Developing a framework for assessing the impact of test measurement uncertainty on clinical and health-economic outcomes

Alison Florence Christine Smith

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

Faculty of Medicine and Health

March 2020

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Details of candidate contributions to manuscript publications relating to this thesis:

i. **Smith AF, Messenger MP, Hall PS, Hulme CT. The Role of Measurement Uncertainty in Health Technology Assessments (HTAs) of In Vitro Tests. *PharmacoEconomics 2018;36:823-35***

The candidate (Smith AF) was responsible for planning and designing the study, completing the study analysis and developing and finalising the manuscript (1). The co-authors provided supervisory feedback and oversight throughout the study. In addition, Hulme CT undertook secondary screening of 10% of abstracts identified from the study database search and also independently checked 10% of the data extraction forms for studies included in the review. Research presented in this manuscript is included in **Chapter 2** of this thesis.

ii. **Smith AF, Shinkins B, Hall PS, Hulme CT, Messenger MP. 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 2019:clinchem.2018.300954***

The candidate was responsible for planning and designing the study, completing the study analysis and developing and finalising the manuscript (2). The co-authors provided supervisory feedback and oversight throughout the study. In addition, each of the co-authors conducted data extraction checking on a proportion of studies included in the review (43% were checked by Shinkins B, 20% by Hulme CT, 20% by Messenger MP, and 18% by Hall PS). Research presented in this manuscript is included in **Chapter 3** of this thesis.

Details of candidate contributions to published conference proceedings relating to this thesis:

iii. **<u>Smith AF</u>, Shinkins B, Hulme CT, Hall PS, Michael MP. Beyond the laboratory: A review of indirect methods to assess the impact of test measurement uncertainty on downstream clinical and cost outcomes. *Clinica Chimica Acta. 2019 Jun 1;493:S353.***

This conference proceeding relates to the same study as outlined in publication (ii) above. Research presented in this poster (3) is included in **Chapter 3** of this thesis.

iv. <u>Smith AF</u>, Shinkins B, Hulme CT, Hall PS, Messenger MP. Methods to assess the impact of test measurement uncertainty on downstream clinical outcomes: A case study of faecal calprotectin (FC) for the diagnosis of Inflammatory Bowel Disease (IBD). *Clinica Chimica Acta. 2019 Jun 1;493:S353-4.*

The candidate was responsible for planning and designing the study, completing the study analysis and developing and finalising the conference proceeding (4). The co-authors provided supervisory feedback and oversight throughout the study. Research presented in this poster is included in **Chapter 5** of this thesis.

v. <u>Smith AF</u>. Defining test value based on outcomes, and bridging the HTA-laboratory divide. In: St John A, *et al*. 57th Annual Scientific Conference. *Clin Biochem Rev Supplement 40 (4) 2019.*

The candidate was solely responsible for this conference symposium presentation (5). This talk consisted of a high-level discussion of various aspects of this thesis including discussion of the test evaluation pathway presented in **Chapter 1**, review findings presented in **Chapter 2** and **Chapter 3**, simulation methods presented in **Chapter 5** and **Chapter 6**, and discussion points presented in **Chapter 8**.

Acknowledgements

To each of my supervisors, I would like to express my extreme gratitude for the kindness, support and expertise you have provided to me throughout this PhD, and in life generally. Dr Peter Hall first introduced me to the world of test evaluation, opened new doors for me and encouraged me to pursue a PhD. Dr Mike Messenger helped navigate me through the new and scary world of measurement uncertainty, made it seem possible, and most importantly made it fun. Dr Bethany Shinkins has provided never-ending support, encouragement, expertise and friendship – she is the best supervisor a student could wish for. Finally, Professor Claire Hulme has provided unwavering support, enthusiasm, and optimism throughout my career, and is the most dedicated mentor I have ever had the privilege to know. Pete, Mike, Beth and Claire – thank you all.

To all of my scientific advisory board, thanks for your extra help and support. Particular thanks to Dr James Turvill, Professor John Deeks, Professor Chris Hyde, Hayden Holmes, Professor Chris Bojke and Dr Joy Allen. Thanks also to Professor Chris McCabe for your kindness and mentorship, and to Dr Dan Turnock for all things laboratory related. Special thanks to members of the EFLM Test Evaluation working group who have welcomed me into their world, and provided insightful feedback on my work – in particular thanks to Dr Sally Lord, Professor Rita Horvath, and Dr Tze Ping Loh. Thanks also to Andrew St John for your kindness.

I would also like to thank the NIHR for funding this research.

On a personal note, thanks to my family for your love and support. Thanks to my sister, Jenny, for having two wonderful children to distract me throughout the PhD, and for teaching me what real strength is. Thanks to my Dad and Helen, for always giving me an escape and supporting me. Most of all thanks to my Mum for being my unwavering rock and ultimate inspiration in life.

Finalmente, gracias a mi amor Armando, por darme la fuerza y el amor para consequirlo.

Funding

Alison Smith was funded by an National Institute of Health Research (NIHR) Doctoral Research Fellowship (DRF-2016-09-084) for this project. This thesis presents independent research funded by the NIHR. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Climate care

Several national and international courses and conferences were attended as part of this PhD. Using an online carbon calculator, I have estimated that 10.65 tonnes of carbon emissions were produced as a result of flights taken to attend these activities (https://climatecare.org/calculator/). To account for additional emissions resulting from train and car travel (not included in the carbon calculator) and other potentially hidden emissions, I have rounded this figure up to 12 tonnes. On the recommendation of the University of Leeds's sustainability team, I have purchased 12 tonnes worth of Gold Standard offsets ('Climate+ Portfolio') (https://www.goldstandard.org/), in an attempt to offset the carbon footprint of this PhD.

Abstract

Background: Many factors can introduce uncertainty into medical laboratory test measurements, and this uncertainty can affect downstream clinical and health-economic outcomes. Currently, however, the impact of measurement uncertainty on outcomes is rarely considered, either within laboratory medicine or Health Technology Assessment (HTA) practices.

Aim: To develop a framework for assessing the impact of test measurement uncertainty on clinical and health-economic outcomes.

Methods: Five hypotheses were addressed in this thesis. Hypothesis A – that measurement uncertainty has not been routinely addressed within HTAs – was assessed via a systematic review of HTAs. Hypothesis B – that methods for assessing the impact of measurement uncertainty on outcomes have been used in the broader literature – was assessed via a methodology literature review. The remaining hypotheses – that methods from the literature could be used/adapted to: [C] evaluate the impact of measurement uncertainty on clinical performance, utility and cost-effectiveness; [D] derive outcome-based analytical performance specifications (APS); and [E] accommodate real world evidence on measurement performance – were assessed via a case study analysis, exploring the role of faecal calprotectin for the diagnosis of Inflammatory Bowel Disease (IBD).

Results: The HTA review confirmed that, to date, HTAs have rarely assessed the impact of measurement uncertainty on outcomes. The methodology review, meanwhile, identified various relevant methods from the broader literature (mostly from the laboratory medicine field). Of those, iterative simulation and decision modelling were selected for further exploration based on their ability to be integrated into existing HTA methodology. The subsequent case study demonstrated a framework of analysis building on these methods. Using both hypothetical and real world evidence simulations, the robustness of faecal calprotectin clinical pathway outcomes to increasing measurement uncertainty was assessed, and regions of acceptable measurement uncertainty (i.e. outcome-based APS) were identified.

Conclusions: The presented framework can help to improve HTA-decision making and inform outcome-based laboratory practices.

Table of Contents

List of Figures

List of Tables

Abbreviations

ACR	Albumin-creatinine Ratio
AIC	Akaike information criterion
AHRQ	Agency for Healthcare Research and Quality
AKI	Acute kidney injury
ALTM	All-laboratory trimmed mean
ANOVA	Analysis of variance
APS	Analytical performance specifications
AUHE	Academic Unit of Health Economics
BIC	Bayesian information criterion
BRISQ	Biospecimen Reporting for Improved Study Quality
CADTH	Canadian Agency for Drugs and Technologies in Health
CCG	Clinical Commissioning Group
CD	Crohn's disease
CE	Conformité Européene
CI	Confidence interval
CLIA	chemiluminescence immunoassay
CLSI	Clinical and Laboratory Standards Institute
CRD	Centre for Reviews and Dissemination
CRM	Certified reference material
CRP	C-reactive protein
CV	Coefficient of variation
DAP	Diagnostic Assessment Programme
EAG	External Assessment Group
EEA	European Economic Area
EFLM	European Federation of Clinical Chemistry and Laboratory Medicine
ELISA	Enzyme linked immunosorbent assay
EQA	External quality assessment
ERG	Evidence Review Group
ESR	Erythrocyte sedimentation rate
EU	European Union
FC	Faecal calprotectin

FC1	Initial FC test
FC2	Repeat FC test
FDA	U.S Food and Drug Administration
FEIA	fluorimetric enzyme-lined immunoassay
GBP	Great British pound
GP	General practitioner
GUM	Guide to the expression of measurement uncertainty
HTA	Health technology assessment
IBD	Inflammatory bowel disease
IBS	Irritable bowel syndrome
ICER	Incremental cost-effectiveness ratio
INAHTA	International Network of Agencies for HTA
INMB	Incremental net monetary benefit
IPD	Individual patient-level data
IQC	Internal quality control
IQR	Inter quartile range
IQWiG	Institute for Quality and Efficiency in Health Care
ISO	International Organisation for Standardization
IVDD	In-Vitro Diagnostic Medical Devices Directive
IVDR	In-Vitro Diagnostic Medical Devices Regulations
LGC	Laboratory of the Government Chemist
LOB	Limit of Blank
LOD	Limit of Detection
LOQ_{lower}	Lower Limit of Quantification
LOQ_{upper}	Upper Limit of Quantification
MLE	Maximum likelihood estimation
MSAC	Medical Services Advisory Committee
NEQAS	National EQA Service (UK)
NHS	National Health Service
NICE	National Institute of Health and Care Excellence
NIHR	National Institute for Health Research
NMB	Net monetary benefit
NPV	Negative predictive value
ODX	Oncotype DX

OLS	Ordinary least squares
PETIA	particle-enhanced turbidimetric immunoassay
POCT	Point-of-care test
PPV	Positive predictive value
PSA	Probabilistic sensitivity analysis
PSS	Personal social services
QALY	Quality adjusted life year
RCT	Randomised controlled trial
RIPOSTE	Reducing IrreProducibility in laboratory STudiEs
RWE	Real world evidence
SD	Standard Deviation
STROBE-ME	Strengthening the reporting of Observational studies in Epidemiology-Molecular Epidemiology
TASK	Technology Assessment at The Hospital for Sick Children
TE	Total error
UC	Ulcerative colitis
UK	United Kingdom
UM	Uncertainty of Measurement
USA	United States of America
WHO	World Health Organisation
YFCCP	York FC Care Pathway
YHAHSN	Yorkshire and Humberside Academic Health Science Network
YHEC	York Health Economics Consortium
µg/g	Micrograms per gram

Chapter 1

Measurement uncertainty and the test evaluation pathway

1.1 Chapter outline

This chapter provides an introduction to the key topics addressed in this thesis. The central subject of test *measurement uncertainty* – defined as systematic and/or random deviation in test measurements – is first outlined in section 1.2. Section 1.3 subsequently discusses the role of measurement uncertainty within the broader context of the *test evaluation pathway* – defined as the trajectory of research required to take a new test from the technology discovery phase, to the test adoption phase. Section 1.4 then presents the rationale for the thesis – in particular highlighting the potential utility that a framework for assessing the impact of test measurement uncertainty on outcomes would provide both within the medical laboratory setting, and the test evaluation (e.g. HTA) setting. The final section of this chapter outlines the thesis scope, aim, hypotheses and structure.

1.2 Measurement uncertainty

All measurements are subject to uncertainty – whether it be determining the distance between two objects, the level of CO_2 in the atmosphere, or the pressure exerted within a mechanical system. Medical laboratory tests (i.e. in-vitro tests conducted on patient samples taken from the human body) are no exception. The time of day a test sample is taken, the mode of sample transportation and the time between sample collection and analysis, are just a few examples of a multitude of factors which may introduce systematic and/or random errors into test measurements – known collectively, as *measurement uncertainty*.

This section provides an introduction to measurement uncertainty. An initial note regarding terminology used within this thesis is provided in section 1.2.1 below. This is followed by details of the central components of measurement uncertainty (*precision* and *trueness*), and aggregate measures of measurement uncertainty (*total error [TE]* and *uncertainty of measurement [U_M]*), in sections 1.2.2 and 1.2.3 respectively. Regulatory and accreditation requirements for the assessment of measurement uncertainty are then outlined in section 1.2.4; and the need for

performance specifications for measurement uncertainty – in particular with respect to *outcome-based* approaches – is discussed in section 1.2.5. Section 1.3 then goes on to discuss measurement uncertainty within the broader context of the test evaluation pathway.

1.2.1 A note on nomenclature

Nomenclature in the field of *medical metrology* (the science of measurement, applied to medicine) is notoriously varied. For the purpose of this study, *measurement uncertainty* is used to refer to systematic and/or random deviation in test measurements; whilst *measurement performance* is used to refer to the overall technical performance of a test, including additional performance parameters. This definition of measurement performance is broadly equivalent to that of *analytical validity* commonly used in the clinical sciences literature. The term measurement performance is adopted here, to emphasise the fact that measurement may be influenced by a range of factors, not limited to the analytical phase (i.e. including factors occurring before and after the point of sample analysis; as described in section 1.2.2.3).

A glossary of key terms relating to the content of this introduction (i.e. measurement uncertainty, measurement performance and the test evaluation pathway) is provided in Appendix A. Where possible, definitions have been taken from the Clinical and Laboratory Standards Institute (CLSI) Harmonized Terminology Database, which lists internationally accepted terminology for key medical metrology concepts (6).

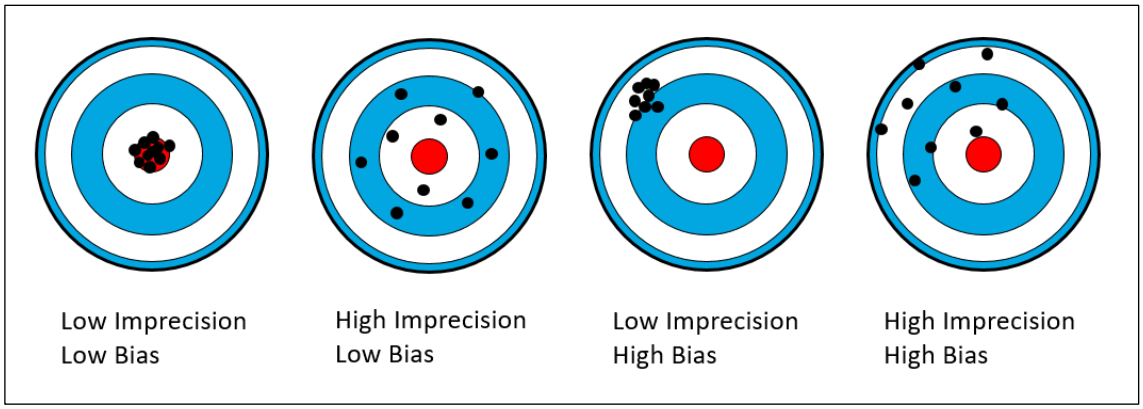
1.2.2 Central components of measurement uncertainty

The central components of measurement uncertainty are *precision*, defined as the closeness of agreement between repeated test results, and *trueness*, defined as the closeness of agreement between the mean of repeated 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.

Figure 1-1 illustrates these concepts using an example of markers on a dart board. A player who exhibits high precision and trueness (i.e. low imprecision and bias) will hit the bullseye target every time, and will thus produce results closely

clustered around this point – as illustrated on the far left dart board. Reducing precision (i.e. increasing imprecision) leads to more widely scattered, but evenly spread, results; whilst reducing trueness (i.e. increasing bias) maintains the close cluster of results, but results in a shift in the central point around which the points are clustered. Introducing imprecision and bias together, leads to an increased spread and a shift in the central positioning of the results – as illustrated in the far right dart board.

Figure 1-1. Bullseye illustration of imprecision and bias



1.2.2.1 Imprecision

Imprecision is assessed by observing the level of dispersion in repeated test measurements conducted on replicate test samples (e.g. identical samples generated by splitting primary test samples into multiple smaller samples or *aliquots*). The level of imprecision captured in a repeated test experiment depends on the conditions under which the analysis is conducted – in particular with respect to five key factors known to affect measurement precision at the point of sample analysis. These are: *time* (i.e. whether the time interval between successive measurements is short or long), *operator* (i.e. whether the same or different operators carry out the successive measurements), *calibration* (i.e. whether the same equipment is or is not recalibrated between successive groups of measurements)[1], *environment* (e.g. whether the temperature and humidity

[1] Calibration refers to the process of testing and adjusting a test instrument or system, to establish a correlation between the measurand (i.e. the substance intended to be measured by a given test) and measurement response. The calibration process is generally based on the analysis of calibration materials of known concentration. This produces a *calibration curve* which expresses the relationship between the measurand quantity and the observed test result, often in the form of a straight line

alters between repeated testing) and *equipment* (i.e. whether the same or different equipment is used in the measurements) (6, 7).

Figure 1-2 illustrates how, as the scope of a repeated test experiment is widened to capture variation in these parameters, the level of imprecision captured similarly expands. For example, a measure of *repeatability* is provided if repeated testing is conducted over a short period of time, in the same location, using the same measurement procedure, the same observer and the same measuring instrument (6). If, however, repeated testing is conducted within the same location but one or more of the factors listed above is allowed to alter (e.g. by analysing samples over a series of days at staggered time intervals), then a measure of *intermediate precision* is provided. Finally, if repeated testing is conducted across multiple laboratories, such that *all* of the listed factors would be expected to vary, then the resulting measure of imprecision will reflect between-laboratory precision – also known as *reproducibility*.

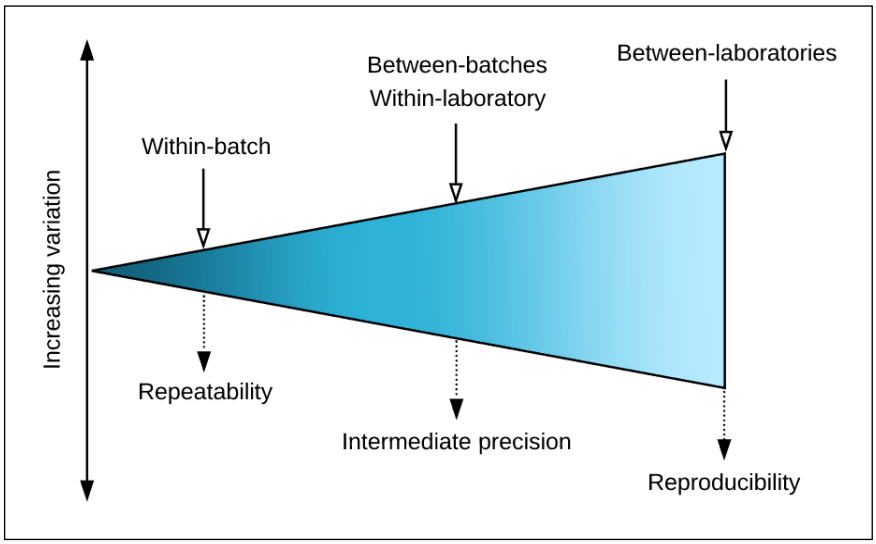


Figure 1-2. Illustration of levels of precision

Since imprecision relates to random variation, it is expressed either as a measure of standard deviation (SD) or coefficient of variation (CV) (i.e. the ratio of the SD to the mean)[2] (8). Although often expressed in terms of the *average* SD or CV,

(e.g. linear regression model) rather than a curve. An example calibration curve is reported in Appendix B.3 (Figure B-3).

[2] For SD = σ and mean = μ, $CV = \frac{\sigma}{\mu} \times 100$. Note: multiplication of the coefficient by 100 is an optional step to express CV as a percentage, rather than a decimal.

imprecision may vary significantly depending on the concentration of the *measurand* (i.e. the substance intended to be measured by a given test). As such, examination of imprecision at various concentrations of the measurand is often required. The results of such an analysis can be illustrated using a *precision profile* plot, in which imprecision is presented against the measurand concentration. Figure 1-3 provides an example of a characteristic U-shaped precision profile, which reflects the fact that precision tends to poorer at lower and higher ends of the measurand concentration range (9).

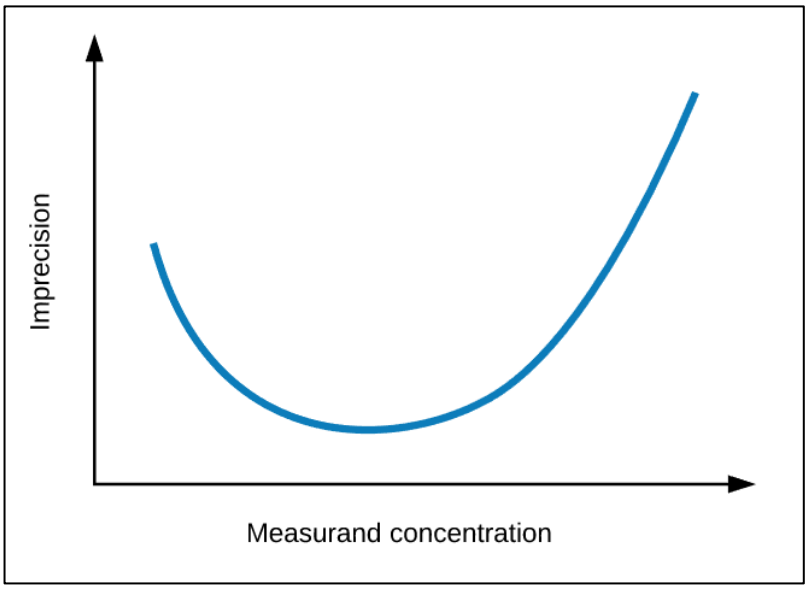


Figure 1-3. Illustration of a precision profile

1.2.2.2 Bias

The assessment of bias requires some form of comparison between measurements from the test of interest (termed the *index test*) versus measurements representative of the truth. Since in reality the 'true' measurand value is unknown, this must be estimated using a specified *reference measurement test*. An ideal reference test in this context is a *reference measurement procedure*, defined as "a thoroughly investigated measurement procedure shown to yield values having an uncertainty of measurement commensurate with its intended use, especially in assessing the trueness of other measurement procedures" (6, 10). If direct use of a reference measurement procedure is not possible, then *certified reference materials (CRMs)* are also useful in this context. CRMs are materials that have been characterised via an unbroken chain of measurement processes, each with a defined measurement

uncertainty, linking back to a reference measurement procedure[3] (6). As such, the value of a CRM (+/- uncertainty) may be considered to be known, and therefore useful in the context of evaluating bias. If neither a reference measurement procedure nor CRM are available, then lower-order reference measurements may also be used – for example, bias may be assessed against an established gold standard measurement procedure.

A routine study undertaken in the assessment of bias is the *method-comparison study,* in which test samples are split and independently analysed using the index test and reference measurement test[4]. Paired measurements from method-comparison studies may then be examined to assess the level of agreement between the two test methods. Common approaches to this end include the Bland-Altman plot (also called a difference plot; see Appendix B.1 for an example); regression analysis (e.g. ordinary least squares [OLS]); and the evaluation of inter-rater reliability metrics (e.g. Cohen's kappa statistic) (9, 11, 12). Bland-Altman and regression analysis techniques are particularly useful in this context, since these enable key details of the method bias to be extracted. This includes: (a) whether the bias is *fixed* (i.e. remains constant over the range of measurand concentrations evaluated) or *proportional* (i.e. increases or decreases in line with the measurand concentration); and (b) whether variability around the bias increases or decreases in line with the measurand concentration (i.e. heteroscedasticity).

1.2.2.3 A note on pre-analytical and analytical factors

Multiple factors may introduce bias and imprecision into test measurements. These include (i) *pre-analytical factors*, occurring prior to the point of sample analysis (e.g. how the test sample is collected, transported and stored in the laboratory); and (ii) *analytical factors*, occurring at the point of sample analysis (e.g. the laboratory environment, testing equipment, and the existence of any interfering substances in the test sample). In addition, *within-subject biological variation* – defined as the fluctuation of measurand concentrations in the body

[3] The sequence of measurement processes linking a CRM to the reference measurement procedures is known as the *metrological traceability chain*.

[4] Alternatively, if using CRMs, then the index test measurements (using the CRMs) may simply be compared to the designated CRM values.

over time – may further contribute to imprecision (6, 13). Consideration of each of these factors is central to any evaluation of measurement uncertainty. A full discussion of pre-analytical and analytical factors, including an example 'feather map' of key factors occurring along the total testing process, is provided in Appendix B.2.

1.2.3 Aggregate measures of measurement uncertainty

Elements of bias and imprecision, as described in section 1.2.2, may be combined to provide an aggregate estimate of total measurement uncertainty. To this end, two main approaches 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 (14). Briefly, TE is calculated as the linear sum of bias and imprecision, as illustrated in Figure 1-4. Assuming that random error can be approximated by a normal distribution, the estimate of imprecision (expressed as an SD) is multiplied by a chosen 'z factor' to cover a required level of confidence. In order to cover a 95% confidence interval (CI), for example, a z value of 1.96 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.

