputs explored: varying sensitivity levels as a function of bias and imprecision are represented by the solid contour lines, whilst varying specificity levels are represented by the dashed contour lines. The grey region represents an "acceptability region" for bias and imprecision, which maintains sensitivity ≥ 90% and specificity ≥ 80%.