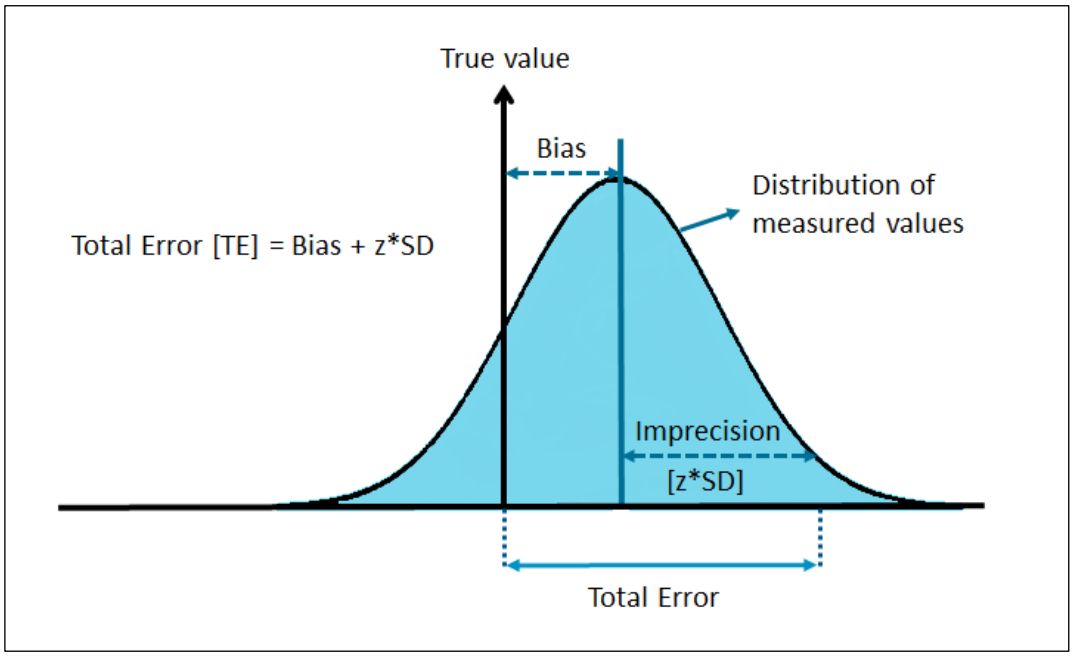


Figure 1-4. Illustration of TE calculation

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 (15). 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" (15). In the original GUM document, a "bottom-up" procedure was proposed for quantifying U_M, via the following four-step procedure[5]:

1. Identify all elements associated with measurement uncertainty along the testing pathway.
2. Determine the *standard uncertainty* (expressed as an SD) around each element from point 1.
3. Determine the *combined standard uncertainty* of all elements from point 1, by combining the associated standard uncertainties from point 2 (using, for example, the sum of squares rule[6] or computer simulation).
4. Determine the *expanded measurement uncertainty* by assigning a coverage factor, *k*, to the combined standard uncertainty from point 3. For example, *k* = 2 produces an expanded uncertainty of measurement region which may be attributable to the measurand with ~95% confidence.

Based on the procedure outlined above, the resulting measure of U_M (expressed as an SD) represents a region around the measured test value, which is expected to include the underlying true measurand value to a specified degree of probability (i.e. depending on the value of *k* selected).

---

[5] According to the "bottom up" procedure, all individual components of uncertainty along a testing pathway are required to be identified and separately measured. More recently, owing to the recognized impracticalities of this approach, an alternative "top down" procedure has been suggested: this method recognises that high level quality assurance and/or method validation data will capture multiple components of measurement uncertainty, and can therefore be used in place of having to individually assess multiple elements [16. ISO. 15189: 2012 Medical laboratories–Requirements for quality and competence. Geneva: International Standardisation Organisation. 2012.].

[6] The sum of squares rule follows Pythagoras' theorem i.e. $a^2 + b^2 = c^2$. Thus, if four factors (a, b, c and d) contributing to measurement uncertainty have been identified, and are associated with standard uncertainties, s_a, s_b, s_c, and s_d, then the combined standard uncertainty is equal to: $\sqrt{{s_a}^2 + {s_b}^2 + {s_c}^2 + {s_d}^2}$.

It should be noted that there has been continued debate in the literature regarding the differences and relative merits of TE vs. U_M (17-20). From a technical point of view, a key difference relates to the handling of bias in each case. In the calculation of U_M, for example, it is assumed that if bias is known then steps should be taken to resolve or remove it (e.g. via recalibration) – as such, this approach incorporates bias only insofar as imprecision introduced from the process of removing bias may be captured (19). TE, on the other hand, assumes that not all components of bias are necessarily resolvable in practice, and thus explicitly incorporates bias into the calculation (20). In addition, from a conceptual point of view, a further key difference between the two approaches relates to the notion of true measurement. On the one hand, the TE calculation assumes that the true measurand value can be known; whilst on the other hand, U_M makes no direct assumptions about the truth, but rather represents an expression of a *lack* of knowledge about the true measurand value (17, 19).

Whilst debate around the use of TE vs. U_M is ongoing, both methods have maintained widespread adoption. Interestingly, in a recent opinion paper from the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Task and Finish Group on Total Error, it was argued that each approach has particular merits in different scenarios (19). TE, for example, may be considered to be of most relevance in scenarios where a reference measurement of known quantity is in use – for example within quality assurance/ proficiency testing schemes utilising a reference measurement procedure or CRMs. U_M, meanwhile, may be considered of greater applicability in scenarios where the evaluation of measurement uncertainty is based on patient samples, wherein the “true” measurand value remains essentially unknown. As such, the EFLM recommend that the two approaches should be considered as complimentary rather than conflicting methods (19).

1.2.4 Test regulation and laboratory accreditation

Within the UK (and internationally), the evaluation and monitoring of measurement uncertainty is a requirement of two key processes: (i) test regulation, which is a legal undertaking required prior to the marketing of new tests; and (ii) laboratory accreditation, which is a voluntary process undertaken by the majority of medical laboratories in the UK, as a means of quality

assurance. This section provides a summary of these two processes, focusing on aspects relating to the evaluation and monitoring of measurement uncertainty.

1.2.4.1 Regulation

From a UK and EU perspective, anyone wishing to market a new test within the European Economic Area (EEA) must, by law, obtain a Conformité Européene (CE) mark to confirm compliance with regulatory standards. Until recently these standards were defined according to the European In-Vitro Diagnostic Directive (IVDD 98/79/EC). In 2017, however, the IVDD was superseded with new regulation in the form of the European In-vitro Diagnostics Medical Devices Regulation (IVDR 2017/746) (21). The IVDR is set to come in to full force in 2022 (21).

In keeping with the IVDD, the IVDR requires manufacturers to provide evidence on analytical validity (i.e. measurement performance), in addition to safety and scientific validity requirements[7]. In particular, tests are required to undergo complete *method validation* – defined as the demonstration via objective evidence (i.e. experimentation) that the test is appropriate[8] for its intended use (22). Assessment of a range of performance characteristics is required for method validation, including the central components of measurement uncertainty (i.e. bias and imprecision, as described in section 1.2.2), as well as additional metrics falling under the wider remit of measurement performance. These include:

- test *selectivity*[9] (the ability of a test to measure the specified measurand as opposed to other interfering substances in the test sample);
- *detection and quantitation limits* (limits describing the smallest and highest concentration of a measurand that can be reliably measured by the test);

[7] *Scientific validity* is defined as the association between a measurand and a clinical condition or disease state.

[8] Note that, benchmarks against which to judge the "appropriateness" of a test's performance are not provided in the IVDR – rather, the onus is on test manufacturers to pre-define their own acceptance criteria. Further discussion of international guidelines related to the definition of acceptance criteria is provided in section 1.2.5.

[9] Sometimes referred to as *analytical specificity*.

- *analytical sensitivity* (the rate of change in the measured test value, in relation to a given increase in the measurand concentration);
- *linearity* (the ability of a test within a given range to provide results that are directly proportion to the measurand concentration); and
- *measurement range* (the range of measurand values over which meaningful test results can be acquired) (6, 21).

Further description of the above performance metrics falling under the remit of measurement performance, is provided in Appendix B.3.

In addition to requirements concerning test measurement performance, the IVDR also includes new requirements for test manufacturers to provide evidence relating to *clinical performance* – defined as the ability of a test to detect patients with a particular clinical condition or in a physiological state (23). As discussed later in section 1.3, the clinical performance of a test is of central importance to the overall value of a test, since this determines the knock-on impact that a test has on patient health outcomes and healthcare costs. The requirement for evidence on clinical performance within the IVDR is, therefore, a welcome addition to the new regulation in terms of securing patient and system-wide benefits. A potential limitation with this aspect, however, concerns the lack of specific guidance provided within the IVDR as to how clinical performance studies should be undertaken (21).

1.2.4.2 Laboratory accreditation

In addition to test manufacturers being required to conform to regulatory standards in order to market tests across the EEA, further requirements are placed on testing laboratories in order to achieve laboratory *accreditation* – defined as independent third-party assurance of the competence, impartiality and performance capability of testing centres (24). Internationally, medical laboratories are required to show compliance to standards set out by the International Organization for Standardization (ISO) in order to achieve accreditation (16). In the UK, the United Kingdom Accreditation Service (UKAS) is the national accreditation body responsible for granting laboratory accreditation (as per ISO standards). Under this scheme, laboratories must undergo an initial assessment to obtain their accreditation certificate, and thereafter receive annual

surveillance visits with a full re-assessment every fourth year (25). Although laboratory accreditation under ISO is voluntary, the majority of UK medical laboratories are accredited (26).

Key requirements for medical laboratories in relation to the assessment of measurement uncertainty are outlined in ISO 15189:2012 (16). Under this standard, laboratories must conduct appropriate *method validation, verification* and *quality assurance* of testing procedures. Validation and verification processes in this context relate to the initial implementation of a new or modified testing procedure: validation is required for tests developed in-house, or for previously validated tests which are being used outside of their intended use; whilst verification is required to confirm the appropriateness of an already validated test (22). Subsequent quality assurance of tests describes the comprehensive set of practices used to monitor testing process and ensure that the testing site's results maintain a required level of performance (6). This includes: (i) internal quality assurance, which describes procedures run in association with the measurement of patients' samples (e.g. based on running CRMs or internal quality control samples of known concentration, to check that patients' results are expected to be valid); and (ii) *external quality assessment (EQA)* (also known as proficiency testing), which describes the evaluation of the laboratory's performance via examination of samples of external origin, to establish between-laboratory and between-instrument comparability (e.g. via regional, national, or international EQA schemes) (6).

As with test regulation, the definition of 'appropriate' performance for a test within the context of accreditation-related exercises (i.e. validation, verification and quality assurance) is similarly expected to be pre-defined by the testing laboratory (rather than being directly stipulated by ISO). Under current ISO standards, therefore, the focus is on ensuring that testing centres take measures to *quantify* and *monitor* measurement uncertainty. How this measurement uncertainty is then judged as appropriate or not, is left largely to the discretion of the test manufacturers and testing centres. Further discussion of international guidelines related to the definition of acceptance criteria is provided in section 1.2.5.

A final note with respect to regulation and accreditation, concerns the withdrawal of the UK from the EU. Although the UK left the EU on 31 January 2020, an

ongoing transition period is in effect until 31 December 2020. During this transition period, regulation and accreditation processes are expected to remain unchanged (i.e. tests marketed in the UK will continue to be required to conform to IVDR regulation; and laboratory accreditation will continue to be determined in line with ISO standards) (27, 28). What may transpire after this transition period, however, is currently unclear. Whilst there appears to be agreement among key stakeholders that the UK should align regulatory and accreditation practices with the rest of the EU (29), only time will tell how UK regulation and accreditation practices will evolve in the post-Brexit era.

1.2.5 Analytical performance specifications (APS)

1.2.5.1 EFLM Milan criteria

As previously highlighted, the results of any validation, verification or quality assurance procedure need to be judged against a specified requirement of analytical performance, in order to determine whether the test may be considered "fit for purpose". These requirements are known as *analytical performance specifications (APS)*. In the context of measurement uncertainty, APS are typically presented as maximum allowable levels of bias, imprecision and/or TE. Whilst APS are a necessary requirement of current regulatory and laboratory accreditation practices, specific levels of performance are not mandated in the IVDR legislation or ISO standards (16, 21). Rather, test manufacturers and laboratories are expected to pre-define appropriate APS for tests within any given validation, verification or quality assurance exercise.

Best practice methods for setting APS are outlined in international guidance documents, most notably in a 2015 EFLM consensus statement on the topic, following the 1st EFLM Strategic Conference held in Milan in 2014 (30). Building on previous guidance in this area (31) the EFLM present three models for setting APS (henceforth referred to as the 'EFLM Milan criteria'), outlined below (30).

- *Model 1. Based on the effect of analytical performance on clinical outcomes*

Under Model 1, minimum requirements for analytical performance (i.e. measurement performance) should be set based on the effect of measurement

uncertainty on clinical outcomes. Two types of studies may be undertaken in this context: under *Model 1a* of the EFLM Milan criteria, the impact of measurement uncertainty on clinical outcomes is established via *direct outcome studies* (i.e. empirical-based analyses, such as randomised controlled trials [RCTs]); whilst under *Model 1b*, the impact of measurement uncertainty on outcomes is established via *indirect outcome studies* (i.e. non-empirical-based analyses, such as simulation or decision analytic modelling). Note that, whilst the emphasis within the EFLM Milan criteria is on clinical outcomes, other end-stage outcomes – such as cost-effectiveness (see section 1.3) – may also be considered of relevance (32). APS defined on the basis of direct or indirect outcome studies are known as *outcome-based APS*.

- *Model 2. Based on components of biological variation of the measurand*

Under Model 2, APS are derived from assessment of the biological variation of the measurand. The underlying premise of this approach is that the 'analytical noise' of a test (i.e. bias and imprecision) should not add significantly to the noise created by biological variation (30, 33). The prevailing method applied in this context follows a series of rules popularised by Fraser and colleagues in the 1990's (34, 35). Via this approach, bias should be maintained under an *eighth*, *a quarter* or *three-eighths* of the total within- and between-subject biological variation, in order to maintain *optimum*, *desirable* or *minimum* levels of bias, respectively[10]. Analytical imprecision, meanwhile, should be maintained under *a quarter*, *a half* or *three-quarters* the level of within-subject biological variation, in order to maintain *optimum*, *desirable* or *minimum* levels of analytical performance, respectively[11] (34, 35).

- *Model 3. Based on state-of-the-art*

---

[10] i.e. $\boldsymbol{Bias} < \boldsymbol{x} \times \sqrt{\boldsymbol{CV_I + CV_G}}$; where CV_I is within-individual biological variation and CV_G is between-individual (i.e. group) biological variation; $x = 0.125$ provides optimum performance; $x = 0.25$ provides desirable performance; and $x = 0.375$ provides minimum performance.

[11] i.e. $\boldsymbol{CV_A < x \times CV_I}$; where CV_A is analytical variation and CV_I is within-individual biological variation; $x = 0.25$ provides optimum performance; $x = 0.5$ provides desirable performance; and $x = 0.75$ provides minimum performance.

Under Model 3, APS are set in relation to what measurement performance is achieved by state-of-the-art tests (i.e. the highest level of measurement performance technically achievable by field methods), or by other laboratories. For example, if most laboratories within an EQA program can achieve a certain level of performance, then laboratories achieving significantly below this level should be required to change their practice (30).

1.2.5.2 Outcome-based APS

Following publication of the EFLM Milan criteria, further EFLM guidance was released in 2017, in which criteria for assigning measurands to the EFLM Milan Criteria were presented (33). This guidance suggests 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 (i.e. *outcome-based APS*, under Model 1 of the EFLM Milan criteria). Model 2, meanwhile, is presented as the relevant choice for measurands that do not have a central role in a specific disease or condition; whilst Model 3 is reserved as a backstop for measurands that cannot be included in Models 1 or 2 (e.g. while awaiting outcome or biological variability data).

The justification for the above guidance is clear: if a test is expected to have an impact on the clinical pathway, and the goal of the health service is to maximise patient health outcomes, then outcome-based APS represent the best approach to ensuring this objective. Despite this clear rationale, outcome-based APS have remained, thus far, elusive (32). In particular, the conduct of direct outcome studies in this context is expected to be limited by challenges related to ethical, financial and time constraints (32). In this respect, indirect outcome studies present a more pragmatic approach to the derivation of outcome-based APS – as of yet, however, there appears to have been limited uptake of indirect methods in this context. Whilst the reason for this is unclear, it seems reasonable to attribute part of this paucity to a limited awareness and/or expertise amongst the clinical sciences community, as to how exactly to undertake indirect outcome studies. The lack of methods guidance in this area, in particular, is expected to be a key barrier to the effective implementation of outcome-based APS.

1.3 The test evaluation pathway

Section 1.2 provided an introduction to the topic of measurement uncertainty. This section discusses the broader context of the *test evaluation pathway* – defined as the trajectory of research required to take a new technology from the biomarker discovery phase, to the test adoption phase. Key components of the pathway are first outlined below (section 1.3.1); followed by a discussion of end-stage outcomes of most relevance to policy decision making (i.e. clinical utility and cost-effectiveness) (section 1.3.2). Section 1.3.3 then discusses the role of HTAs in the context of test reimbursement, and highlights the current lack of guidance concerning the evaluation of measurement uncertainty in this context.

1.3.1 Key components of the test evaluation pathway

Various authors have provided alternative characterisations of the test evaluation pathway. In particular, a systematic review conducted in 2008 identified 19 such frameworks (36), and numerous additional pathways (23, 37-47) and supplemental tools (48) have since been published. Whilst variation exists in the exact detail of the proposed pathways, there is general agreement regarding the key performance domains requiring evidence, as illustrated in Figure 1-5.[12] A brief summary of these elements is provided below.

[12] This figure is an adapted version of the figure provided in: 23. 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.

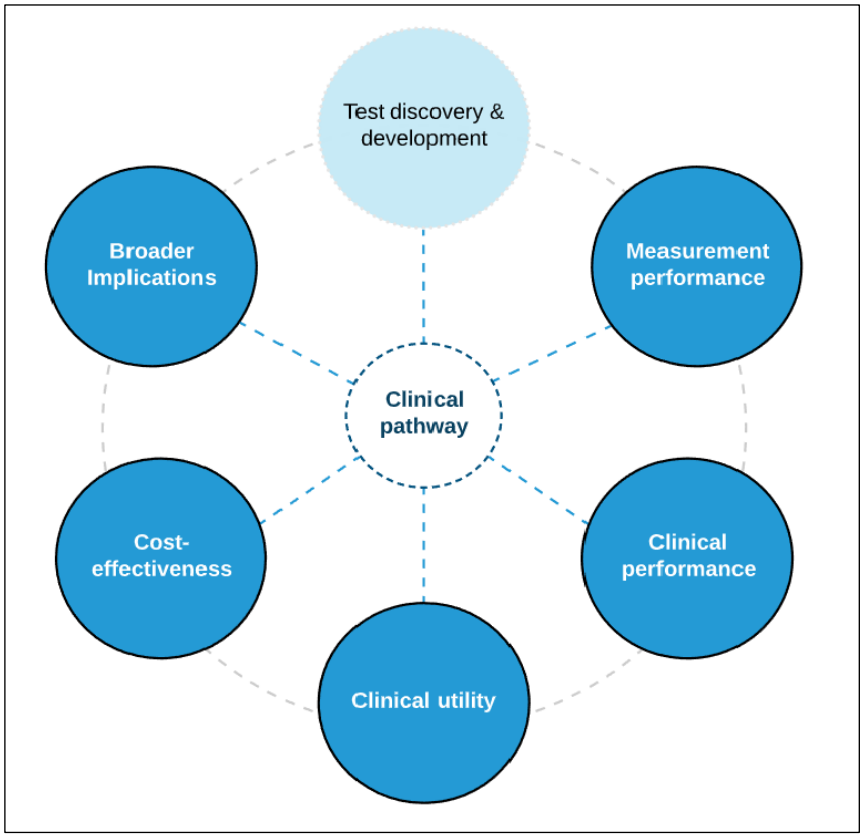


Figure 1-5. The test evaluation pathway

1. The first phase of the pathway is ***test discovery and development***. Essentially this relates to confirming the *scientific validity* of the test – that is, establishing that the test measurand is associated with a given clinical condition or disease state, such that there is a mechanism by which the test may provide useful information (and is therefore worth evaluating further).

2. The second element in the pathway is ***measurement performance*****.** Demonstration of measurement performance relates to the process of test validation and verification (discussed in section 1.2), including assessment of the central components of measurement uncertainty plus additional metrics of measurement performance (further outlined in Appendix B.3).

3. The third element in the pathway is ***clinical performance***. Focusing on diagnostic and screening tests, clinical performance is assessed in terms of *diagnostic accuracy* – defined as the ability of a test to discriminate between diseased and non-diseased subjects, or between two or more clinical states (23). Diagnostic accuracy is evaluated by comparing test-directed diagnoses against "true" diagnoses (based, for example, on a gold standard

diagnostic reference test); this enables calculation of diagnostic accuracy estimates (including diagnostic sensitivity and specificity, and positive and negative predictive values), as demonstrated in Appendix C. For predictive or prognostic tests, clinical performance may alternatively be evaluated in terms of the strength of association between the prognostic/predictive marker and the event or health state of interest; whilst for monitoring tests, clinical performance generally relates to the ability of changes between or trends in serial test values to identify or predict a change in health state.

4. The fourth element of the pathway is ***clinical utility***, which describes the clinical value that can be derived from a test. This includes both *intermediate clinical utility*, relating to the impact of test results on patient management decisions (e.g. the decision to treat or not treat); and *end-stage clinical utility*, relating to the impact of test results on patient health outcomes (e.g. patient mortality and morbidity).
5. The fifth element of the pathway is ***cost-effectiveness***, defined as the ability of an intervention to produce an efficient impact on patient health outcomes in relation to costs. In the context of a publicly funded healthcare system (such as the National Health Service [NHS]), the efficiency of additional spending is assessed in terms of whether or not the clinical value gained from that additional spending (e.g. life years gained), outweighs the associated opportunity cost (e.g. life years that could have been gained, had the money been spent elsewhere in the healthcare system). A summary of key measures of cost-effectiveness (i.e. the incremental cost-effectiveness ratio [ICER] and net monetary benefit [NMB]) is provided in Appendix D.
6. The final element of the pathway concerns ***broader implications*** of the adoption decision. This includes potential social, psychological, legal, ethical, societal and organizational consequences which may result from implementing a new test (23).

1.3.2 End stage outcomes: clinical utility and cost-effectiveness

In the context of deciding whether or not to adopt a new test into routine clinical practice, the primary concern for clinical decision makers and commissioners is

establishing the impact of testing strategies on end-stage outcomes – that is, *end-stage clinical utility (i.e. health outcomes)* and *cost-effectiveness*.

With respect to end-stage clinical utility, the gold standard method of evaluation is the randomized controlled trial (RCT)[13] (23, 49). However, whilst RCTs are well established in the evaluation pathway for pharmaceutical interventions, they are less common in the context of test evaluations – primarily due to the complex nature of testing pathways. In particular, tests do not have a *direct* impact on patient health, but rather exert an *indirect influence* by informing clinical management decisions. As such, the value of a test depends first on its ability to provide correct information on patients' health status; second on the potential for that information to produce a change in healthcare management; and third on the resulting impact of healthcare management changes on patient outcomes. Many other aspects of testing may further impact on clinical utility – including cognitive, emotional, social and behavioural effects of testing (50, 51). The design of test-treatment RCTs is therefore complicated by the need to appropriately capture each of these considerations. As a result RCTs in this context are rare, and the overwhelming majority of test evaluations instead focus on the intermediate outcome of clinical performance (52-54).

With respect to cost-effectiveness, test evaluations rely on obtaining evidence relating to the overall *cost*[14] and *effect*[15] of both the standard care clinical pathway, and the clinical pathway including the test intervention. Two overarching approaches are possible: (i) *trial-based analyses*, in which the required cost and effect estimates are derived directly from an RCT (with cost-effectiveness evaluated over the time horizon of the RCT); and (ii) *model-based analyses*, in which costs and effects are estimated via a mathematical model representative

---

[13] Or, where possible, meta-analysis of multiple RCTs.

[14] Depending on the perspective of the cost-effectiveness evaluation, different costs may be included. For example if an NHS perspective is adopted, then any costs relating to the consumption of NHS resources (e.g. primary care and secondary care appointments, tests, treatments, overheads etc.) should be captured.

[15] Different effects may be evaluated within cost-effectiveness analyses depending on the expected impact of the test: effectiveness may be assessed in terms of life years saved, for example, or according to quality-adjusted life years (QALYs) gained. QALYs provide a composite measure of patient survival weighted by quality of life (utility) over time. Cost-effectiveness analyses based on QALYs are also referred to as *cost-utility analyses*.

of the clinical pathway, and model input parameters (e.g. clinical performance, costs, and treatment efficacy) may be derived from various different sources (with cost-effectiveness evaluated over an assigned model time horizon) (55).

Due to the lack of test-treatment RCTs previously mentioned, cost-effectiveness analyses of tests commonly rely on model-based assessments (54, 56). In addition, since the majority of direct evidence in this field relates to clinical performance, a common approach to modelling in this context is to utilise "linked-evidence" decision models (54, 57). Essentially, linked-evidence models work on the basis of linking data on clinical performance (e.g. diagnostic accuracy) with data on (i) clinical decision making (e.g. treatment protocols) and (ii) treatment effectiveness (e.g. based on historic treatment RCT data; assuming transferability of this data to the tested population) (57). In this way, the use of linked-evidence models enable the test-treatment pathway to be modelled without the need for test-treatment RCT data. Whilst the potential limitations of this approach should be noted (in particular the required assumption of transferability of linked data), the pragmatic utility of this approach has resulted in linked-evidence models becoming widely endorsed by key technology appraisal and reimbursement bodies (discussed in section 1.3.3 below) (58-60).

1.3.3 Test reimbursement and HTA

As discussed in section 1.2.4, test manufacturers are required to show compliance to regulatory standards before being lawfully entitled to market tests and devices across the EEA; securing a CE mark, however, does not secure a test's place into routine clinical practice. Rather, test manufacturers must further illustrate to relevant local and/or national technology appraisal and reimbursement authorities that their test is of value both to the tested patient population, and to the health service as a whole.

The internationally accepted gold standard tool for informing test adoption and reimbursement decisions, is the health technology assessment (HTA). The World Health Organisation (WHO) defines HTA as follows:

> *"[HTA] refers to the systematic evaluation of properties, effects, and/or impacts of health technology. It is a multidisciplinary process to evaluate the social, economic, organizational and ethical issues of a*

health intervention or health technology. The main purpose of conducting an assessment is to inform a policy decision making." (61)

According to the International Network of Agencies for Health Technology Assessment (INAHTA), there are now over 50 HTA agencies registered worldwide, affecting decision making for over 1 billion people across 32 countries (62). Focusing on the UK, there are two key national bodies relating to the conduct of HTAs: (1) the National Institute for Health Research (NIHR) – the country's largest funder of health and care research, which includes a specific HTA programme intended to help inform policy decision making (63); and (2) the National Institute of Health and Care Research (NICE) – the nation's primary technology reimbursement authority, which makes recommendations on the adoption of new health technologies based on HTA-style assessments, and produces clinical guideline documents (64).

In response to the growing importance of tests, many HTA and reimbursement authorities now include such technologies within their remit. NICE in particular has three schemes under which test evaluations may fall: (i) the Technology Appraisal Programme (TAP), primarily for the assessment of pharmaceutical interventions but also including companion diagnostics; (ii) the diagnostics assessment programme (DAP), for the assessment of stand-alone tests expected to increase costs and/or disrupt current care pathways; and (iii) the medical technologies guidance (MTG) stream, for the assessment of stand-alone tests expected to be cost saving and with limited impact on care pathways (65). In addition, population screening tests may be separately evaluated by the UK National Screening Committee (NSC) (66). Several other national reimbursement authorities have also begun to issue test-specific recommendations over the past two decades – most notably the Medical Services Advisory Committee (MSAC) in Australia, and the Canadian Agency for Drugs and Technologies in Health (CADTH) (59, 67).

With respect to the test evaluation pathway, the predominant focus of HTAs is on the evaluation of intermediate and end-stage outcomes i.e. clinical performance, clinical utility, and cost-effectiveness; whilst the assessment of measurement uncertainty (or measurement performance) appears to have been largely overlooked. For example, each of the UK NICE and NSC programmes highlighted

above demands evidence on the clinical performance and clinical utility of testing strategies, with the TAP, DAP and NSC schemes also requiring a full economic evaluation (65, 68). No requirement for evidence on test measurement uncertainty or measurement performance-related concepts, however, is listed in any of the associated NICE or NSC programme manuals (58, 68-70). Based on an informal review of documentation from other national reimbursement/HTA authorities, this stance appears to be largely mirrored on the international stage (60, 71-76). The current state of play in the HTA field, therefore, seems to disregard the role of measurement uncertainty, and instead focuses on subsequent domains of the test evaluation pathway (as per Figure 1-5).

A notable exception to the above observation relates to the Australian MSAC programme: under this scheme, measurement uncertainty is included within the requested evidentiary requirements, focusing on the assessment of imprecision (in particular reproducibility) as reported in diagnostic accuracy studies included in the HTA clinical performance assessment (59). Nevertheless, the assessment of measurement uncertainty under this programme is limited in two key respects: first, measurement uncertainty is evaluated as a secondary outcome (since the evidence review relates only to what happens to have been reported within identified diagnostic accuracy studies); and second, the potential impact of measurement uncertainty on outcomes is not evaluated. Thus, whilst the MSAC programme currently sets the highest bar in terms of the evaluation of measurement uncertainty within HTAs, one can argue that this programme does not go far enough. Section 1.4 below further expands on the justification for why the broad omission of measurement uncertainty should be considered a key limitation of prevailing test evaluation methodology.

1.4 Thesis rationale

The existence of measurement uncertainty means that any observed test value will be different, to some degree, from the underlying true target value one wishes to measure. If, as a result, test values are incorrectly observed as lying above or below key test decision thresholds, then this uncertainty will – in the first instance – affect the *clinical performance* of testing strategies.

This impact can be illustrated via a series of simple simulations. Consider the case of a diagnostic test which aims to distinguish between 'diseased' and 'healthy' patients. Suppose that both populations exhibit normal distributions with respect to the measurand, and that, even in the face of perfect measurement a proportion of diagnostic errors occur due to a natural overlap between the two distributions. This scenario is illustrated in Panel A of Figure 1-6.[16] In this case, the placement of the diagnostic cut-off threshold (set at 45) produces equal proportions (6.67%) of false positive and false negative cases.

Introducing a fixed positive bias (+5) leads to an upward shift in both populations, resulting in an increased false positive rate (15.86%) and a decreased false negative rate (2.28%) (Panel B of Figure 1-6); while introducing imprecision (10% CV) increases the spread of the distributions resulting in an increased false positive rate (7.95%) *and* false negative rate (9.46%) (Panel C of Figure 1-6). Introducing both imprecision and bias together, meanwhile, leads to both distributions being shifted upwards and more widely dispersed, resulting in an increased false positive rate (16.81%) and decreased false negative rate (3.63%) (Panel D of Figure 1-6).

[16] The distributions of healthy (H) and diseased (D) populations here are based on simulations drawn from normal distributions: H~N(30,10) and D~N(60,10). Bias (α) is applied by adding α to the population means: H′~N(30+α,10) and D′~N(60+α,10). Imprecision [CV=β%] is applied at the individual simulation level: for the ith simulation from the H and D populations (i.e. H_i and D_i), imprecision is applied as an additional random draw from $N(0,H_i*(\beta/100))$ and $N(0,D_i*(\beta/100))$ respectively.

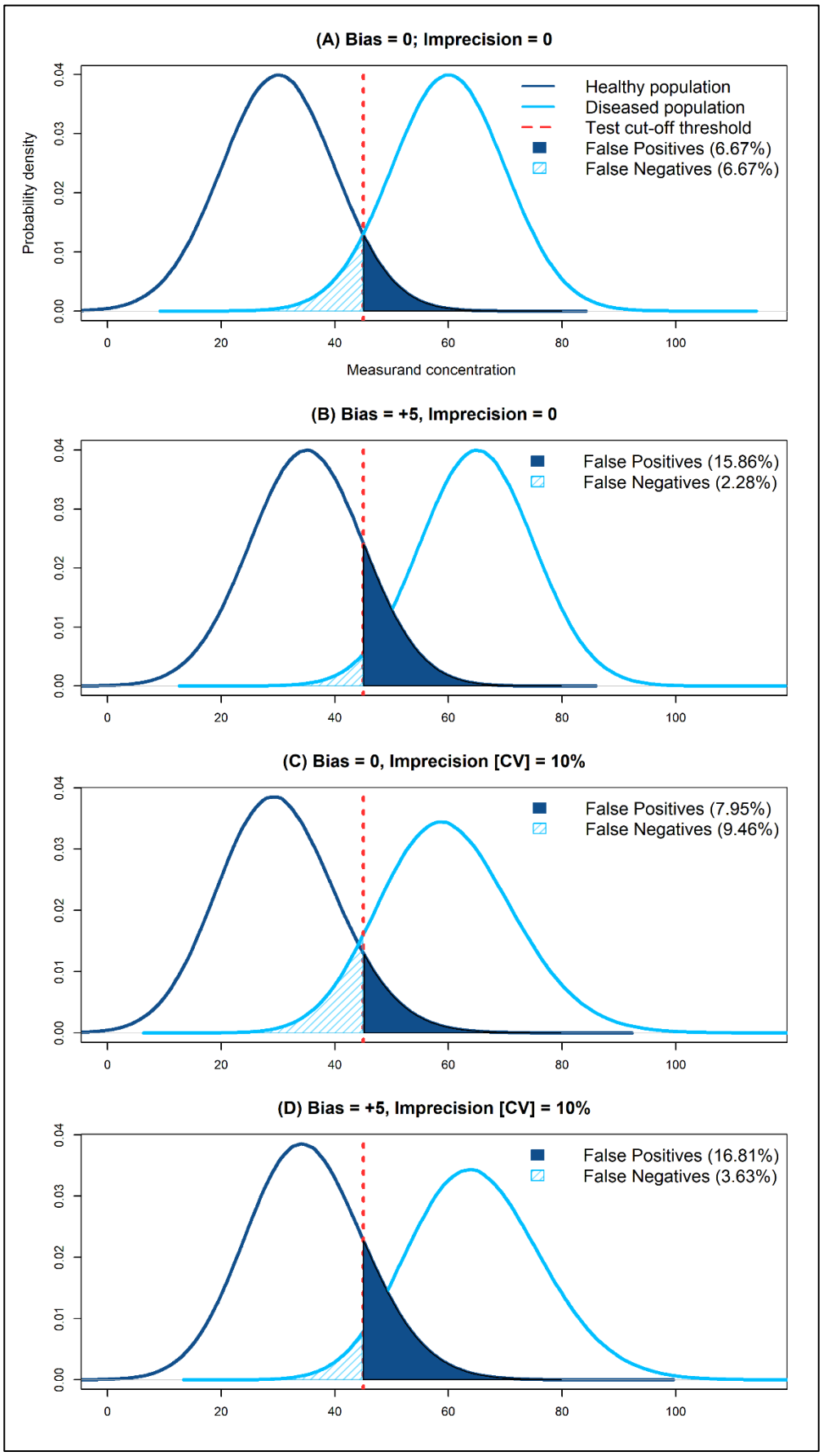


Figure 1-6. Hypothetical simulation results showing the impact of bias and imprecision on test diagnostic accuracy

Figure 1-6 illustrates how test measurement uncertainty can affect the clinical performance of testing strategies. Since the clinical and health-economic value of a test depends crucially on its clinical performance, it stands to reason that shifts in clinical performance caused by measurement uncertainty, if uncorrected, will lead to associated shifts in clinical utility and cost-effectiveness. In the scenario illustrated in Figure 1-6, for example, if an unknown bias of +5 were to occur, then the false positive rate is expected to increase from ~7% to ~16%; this could have serious consequences for patient and health service outcomes if a positive diagnosis initiates a course of expensive and/or risky treatment. The seriousness of this effect will clearly be context dependent; nevertheless, this example illustrates the mechanisms by which test measurement uncertainty may impact on clinical and health-economic outcomes.

Although the importance of measurement uncertainty has long been appreciated in the medical laboratory setting, associated activities in this field (e.g. validation, verification and quality assurance procedures) focus primarily on the quantification and monitoring of measurement uncertainty, without formal consideration of downstream clinical or health-economic effects (see section 1.2). In addition, whilst current international guidelines encourage the use of outcome-based APS, these have – as of yet – been largely overlooked in favour of more pragmatically appealing approaches (i.e. Models 2 and 3 of the EFLM Milan criteria) (see section 1.2.5). Ultimately this is damaging to patients, since the measurement performance of tests is not optimised against patient outcomes.

In the HTA field meanwhile, where clinical and health-economic outcomes are of primary concern, the potentially influential role of measurement uncertainty on outcomes appears to have been largely overlooked. Assuming that HTA guidance documents are representative of methods used in HTA practice, then it can be hypothesised that little to no consideration of measurement uncertainty is currently taken within HTA studies (see section 1.3.3). If true, this would mean that test adoption decisions based on HTAs are currently being made without an understanding of how measurement uncertainty might affect tests' real-world performance. In the worst case scenario this means that test adoption decisions may simply be wrong: for example, if the positive assessment of a test is based around diagnostic accuracy estimates drawn from clinical studies that have failed

to capture real-world measurement uncertainty, and measurement uncertainty is also unknowingly a key driver of clinical utility and cost-effectiveness, then the expected clinical and economic benefits associated with a testing strategy will fail to be met in the routine testing environment. At a less extreme level, the omission of measurement uncertainty within the HTA setting represents a lost opportunity to inform evidence-based laboratory implementation and monitoring procedures: for example, if the outcomes for a given testing strategy are shown to be volatile to measurement uncertainty, then this information can be used to inform the need for, and design of, a national EQA scheme. Overall, the apparent disregarding of measurement uncertainty in the HTA context means that the validity of test-adoption decisions may be called into question. Based on this, and the lack of formal consideration of outcomes within the medical laboratory setting, there is clearly potential utility in exploring methods for the assessment of the impact of measurement uncertainty on clinical and health-economic outcomes.

1.5 Scope, aim, hypotheses and structure

1.5.1 Scope

In the broadest sense, a medical test comprises any piece of information which can inform the presence, nature and/or future trajectory of a patient's disease, from which a clinical course of action can be determined. This information may range from a basic review of patient history and presenting signs and symptoms, through to more complex and invasive testing such as imaging tests, laboratory tests, or biopsies.

Whilst all of the above tests are clearly important, the focus of this thesis is on medical laboratory (i.e. in-vitro) tests. As can be seen from the prior discussion of measurement uncertainty (section 1.2), laboratory tests are associated with their own specific set of factors which contribute to measurement uncertainty – factors which would be expected to differ across different types of medical tests. Thus, whilst many of the general concepts and ideas discussed in this thesis in relation to measurement uncertainty are likely to be relevant to other tests, the

remit of this research is restricted to medical laboratory tests[17]. Henceforth, 'medical laboratory tests' and 'tests' are used interchangeably.

Further to the above, the introduction to measurement uncertainty provided in section 1.2 also centred on the case of *quantitative tests* – that is, tests which measure the quantity of a given measurand on a continuous scale, and which report this quantity as a numerical result. Other types of tests include *semi-quantitative tests* (which report a range within which the numerical result is expected to fall) and *qualitative tests* (which indicate a binary result [e.g. absent/ present] or an ordinal result [e.g. low/ moderate/ high]). Whilst semi-quantitative and qualitative tests are not excluded from the remit of the thesis[18], the preceding introduction to measurement uncertainty focused on quantitative tests due to the fact that methods for evaluating test measurement uncertainty in the clinical sciences field have largely focused on quantitative tests (77, 78). The thesis case study (Chapter 4 to Chapter 7) also relates to a quantitative test.

In addition to there being many different types of tests, there are also many ways in which tests may be used. As well as informing clinical diagnoses, tests may be used to screen asymptomatic patients; provide a prognosis or prediction regarding the future course of disease; or monitor disease status and/or risk of disease. Where possible, this thesis considers each of these different test roles. In particular, two literature reviews conducted as part of this research (presented in Chapter 2 and Chapter 3), do not exclude any studies on the basis of the specific role of the test or tests evaluated. However, the thesis case study (presented in Chapter 4 to Chapter 7) necessarily focuses on a particular example – in this case, a diagnostic test.

1.5.2 Aim

In order to help address key issues previously highlighted in section 1.4, the aim of this thesis is to develop a framework for assessing the impact of test

[17] Note however that the literature reviews presented in Chapter 2 and Chapter 3 of this thesis also include point-of-care tests (POCTs) (used outside of the traditional laboratory setting) within their remit.

[18] In particular two literature reviews conducted as part of this research (see Chapter 2 and Chapter 3) do not exclude any studies on the basis of this feature of the evaluated tests.

measurement uncertainty on clinical and health-economic outcomes (including clinical performance, clinical utility and cost-effectiveness). This framework is intended to provide utility both in the medical laboratory setting, as a means of aligning test measurement performance with clinical and health-economic outcomes; and in the HTA setting, as a means of capturing the impact of test measurement uncertainty on evaluated outcomes to inform appropriate test adoption and reimbursement decisions.

1.5.3 Hypotheses

Specific hypotheses assessed in this thesis are listed below:

- **Hypothesis A**: That measurement uncertainty has not, to date, been routinely addressed within HTAs.
- **Hypothesis B**: That methods for assessing the impact of measurement uncertainty on outcomes have been used in the broader literature (e.g. in laboratory medicine studies).
- **Hypothesis C**: That methods from the broader literature (e.g. the medical laboratory field) may be applied within HTA-style assessments, to evaluate the impact of measurement uncertainty on clinical performance, clinical utility and cost-effectiveness outcomes.
- **Hypothesis D**: That the application of methods from the broader literature to HTA-style assessments (as outlined in Hypothesis C) could enable outcome-based APS to be derived.
- **Hypothesis E**: That methods from the broader literature may be applied or adapted to allow real world evidence (RWE) (relating to test measurement performance data) to be utilised within outcome-based assessments.

Section 1.5.4 below highlights which chapters of this thesis address each of these hypotheses.

1.5.4 Structure

In order to address the thesis aim, the first part of this research focuses on evaluating the current methods landscape in this area: first by reviewing methods

applied in the HTA setting; and second by reviewing methods used in the wider literature. The second part of this thesis is aimed at developing key methods identified from the two reviews via a case study analysis. Overall this thesis is divided into eight chapters, outlined below.

- **Chapter 1** (the current chapter) provides an introduction to measurement uncertainty within the context of the test evaluation pathway.
- **Chapter 2** reports on a systematic review of international HTAs, which aims to identify if and how test measurement uncertainty has been assessed within HTAs to date. This chapter addresses hypothesis A.
- **Chapter 3** presents a methodology review of the wider literature, which aims to identify studies using indirect methods (i.e. excluding trial-based analyses) to incorporate or explore the impact of test measurement uncertainty on clinical and/or health-economic outcomes. This chapter addresses hypothesis B.
- **Chapter 4** introduces the case study used in this thesis: faecal calprotectin (FC) for diagnosing Inflammatory Bowel Disease (IBD) in primary care. Two FC primary care pathways are introduced: the 'NICE FC pathway' (based on a single FC test), and the 'York FC Care Pathway' (YFCCP) (based on a repeat-test strategy).
- **Chapter 5** applies simulation techniques identified in Chapter 3, to evaluate the impact of increasing FC measurement uncertainty on the diagnostic accuracy of the two FC pathways. Based on the simulated results, APS are presented based on assumed diagnostic accuracy requirements. This chapter addresses hypotheses C and D.
- **Chapter 6** extends the evaluation outlined in Chapter 5 to cost-effectiveness outcomes, using an adaptation of a previously constructed YFCCP economic model. Based on the simulated results, APS are presented based on achieving cost-effectiveness benchmarks. This chapter addresses hypotheses C and D.
- Building on Chapter 5 and Chapter 6, **Chapter 7** presents an analysis of how real-world measurement performance data may be applied within the simulation framework. In this case, EQA data is used to evaluate the

performance of alternative FC assays within the YFCCP, in order to address the question of how between-assay measurement differences may affect clinical and health-economic outcomes. This chapter addresses hypothesis E.

- **Chapter 8** provides a summary and discussion of the presented research. Key findings and limitations of the research are outlined, and possible areas for future applications and development of the methods are discussed.

Following on from this introduction, Chapter 2 subsequently aims to review methods applied in the HTA context, with respect to the evaluation of measurement uncertainty.

Chapter 2

The role of measurement uncertainty in HTAs of tests: a systematic review

2.1 Chapter outline

Chapter 1 provided an introduction to the thesis topic of test measurement uncertainty within the context of the test evaluation pathway. The hypothesis stated was that, to date, measurement uncertainty has rarely been considered within downstream evaluations of tests – such as HTAs – which drive test reimbursement and adoption decisions (hypothesis A; sections 1.3.3 and 1.5.3). The aim of this chapter, therefore, is to formally identify if and how test measurement uncertainty has previously been evaluated within HTA's, focusing on those assessments including a model-based economic evaluation. To that end, a systematic review of HTAs of tests was conducted. This chapter first outlines the review methods (section 2.2), followed by the study findings (section 2.3), discussion (section 2.4) and summary (section 2.5).

The work presented in this chapter has also been published as part of a jointly-authored peer-reviewed publication in *PharmacoEconomics* (Smith AF *et al*. [2018]) (1).

2.2 Methods

A systematic review was conducted to evaluate if and how measurement uncertainty has previously been evaluated within HTAs of tests, focusing on studies including a model-based economic evaluation. The review protocol was registered in advance on the PROSPERO database and can be accessed at: www.crd.york.ac.uk/PROSPERO (ID=CRD42017056778).

The primary source for this review was the Centre for Reviews and Dissemination (CRD) HTA database (79)[19]. At the time of conducting the review, this database consisted of completed and ongoing HTA's from authorities registered with the

[19] Whilst the maintenance of other CRD databases (the Database of Abstracts of Reviews of Effects [DARE] and the NHS Economic Evaluation Database [NHS EED]) ceased in March 2015, the HTA database continued to be updated and maintained at the time of conducting this review (March 2017).

International Network of Agencies for HTA (INAHTA) (n=49), in addition to other regional and national authorities (n=18). As per the INAHTA membership eligibility criteria, INAHTA members are non-profit organisations assessing healthcare technologies, relating to a regional or national government, funded at least 50% by public sources and providing free access to reports on request (http://www.inahta.org/). The additional listed authorities, meanwhile, consisted of regional or national bodies who had submitted a request to CRD to be included in the database[20]. The database therefore included reports from: (i) national assessment agencies, such as the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany; (ii) regional centres, such as Technology Assessment at The Hospital for Sick Children (SickKids) (TASK) in Canada; and (iii) publically funded research councils such as the UK NIHR. Appendix E provides a full list of the included authorities.

A search was conducted on the CRD HTA database in March 2017. The search strategy (provided in Appendix F) aimed to identify HTA reports which had evaluated an in-vitro test (including medical laboratory tests and POCTs used outside of the laboratory) and included an economic decision model. The strategy was developed with support from information specialists at the university of Leeds, and combined two elements: (i) MeSH heading and free-text terms related to *in-vitro tests* (lines #1-11); and (ii) a search filter consisting of MeSH heading and free-text terms for identifying *economic decision models* (lines #12-28).

In addition to the CRD HTA database search, online records of key reimbursement authorities expected to be the largest contributors of relevant HTAs were cross-checked. These included: NICE in the UK; CADTH in Canada; and MSAC in Australia (80-82). Backwards citation checking of all included HTAs (in which the bibliographies of included reports were electronically checked) was also conducted to identify any additional relevant studies.

HTA reports were included in the review if they met the inclusion criteria listed in Table 2-1. Note that, whilst the CRD HTA database records extend back to 1989 (83), technology evaluations of tests are a relatively recent phenomenon – the

[20] Note: no formal selection process was conducted by the CRD to identify or screen these additional included authorities.

pragmatic decision was therefore made to restrict the search to reports published from 1999 onwards (i.e. the year in which NICE was established). As highlighted in Table 2-1, only HTAs including a model-based health-economic evaluation were included. This restriction was applied so as to focus on reports that: (a) would provide information on if and how measurement uncertainty has been evaluated within both clinical and health-economic components of HTAs; and (b) would be most likely to have attempted to evaluate of the impact of measurement uncertainty on outcomes (i.e. focusing on indirect studies, such as model-based analyses, and excluding direct [trial-based] analyses unlikely to have attempted such an evaluation [as discussed in section 1.2.5.2]).

Table 2-1. HTA systematic review: inclusion criteria

Item	Inclusion criteria
Study design	HTA report including a model-based economic evaluation
Intervention	In-vitro laboratory medical test or POCT (including diagnostic, screening, prognostic, predictive and/or monitoring tests)
Population	Any human population
Setting	Any
Indication	Any
Date	1999 onwards
Language	Full HTA report available in English

A two-stage screening process was conducted to identify eligible reports. First, all titles and abstracts were screened by the primary reviewer (Smith AF) and 10% were independently screened by a secondary reviewer (Hulme CT). Studies judged as potentially meeting the inclusion criteria were included in the second round of screening, in which full HTA reports were reviewed by the primary reviewer only. Any uncertainties regarding final inclusions were checked with the secondary reviewer and, where necessary, additional project supervisors (Messenger MP and Hall PS). Identified records were managed using Endnote V 7.2 (Thompson Reuters).

A data extraction table was constructed and piloted on the first 10% of included HTAs by the primary reviewer. Details relating to the study and test characteristics, components of measurement uncertainty assessed (if any), and methods used within such assessments (where applicable) were included in the final data extraction table. All data extraction was conducted by the primary reviewer, with 10% independently checked by the secondary reviewer.

It should be noted that a wide remit of what constituted a relevant metric of 'measurement uncertainty' was considered, due to the hypothesis that few HTAs would have addressed this topic. This included the central components of imprecision and bias, as well as summary metrics (TE and U_M), and other components related to the broader topic of measurement performance (including detection and quantification limits, pre-analytical and analytical affects, linearity and test failure[21] rates). In addition, in order for a study to be classified as having conducted an *assessment* of measurement uncertainty, relevant components of measurement uncertainty had to feature in both the methods and results section of the HTA report. This means that studies which mentioned measurement uncertainty in the introductory and/or discussion sections of the report, but where measurement uncertainty did not clearly feature in the study analysis, were not included under the banner of studies having conducted an assessment measurement uncertainty.

[21] A *test failure* relates to instances wherein no quantitative, semi-quantitative or qualitative result is able to be provided for a given test, due to some form of failure occurring during the testing pathway (e.g. insufficient sample to run the test, sample spillage or equipment failure).

2.3 Results

All data generated from this study (including the endnote library, data extraction form and associated analyses) are available in the Research Data Leeds Repository (https://doi.org/10.5518/324). From the CRD HTA database search, 1908 citations were retrieved and one duplicate study was subsequently removed. After conducting the two-stage screening process, 90 studies were included. Agreement between the primary and secondary reviewers during abstract screening was good (k=0.85)[22]. A further 17 studies were identified via checking the online records of key HTA authorities (n=10) and citation tracking (n=7), resulting in a total of 107 included HTA reports (see Figure 2-1). A summary of included study characteristics is provided in Table 2-2.

Of the 107 identified HTAs, 71 (66%) did not evaluate measurement uncertainty or any of the additional components of measurement performance considered. Sixteen (15%) studies incorporated data on test failure rates only – for example, including 'test failures' as an item within the HTA literature review or as a parameter within the economic model. The isolated inclusion of test failure rates was considered to be of limited interest for the purposes of this review. This is because test failures – whilst clearly important in terms of determining the overall clinical performance and utility of a test – do not represent a component of measurement *uncertainty* per se, since quantification of uncertainty around a measurement first requires a measurement to be successfully obtained. These studies are therefore not included in the subsequent narrative review and are henceforth included under the banner of studies not addressing measurement uncertainty.

[22] Note that all discrepancies were a result of the primary reviewer being more inclusive than the secondary reviewer at the initial screening stage.

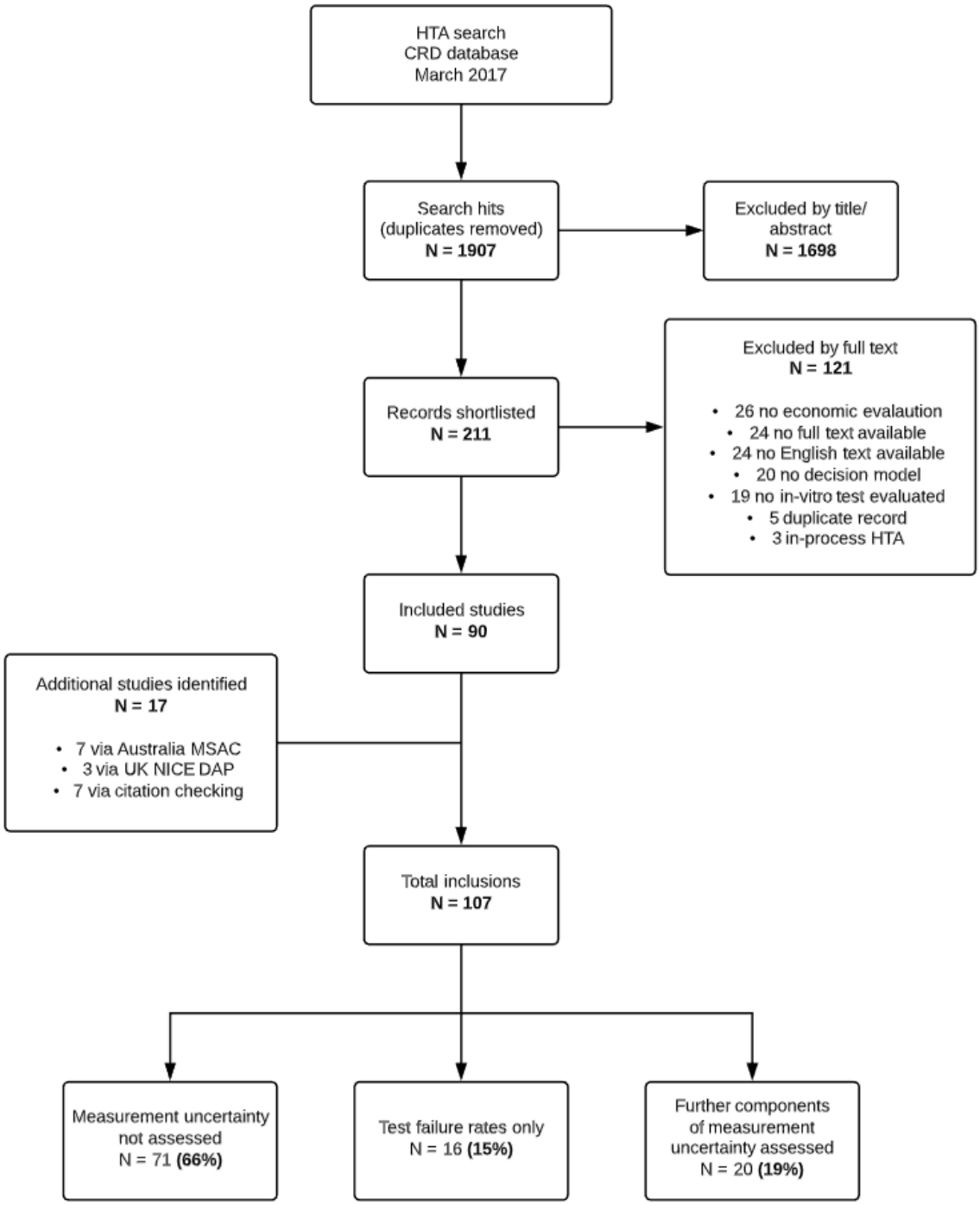


Figure 2-1. HTA systematic review: PRISMA diagram of included studies

Table 2-2. HTA systematic review: summary of study characteristics

	All studies (N=107)		Studies including measurement uncertainty (n=20)	
	Number	% (out of 107)	Number	% (out of 20)
Country				
UK	66	62%	15	75%
Canada	17	16%	1	5%
Australia	15	14%	3	15%
Belgium	3	3%	1	5%
USA	3	3%	0	0%
Ireland	2	2%	0	0%
Italy	1	1%	0	0%
Disease Area				
Cancer	36	34%	8	40%
Pregnancy care & screening	14	13%	1	5%
Cardiology	12	11%	3	15%
Haematology	12	11%	2	10%
Infections	13	12%	1	5%
Diabetes	6	6%	1	5%
Gastroenterology	5	5%	2	10%
Other	9	8%	2	10%
Type of test(s) Evaluated				
Laboratory tests only	85	79%	14	70%
POCT- clinician led	18	17%	3	15%
POCT- self led	5	5%	3	15%
Primary Role of Test(s)				
Diagnosis	39	36%	5	25%
Screening	37	35%	2	10%
Prognosis	14	13%	5	25%
Monitoring	9	8%	5	25%
Predictive	6	6%	2	10%
Other	2	2%	1	5%
Type of Evaluation				
Cost-utility	53	50%	13	65%
Cost-effectiveness	36	34%	5	25%
Cost-utility & cost-effectiveness	17	16%	2	10%
Cost-consequences	1	1%	0	0%
Type of Economic Model				
Decision tree	48	45%	7	35%
Cohort Markov	22	21%	4	20%
Decision tree + Markov	17	16%	6	30%
Patient level simulation	12	11%	2	10%
Infectious disease/ dynamic	6	6%	1	5%
Not reported	2	2%	0	0%

2.3.1 Studies including measurement uncertainty

Twenty HTAs (19%), summarized in 22 reports (84-105), considered further components of measurement uncertainty (i.e. not limited to inclusion of test failure rates alone). The majority were UK studies (n=15, 75%), which evaluated laboratory tests (n=14, 70%) (Table 2-2) and were published from 2009 onwards (n=18, 90%) (Figure 2-2).

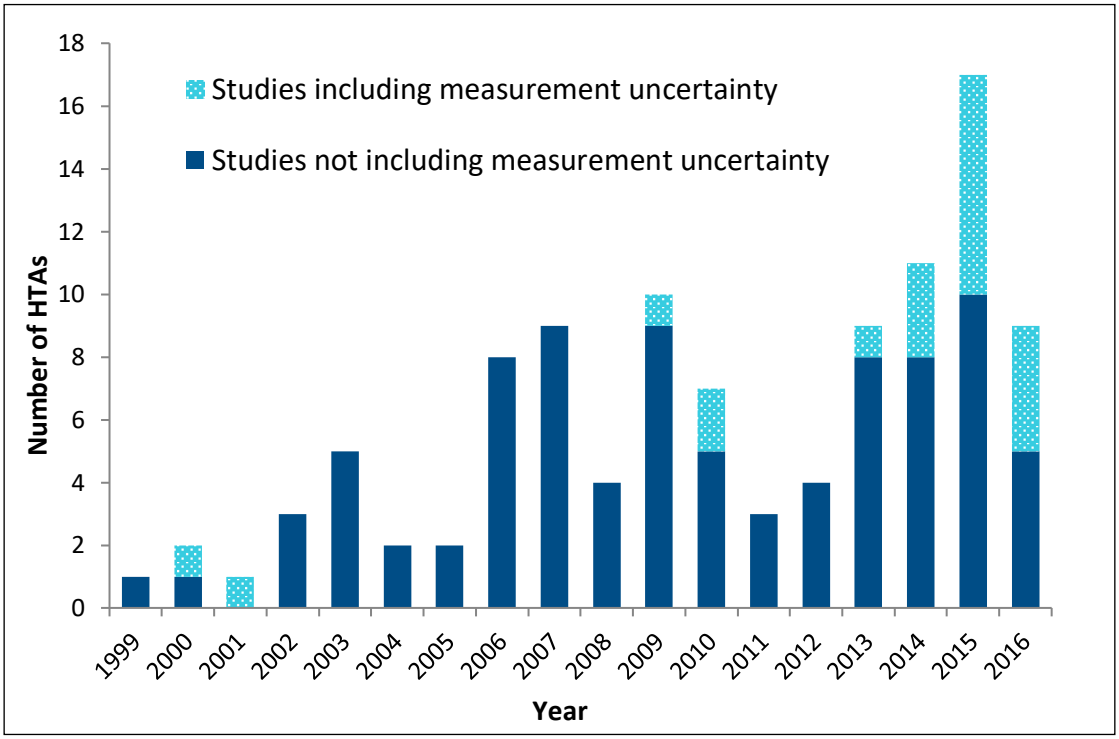


Figure 2-2. HTA systematic review: frequency of HTA reports by year of publication and inclusion of measurement uncertainty

Nineteen of the 20 studies assessed the specified element(s) of measurement uncertainty via some form of assessment prior to the economic model, henceforth referred to as "*pre-model assessments*". Details of the methods used within these studies are summarised in Table 2-3.

The majority of pre-model assessments included elements of measurement uncertainty within a systematic (n=13) (85-92, 96, 99-101, 104, 105) or non-systematic (n=2) (95, 97, 98) literature review. In these cases, components of measurement uncertainty were typically included as an additional outcome within the primary HTA systematic review alongside clinical performance and/or clinical utility outcomes, using a single overarching search strategy (86, 89-92, 99, 101, 104, 105). Alternatively, a handful of studies conducted a *separate* review for

measurement uncertainty (85, 87, 88, 95-98, 100). Of those, three studies applied an outcome-specific search filter: Pearson *et al*. (2010) simply combined the test name with the term 'measurement' (87); while in the MSAC (2001) study, several key words ('Precision', 'Accuracy', 'Quality control' and 'Quality assurance') were combined (85); and in the Nicholson *et al*. (2015) study, key MeSH terms ('Accuracy', 'Diagnostic Errors', 'Sensitivity and Specificity' and 'reproducibility of results') were combined with a title and abstract search for 'analytic validity' or '(repeatability or reproducibility)' (96).

Whether or not a single or separate searches were conducted, most studies either did not attempt to conduct a quality assessment of the included measurement literature, or, when quality assessment was undertaken, used checklists developed for other primary purposes (e.g. QUADAS/ QUADAS-2 [Quality Assessment of Diagnostic Accuracy Studies]; a checklist intended to assess the risk of bias in diagnostic accuracy studies included in a systematic review) (106, 107). One exception was the Nicholson *et al*. (2015) study, which used an adapted version of the Evaluation of Genomic Applications in Prevention (EGAPP) initiative checklist – a tool developed for the purpose of evaluating the internal validity of analytical validity studies (96, 108).

A handful of pre-model assessments used alternative/additional methods to reviewing the literature. Two HTAs supplemented their systematic review with an online survey of laboratories participating in a national EQA scheme, collating data on test methods, logistics, technical performance and costs (95, 97, 98). In addition, one HTA included a primary pathology study, in which test sample data was used to evaluate the measurement agreement between alternative index tests (102); and a further study included a clinical trial to evaluate the measurement agreement between the same test conducted across alternative laboratory sites (103). Finally, two studies used individual patient-level data (IPD) datasets to construct statistical models describing the trajectory of test values over time, accounting for analytical and biological variation (91, 92).

Table 2-3. HTA systematic review: summary of methods used in pre-model assessments

Study	Test characteristics			Pre-model Assessments of Measurement Uncertainty			Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Components of MU assessed	Method (general)	Method (details)	
Marks *et al.* 2000 (UK) (84)	-	Cardiology	Screening	NA (no pre-model assessment conducted)	NA	NA (no pre-model assessment conducted)	Yes
MSAC 2001 (AUS) (85)	POCT: clinician-led	Cardiology	Prognosis	Trueness (% bias); precision (repeatability and reproducibility); TE; analytical effects (site, operator and sample type)	Systematic review	• Separate search for evidence on measurement uncertainty (search strategy included an outcome search filter) • Quantitative synthesis conducted to derive pooled CV% and % bias estimates • Quality assessment using: (i) an NHMRC grading system *(109)* and (ii) a Cochrane study validity checklist (reference NR)	Yes
Gailly *et al.* 2009 (BEL) (86)	POCT: self-led	Haematology	Monitoring	Precision (repeatability and intermediate); test failures	Systematic review	• Combined search for evidence on measurement uncertainty and clinical performance (searching on test name only) • Measurement uncertainty results were narratively synthesised • Quality assessment using an INAHTA checklist for HTAs *(110)* and the QUADAS checklist for primary studies *(106)*	-

Study	Test characteristics			Pre-model Assessments of Measurement Uncertainty			Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Components of MU assessed	Method (general)	Method (details)	
Pearson *et al.* 2010 (UK) (87, 88)	POCT: clinician-led	Gastro-enterology	Diagnosis	Biological variability; distribution in faeces; faecal matrix; interference; stability; patient compliance; normal range	Systematic review	• Separate search for evidence on measurement uncertainty (searching by test name + 'measurement') • Measurement uncertainty results were narratively synthesised • Quality assessment stated to be based on Oxford Centre for Evidence-Based Medicine criteria (reference NR); however it was unclear if this was used for measurement uncertainty studies	-
M.A.S 2010 (CA) (89)	-	Cancer	Prognosis	Precision (intermediate and reproducibility); test failures	Systematic review	• Combined search for evidence on measurement uncertainty, clinical performance and utility (searching on test name and condition only) • Measurement uncertainty results were narratively synthesised • No quality assessment of measurement uncertainty studies was conducted	-

Study	Test characteristics			Pre-model Assessments of Measurement Uncertainty			Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Components of MU assessed	Method (general)	Method (details)	
Ward *et al.* 2013 (UK) (90)	-	Cancer	Prognosis	Precision (intermediate and reproducibility); trueness (concordance)	Systematic review	• Combined search for evidence on measurement uncertainty, clinical performance and utility (searching on test name and condition only) • Measurement uncertainty results were narratively synthesised • Quality assessment of prognostic studies was conducted according to Altman et al. (2001) *(111)*	-
Westwood *et al.* 2014 (UK) (91)	-	Cancer	Predictive	Proportion of tumour cells needed; test failures	Systematic review + survey	• Combined search for evidence on measurement uncertainty, clinical performance and utility (searching on test name and condition only) • No measurement uncertainty studies were identified • Additional data obtained from an online survey of laboratories (n=31) participating in a UK NEQAS EQA scheme	-
Westwood *et al.* 2014 (UK) (92)	-	Cancer	Predictive	Proportion of tumour cells needed; LOD; test failures	Systematic review + survey	• Combined search for evidence on measurement uncertainty, clinical performance and utility (searching on test name and condition only) • No quality assessment of measurement uncertainty studies was conducted • Additional data obtained from an online survey of laboratories (n=13)	-

Study	Test characteristics			Pre-model Assessments of Measurement Uncertainty			Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Components of MU assessed	Method (general)	Method (details)	
						participating in a UK NEQAS EQA scheme	
Farmer *et al.* 2014 (UK) (93)	-	Diabetes	Screening	Biological and analytical variation	Analysis of IPD	• A longitudinal hierarchical linear model was constructed from IPD to model longitudinal test values, incorporating biological and analytical CV (further details in Table 2-4)	Yes
Perera *et al.* 2015 (UK) (94)	-	Cardiology	Monitoring	Biological and analytical variation	Analysis of IPD	• A longitudinal hierarchical linear model was constructed from IPD to model longitudinal test values, incorporating biological and analytical CV (further details in Table 2-4)	Yes
Sharma *et al.* 2015 (UK) (95)	POCT: self-led	Haematology	Monitoring	Precision (reproducibility); trueness (r correlation coefficient)	Literature review	• A table of studies reporting precision and bias outcomes was provided, stated to be based on FDA documentation and relevant published papers (review methods NR)	-

Study	Test characteristics			Pre-model Assessments of Measurement Uncertainty			Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Components of MU assessed	Method (general)	Method (details)	
Nicholson et al. 2015 (UK) (96)	-	Cancer	Diagnosis	Precision (intermediate and reproducibility); trueness (recovery); LOB, LOD, LOQ; interference; linearity; range; pre-analytical effects; stability; test failures	Systematic review	• Separate search for evidence on measurement uncertainty (search strategy included an outcome search filter) • Measurement uncertainty results were narratively synthesised • Quality assessment of measurement uncertainty studies using a modified version of a published checklist (Teutsch *et al*. 2009) *(108)*	-
MSAC 2015 (AUS) (97, 98)	-	Cancer	Prognosis	Analytical sensitivity and specificity (i.e. selectivity)	Literature review	• A table of studies reporting analytical sensitivity and specificity was provided, stated to be based on a recent review of these outcomes (review methods NR) • No quality assessment of the MU studies was conducted	-
Kessels *et al*. 2015 (AUS) (99)	-	Pregnancy care & screening	Diagnosis	Imprecision (test-retest reliability); analytical sensitivity; test failures	Systematic review	• Combined search for evidence on measurement uncertainty and clinical performance (searching on test name and condition only) • Measurement uncertainty results were narratively synthesised • Quality assessment using the QUADAS-2 checklist *(107)*	-

Study	Test characteristics			Pre-model Assessments of Measurement Uncertainty			Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Components of MU assessed	Method (general)	Method (details)	
Harnan *et al.* 2015 (UK) (100)	POCT: self-led	Other (asthma)	i) Diagnosis ii) Monitoring	Trueness (Bland-Altman analysis, correlation coefficients); test failures	Systematic review	• Separate search for evidence on measurement uncertainty (searching on test name only) • Measurement uncertainty results were narratively synthesised • No quality assessment of measurement uncertainty studies was conducted	-
Freeman *et al.* 2015 (UK) (101)	-	Cancer	Monitoring	Trueness (Bland-Altman analysis, Deming regression); test failures	Systematic review	• Combined search for evidence on measurement uncertainty and clinical utility, (searching by test name, condition, and outcome) • Measurement uncertainty results were narratively synthesised • Quality assessment using an adapted version of the QUADAS-2 checklist *(107)*	-
Stein *et al.* 2016 (UK) (102)	-	Cancer	Prognosis	Trueness (Kappa statistic, discordance)	Pathology study	• A pathology study (n=302 samples) was conducted, within a wider RCT study, to evaluate the agreement between alternative tests evaluated	Yes
Hay *et al.* 2016 (UK) (103)	POCT: clinician-led	Other (urology)	Diagnosis	Trueness (Kappa statistic); test failures	Clinical study	• Within a prospective diagnostic cohort study, tests with sufficient samples (n=4808) were analysed at both a central research laboratory and NHS laboratory, to assesses concordance	-

Study	Test characteristics			Pre-model Assessments of Measurement Uncertainty			Measurement uncertainty included in economic model?
	POCT?	Disease area	Primary role of test	Components of MU assessed	Method (general)	Method (details)	
Freeman *et al.* 2016 (UK) (104)	-	Gastro-enterology	Monitoring	Trueness (Bland-Altman analysis, Cohen's Kappa); test failures	Systematic review	• Combined search for evidence on measurement uncertainty, clinical performance and clinical utility (searching by test names, condition, and general outcome terms) • Measurement uncertainty results were narratively synthesised • Quality assessment using an adapted version of the QUADAS-2 checklist *(107)*	-
Auguste *et al.* 2016 (UK) (105)	-	Infection (TB)	Diagnosis	Trueness (Kappa statistic, discordance); test failures	Systematic review	• Combined search for evidence on measurement uncertainty and clinical performance (searching by test name and condition only) • Measurement uncertainty results were narratively synthesised • Quality assessment of clinical performance studies using the QUIPs tool *(112)*	-

AUS = Australia; BEL = Belgium; CA = Canada; CV = coefficient of variation; EQA = external quality assessment; FDA = U.S Food and Drug Administration; LOB = limit of blank; LOD = limit of detection; LOQ = limits of quantification; M.A.S = Medical Advisory Secretariat; NA = not applicable; NEQAS = National EQA Service (UK); NHMRC = (Australian) National Health and Medical Research Council; NR = not reported; POCT = point of care test; QUADAS = Quality Assessment of Diagnostic Accuracy Studies; QUIPs = Quality In Prognosis Studies; TE = total error; UK = United Kingdom.

Further to the identified pre-model assessments, five studies incorporated measurement uncertainty within the economic model itself (84, 85, 93, 94, 102). These reports are summarised in Table 2-4 and outlined below. Further discussion of these studies is provided in section 2.4.

The earliest of these studies, Marks *et al.* (2000), was the only HTA to consider measurement uncertainty in the economic model alone (i.e. without an accompanying pre-model assessment) (84). In this study, a measure of analytical and biological variation (CV%) was taken from a single published article, and used within the model as an estimate of the test's false negative rate. In the MSAC (2001) study, meanwhile, a more complex simulation process was used in which the addition of TE (derived from the HTA systematic review) was iteratively applied to baseline "true" test values (sampled from national survey data), with the resulting probability of disease misclassifications calculated for four levels of TE (0%, 4%, 8% and 11%) (85). The associated misclassification rates were then applied within the model to assess the cost-effectiveness of the index test at the varying TE levels.

Of the more recent model-based assessments, both the Farmer *et al.* (2014) and Perera *et al.* (2015) HTAs included statistical modelling of IPD datasets to estimate the longitudinal trajectory of test values, accounting for analytical and biological variation (93, 94). These statistical models subsequently formed the foundation of the economic decision models used in each study to assess the cost-effectiveness of repeated-testing strategies.

Finally, in the most recent study – Stein et al. (2016) – a pathology sub-study was conducted to assess the concordance between multiple alternative index tests (102). Within this analysis, the primary test under evaluation (Oncotype DX [ODX]) was used as the reference test, against which a series of competitor tests were assessed. In the subsequent economic model, the predictive utility of ODX was set equal to that observed in a previous clinical trial; the alternative tests were then evaluated by applying added uncertainty to the ODX level of predictive utility, proportional to the level of discordance observed in the pathology sub-study.

Table 2-4. HTA systematic review: summary of methods used in economic model assessments

Study	Model details			Assessment of measurement uncertainty				
	Tests evaluated	Type of model	Base case results	Components included	Source of evidence	Value(s) used	Method of incorporation	Impact on cost-effectiveness results
Marks *et al.* 2000 (UK) (84)	Screening test for hyper-cholesterol-aemia (universal, opportunistic & case finding strategies)	Decision Tree	Cost per LYG: £14,842 - £78,060 (universal strategies); £21,106 - £70,009 (opportunistic strategies); £3,300 - £4,914 (case finding strategies).	Biological and analytical variation	Individual cited paper (no formal review conducted)	Base case: co-efficient of biological and analytical variation = 6.5%.	Rate of false negatives in the model set equal to the reported coefficient of biological and analytical variability.	Not assessed
MSAC 2001 (AUS) (85)	Cholesterol screening POCT for coronary heart disease (vs. standard laboratory test)	Decision Tree	Incremental cost per LYG: AUS$133,934.	TE (% bias + 1.96*%CV)	Systematic review. Calculation used average of reported CV's and total % biases.	Base case: TE = 8%. Sensitivity analysis: TE = 0%, 4%, 11%.	10,0000 Monte Carlo simulations: (i) patients assigned a 'true' cholesterol level based on population survey data; (ii) two observed results generated based on CI of +/- 8% (i.e. TE); (iii) diagnosis based on average of the two results against threshold of 6.5 mmoL/L; (iv) probability of misclassifications based on weighted average across cholesterol range assessed (2.5 – 9.4 mmol/L).	Incremental cost (AUS$) per LYG: $101,419 (TE=0%); $115,615 (TE= 4%); $151,378 (TE=11%). Cost-effectiveness threshold = $100,000 per LYG.

Farmer *et al.* 2014 (UK) (93)	Screening test (ACR) for kidney disease in diabetes patients (1, 2, 3, 4 and 5-yearly intervals).	Individual patient simulation	Incremental cost per QALY (2-year vs. 1-year screening): £9,601 (Type 1 diabetes; SD = 34,112); £606 (Type 2 diabetes; SD = 1,782).	Biological and analytical variation	Retrospective analysis of longitudinal IPD databases	Estimated SD of variability: Type 1 diabetes = 0.79 (95% CI 0.73 to 0.86); Type 2 diabetes = 0.85 (0.74 to 1.00). Both correspond to >100% CV.	A longitudinal hierarchical linear model for log ACR was constructed from the IPD. Individual simulations as follows: (i) a representative population (n=75,000) was generated; (ii) baseline log ACR and progression rates simulated and used to calculate annual true log ACR values post-diagnosis; (iv) observed ACR values derived by adding biological & analytical variation to the true ACR values; (v) clinical performance determined using gender-specific threshold values.	Not assessed
Perera *et al.* 2015 (UK) (94)	Lipid monitoring tests for patients at risk or with cardiovascular disease.	Individual patient simulation	Annual monitoring dominated (was less costly and more effective than) all other strategies.	Biological and analytical variation	Retrospective analysis of longitudinal IPD databases	Estimated SD of variability across tests: 0.12 - 0.35 (male population); 0.14 - 0.37 (females).	Same method as above [longitudinal regression of IPD + individual simulations to model impact of progression and biological and analytical variation over time].	Not assessed
Stein *et al.* 2016 (UK) (102)	ODX (+ additional tests) to guide use of adjuvant chemotherapy in breast cancer patients (vs. chemotherapy for all).	Decision tree + cohort Markov model	Net Health Benefit (QALYs) for tests vs. chemotherapy for all: 6.99 QALYs (ODX); 7.16 - 7.20 (alternative tests).	Test discordance	*De novo* clinical pathology study	Kappa statistics for tests vs. ODX: 0.40 - 0.53. Agreement with ODX ranged from all tests agreeing in 39% of cases, to no test agreeing in 4% of cases.	Predictive effect of ODX for recurrence-free survival in the model was derived from a historic ODX clinical trial. For the alternative tests, extra uncertainty was introduced in the model according to the degree of discordance for each test vs. ODX. Tests were only included in the model if they met inclusion criteria, including a requirement of "sufficient evidence of analytical validity in support of an achievable rollout into routine care in the NHS".	Not assessed

ACR = Albumin-to-creatinine ratio; AUS = Australia; CI = confidence interval; CV= coefficient of variation; IPD = individual patient data; LYG = life year gained; ODX = Oncotype DX; POCT = point of care test; QALY = quality-adjusted life year; SD = standard deviation; TE = total error; Net Health Benefit (QALYs) = Incremental QALYs - (Incremental costs/ cost-effectiveness threshold); UK = United Kingdom.

2.4 Discussion

2.4.1 Review findings

The findings of this review verify the introductory claim that measurement uncertainty has not, to date, been routinely considered within HTAs of in-vitro tests: of 107 identified HTAs, most either did not assess measurement uncertainty (n=71, 66%), or only considered test failure rates (n=16, 15%). Nevertheless, despite limited guidance in this area, assessment of test measurement uncertainty was attempted in a minority of HTAs (n=20; 19%), indicating that such analyses are feasible. Indeed the inclusion of measurement uncertainty appears to be a predominantly recent phenomenon: of those studies which addressed measurement uncertainty, 75% (n=15) were published in the last 5 years alone (2012-2017), making up 30% (15/50) of the total HTAs over this time period (Figure 2-2). Although identification of the reasons driving the inclusion of measurement uncertainty within HTAs was beyond the scope of this review, the observation of this recent trend is encouraging.

The majority of identified HTAs including measurement uncertainty did so via some form of pre-model assessment (n=19; 95%). The typical method was to include aspects of measurement uncertainty within the primary literature review, using one overarching search strategy to identify evidence on multiple outcomes (e.g. measurement uncertainty and clinical performance). Although this approach is efficient in terms of utilising a single search, it nevertheless requires the use of a sensitive (and likely non-specific) search strategy, in order to ensure that studies reporting on separate outcomes are identified – for example, searching on the test name +/- the clinical condition alone. An alternative approach, taken in three of the identified studies, was to conduct a separate review for measurement uncertainty applying an outcome-specific search filter (85, 87, 96). The concern here is whether or not the adoption of such filters can safely improve the efficiency of the overall review whilst maintaining high sensitivity. As of yet, this question does not appear to have been addressed within the methodology literature.

There are two further notable aspects relating to the conduct of systematic reviews of measurement uncertainty, for which there is a clear lack of current

consensus and/or guidance. First, the evaluation of review findings has thus far been limited to narrative syntheses. Only one of the identified studies attempted to conduct a quantitative analysis, basing pooled estimates of imprecision and bias on simple arithmetic means of the individual study values identified *(85)*. Whether or not more sophisticated methods of quantitative synthesis may be warranted or feasible for measurement uncertainty outcomes is currently unclear. Second, in the handful of studies where quality assessment of the measurement literature was attempted, only one applied a tool specifically intended for this task (the EGAPP checklist) (108). Several other related quality and reporting frameworks are available in the literature, including: BRISQ [Biospecimen Reporting for Improved Study Quality]; STROBE-ME [Strengthening the reporting of Observational studies in Epidemiology-Molecular Epidemiology]; and RIPOSTE [Reducing IrreProducibility in laboratory STudiEs]) (113-115). The reason for the lack of uptake of these tools is unclear: it may be due to a lack of awareness or understanding as to which tool(s) to apply within the HTA community, or a lack of direct applicability of these tools to the HTA context.

A small minority of HTAs (n=5, <5%) included data on test measurement uncertainty within the economic model. Of those, the most recent study by Stein and colleagues (2016) was not a direct attempt to account for measurement uncertainty, but rather the authors here utilised data on between-test discordance for a group of prognostic tests, as a means of evaluating additional tests in the economic model for which no prognostic utility data was available (102). This approach presents an interesting means of evaluating additional tests, which at least recognises the fact that measurement discrepancies do impact on clinical performance. Nevertheless, meaningful assessment of this impact requires knowledge of patients' true clinical status, in order to ascertain if diagnostic/ prognostic classification changes resulting from measurement differences should be considered appropriate or inappropriate. The approach adopted herein, therefore, should only be considered in the absence of clinical performance or utility data, with the results interpreted with due caution.

The Marks *et al.* (2000) HTA similarly oversimplified the relationship between measurement uncertainty and clinical outcomes, this time in relation to a diagnostic biomarker (84). Here the authors set the proportion of false negative

results in the model equal to a given level of biological and analytical variability (CV%). This approach (similar to the *Stein et al*. (2016) HTA) fails to account for the fact that clinical performance depends on three factors: the true distribution of test values, the placement of diagnostic/decision threshold(s), *and* measurement uncertainty. This means that applying 10% imprecision to a healthy population distribution, for example, will result in different numbers of false positive cases depending on the exact distribution of test values in relation to the diagnostic cut-off threshold. As such, one cannot draw conclusions regarding the level of clinical performance achieved with a given test based on measurement uncertainty data alone: data on the underlying true distribution of measurand values to which the measurement uncertainty is applied, and the position of any clinical decision thresholds, is also required.

In contrast to above, the approach taken in the MSAC (2001) HTA (which evaluated a cholesterol screening POCT for coronary heart disease) correctly accounted for this relationship. In this study, “true” test values were first assigned (based on a distribution observed in a published national survey), and ‘measured’ test values (i.e. including measurement uncertainty) were simulated assuming a 95% confidence interval (CI) of +/- 8% around the “true” test value (i.e. TE = 8%). For each “true” cholesterol level, the simulation was repeated 10,000 times and the probability of an incorrect classification at each cholesterol level (and subsequently across the total population distribution) was determined. The essential advantage of this approach is that, by first sampling baseline “true” values and subsequently simulating error on top of these values, one can determine the proportion of test values incorrectly pushed above (or below) the test’s diagnostic cut-off threshold, and thereby calculate clinical performance for a given level of measurement uncertainty. Nevertheless, there are limitations with the approach taken in this study with respect to the data used to inform “true” test values (which will likely be subject to high baseline levels of measurement uncertainty); and the use of TE to inform CIs (since bias would be expected to act in one direction only).

The MSAC HTA is of further interest, due to the fact that it was the only study identified which explored the impact of increasing measurement uncertainty (in this case, in the form of TE) on cost-effectiveness (i.e. rather than simply

accounting for a baseline level of measurement uncertainty as in the other economic evaluations). Here the authors found that, whilst variation in TE was not expected to alter the overall decision uncertainty (since all results remained above the specified AUS$100,000 cost-effectiveness threshold), it was expected to have a significant impact on the base case results (resulting in a 24% drop from $133,934 to $101,419 per life year gained when reducing TE from 8% to 0%). This example therefore clearly illustrates the potential impact that varying measurement uncertainty can have on outcomes.

The final two studies, Farmer *et al*. (2014) and Perera *et al*. (2015), also simulated the addition of uncertainty on top of "true" baseline values; in this case accounting for the impact of uncertainty within repeated testing scenarios, based on regression analysis of longitudinal IPD (93, 94). Although the impact of varying measurement uncertainty was not explored in these analyses (rather a single level of analytical and biological variation was applied to the underlying "true" test values), it is possible that a similar approach to that taken in the previous MSAC HTA could also be applied within evaluations of repeated test strategies or monitoring scenarios. Whilst such analyses would likely impart a higher computational burden (since iterative simulations would need to be run over a series of test values), this is increasingly feasible with the availability of high level performance computing.

2.4.2 Limitations

The scope of this review was limited to HTA reports including an economic decision model. It is expected that additional findings of interested may have been retrieved if a broader perspective had been adopted, for example considering: (a) any form of HTA, with or without an economic evaluation and including both within-trial and model-based economic analyses; and/or (b) other forms of evidence which could inform healthcare decision making, such as stand-alone literature reviews, clinical trials and cost-effective analyses. The aim of this review, however, was to determine how measurement uncertainty has been considered at the test adoption and reimbursement decision point. The remit was therefore limited to HTAs so as to focus on gold standard technology assessments most likely to have directly informed technology adoption decisions. The further restriction to studies including an economic evaluation was taken so

as to provide a pragmatic and overarching review of both clinical and economic HTA components; whilst the exclusion of trial-based economic analyses was based on the reasoning that direct (i.e. trial-based) assessments of the impact of measurement uncertainty on outcomes are expected to be extremely rare in light of pragmatic and ethical concerns associated with such analyses (previously highlighted in section 1.2.5).

The identification of HTA reports within this review was based primarily on a search of the CRD HTA database. There are two key limitations to note with respect to this database. First, although the CRD HTA database is the only database currently available which provides a collation of international HTA reports, it does not necessarily provide a complete account of HTA activities. It can be seen from the search results, for example, that few HTA reports were identified from the USA, which does not appear to reflect the number of economic evaluations undertaken there (in particular by key authorities such as the Agency for Healthcare Research and Quality (AHRQ)). It is expected that relevant activities from the AHRQ in particular have been excluded from this review due to them being published as a series of separate analyses (i.e. a systematic review report followed by a separate economic evaluation report) as opposed to a single unified HTA report. This reiterates the previous assertion that broadening the scope to include individual literature reviews and economic evaluations, may have identified further relevant findings. In particular, it should be noted that one of the only HTA test evaluation frameworks produced to date that has specified the need to review evidence on analytical performance (the EGAPP framework), was produced from the USA (108); broadening the scope of the review to ensure USA activities were captured, may therefore have identified further studies evaluating test analytical performance. Nevertheless, the EGAPP framework does not address the question of how to assess the potential impact of measurement uncertainty on clinical performance or cost-effectiveness outcomes – rather it provides piecewise guidance on appropriate methods for evaluating the quality of evidence on analytical performance, clinical performance and cost-effectiveness *separately*. It is likely therefore that any assessments of measurement uncertainty triggered by this guidance will have been limited to the

type of pre-model assessments identified in this study (i.e. literature reviews), rather than any novel model-based assessments.

A second potential limitation with the CRD HTA database concerns the fact that HTA reports uploaded onto this database do not undergo any process of quality assessment or critical appraisal, and no attempt was made to conduct such an assessment in this study. Nevertheless, all of the items included on the CRD database were HTA reports conducted by INAHTA members and other recognized HTA organizations: as such, this database represents a principle resource for international HTAs expected to directly influence regional and national healthcare decisions, and should reflect best practice test evaluation methodologies. In addition, since the goal of this review was not to inform clinical guidelines, but rather to assess the state-of-play in terms of HTA methodology, the quality of the included studies (albeit expected to be high in general) is of secondary relevance.

A further limitation with this review concerns the fact that not all screening and data extraction was independently checked by a second reviewer. Instead, a pragmatic review process was adopted, wherein all initial screening and data extraction was conducted by the primary reviewer and a subset of 10% of records in each case was independently checked by the secondary reviewer. Note that with respect to the abstract screening, all disagreements between the two reviewers resulted from the primary reviewer being more cautious (i.e. inclusive) than the secondary reviewer, with all of the additional abstract inclusions being excluded upon full text reviewing. In addition, none of the disagreements in this case were considered to require further clarification of the inclusion criteria, and therefore no further abstract checking was deemed necessary[23]. Similarly for the data extraction check, no disagreements were identified and the 10% check was therefore deemed sufficient. Whilst this approach should have ensured that issues within the review process were identified and corrected, it is possible that some screening and extraction errors may have gone undetected.

[23] Had any disagreements resulted in an amendment of the inclusion criteria, a further 10% check of abstract screening would have been conducted.

A final limitation with this review concerns the fact that, at the time of writing this thesis (March 2020), the searches undertaken for this review were three years old. It is likely that several studies published since March 2017 will have incorporated elements of measurement uncertainty and would therefore be of interest. However, given that HTA guidelines have remained largely unchanged since 2017, it is not expected that there will have been any meaningful shift in the methods used to assess measurement uncertainty in this context, and the results of this study should therefore remain valid.

2.5 Summary

- This study has verified the introductory claim that measurement uncertainty has not, to date, been routinely considered within HTAs of in-vitro tests (i.e. the findings support hypothesis A).
- In the minority of identified HTAs that did include measurement uncertainty, most consisted of a narrative review of the measurement literature in which the potential influence of measurement uncertainty on outcomes was not considered.
- Five of the identified HTAs included measurement uncertainty within the economic model itself; however of those, only one study explored the impact of increasing measurement uncertainty on cost-effectiveness outcomes. Whilst a potentially useful simulation approach was identified from this study, little can be concluded on the basis of a single example.

Given the paucity of applications identified in this review, a methodology review – reported in Chapter 3 – was conducted to explore methods used in the broader literature to assess the impact of measurement uncertainty on outcomes.

Chapter 3
Indirect methods for evaluating the impact of test measurement uncertainty on clinical and economic outcomes: a methodology review

3.1 Chapter outline

In Chapter 2 a systematic review was conducted to identify methods applied within HTAs to evaluate test measurement uncertainty. The findings confirmed the hypothesis that the impact of measurement uncertainty on clinical and health-economic outcomes has rarely been considered within HTAs to date. It is likely, however, that this topic has been considered elsewhere – in particular within the laboratory sciences community within the context of deriving outcome-based APS (see Chapter 1, section 1.2.5). The aim of this chapter, therefore, was to conduct a methodology review to identify studies using indirect methods (i.e. excluding purely empirical, clinical-trial-style analyses) to assess the impact of measurement uncertainty on downstream clinical and health-economic outcomes. This study addressed hypothesis B of the thesis: that methods for assessing the impact of measurement uncertainty on outcomes have been used in the broader literature (e.g. in laboratory medicine studies). The review methods are first outlined in section 3.2, followed by the study findings (section 3.3), discussion (section 3.4) and summary (section 3.5).

The work presented in this chapter has also been published as part of a jointly-authored peer-reviewed publication in *Clinical Chemistry* (Smith AF *et al*. [2019]) (2).

3.2 Methods

A methodology review was conducted to identify indirect studies assessing the impact of test measurement uncertainty on clinical and economic outcomes. This review addresses hypothesis B of the thesis: that methods for assessing the impact of measurement uncertainty on outcomes have been used in the broader literature (e.g. in laboratory medicine studies) (see section 1.5.3). The restriction to *indirect* methods of assessment (i.e. excluding purely empirical analyses, such as clinical trial-based assessments) was set for two primary reasons. First, from

a pragmatic standpoint, direct assessments are expected to be rare due primarily to ethical barriers associated with this form of analysis. The deliberate exploration of measurement errors on downstream clinical outcomes, for example, would not be expected to pass current ethical review standards. Second, in the context of exploring methods for use in HTA-style assessments, it is expected that simulation-based approaches to evaluating the impact of measurement uncertainty on outcomes would be easier to integrate within model-based health-economic evaluations, as frequently used within HTAs.

The review consisted of two key components: (i) a central database search, which aimed to identify contemporary methods of analysis published in the last 10 years; and (ii) extensive citation tracking of included studies, published on any date, to identify key seminal papers informing modern practices. The central database search was conducted in November 2017 across four databases: Embase, Ovid Medline(R), Web of Science (core collection) and Biosis Citation Index. Based on the advice of an experienced information specialist, these databases were chosen to provide coverage of a wide cross-section of clinical and laboratory journals expected to be the primary contributors of relevant studies. The database searches focused on identifying relevant material across these four databases over a pragmatic 10-year period: from January 2008 to November 2017. A subsequent update of the searches was conducted in March 2019 (covering the period January 2008 to March 2019). The central 10-year search was then supplemented with extensive citation checking to ensure that any key methods missed in the database searches would be identified. This included: (a) backwards citation checking (in which the bibliographies of included studies were electronically checked) and (b) forwards citation checking (in which subsequent studies referencing the included studies were electronically checked).

The database search strategies (provided in Appendix G) were developed via consultation with expert information specialists. The searches combined key terms relating to: (i) in-vitro tests; (ii) measurement uncertainty and related

performance metrics (including biological variation and quantification limits)[24]; and (iii) simulation/ methodology identifiers. All identified records were managed using Endnote V 7.2 (Thompson Reuters).

Included studies were required to meet the criteria listed in Table 3-1. In particular, studies had to incorporate or evaluate the impact of test measurement uncertainty on downstream outcomes (including clinical performance, clinical utility, costs and/or cost-effectiveness) using indirect methods of assessment (i.e. excluding purely empirical-based analyses, such as RCTs). Note that studies using indirect methods *at any stage* of the analysis were eligible for inclusion. This means that several method-comparison studies (an essentially empirical study design, in which the agreement between index and reference test measurements is assessed) were included when an indirect method was applied to assess the impact of identified measurement discrepancies on one of the listed outcomes (such as in error grid analyses, discussed in section 3.3.3). The following studies were excluded from the review: animal (i.e. non-human) studies; studies not evaluating an in-vitro test or device (e.g. pharmacological studies); studies evaluating non-clinical or non-cost outcomes; studies conducting a *direct* assessment of the impact of measurement uncertainty on outcomes (i.e. clinical trial-based analyses); studies not reporting an original analysis (e.g. reviews and editorials); and non-English language studies.

A two-stage screening process – consisting of initial title/abstract screening, followed by full-text screening – was conducted by the primary reviewer (Smith AF). Uncertainties regarding final inclusions were resolved via discussion with the study secondary reviewers (Shinkins B, Messenger MP, Hulme CT and Hall PS). A data extraction table (including details relating to the study and test characteristics, outcomes and method of assessment) was constructed and piloted on the first 10% of included studies. Subsequent full data extraction of included studies was conducted by the primary reviewer, with each study double-checked by one of the four secondary reviewers. Any disagreements with regards

[24] Note that these related measurement performance metrics were included in the search strategy as possible identifiers for studies including measurement uncertainty. To be included in the review, studies had to include assessment of measurement uncertainty as defined in Table 3-1.

to data extraction were resolved via group consensus. The findings of the review were narratively synthesised. No formal quality assessment of identified methods was conducted, due to the fact that no relevant quality appraisal checklists currently exist in this setting.

Table 3-1. Methodology review: inclusion criteria

Population	Any human population with any indication
Intervention	In-vitro test or device (including medical laboratory tests and POCTs; excluding imaging) used for the purpose of screening, diagnosis, prognosis, monitoring or predicting treatment response
Comparator	Any
Outcomes	(a) Clinical performance e.g.: - Diagnostic sensitivity and/or specificity - Positive/negative predictive values - ROC curve/ AUC analysis - 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 - 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 date
ROC = Receiver operator characteristic; AUC = Area under the curve	

3.3 Results

3.3.1 Study characteristics

From the initial database searches 1709 citations were retrieved, followed by a further 362 citations from the review update. After conducting the two-stage screening process, 26 studies were included from the database searches. A further 56 studies (25 of which were published prior to 2008) were identified via citation checking, resulting in a total of 82 included studies (see Figure 3-1). At the data extraction checking stage, 35 papers (43%) were checked by Shinkins B; 16 (20%) by Hulme CT; 16 (20%) by Messenger MP; and 15 (18%) by Hall PS. Agreement between reviewers across extraction items was >99%[25].

A summary of the included study characteristics is provided in Table 3-2, and details of the measurement uncertainty components included and outcomes evaluated are provided in Table 3-3. The majority of studies focused on evaluating technologies used for the purposes of monitoring (n=44, 54%), diagnosis (n=24, 29%) and/or screening (n=11, 13%). Imprecision was most commonly addressed (n=50, 61%), followed by bias (n=39, 48%) and total error (n=26, 32%), and studies primarily evaluated clinical performance outcomes (n=45, 55%).

[25] This statistic is calculated based on the fact that there were 10 data items listed in the data extraction table, across 82 papers, giving a total of 820 data extraction items. Seven disagreements were identified within the data extraction check, giving an agreement rate of >99% (813/820).

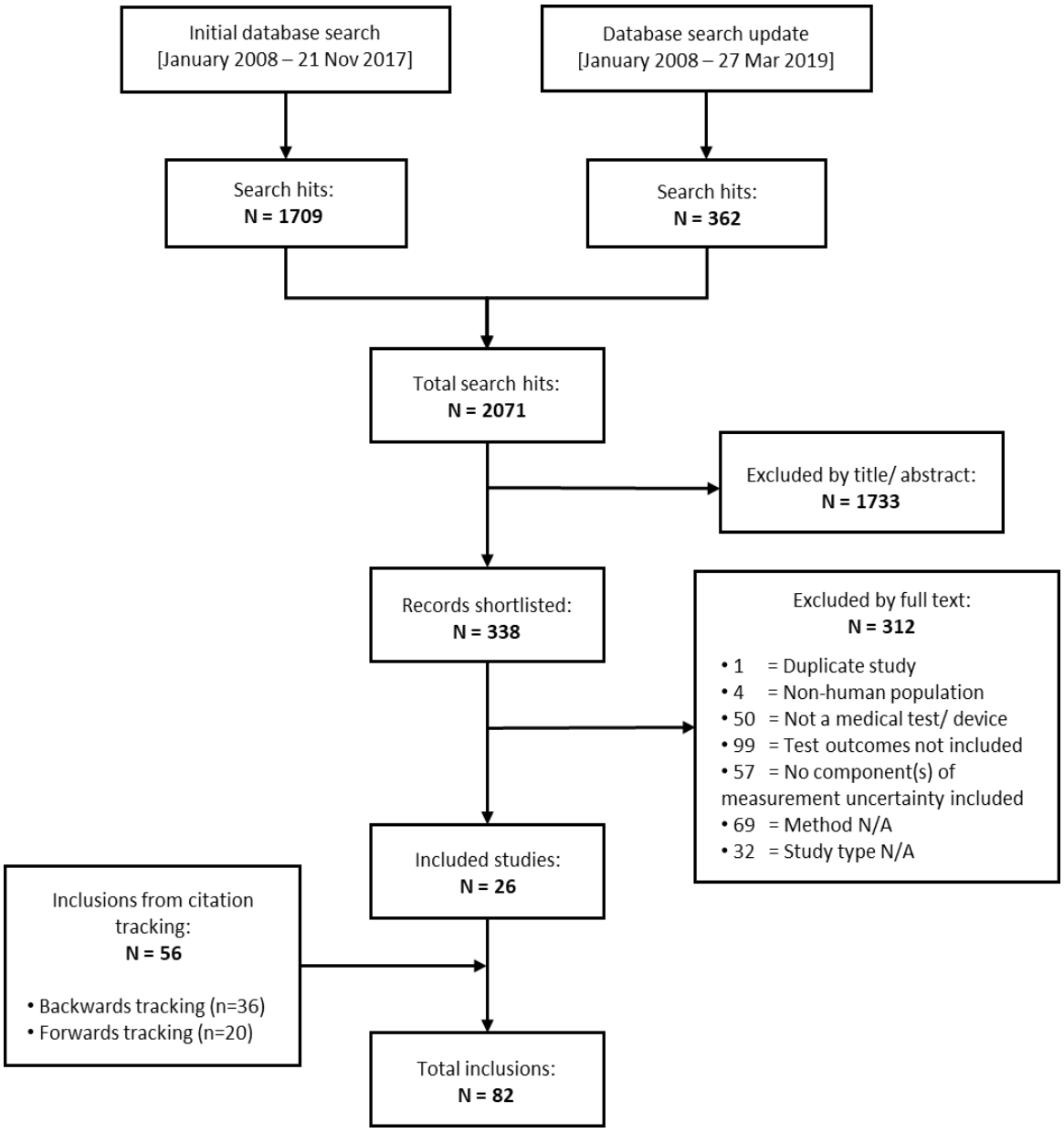


Figure 3-1. Methodology review: PRISMA diagram

Table 3-2. Methodology review: study characteristics

	Number	% (out of 82)
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 & glycaemic 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 3-3. Methodology review: components of measurement uncertainty included and outcomes assessed

	Number	% (out of 82)
Component(s) of measurement uncertainty included[a]		
Imprecision:		
Analytical[b]	31	38%
Pre-analytical[c] / combined pre-analytical and analytical	8	10%
Non-specific[d]	11	13%
Total	**50**	**61%**
Bias:		
Analytical[b]	18	22%
Calibration bias[e]	9	11%
Non-specific[d]	9	11%
Pre-analytical[c] / combined pre-analytical and analytical	2	2%
Between-method bias[f]	1	1%
Total	**39**	**48%**
Total error:		
Method-comparison study[g]	18	22%
EQA study[h]	2	2%
Other	6	7%
Total	**26**	**32%**
Biological variation[i] included?		
Yes - included as a separate element	13	16%
Yes - combined with imprecision	5	6%
Total	**18**	**22%**
Primary test outcome assessed[a]		
Clinical performance	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%).

[b] *Analytical* bias or imprecision, relates to bias or imprecision occurring at the point of sample analysis (see section 1.2.2 and Appendix B.2).
[c] *Pre-analytical* bias or imprecision, relates to bias or imprecision occurring prior to the point of sample analysis (see section 1.2.2 and Appendix B.2).
[d] *Non-specific* bias or imprecision, relates to bias or imprecision which has not been clearly specified as either analytical or pre-analytical in the associated study.
[e] *Calibration bias* relates to bias resulting from the process of assay calibration (see section 1.2.2 and Appendix A).
[f] *Between-method bias* relates to systematic differences in measurement resulting from the use of different assay methods (e.g. different manufacturer assays or test platforms), typically estimated using a method comparison study (see below).
[g] *Method comparison study* refers to any study which aims to determine if two methods for measuring the same measurand are equivalent. Typically the assessment of equivalence is based on the comparison of paired measurements (e.g. split samples analysed using two different assays of interest), using statistical methods of analysis such as regression analysis or Bland-Altman plots (see Appendix B.1) to determine the level of between-method bias and variability.
[h] *EQA (External Quality Assessment) study* refers to any study conducted as part of a regional, national or international EQA scheme, wherein the aim is to determine between-laboratory and/or between-method comparability using samples of external origin (i.e. samples distributed to each participating laboratory, as opposed to patient samples).

3.3.2 Aim of analyses

Most of the identified evaluations were conducted with the objective of either: (i) determining or informing APS (116-134); (ii) exploring the impact of uncertainty allowed by *current* APS (135-146); or (iii) evaluating the potential impact of measurement uncertainty on outcomes (without explicitly defining or mentioning APS) (147-190). 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 (191-197).

3.3.3 Methodology framework

A data extraction summary table, detailing the methods used within each individual study identified from the methodology review, is provided within supplemental table accompanying the publication of this study (2). Although there was variation across the included studies in terms of specific methods adopted, a common analytical framework underpinning the various approaches was identified. This framework consists of three fundamental steps:

1. calculation of the “*true*” test values;
2. calculation of the *measured* test values (i.e. incorporating measurement uncertainty); and
3. calculation of the *impact* of differences between (1) and (2) on the outcome(s) under consideration.

A high-level summary of the methods adopted within this analytical framework across the range of studies identified is provided in Figure 3-2 below, followed by a detailed narrative review.

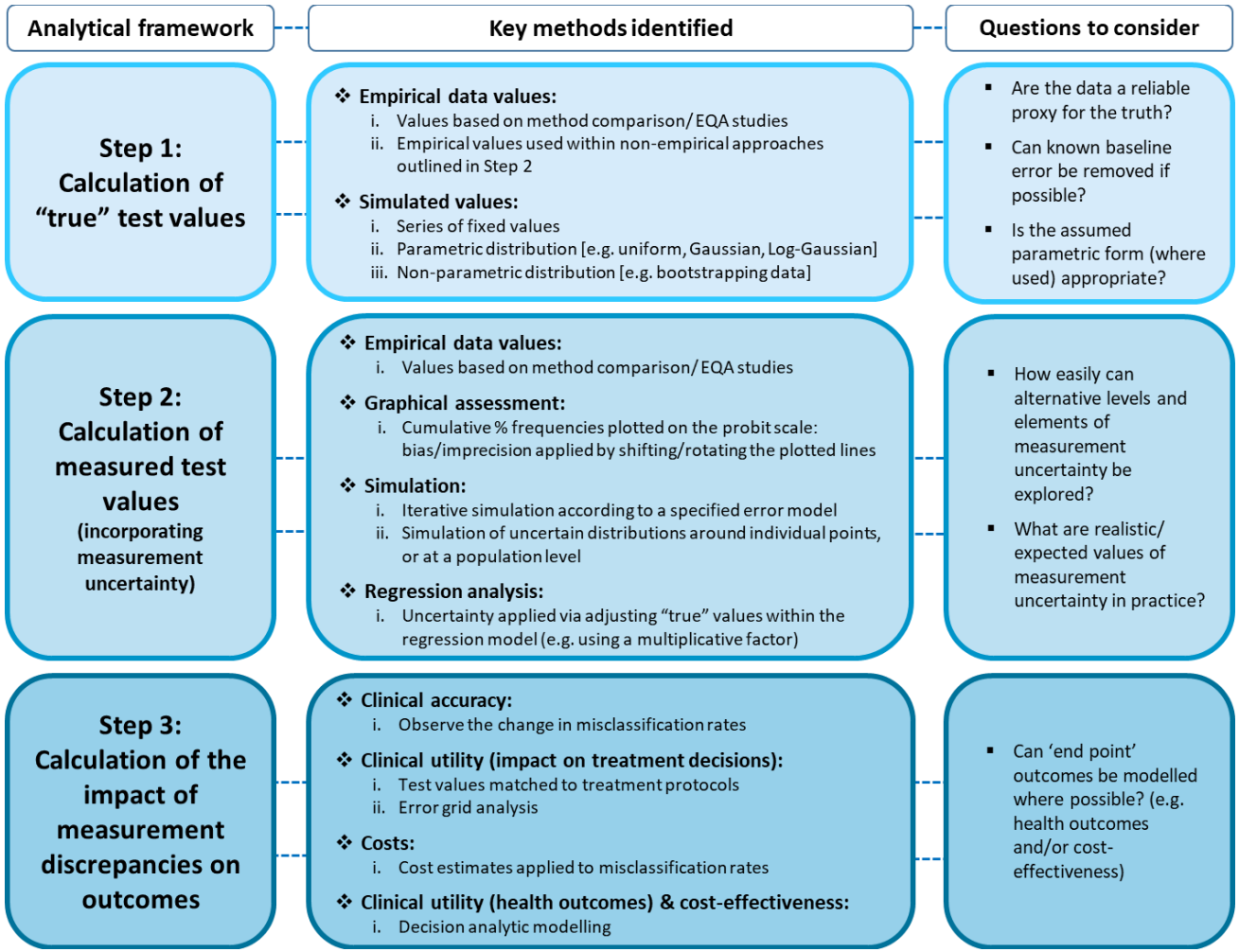


Figure 3-2. Methodology review: summary of the three-step analytical framework

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

The first step within the framework is to calculate "true" measurand values for the test under assessment. Across the included studies, "true" values were based either on *empirical* and/or *simulated* data.

Studies using empirical data here included method comparison and EQA studies, which based both the "true" *and* measured values on empirical data, but used an indirect method to determine the impact of between-method measurement discrepancies on the specified outcome in Step 3 (for example using the "error grid" approach, outlined in Step 3) (147, 149, 153, 154, 163, 165, 168-170, 172, 176, 178-181, 183, 187, 190). Alternatively, several studies based "true" test values on empirical data, but used various non-empirical approaches to derive measured test values in Step 2 (including studies using the graphical, simulation or regression-based methods, outlined in Step 2) (117, 119, 121-123, 130, 133, 138, 142-144, 146, 148, 151, 152, 157, 160-162, 164, 173, 189, 197).

Of those studies using simulation methods, the simplest approach was to assume a *fixed set*[26] of individual "true" values specified along the measurement range, with uncertainty simulated around these individual points in Step 2 (124, 128, 139, 145, 148, 150, 191, 195, 196). A handful of studies instead simulated "true" values from a specified parametric distribution for a given population: most frequently using the uniform distribution (126, 129, 131, 156, 167); or the *Gaussian* or *log-Gaussian* distribution (116-118, 120, 125-127, 132, 158, 159, 171, 175, 177). Other parameterisations included mixed Gaussian distributions (166, 174), multivariate Gaussian distributions (where data on between-test correlations was available) (155), and the exponential distribution (194). Non-parametric simulation approaches were also used, whereby test values were randomly sampled with replacement from an empirical dataset (i.e. bootstrap sampling) (130, 142). A final set of studies appeared to base "true" values on some form of simulation (or findings from a previous simulation study), but reported incomplete details of the data or method used (134-137, 182, 185-188).

[26] 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.

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

Approaches to the calculation of measured test values predominantly fell into four categories: empirical assessment, graphical assessment, simulation or regression analysis.

Studies using empirical assessment here included method-comparison studies (147, 149, 153, 154, 165, 168-170, 172, 176, 178-181, 183, 187, 190) and an EQA study (163) which based “true” test values on the specified reference test and measured values on the index test measurements, as previously discussed in Step 1.

Several of the identified studies, dating back to 1980, used a graphical method of assessment *(117, 119, 121-123, 148)*. This approach centres on plotting the cumulative percentage frequency of “true” test values on the probit scale (y-axis) as a function of “true” values on the logarithmic scale (x-axis). Assuming that the log-transformed data are normally distributed, then in the bimodal case (where healthy and diseased populations are modelled 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 bias is then explored by shifting the straight lines along the x-axis; whilst the addition of imprecision is explored by rotating the straight lines around their mean value (i.e. broadening the 95% CI of the values on the probit scale). Given a specified diagnostic cut-off threshold, the proportion of false positives and negatives relating to a given level of bias and imprecision can then be derived 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 since been superseded by computer simulation approaches which can better accommodate complex specifications of the measurand distribution and measurement uncertainty. The most flexible and widely adopted approach in the identified literature is based on a process of 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, incorporating specified components of measurement uncertainty (126, 129-131, 140-142, 146, 166, 174,

191, 194-196). This method is largely attributed to a seminal paper published in 2001 by Boyd and Bruns (126), which was the first study of this kind to clearly specify the error model as a mathematical function (as opposed to earlier (116-118) and later (133-137, 156, 161, 164, 182, 184, 185, 188, 189, 192, 193, 197) studies limited to textual descriptions or indirect referencing).

The most widely adopted error model in the identified literature specifies imprecision as a normally distributed term and bias as a fixed absolute value, as shown in Equation 3.1:

$$\mathbf{Test}_{sim} = \mathbf{Test}_{true} + [\,\mathbf{Test}_{true} * \mathbf{N(0,1)} * \mathbf{CV}\,] + \mathbf{Bias} \qquad \mathbf{(3.1)}$$

where $Test_{true}$ is the "true" measurement value; $Test_{sim}$ is the measured test value incorporating the specified level of imprecision (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 $Test_{true}$. Figure 3-3 provides an example of a generic error model simulation approach, which may be applied to the case of a single-test continuous diagnostic biomarker – that is, where a single test is used to distinguish between healthy and diseased populations on a continuous biomarker scale. Where data is available on the longitudinal trajectory of test values, the error model approach can also be applied to evaluate repeat-test or monitoring scenarios, via repeated application of the error model to the series of "true" test values in question. In either case, the iterative simulation process can be efficiently implemented using standard statistical software, such as Excel or R.

i. A sample of $Test_{true}$ values is assigned for the healthy and diseased populations (e.g. based on empirical or simulated data)

ii. For each $Test_{true}$ value assigned in step (i), a corresponding $Test_{sim}$ value is derived according to a specified error model, for a given level of Bias and CV e.g.:

$$\mathbf{Test_{sim} = Test_{true} + [Test_{true} \times \mathit{N}(0,1) \times CV] + Bias} \quad \mathbf{(3.2)}$$

iii. The diagnostic accuracy of the simulated data is calculated according to the proportion of $Test_{sim}$ values for the healthy and diseased populations falling the correct side of the diagnostic cut-off threshold

iv. Steps (ii) to (iii) are repeated for a range of CV and bias values (e.g. CV% ranging from 0-20% and Bias ranging from +/-10% in 1% increments)

Figure 3-3. Methodology review: error model simulation approach for a single-test diagnostic strategy

Rather than iteratively applying uncertainty via the error model simulation, as in Figure 3-3, an alternative approach is to incorporate uncertainty directly within a specified probability distribution. Several studies, for example, applied uncertain distributions around individual "true" values selected along the measurement range (124, 128, 130, 139, 142, 150, 158, 171, 173); whilst others applied added bias and/or imprecision to population-level distributions assumed to be representative of "true" measurement (120, 127, 143, 175, 177). In the same way as with the error model simulation, these distributional simulations can be iteratively run applying varying levels of bias and/or imprecision, to establish how the clinical performance changes in line with increasing or decreasing measurement uncertainty.

Of the remaining studies, a handful utilised regression analyses *(138, 144, 155, 159)*. Within these assessments, bias or TE was applied as a multiplicative factor to baseline "true" measurements within a specified regression model, with the resulting impact on the regression output (e.g. likelihood ratio) subsequently determined. The final set of studies used other one-off methods (125, 145, 152, 157, 160), or reported insufficient details of the simulation technique to identify

the exact method used (186, 187). Details of these individual studies can be found in the published data extraction summary table (2).

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

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

Most studies focused on evaluating clinical performance (116-125, 127, 128, 132, 138-141, 143-145, 150, 151, 155, 157-164, 167, 171, 173-175, 177, 191-197). In this case the calculation is generally straightforward. For a diagnostic test, for example, as long as each individual patient's clinical diagnosis is known (e.g. "true" test values have been parametrically sampled for diseased vs. healthy patients separately; or drawn from an empirical dataset where the confirmed clinical diagnosis for each patient is known), then the diagnostic accuracy of the simulated test values can be calculated in the usual way (i.e. by comparing the diagnoses based on the measured test values with patients' true clinical diagnoses). 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 (126, 130, 133, 142, 147, 149, 153, 154, 163, 165, 168-170, 172, 176, 178-181, 183, 186, 187, 190). Most of these were method-comparison studies which determined the impact of measurement deviations on treatment decisions using *error grid analysis* (147, 149, 153, 154, 165, 168-170, 172, 176, 178-181, 183, 186, 187, 190). First developed in the 1980s, the original error grid aimed to evaluate the impact of measurement discrepancies between self-monitoring blood glucose devices and laboratory reference measurements in terms of insulin dosing errors (147). Using a scatter plot of reference vs. index test measurements, the plotted region was divided into five error grid "zones", based on expert consensus on the assumed severity of dosing errors resulting from the measurement discrepancies. These error zones spanned from zone A, depicting clinically accurate results; to zone E, depicting erroneous results expected to lead to a dangerous (potentially life-threatening) failure to detect and treat. Error grid analysis remains common today, with recent studies further developing the method – for example by expanding on the small sample of

experts used within the original error grid study (149, 186, 187); accounting for temporal aspects of measurement (153); or applying the same methodology to alternative clinical settings (176).

Others have attempted to incorporate the impact of measurement uncertainty on patient health outcomes *(129, 131, 134, 135, 156, 166, 182, 184)*. All of these studies related to evaluations of monitoring devices for glycaemic control, in which health outcomes such as hypoglycaemia and hyperglycaemia were determined using decision analytic models based on 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, these models were used to track patients' response to administered doses and resulting health outcomes.

A final group of studies included an assessment of costs or cost-effectiveness (119, 120, 123, 136, 137, 152, 185, 188, 189). Of these, one study focused solely on costs, with the aim of exploring the potential financial implications of calibration bias in serum calcium testing for the diagnosis of hypercalcemia (152). This assessment centred around an estimated 'cost curve', which related baseline serum calcium measurements to expected 12-month hospital follow-up costs. This curve was constructed using data on the population frequency distribution of calcium (based on a laboratory clinical dataset), linked with data on subsequent tests and procedures associated with hypercalcaemia (determined via regression analysis), and published costs for each of the included activities. Using the constructed cost curve, the impact of bias was explored by shifting the curve to reflect the associated shift in observed values that would result from a given magnitude of error, and reading off the difference in the annual follow-up costs.

The remaining eight studies considered cost outcomes alongside clinical outcomes. Of these, half were based on a simple assignment of expected costs of misdiagnoses to rates of false positive/negative results (119, 120, 123), or expected costs of adverse events applied to simulated health outcomes data (189). The other half all utilised findings from a previous study by Breton and Kovatchev (2010), in which the impact of reduced glucose meter imprecision on glycaemic events for patients with Type 1 diabetes was simulated using a

published simulation platform (135)[27]. Using these results, two studies constructed simple cost-consequence decision models, combining the data on reduced glycaemic events with data on patient population numbers, glucose meter costs, and the rate of myocardial infarctions resulting from glycaemic outcomes, to estimate annual cost savings associated with improved meter precision (185, 188). More recently, the two remaining studies conducted full cost-utility analyses. These used cohort Markov (i.e. state-transition) models to link the data on improved glycaemic control and reduced glycaemic event rates, with data on diabetes complication rates, patient health-related quality of life and health service costs (136, 137). Using these models the authors were able to estimate ICERs for the incremental cost per additional QALY associated with reduced device error. The most recent of these studies, for example, found that when accounting for hypoglycaemic events, a self-monitoring blood glucose device with an imprecision of 8.4% was cost-saving and more effective compared to a device with 15% imprecision, from an English NHS perspective (136). These results were similar to the earlier study from the same authors, in which a comparable assessment was conducted from a Canadian health care payer perspective (137).

[27] The methods used within the original Breton and Kovatchev (2010) study were only partially reported. Based on the details provided in this study, different levels of imprecision appear to have been applied to baseline glucose values using the error model simulation approach, applied within an existing simulation platform for glucose and insulin metabolism.

3.4 Discussion

3.4.1 Review findings

Based on the methodology review findings, a three-step analytical framework for determining the impact of measurement uncertainty on outcomes was identified (see Figure 3-2). 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 test values along the entire test measurement range, population-level conclusions (such as APS) may be derived. In contrast, the primary drawback of the fixed-values approach (i.e. taking a selection of fixed values along the measurement range), and by extension the uniform distribution approach (assuming this is not a realistic parameterisation), 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.

For the majority of studies which assigned an informative sample of “true” test values in Step 1 (i.e. as opposed to the fixed-values approach), a key issue for consideration relates to how well the data may be considered to be a reliable proxy for the truth. If values used to inform the “true” distributions are themselves subject to measurement uncertainty, then all subsequent analyses may be affected by this confounding factor. If this is the case, then care should be taken when asserting absolute maximum bounds for imprecision and bias. In general, the likelihood that the adopted “true” test values would in fact be representative of the truth was either implicitly assumed or not discussed within the identified studies. A handful of authors did attempt to address this issue, by “stripping” known analytical variation from estimates of total imprecision via statistical adjustment to isolate the “pure biologic distribution” (119-122, 125, 127, 143). An example of this method is provided in Appendix H. This approach, however, depends on having reliable information on the measurement uncertainty

contained in the baseline “true” measurement values, which in many cases may not be available.

Within Step 2 (calculation of *measured* test values) computer simulation methods appear to offer the most flexible approach for exploring alternative specifications and levels of measurement uncertainty. Studies based on method-comparison analyses are of limited utility in this context, given that alternative levels of measurement uncertainty cannot be efficiently explored via this method; similarly, analyses using the graphical method suffer from the issue that non-Gaussian parameterisations or non-constant/ non-linear specifications of imprecision or bias cannot be accommodated. The error model simulation approach was found to be particularly useful in this respect. While the formula provided in Equation 3.1 specifies one CV% element representing total imprecision, separate components of imprecision (e.g. pre-analytical and analytical) may be specified, and alternative characterisations of imprecision may be defined (e.g. using a fixed SD; different SD/CV values for different sections of the measurement range; or imprecision defined as a linear/ non-linear function of $Test_{true}$). In the same way, bias may be characterised in numerous ways.

The majority of identified error model studies applied bias and imprecision as separate elements within the simulation. An alternative method is to apply an aggregate metric of measurement uncertainty: this was the approach taken in the MSAC (2001) HTA highlighted in Chapter 2 for example. In this case, measured test values were derived by applying a 95% CI around the sampled “true” test values, based on an estimate of TE, and simulating values from this region (85). Although no explicit error model was reported in the MSAC HTA, this approach relies on the same concept of iteratively applying varying levels of measurement uncertainty onto baseline “true” test values. Nevertheless, it is arguably preferable to delineate between systematic and random components of measurement uncertainty within the simulation, since each can have a markedly different impact on clinical performance. Separately specifying these elements (e.g. as in Equation 3.1) is the only way that such variable impacts can be appropriately identified and explored.

When imprecision and bias are separately specified, an effective and useful means of illustrating the simulation results is via presentation on a *contour plot*,

as in several of the identified studies (126-131, 133, 142, 146, 174). An illustrative example is provided in Figure 3-4. This presents a hypothetical case in which bias and imprecision have been iteratively applied (according to the error model provided in equation (3.1)) to “true” test values randomly drawn from normally distributed healthy [N(30,5)] and diseased [N(60,10)] populations, with a diagnostic cut-off threshold of 45 and exploring a range of hypothetical bias (-40 to +40, in 1 unit increments) and imprecision values (0 to 80%, in 1% increments). Panel A presents the contour plot for diagnostic sensitivity; Panel B presents a similar plot but for diagnostic specificity; and Panel C presents a joint plot of both outcomes together.

For each panel presented in Figure 3-4, the presented contour lines illustrate for what values of bias and imprecision the given level (i.e. ‘contour’) of diagnostic sensitivity/specificity is maintained. By inspecting these plots, one can observe how increasing measurement uncertainty affects each outcome. For example in panel A, holding imprecision at 0% and applying negative bias (i.e. moving horizontally to the left of the (0,0) point) leads to decreased diagnostic sensitivity, whilst applying positive bias has the opposite (albeit marginal) effect; and holding bias at 0 and increasing imprecision (i.e. moving vertically upwards from the (0,0) point) has less of a marked effect, resulting in a gradual reduction in sensitivity. Inspection of panel B meanwhile indicates that positive bias decreases diagnostic specificity and negative bias increases specificity; whilst increasing imprecision again has a limited impact. These plots are therefore useful as a means of exploring the impact of measurement uncertainty on outcomes (i.e. relating to hypothesis C; see section 1.5.3). In addition, it is also possible that these plots could be utilised to enable outcome-based APS to be extracted (i.e. relating to hypothesis D), by setting a minimum criteria for clinical performance (or whatever outcome is being modelled), and identifying the region of analytical performance on the contour plot which achieves this goal. The use of contour plots for this aim is explored further in Chapter 5 and Chapter 6.

Note that, whilst the focus of these plots is on showing the differential impact of bias and imprecision on the given outcome, the concept of TE can also be incorporated by overlaying TE% ‘bands’ onto the plot. This additional feature was included in a handful of the identified studies, for example (130, 133, 142, 146,

174), and is also explored in the subsequent case study analysis in Chapter 5 and Chapter 6 (see, for example, section 5.3.2.2, Figure 5-5).

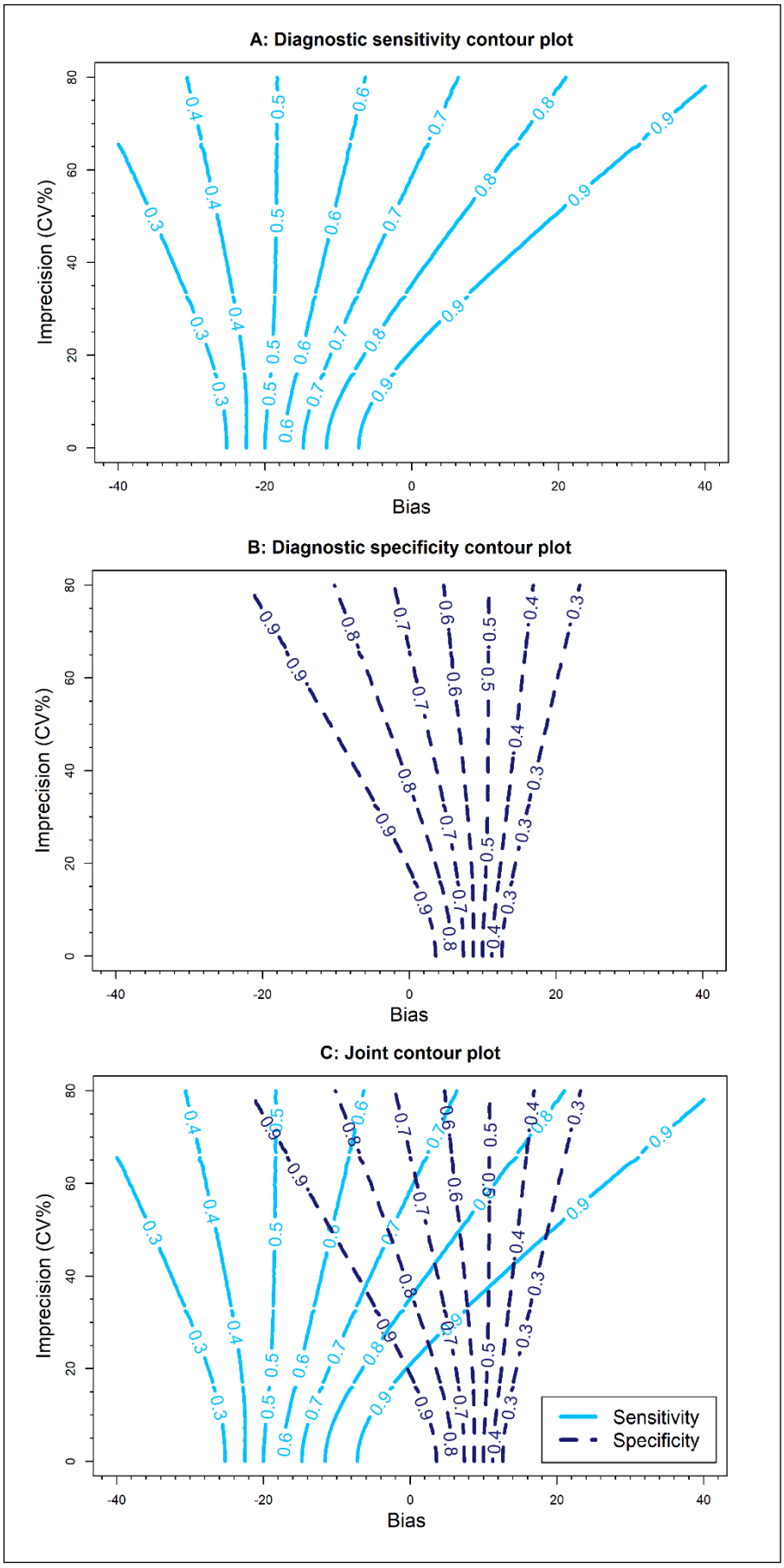


Figure 3-4. Methodology review: example contour plots for diagnostic accuracy

With regards to Step 3 (calculation of the impact of differences between “true” and measured test values on the outcome(s) under consideration), most studies focused on the intermediate outcome of clinical performance. Ideally, however, technologies should be evaluated in terms of their influence on “end-point” outcomes, such as health outcomes and/or cost-effectiveness. Several of the identified studies, for example, used analytic decision models to track the impact of measurement uncertainty (e.g. using error model simulation) on treatment administration and resulting health impacts. While all of these studies related to the context of glycaemic control, decision models can feasibly be used to explore any clinical pathway of interest, subject to data availability. In addition, such models are a key tool used in health-economic evaluations, and as such can be used to model cost-effectiveness outcomes. Of the two studies identified in this review which did evaluate cost-effectiveness, for example, both assessments were based on analytical decision models. Nevertheless, these evaluations were considered to be limited in two key respects: first, they were conducted as separate analyses from the original error model simulation; and second, they only explored a limited set of fixed imprecision levels relating to pre-existing APS (136, 137). Both of these limitations are addressed in the subsequent case study analysis (Chapter 4 to Chapter 7).

3.4.2 Limitations

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. Thus, while the database searches were intended to be as sensitive as possible, they were necessarily focused in certain aspects due to the vast volume of literature in this area. In particular, the database searches were restricted to literature published across four key databases over the past 10 years, and a simulation and methodology filter was included in an attempt to screen out routine measurement performance studies. It is likely that extending the search to include additional databases, a longer search period and a more sensitive and less specific search strategy could have retrieved further relevant material. However, extensive forwards and backwards citation tracking was conducted to ensure that all key methods were identified – including both current state-of-the-art methods and seminal methods informing modern practices. The

primary limitation of the pragmatic methods adopted, therefore, concerns the fact that not all relevant material relating to each individual method will have been identified – as such, no definitive conclusions can be drawn from this study regarding the *frequency* that each method has been used.

An additional limitation with this review concerns the fact that citation screening was conducted solely by the primary reviewer. There is a further risk, therefore, that relevant methods may have been missed during the screening stages. Nevertheless, regular discussion of the review inclusions and exclusions with the secondary reviewers was conducted throughout the screening process. The risk of excluding relevant methods is therefore expected to be small.

3.5 Summary

- This chapter has confirmed hypothesis B: that methods for assessing the impact of measurement uncertainty on outcomes have been used in the broader literature.
- Based on 82 indirect outcome studies, this review identified a three-step analytical framework underpinning the various approaches adopted (summarised in Figure 3-2).
- Within this framework, several key methods were highlighted as having particular utility with respect to facilitating evaluations of the impact of measurement uncertainty on outcomes, using methods which it is expected could be straightforwardly integrated into existing HTA methodology. In particular:
    - Iterative simulation using the *error model simulation approach* offers a flexible and efficient method for exploring the impact of measurement uncertainty on clinical performance;
    - *Decision analytic modelling* provides a useful tool for linking clinical performance outputs to downstream clinical and health-economic outcomes; and
    - *Contour plots* provide a useful means of presenting simulation results and could possibly be used to identify outcome-based APS.

The remainder of this thesis is focused on exploring and developing the above methods within a *de novo* test case study, relating to a diagnostic setting. The clinical context of this case study is first introduced in **Chapter 4,** and **Chapter 5 to Chapter 7** present the case study analyses.

Chapter 4
Case study introduction

4.1 Chapter outline

In Chapter 3 a methodology review was conducted to identify methods for assessing the impact of measurement uncertainty on outcomes. In the following case study, key methods identified from that review are applied within an analysis of faecal calprotectin (FC) as a diagnostic rule-out test for Inflammatory Bowel Disease (IBD) in primary care in the UK.

This chapter provides an introduction to the thesis case study. The clinical context is first outlined (section 4.2), highlighting the key challenge for GPs when faced with patients presenting with non-specific lower gastrointestinal symptoms: how to distinguish between those with serious organic bowel conditions requiring secondary care management, such as IBD, from those with functional conditions that can be routinely managed within primary care, such as Irritable Bowel Syndrome (IBS). An overview of FC in this setting is then provided (section 4.3). In particular, two primary care clinical pathways (explored in the following case study analyses) are introduced: the 'NICE FC pathway' (section 4.3.2) and 'York FC Care Pathway' (YFCCP) (section 4.3.3). A summary of recent studies reporting on the measurement performance of FC is also presented (section 4.3.4), before outlining the case study analysis (section 4.4).

4.2 Clinical context

4.2.1 Inflammatory Bowel Disease (IBD)

IBD is a chronic, lifelong condition which causes periodic inflammation of the gut, and can result in complications including fistulas, abdominal abscesses, malignancy and possible premature mortality (198). The highest rates of IBD are found across Europe and North America, where prevalence exceeds 0.3% (199). In the UK, the annual incidence of IBD is estimated at 18-20 per 100,000 in people aged ≥15 years (200), with the overall prevalence of IBD estimated at 397 per 100,000 people (0.4%) (201, 202). The exact cause of the disease is currently unclear, however unfolding incidence patterns appear to support the argument

that a Western lifestyle, urbanization and/or industrialisation may play a part in the disease aetiology (203-206).

The two most common forms of IBD are Ulcerative Colitis (UC) and Crohn's Disease (CD). Each can affect the gut in different ways: CD may affect any part of the gastrointestinal tract, from the oesophagus to the anus, but is most commonly associated with patchy inflammation of the ileum (the last part of the small intestine) or the colon; UC, meanwhile, causes inflammation and ulceration of the inner lining of the rectum and/or colon (207, 208). UC is the more prevalent form of IBD, affecting ~240 per 100,000 people in the UK, compared to ~157 per 100,000 for CD (201, 202). In total, over 265,000 people in the UK are estimated to be affected by IBD – 160,000 with UC and 105,000 with CD[28].

Symptoms of IBD commonly emerge in patients during their late teens to early 30's, but onset may occur at any age (209). Presenting symptoms typically include: diarrhoea, abdominal pains, tiredness/ fatigue, feeling generally unwell or feverish, loss of appetite and weight loss and anaemia (210). When faced with these symptoms, the key challenge for clinicians is distinguishing IBD cases from other gastrointestinal conditions – in particular, the largely non-specific symptoms associated with IBD are also typical of patients suffering from the more common condition, IBS.

4.2.2 Irritable Bowel Syndrome (IBS)

Estimated to affect over 10% of the population in the UK and worldwide, IBS is the most common gastrointestinal disorder seen by primary and secondary care practitioners (211). Although IBS is associated with many of the same symptoms as IBD, it is a functional bowel disorder, meaning that there is no evidence that IBS symptoms are caused by any underlying inflammation or physical damage within the gut (212). As such, although IBS is associated with decreased quality of life and increased incidence of mental health issues (such as anxiety and depression) (213), it does not harbour the same risks of severe complications as IBD. In most cases therefore, IBS can be safely managed within primary care,

[28] These figures are based on an estimated UK 2019 population of >66.8 million (ONS: Overview of the UK population: November 2018).

whilst IBD requires referral to gastroenterology specialists for definitive diagnosis and treatment management (214-216).

4.2.3 Differentiating between IBD and IBS

Historically, diagnosis of IBD has relied on clinical assessment, together with endoscopic investigations and other imaging modalities (217). Routine blood tests, including C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR) have also been used as markers indicative of inflammation. However these tests are known to have low sensitivity and specificity for IBD, since raised CRP and ESR levels may occur as a result of other non-gastrointestinal conditions, and normal levels of these markers may occur in patients with IBD (218-220). Previously therefore, the only reliable means of diagnosis has been to refer all patients with suspected IBD to secondary care for invasive gastrointestinal investigations.

The similarities between IBS and IBD presentation may result in delayed diagnosis for patients with IBD (221), as well as exposure of patients with IBS to unnecessary and costly secondary care consultations and endoscopic investigations (most commonly colonoscopy) (222). In addition to being generally uncomfortable and, for some patients, painful, colonoscopy is associated with a small risk of bowel bleeding or perforation, and in rare cases, mortality (223). Endoscopy units in the UK are also under increasing pressure to deliver expedited care for patients referred with suspected cancer (224); minimising the volume of unnecessary endoscopic investigations in patients with IBS is therefore vital to ensuring that both patients with IBD and cancer can receive timely diagnoses and access to appropriate treatments.

Given the importance of correctly identifying IBD, there has been significant interest in the development of new biomarkers to help clinicians identify the minority of patients presenting with lower gastrointestinal symptoms who have IBD. Over the past decade, the majority of research has centred on the stool test, FC, which forms the subject of this case study.

4.3 Faecal Calprotectin (FC)

Calprotectin is a protein (belonging to the S100-protein family) found in human blood, saliva, cerebrospinal fluid, and urine. It is secreted by white blood cells,

called neutrophils, in response to inflammation in the body. When inflammation is detected in the gastrointestinal tract, neutrophils migrate to the intestinal lining, resulting in elevated levels of calprotectin in faeces. FC, therefore, acts as a biomarker for gastrointestinal inflammation (225, 226).

4.3.1 FC assays

Several FC assays have been developed and marketed as a tool to help rule out IBD and other organic enteric diseases in patients presenting with gastrointestinal symptoms. FC assays available in the UK include: fully quantitative laboratory-based tests, which provide a numerical value for the amount of calprotectin detected in a given stool sample (e. g. '100 micrograms/ gram [µg/g]'); and quantitative or semi-quantitative point of care tests (POCTs), which may be used inside or outside of the laboratory setting. In the case of semi-quantitative assays, these provide an indication of the region within which the underlying quantitative value lies (e.g. '50-100 µg/g'), rather reporting a specific numerical result as with quantitative tests (227).

The most commonly used FC assays are based on laboratory enzyme-linked immunosorbent assay (ELISA) platforms, the basic principle of which centres on a process of antibody-to-antigen binding (227). The BÜHLMANN fCAL® ELISA test, for example, uses a sandwich ELISA method consisting of the following steps: (1) the assay plate is first coated with a 'capture antibody', (2) the faecal sample is applied to the plate, resulting in calprotectin protein molecules in the sample binding to the capture antibodies, (3) the plates are 'washed' to leave the antibody-protein elements, and (4) a 'detection antibody' is applied, which detects the calprotectin molecules bound to the capture antibodies on the plate, and informs the resulting quantitative measure of calprotectin in the sample (228).

A summary of FC assays available in the UK is provided in Table 4-1. This list is based on three sources: assays listed within the 2013 NICE DG11 guidance on FC (discussed in section 4.3.2) (227); an updated list of technologies provided within a NICE 2017 review of DG11 (229); and assays included in national EQA reports for FC within the year 2018 (which were used to inform an analysis presented in Chapter 7 [data outlined in section 7.2]) (230). Where necessary, specific assay details not provided in either of these documentations were

identified via FC manufacturer websites or US FDA premarket notification [510(k)] documents (231-236). For several tests, the measurement range could not be identified from any of the above sources – in those cases, this detail has been left blank in the table. Footnotes are provided in the table to provide brief explanations of the different assay methods used.

Table 4-1. FC assays available in the UK

Manufacturer & platform	Assay method	Measurement range
Tests included in both the NICE guidance lists and national EQA reports:		
BÜHLMANN fCAL® ELISA	ELISA[a] – quantitative	Range 1: 10-600 µg/g Range 2: 30-1800 µg/g
CALPRO® Calprotectin ELISA Test (ALP)	ELISA[a] – quantitative	Range 1: up to 1250 mg/kg Range 2: up to 2500 mg/kg
Immundiagnostik (IDK®) Calprotectin ELISA	ELISA[a] – quantitative	Range: up to 2100 µg/g
(Thermo Fisher) EliA™ Calprotectin	Fluorescence enzyme immunoassay[b] – quantitative	Range: 15-3000 mg/kg
BÜHLMANN Quantum Blue® fCAL	Immunoassay (lateral flow)[c] – rapid quantitative test	Range 1: 30-300 µg/g Range 2: 100-1800 µg/g
Tests included the NICE guidance lists only:		
Eurospital Calprest	ELISA[a] – quantitative	-
Eurospital CalFast	ELISA[a] – rapid quantitative test (with dedicated reader)	-
(Preventis) PreventID Cal Screen	Immunochromatographic[d] POCT rapid test – semi-quantitative	-
(Preventis) PreventiD Cal Detect 50/200	Immunochromatographic[d] POCT rapid test – semi-quantitative	-
Tests included national EQA reports only:		
BÜHLMANN fCAL® turbo	Immunoassay (turbidimetric)[e] - quantitative	Range: 30 - 2000 µg/g
(Launch Diagnostics) Accusay Calprotectin™	ELISA[a] – quantitative	-
Orgentec Alegria®	ELISA[a] – quantitative	Range: 0 - 1000 µg/g
Inova QUANTA Flash®	Chemiluminescent immunoassay[f] - quantitative	Range: 16.1 – 3500.0 mg/kg
DiaSorin LIAISON® Calprotectin	Chemiluminescent immunoassay[f] - quantitative	Range: 5 - 800 µg/g

(Thermo Fisher) EliA™ Calprotectin 2	Fluorescence enzyme immunoassay[b] – quantitative	-

ELISA = Enzyme-linked immunosorbent assay. *POCT = point of care test.*

[a] *ELISA tests* are plate-based assays which are based on the central principle of antibody-to-antigen binding (see section 4.3.1 above). The amount of antibody-to-antigen complexes formed within ELISA tests is typically quantified by adding a chromogenic substrate to the sample, which reacts with an enzyme bound to the antibody-antigen complexes to form a detectable coloured compound, which can be quantified using spectrophotometric absorbance techniques. Quantitative ELISA assays provide a numerical result, calibrated against a reference material. 'Rapid' ELISA assays provide results in a short time frame (usually minutes) compared to standard laboratory-based ELISA assays which may take hours or days.

[b] *Fluorescence enzyme immunoassays* are similar to ELISA tests, but the antibodies in this case are labelled with fluorescent probes (rather than enzymes) to enable the amount of antibody-to-antigen complexes to be quantified based on the measurement of fluorescent intensity.

[c] *Lateral flow immunoassays* are rapid tests in which labelled capture antibodies are immobilised across an absorbent strip of material; the test sample is then added to one side of the strip, the sample flows over the capture antibody line (driven by lateral capillary force), and the target antigens are captured by the capture antibodies. As the captured antibody-to-antigen complexes accumulate, they can typically be viewed directly by the naked eye to provide qualitative (i.e. yes/no) or semi-quantitative (i.e. numeric range) results.

[d] *Immunochromatographic* tests use the same technology as lateral flow immunoassays. The terminology provided for each assay in this table has been based on the terminology used by the individual manufacturers.

[e] *Turbidimetric* assays depend on the process of antibody-to-antigen binding. The quantification process in this case is based on measuring the loss of intensity of light (of a known wavelength) passed through the sample, which occurs due to the effect of light-scattering caused by passing light through the antibody-to-antigen complexes in the solution.

[f] *Chemiluminescent immunoassays* are similar to ELISA and fluorescence enzyme immunoassays, but the detection antibody has a chemiluminescent label, rather than a chromogenic or fluorescent label. In this case, the addition of the substrate to the sample causes a chemiluminescent reaction, which allows the concentration of measurand to be measured according to the units of light emitted.

4.3.2 NICE assessment (DG11)

In 2013, NICE issued guidance (DG11) under its DAP scheme recommending FC as a test to help distinguish between IBD and IBS (227). This section summarises the clinical and economic evidence used to inform that NICE recommendation.

The primary source of evidence underpinning the NICE recommendation was an independently commissioned External Assessment Group (EAG) report (220). This included a systematic review to identify and synthesise data on the diagnostic accuracy of FC, and an economic evaluation to determine the expected cost-effectiveness of FC compared to standard care. Twelve FC technologies were included in the assessment scope, consisting of laboratory

based ELISA tests (n=6), immunochromatographic POCTs (n=2), other rapid[29] tests (n=3) and a laboratory quantitative fluorescence enzyme immunoassay test (see Table 4-1)[30]. The reference standard, used to classify the diagnostic accuracy of FC assays within the diagnostic accuracy studies, was histology after endoscopy.

Within the EAG systematic review, several scenarios were considered relating to paediatric and adult populations. Focusing on the adult population, a total of 7 studies were identified which assessed the diagnostic accuracy of FC for distinguishing between IBS and IBD. All of these studies were conducted within secondary care settings and reported diagnostic accuracy across 8 different cut-off thresholds (ranging from 8–150 µg/g). Most (5/8 studies) evaluated ELISA assays and reported diagnostic accuracy using a 50 µg/g cut-off threshold: a meta-analysis of these five studies (based on a total pool of 596 people) found a combined sensitivity of 93% and a specificity of 94%. In addition, one study (Otten et al. 2008) was identified which assessed the accuracy of a POCT, CalDetect, reporting a sensitivity of 100% and specificity of 95% at a cut-off threshold of 15 µg/g, based on a sample of 114 patients (237).

Based on the above review findings, the EAG conducted an economic evaluation to assess the cost-effectiveness of FC from an NHS and PSS perspective. Two testing strategies were evaluated, both of which assumed that a single FC test would be conducted: (1) ELISA FC testing using a 50 µg/g cut-off (with diagnostic accuracy as reported in the EAG meta-analysis), and (2) POCT CalDetect using a 15 µg/g cut-off off (with diagnostic accuracy as reported in the Otten et al. 2008 study). Under the standard care comparator strategy, GPs were assumed to have a 100% sensitivity and 79% specificity for detecting IBD based on clinical data (220). In the EAG economic model, an initial test sub-model was used to combine FC diagnostic accuracy estimates with IBD prevalence (assumed to be 6.3%),

[29] Note: NICE separately classified three tests as “rapid tests” rather than POCTs, stating that these tests needed a dedicated reader to process the tests, but that with appropriate training and quality assurance processes may be appropriate for future use in point-of-care settings.

[30] Note: NICE listed tests from the same manufacturer but with different measurement ranges as separate assays. For example, the BÜHLMANN fCAL® ELISA test was counted as two different assays in the NICE list of technologies, one for each of the measurement ranges listed in Table 4-1.

and three long-term Markov models were used to track downstream outcomes for CD, UC and IBS patients separately, adopting a 10-year time horizon.

The results of the EAG economic evaluation found FC plus standard care to dominate standard care alone, resulting in cost savings of £83 and £82 per patient (for ELISA FC and POCT CalDetect and strategies respectively) and a marginal QALY gain of 0.0007 for both strategies[31] (220). The cost savings were stated to be driven by the lower number of referrals to secondary care when using FC: within the modelled population, 19.8% of patients were referred to secondary care for colonoscopy under the standard care strategy, compared to 5.6% (FC ELISA strategy) and 5.1% (FC CalDetect strategy). The QALY gains meanwhile were stated to be driven by a slight mortality reduction resulting from fewer colonoscopies.

Based on the EAG analyses, the NICE DAP committee noted a paucity of evidence relating to three key areas for FC:

1. *Primary care performance*

The committee noted that, whilst there was a growing focus on the use of FC in primary care, there was a lack of data identified in the EAG analysis relating to this setting. In particular, all of the studies informing the estimates of FC diagnostic sensitivity and specificity used within the EAG adult cohort economic evaluation were conducted in secondary care settings. Interestingly, the committee concluded that the estimated benefits related to FC testing identified from the EAG analysis would nevertheless be expected to generalise to the primary care setting. This assumption has since been contradicted by recent study findings (238, 239) (see section 4.3.3).

2. *FC assay comparative performance*

The committee noted a lack of data on the head-to-head performance of alternative FC assays. In the absence of such data, the committee stated that preferred FC tests may be selected locally in the NHS but that potential differences between tests should be considered.

[31] Health care costs were measured in 2011 GBP (£).

3. FC cut-off thresholds

The committee highlighted the range of cut-off thresholds used across FC tests and concluded that further research was needed before recommendations regarding particular cut-offs could be made. In addition to optimal and assay-specific cut-offs, the committee recommended research into repeat-testing strategies in people with intermediate FC levels. The committee recommended that test cut-offs should be discussed and agreed locally as part of the FC implementation process.

In addition to the above, the committee also highlighted the potential influence of pre-analytical and analytical factors on the performance of FC tests, including: stool sampling, stool consistency, sample storage, extraction and extract dilution. It was noted that a national EQA scheme run by the National EQA Service (NEQAS) had been set up for FC, and participation in this scheme as well as standardisation of sample preparation methodology, where possible, was encouraged. Of note, however, the committee did not include the potential influence of pre-analytical and analytical factors on the performance of FC tests under their suggested areas for further research; this aligns with the historical lack of consideration of these matters within HTA settings, as demonstrated in Chapter 2 (1).

Whilst noting the above considerations, the NICE committee concluded that, based on the available evidence, FC testing was a cost-effective use of NHS resources (227).

4.3.3 The York Faecal Calprotectin Care Pathway (YFCCP)

Implementation of FC within primary care settings in the UK was initially limited following the NICE 2013 DG11 guidance, at least in part because the majority of evidence on FC at that time came from secondary care settings using a low (50 µg/g faeces) cut-off threshold (220, 238). A primary care diagnostic accuracy study conducted at the Sheffield Teaching Hospitals NHS Foundation Trust following the NICE recommendation, for example, found FC to have significantly lower sensitivity (72.7%) and specificity (64.9%) for detecting IBD at the 50 µg/g faeces threshold, compared to the previous NICE estimates (238). It is expected that this drop in performance is driven by the fact that: (a) the prevalence of

disease is lower in primary care, and (b) the spectrum of disease is different, with less severe cases likely to be more prevalent in primary vs. secondary care.

In an attempt to address this issue, led by gastroenterologist Dr Turvill at the York Teaching Hospital NHS Foundation Trust, a tailored pathway for the use of FC in primary care was established in 2014, called the 'York FC Care Pathway' (YFCCP). In response to fears over increased false-positive rates when using FC in primary care, this pathway applies a raised FC cut-off threshold of 100 µg/g (compared to the "standard" cut-off of 50 µg/g), and includes a second FC test to confirm suspected IBD following an initially elevated result.

The YFCCP protocol is outlined in Figure 4-1. Patients entering the pathway are required to meet the following eligibility criteria:

- aged 18-60 years;
- presenting to primary care with new lower gastrointestinal symptoms;
- not suspected of having cancer (these patients would be urgently referred to secondary care under current NICE guideline NG12 (224)); and
- having received non-diagnostic standard initial GP investigations (e.g. full blood count, renal function tests, CRP with stool culture [and Clostridium difficile screen], coeliac screen and thyroid function tests, as indicated).

Under the YFCCP, patients meeting the above criteria are asked to return an initial stool sample for FC testing (henceforth referred to as 'FC1') within 24 hours. Those with a low FC1 value (<100 µg/g) are treated as having likely IBS and managed in primary care, with treatment as per NICE guidance [CG61] (214). As a safeguard, these patients are reviewed at six weeks within primary care and may be referred if symptoms persist – this enables both false negative IBD cases, as well as severe IBS cases, to benefit from subsequent secondary care referral. Patients with an elevated FC1 value (≥100 µg/g) have a second test conducted (henceforth referred to as 'FC2'). If the FC2 result is <100 µg/g the patient is managed within primary care with suspected IBS, as outlined above. If the result is ≥100 µg/g patients are treated as having suspected IBD and referred to secondary care for endoscopic investigations: patients with intermediate results (100-250 µg/g) receive a routine referral; whilst those with a high result (≥250 µg/g) receive an urgent referral (see Figure 4-1).

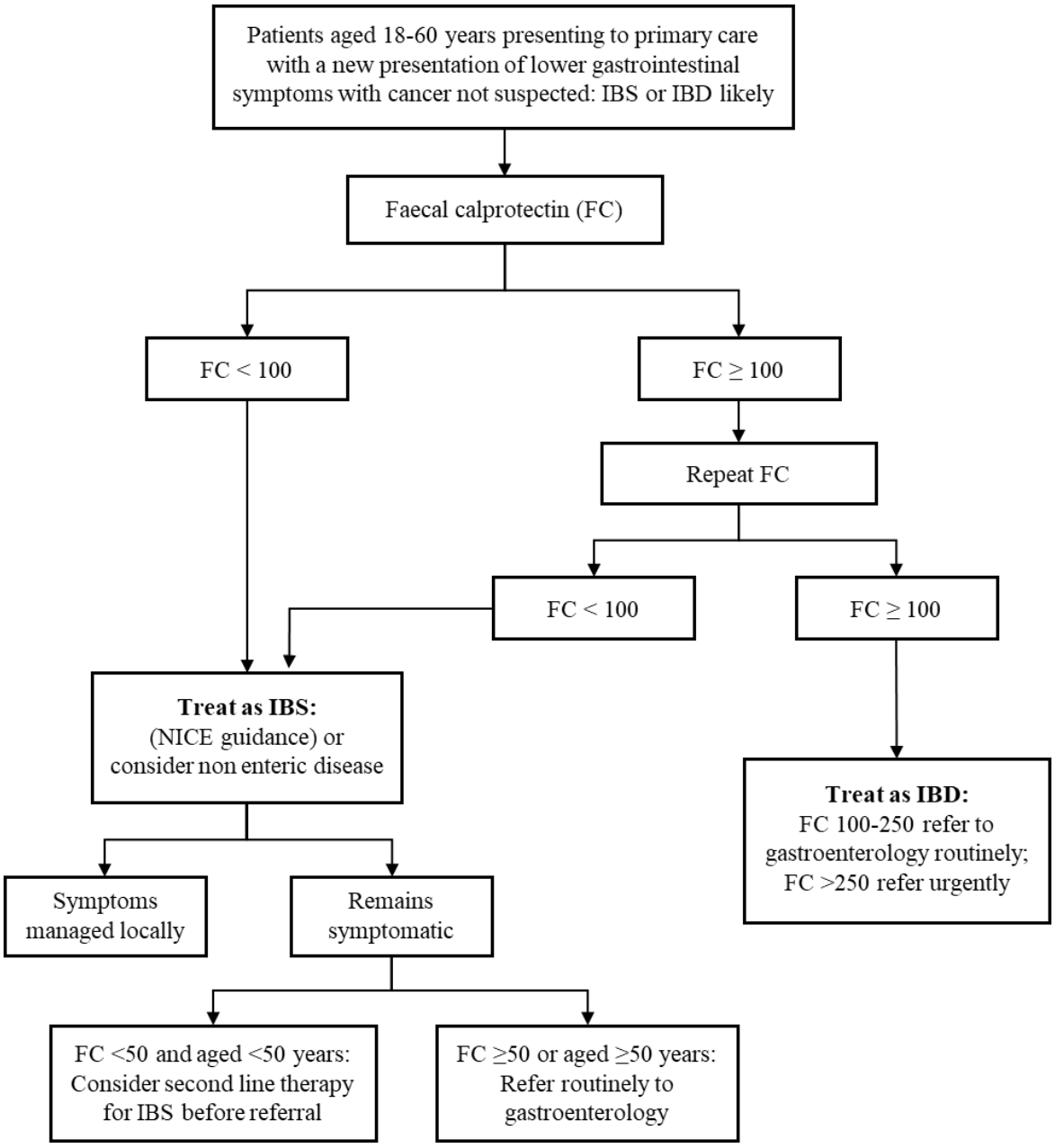


Figure 4-1. The York Faecal Calprotectin Care Pathway (YFCCP)

The diagnostic accuracy of the YFCCP was first assessed based on a pilot scheme run within the York region in 2014, using the BÜHLMANN fCAL® ELISA assay with a measurement range of 10-600 µg/g (240). Using data from 262 patients, FC was found to have a negative predictive value (NPV) of 0.97 and a positive predictive value (PPV) of 0.40 when using the YFCCP, which doubled the PPV of the test compared to using the standard 50 µg/g cut-off (240). Based on these positive findings, the pathway was rolled-out across the York region in 2016, using the same FC assay.

An audit evaluation of the first 951 patients treated according to the rolled-out YFCCP pathway (between August 2016 and April 2017) was subsequently conducted (239). This assessment reported a NPV of 0.99 (95% CI: 0.98 to 1.0) and PPV of 0.51 (95% CI: 0.43 to 0.59), with a diagnostic sensitivity of 0.94 (95% CI: 0.85 to 0.98) and specificity of 0.92 (95% CI: 0.90 to 0.94). A companion economic evaluation, applying these diagnostic accuracy results within a one-year decision tree model, was also conducted (239). This evaluation found the YFCCP to be cost-effective, being less costly and more effective than most of the comparators considered in the analysis, which included both standard GP referral comparator strategies as well as single-test FC comparator strategies (239, 241). Further description of this study and a recent update of the economic evaluation is provided in Chapter 6.

Based on the latest evidence relating to the diagnostic accuracy and cost-effectiveness of the YFCCP, this pathway has since been introduced within 240 practices across 9 Clinical Commissioning Groups (CCGs) spanning the York region and beyond, and scaling up of implementation of the YFCCP has begun in other regions in the UK including South Tees, Oxford, Bristol and Exeter (242). In addition, in 2018 NICE endorsed the YFCCP within a new consensus document on FC, which advocates a national algorithm for FC as a rule-out test for IBD in primary care based on the YFCCP repeat-test strategy (222, 242). The YFCCP therefore presents the current best practice for the differential diagnosis of IBD and IBS in primary care, supplementing the previous NICE 2013 DG11 guidance (227).

4.3.4 Measurement performance

Although the NICE DG11 assessment did not include an assessment of FC measurement performance, the NICE Committee did discuss pre-analytical factors that may affect FC results, and encouraged participation in the UK NEQAS scheme for FC (established in early 2012) (227). In the years following DG11, research around pre-analytical and analytical factors affecting FC measurement performance has continued. A brief description of recent studies in this area is provided below; in addition, an analysis of between-assay measurement performance based on data from the NEQAS FC EQA scheme is provided separately in Chapter 7.

4.3.4.1 Pre-analytical factors

With respect to pre-analytical factors affecting the performance of FC in the diagnostic setting, researchers have explored a range of factors including sample timing, storage procedures and sample extraction method, as outlined below.

Sample timing

For patients with IBD, calprotectin levels have been shown to vary significantly within individuals over a single day, with levels appearing to increase in line with the length of interval between bowel movements (243, 244). This has led to calls for a standardised approach to sampling, based on sampling from patients' first bowel action in the morning (245). However, contradictory evidence has indicated that FC levels may not be consistently elevated with the first morning bowel movement (246). Therefore, in the absence of clear evidence to inform a standardised sampling approach, the collection of multiple samples for individuals has been recently suggested as an alternative route to minimising variability – in both the diagnostic and monitoring contexts (247, 248).

Sample storage

A series of studies spanning back to the 1990's suggested that faecal samples could be safely stored at room temperature for up to a week (243, 249-251). More recently, however, Padoan and colleagues (2018) found that FC levels fell on average by 12% within the first 24 hours following stool collection irrespective of storage temperature, whilst samples stored for longer than 24 hours at room temperature exhibited further degradation (252). The authors therefore

recommend room temperature storage for no longer than 24 hours; refrigeration for up to 48 hours; and longer-term storage at -20ºC (252).

Sample extraction

The standard faecal sample extraction procedure involves a series of processes including: (i) homogenization (i.e. to obtain an even concentration of calprotectin); (ii) extraction of a sample; (iii) weighting; and (iv) dilution (248). Due to the time-consuming nature of this process, test manufacturers have developed novel extraction devices which avoid the need to manually weigh each sample – such as the Inova Diagnostics Fecal Extraction Device, or the Bühlmann CALEX® Cap Calprotectin Stool Extraction device (253, 254). Despite efficiency gains associated with these devices, several studies have highlighted significant quantitative differences in FC results depending on the extraction method and/or device adopted (248, 255-258).

4.3.4.2 Analytical factors

Bias (method comparison)

With respect to the analytical performance of FC, several recent method-comparison studies have explored the level of agreement between alternative FC assays (248, 259, 260). In particular a 2017 UK study of four FC assays found that, depending on the assay manufacturer, up to 3.9-fold differences may occur between quantitative FC results, with all inter-assay differences found to be significant in all cases (259). Subsequent international studies have also demonstrated significant differences depending on the FC assay and platform (248, 260). Possible reasons for these differences include the application of different antibodies across FC assays, and the use of different immunoassay techniques[32] (248). Interestingly in the aforementioned 2017 UK study (which also included a diagnostic assessment), inter-assay differences were concluded to have limited impact on the clinical performance of the assays (with 94-100% sensitivity and 82-100% specificity reported across assays, at a 50 µg/g cut-off

[32] E.g. ELISA vs. chemiluminescence immunoassay (CLIA) vs. fluorimetric enzyme-lined immunoassay (FEIA) vs. particle-enhanced turbidimetric immunoassay (PETIA).

threshold) (259). It should be noted, however, that this study considered a selected population of only IBS and IBD patients.

Analytical variation

Several studies have quantified the precision of FC assays (248, 252, 260). In particular in a 2018 Italian study, repeatability and intermediate precision were evaluated for three FC assays (using assay-specific extraction devices) at low, intermediate and high FC values, based on samples from 110 patients with IBD. This study found that repeatability (CV%) of the assays ranged from 6.1% to 15.7% at low FC, 9.7% to 14.7% at intermediate FC, and 9.6% to 25.3% at high FC; whilst intermediate precision ranged from 8.1% to 15.6% at low FC, 13% to 16.4% at intermediate FC, and 11.8% to 27.6% at high FC (252). Subsequently in a 2019 Dutch study, an assessment of intermediate precision was conducted on four alternative assays (using the manufacturers' recommended extraction devices), based on manufacturer-provided low and high quality control samples[33] run over 20 different days (260). In this case for low FC samples, imprecision (CV%) ranged from 4.9% to 52.4% across the different assays; and at high FC samples, CV% ranged from 5.6% to 23.8% (260). Interestingly, in most cases, the observed CV% levels were substantially higher than the associated manufacturer imprecision claims (260).

4.3.4.3 Towards harmonisation

Based on the findings highlighted in sections 4.3.4.1 and 4.3.4.2 above, there have been continued calls for greater standardisation of FC pre-analytical and analytical processes as a means of achieving assay *harmonisation* (248, 255, 261-264). *Harmonisation* in this context means that test results may be considered comparable irrespective of the measurement procedure used, and where or when a measurement was made (265). For FC, a current barrier to harmonisation concerns the lack of FC reference measurement procedure or reference materials against which to "anchor" test results: only with a reliable

[33] Quality control samples are special specimens (in this case developed by the test manufacturers), which are intended to represent a known, stable level of the measurand (+/- uncertainty). These samples are treated as if they were patient samples, undergoing the same pre-analytical and analytical processes. By virtue of being "known" quantities, they enable assay imprecision and bias to be monitored.

reference can manufacturers make their assay results metrologically traceable to the same measurement unit (266). As such, in the absence of standardisation or harmonisation, there have been repeated calls for assay-specific cut-off thresholds for FC to ensure that diagnostic performance is individually optimised (252, 255, 259, 267). As of yet, however, assay-specific thresholds have yet to be widely researched or adopted into clinical guidelines, and concerns around the lack of FC standardisation and associated potential misinterpretation of FC results persists (268).

4.4 FC case study

4.4.1 Case study motivation and objectives

The primary aim of the thesis case study is to explore and develop methods for assessing the impact of measurement uncertainty on clinical and health-economic outcomes. Key objectives with respect to this primary aim are outlined in section 4.4.1.1 below. In addition, FC was chosen as the case study test for several reasons relating to clinical need, and data characteristics and availability. These aspects are outlined in sections 4.4.1.2 and 4.4.1.3.

4.4.1.1 Methods

In the methodology review presented in Chapter 3, three methods were highlighted as having particular utility with respect to the thesis aim and hypotheses (see section 1.5): (i) error model simulation, (ii) decision analytic modelling, and (iii) contour plotting (see section 3.4). The focus of the thesis case study is therefore on exploring and developing these methods, with the following objectives:

A. To explore the use of the *error model simulation approach* for (i) assessing the impact of increasing measurement uncertainty on the clinical performance of testing strategies (i.e. hypothesis C), and (ii) identifying outcome-based APS (i.e. hypothesis D).

B. To explore the use of *decision analytic modelling* as a means of (i) assessing the impact of increasing measurement uncertainty on clinical-utility and cost-effectiveness outcomes (by linking simulated clinical

performance outputs from (A) to downstream outcomes) (i.e. hypothesis C) and (ii) identifying outcome-based APS (i.e. hypothesis D).

C. To explore the use of *contour plotting* as a means of illustrating simulation/model outputs and identifying outcome-based APS (i.e. hypotheses C and D).

D. To explore how real world evidence (RWE) may be utilised within the *error model* and *decision analytic model* simulation framework, to evaluate real-world scenarios – in particular to assess the impact of between-assay differences on outcomes (i.e. hypothesis E).

In addition to the points above, the case study analysis is intended to provide general utility as a demonstrative example of the impact that measurement uncertainty may have on downstream outcomes.

4.4.1.2 Clinical need

Whilst the potential influence of pre-analytical and analytical factors on the clinical performance of FC tests was highlighted by the NICE committee in 2013 (227), no formal assessment was undertaken. A study evaluating the impact of FC measurement uncertainty on clinical and health-economic outcomes would therefore be of use, to help inform clinical decision makers as to the expected robustness of FC testing strategies to increased measurement uncertainty. In addition, the NICE committee also highlighted a paucity of data on the head-to-head performance of alternative FC assays (227). Based on the availability of EQA method-comparison data for FC assays (see section 4.4.1.3), this case study provides a further opportunity to explore the comparative performance of alternative FC assays in terms of clinical and health-economic outcomes. Finally, no evidence-based APS (outcome-based or otherwise) are currently in place for FC. There is a further opportunity, therefore, for this case study to help inform and direct the future development of APS for FC, as well as more broadly advancing the field of outcome-based APS – of key concern within the laboratory community at the moment (32, 269, 270).

4.4.1.3 Data characteristics and availability

IPD pertaining to the most recent diagnostic accuracy evaluation of the YFCCP was available for this case study (239). Whilst this data relates to patients treated

under the YFCCP, it can also be used to retrospectively evaluate the NICE DG11 single-test FC pathway by determining what patients' FC diagnoses would have been, had only their initial FC1 test results been available. This dataset therefore provides a rich source upon which an evaluation of the two FC pathways can be conducted. Through the provision of IPD, a range of methodological approaches can be explored within the error model simulation analysis – including both bootstrap and parametric sampling (detailed in Chapter 5).

Two further data items were available for this analysis. First, the economic decision model informing the most recent YFCCP economic evaluation was kindly provided for this analysis by the model developers (241). This model provides a means of exploring how simulation techniques may be utilised within decision analytic models, to evaluate the impact of measurement uncertainty on clinical utility and cost-effectiveness outcomes (241). It should be noted, however, that as this economic model is a short-term deterministic (rather than probabilistic) model, the analysis of economic outcomes presented in this thesis (Chapter 6) is limited to short-term deterministic outputs. This limitation means that: (a) caution is required when attempting to extract clinical policy conclusions from the analysis; and (b) the presented analysis does not address questions around how measurement uncertainty in this case study impacts on the probability of cost-effectiveness for FC, and how the impact of measurement uncertainty on clinical-utility and cost-effectiveness outcomes compares to the impact of other sampling (i.e. second-order, parameter) uncertainty. The full implications of this limitation, including suggested future research activities, are discussed in section 6.4.3. In terms of the justification for using this model, the primary concern in this case study was to show how the impact of measurement uncertainty on clinical utility and cost-effectiveness outcomes could be evaluated by embedding the error model simulation approach within a decision analytic model, and this model is sufficient for this purpose.

High-level data on the measurement performance of alternative FC assays was also available for this analysis, via the UK NEQAS EQA scheme for FC (discussed in Chapter 7, section 7.2). This data provides a means of exploring how real world evidence (RWE) may be utilised to evaluate specific testing scenarios using the error model approach, to further develop the methods used

in previous studies in this area. In particular in this case, this data enables an assessment of the impact of between-assay differences on clinical and health-economic outcomes.

4.4.2 Outline of case study analysis chapters

The case study analysis is divided into three chapters, outlined below.

In **Chapter 5**, the error model simulation approach introduced in Chapter 3 is used to assess the impact of additional FC measurement uncertainty on the diagnostic accuracy of two FC testing strategies: the 'NICE FC pathway' (a single-test FC strategy adopting a 50 µg/g faeces threshold, as evaluated in the NICE DG11 assessment); and the YFCCP (a repeat-test FC strategy, adopting a 100 µg/g faeces threshold). The simulation results are used to assess the robustness of each pathway's diagnostic accuracy to increasing measurement uncertainty, and to explore the derivation of outcome-based APS – in particular using contour plots to provide a visual illustration of the findings.

In **Chapter 6**, the framework outlined in Chapter 5 is extended to end-stage outcomes using an existing FC economic model. Within this analysis, for each FC pathway the diagnostic accuracy results from Chapter 5 are applied within the FC economic model, to assess the impact of increasing measurement uncertainty on cost, QALY and cost-utility (NMB) outcomes. Contour plots are again used to assess the robustness of each pathway's outcomes to increased measurement uncertainty, and to explore the derivation of outcome-based APS.

In **Chapter 7** an analysis of RWE is presented, which explores how between-method measurement performance data may be used within the error model framework to evaluate alternative assay outcomes. In this analysis, national EQA data for FC is used to assess the expected impact of switching FC assay within the YFCCP on the pathway's clinical and economic outcomes.

Note that within each of the case study analysis chapters outlined above, a discussion of the specific findings and limitations of the analysis is provided. An overarching discussion of the thesis findings, including suggestions for future research, is subsequently provided in the final thesis chapter (**Chapter 8**).

4.5 Summary

- This chapter has outlined the clinical context underpinning the thesis case study: FC as a primary care diagnostic test for IBD.
- Two diagnostic pathways currently used in the UK primary care setting were highlighted: the NICE FC pathway (a single-test strategy, using a 50 µg/g faeces cut-off threshold); and the YFCCP (a repeat-test strategy, using a 100 µg/g faeces cut-off threshold).
- In the following case study analysis, the impact of FC measurement uncertainty on the diagnostic accuracy, clinical utility and cost-effectiveness of the NICE FC pathway, and the YFCCP, is explored.

In the following chapter (**Chapter 5**) the first part of the thesis case study analysis is presented: an assessment of the impact of increasing FC measurement uncertainty on the diagnostic accuracy of the NICE FC pathway and the YFCCP.

Chapter 5
The impact of measurement uncertainty on the diagnostic accuracy of FC testing strategies

5.1 Chapter outline

The previous chapter outlined the clinical context and motivating factors underpinning the thesis case study. This chapter reports on the first section of the case study analysis: an assessment of the impact of measurement uncertainty on the diagnostic accuracy of FC diagnostic pathways. This part of the case study addresses hypotheses C and D of the thesis – i.e. that methods from the broader literature may be applied within HTA-style assessments to: [C] evaluate the impact of measurement uncertainty on clinical performance outcomes, and [D] identify outcome-based APS.

Two pathways are evaluated in this study: (1) the NICE FC pathway (a single-test FC strategy, adopting a 50 µg/g faeces cut-off threshold); and (2) the YFCCP (a repeat-test FC strategy, adopting a 100 µg/g faeces cut-off threshold). For both pathways, baseline FC values were sampled from an IPD dataset of 951 patients treated under the YFCCP (summarised in section 5.2). For the NICE FC pathway, the impact of measurement uncertainty on the pathway's diagnostic accuracy was explored using the error model simulation approach as outlined in Chapter 3 For the YFFCP, due to the repeat-test nature of this pathway a slightly adapted simulation process was required. The analysis is therefore presented in two parts: part 1 reports on the NICE FC pathway evaluation (section 5.3); and part 2 reports on the YFCCP evaluation (section 5.4). In both cases the simulation results are presented using contour plots. These are used to assess the robustness of each pathway's diagnostic accuracy to increasing measurement uncertainty, and to further explore the possibility of outcome-based APS in the form of "acceptable regions" of bias and imprecision. The final section of this chapter provides a discussion of the key findings and limitations of the analysis (section 5.5).

5.2 Data

An anonymised dataset was obtained relating to 951 patients treated under the YFCCP between August 2016 and April 2017 – the same patient cohort that

informed the latest YFCCP diagnostic accuracy assessment (Turvill *et al*. 2018) (239). Although these patients were treated under the YFCCP, the associated dataset included separate FC1 and FC2 values, in addition to patients' clinical diagnoses. This data could therefore be used to evaluate both the single-test NICE FC pathway (by reference to the FC1 values only – see section 5.3) and the YFCCP (by reference to both the FC1 and FC2 values – see section 5.4). The dataset obtained for this study, henceforth referred to as the 'YFCCP dataset', included information on patients' FC1 results, FC2 results (where applicable), secondary care endoscopic investigations and final clinical diagnoses (discussed further below).

Note that, as this data relates to a routine service evaluation (i.e. audit data) and was fully anonymised, formal ethical approval was not required for this study. This was confirmed via email correspondence with the ethics teams at both the York Teaching Hospital NHS Foundation Trust and the University of Leeds, before obtaining the data.

As per the YFCCP protocol, all patients within the YFCCP dataset were aged 18-60 years and presented to primary care with new lower gastrointestinal symptoms, with initial GP diagnostic workup (e.g. full blood count, CRP, coeliac screen, Clostridium difficile screen etc.) being non-confirmatory (239). Of the first 1005 consecutive patients entering the YFCCP, 52 were subsequently excluded from the 2018 diagnostic analysis due to either incompletion of the pathway before secondary care referral (n=34), or non-referral despite direction from the pathway (n=18). This left a total of 951 patients included in the YFCCP dataset (239). Of those, 63% were women and 47% were men, and the median patient age was 38 years (IQR: 27-48) (239).

All FC testing within the YFCCP was conducted at York Teaching Hospital NHS Foundation Trust. Samples were extracted using Bühlmann CALEX extraction tubes and analysed using the Bühlmann fCAL® ELISA assay (Alpha Laboratories Ltd, UK) with a quantitative measurement range of 10-600 µg/g faeces. Faecal samples were stored for 1-4 days at 4ºC before extraction, and extracts were analysed within 7 days of storage (also at 4ºC). As per the York laboratory's standard operating procedures for FC, two internal quality control (IQC) samples were run with each FC testing batch to monitor imprecision resulting from the

assay platform. Based on the IQC samples run during the data collection period, the results indicated an analytical CV of 7% at a lower measurement range (~50 µg/g), and 4% at a higher measurement range (~150 µg/g) (271).

A summary of patients' final clinical diagnoses within the YFCCP dataset is provided in Table 5-1. These classifications are based on the results of secondary care investigations where available, assuming that colonoscopy is 100% accurate at diagnosing patients. An initial diagnosis of IBS (or other functional disease) was assumed to be correct so long as patients did not return to their GP with persisting symptoms (either at the 6-week GP review, or over a subsequent 6-month follow-up period). The final clinical diagnosis of those patients who remained symptomatic and were referred to secondary care was based on the results of any secondary care investigations.

In line with the original diagnostic accuracy publication (239), clinical diagnoses were classified as either 'IBS' (a proxy for all functional intestinal diseases) or 'IBD' (a proxy for all organic enteric diseases requiring secondary care intervention). Table 5-1 provides details of the various functional and organic conditions included under these dichotomous classifications. Out of the 951 patients included in the YFCCP dataset, 92% (n=873/951) were diagnosed with a functional disease (91% with IBS specifically) and 8% (n=78/951) were diagnosed with an organic enteric disease (5% with IBD specifically).

Table 5-1. YFCCP dataset: patient clinical diagnoses

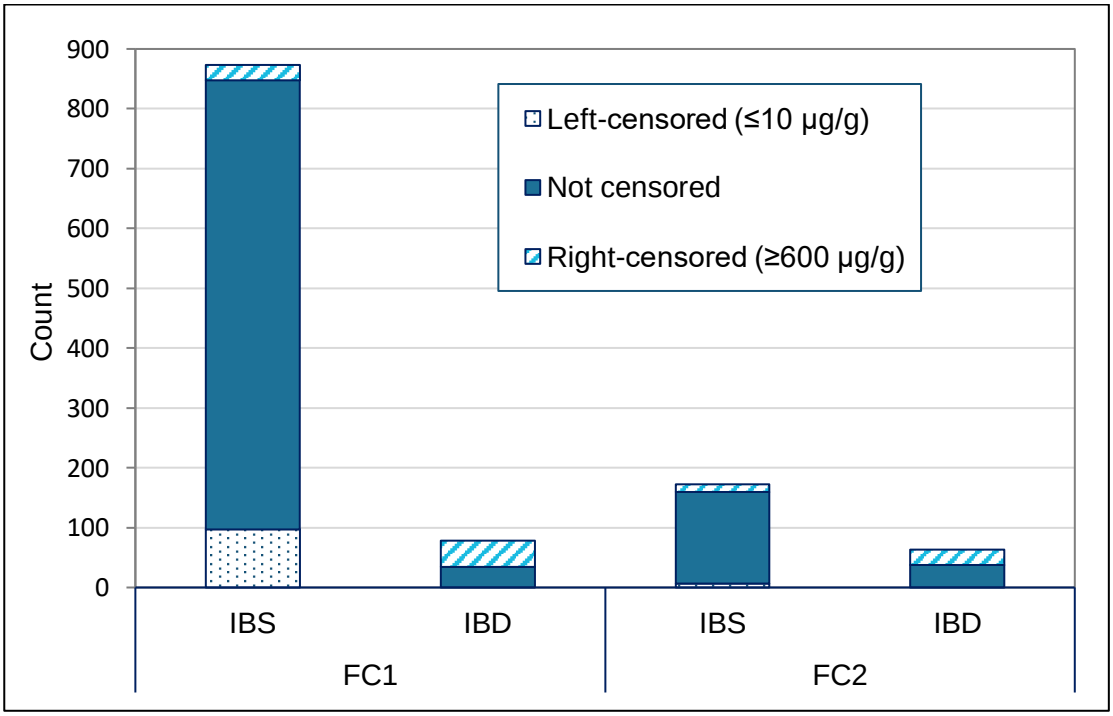
Diagnosis	N (%)
Any functional disease	**873 (92)**
IBS	**862 (91)**
Haemorrhoidal bleeding	8
Coeliac disease	3
Any organic enteric disease	**78 (8)**
IBD	**49 (5)**
Diverticular disease	5
Gastroenteritis	5
Microscopic colitis	3
Adenomatous polyps ≥10 mm	3
Colorectal cancer	2
Threadworm	2
Clostridium difficile	1
Endometrioma	1
Giardiasis	1
Incarcerated hernia	1
Lymphoma	1
Mesenteric ischaemia	1
Non-specific inflammation	1
Small bowel cyst	1
Subcentimetre neuroendocrine tumour	1

5.2.1.1 Censored data

A key feature of the YFFCP dataset, pertinent to the subsequent case study analysis, concerns the issue of censored FC data. Censoring here relates to the analytical measurement range of the Bühlmann fCAL® ELISA assay (10 – 600 µg/g) used within the YFCCP. Within this range the test returns a numerical result, whilst outside of this range the assay returns a semi-quantitative result: i.e. below the lower measurement limit (10 µg/g) a value of "<10" is reported (henceforth referred to as 'left-censored data'), whilst above the upper measurement limit (600 µg/g) a value of ">600" is reported (referred to as 'right-censored data').

The proportion of censored data within the YFCCP dataset is illustrated in Figure 5-1. Left-censored data were confined to the IBS population, with 11% (n=97/873) of FC1 results and 4% (n=7/172) of FC2 results being left-censored in this cohort. Right-censoring was more common in the IBD population, with 56% (n=44/78) and 40% (n=25/63) of IBD patients having right-censored FC1 and FC2 values respectively, compared to 3% (n=25/873) and 7% (n=12/172) in the IBS population.

Figure 5-1. YFCCP dataset: censored data summary



Censored data present a problem for the error model simulation approach, since a quantitative baseline value is required within the error model, upon which

additional hypothetical bias and imprecision is applied. Alternative approaches to dealing with censored data were explored in the analysis, as described in the following sections.

5.3 Part 1: NICE FC pathway evaluation

This section presents the methods and results for the evaluation of the NICE FC pathway. All data analysis and simulation conducted within this assessment, and in the subsequent YFCCP evaluation presented in part 2 (section 5.4), were performed using R software (version 3.4.3) (272).

5.3.1 Methods

5.3.1.1 Baseline diagnostic accuracy

The YFCCP dataset was first used to assess the 'baseline' diagnostic accuracy of the NICE FC pathway: that is, diagnostic accuracy as calculated using the empirical YFCCP dataset FC1 data, without simulating any additional test measurement uncertainty (which is the focus of the next section, section 5.3.1.2).

Using the YFCCP FC1 data, FC diagnoses were assigned according to the NICE FC pathway cut-off threshold i.e. patients were categorised as having suspected IBD if their FC1 value was ≥50 µg/g, or suspected IBS if their FC1 value was <50 µg/g. These FC diagnoses were then compared to the YFCCP clinical diagnoses (Table 5-1) and classified as true positive, true negative, false positive or false negative. Diagnostic sensitivity and specificity was thus calculated based on the proportion of results falling into each of these categories, as illustrated in Appendix C. Count plots, showing the distribution of FC1 values within the IBS and IBD populations, were also produced.

5.3.1.2 Simulated diagnostic accuracy

The impact of additional measurement uncertainty on the baseline diagnostic accuracy of the NICE FC pathway was assessed using the error model simulation approach, introduced in Chapter 3. Figure 3-3 previously outlined the error model

simulation process required for a single-test diagnostic strategy. This procedure was herein applied to the NICE FC pathway, as summarised in Figure 5-2 below.

i. A sample of $FC1_{true}$ values is assigned;

ii. For each $FC1_{true}$ value, the addition of bias and imprecision is simulated according to the specified error model to generate $FC1_{sim}$ values e.g.:

$$\mathbf{FC1_{sim} = FC1_{true} + [FC1_{true} \times N(0,1) \times CV] + Bias} \qquad \mathbf{(5.1)}$$

iii. The diagnostic accuracy of the NICE FC pathway including additional imprecision and bias is calculated by comparing diagnoses based on the $FC1_{sim}$ values (using a 50 µg/g threshold) with patients' clinical diagnoses;

iv. Steps (i) to (iii) are repeated for a range of CV and bias values.

Figure 5-2. NICE FC pathway: error model simulation approach required for a single-test strategy

Baseline "true" FC values required in the error model simulation were derived from the YFCCP dataset. Given that this dataset provided IPD on patients' FC1 results, two approaches were possible for sampling $FC1_{true}$ values within the error model simulation – the 'bootstrap method' (discussed in section 5.3.1.2.1 below), and the 'parametric method' (discussed in section 5.3.1.2.2). An additional analysis, based on applying the error model directly to the 951 FC1 values within the YFCCP dataset (i.e. with no sampling process applied), was also considered within a sensitivity analysis under the bootstrap method (see section 5.3.1.2.3).

5.3.1.2.1 Bootstrap sampling method

Under the bootstrap sampling method, a bootstrap simulation dataset was generated by random sampling with replacement (i.e. bootstrap sampling) from the YFCCP dataset rows (with each row including an individual patients' FC1 value and final clinical diagnosis). In the base case analysis 10,000 bootstrap samples were generated. The FC1 values within each bootstrap sample were then used as $FC1_{true}$ values within the error model simulation (i.e. steps (i) and (ii) in Figure 5-2).

As previously discussed, censored FC data within the YFCCP dataset present a problem for the error model simulation approach, since the error model requires a numerical $FC1_{true}$ value upon which to apply additional bias and imprecision. Note that whilst several studies have explored the performance of bootstrap estimators in scenarios involving censored data, most of these have focused on the context of survival or regression analysis; nevertheless, the bootstrap method has been shown to perform well in these cases (273-276). For the purpose of this analysis, the key concern is whether the specific values assigned to left- and right-censored data within the analysis (see below) have an impact on the estimated impact of measurement uncertainty on the clinical performance outputs. Within the bootstrap method base case analysis, censored FC data were replaced with numerical values equal to the associated limit value: that is, all left-censored data were replaced with a value of 10 µg/g and all right-censored data were replaced with 600 µg/g. In order to explore the potential impact of these replacement values, a range of sensitivity analyses were conducted: in each of these analyses, left-censored data were substituted with a 'half-way point' value (i.e. 5 µg/g) and right-censored data were substituted with values equal to 1.25, 1.5, 2 and 3 times the 600 upper limit value in turn (i.e. 750, 900, 1200 and 1800 µg/g). Note that, based on the results of these sensitivity analyses, censored data did not have a measureable impact on the analysis in this case (see section 5.3.2.2.2).

A final 'complete case' analysis was also conducted, in which all censored FC1 data from the YFCCP dataset were excluded. It should be noted, however, that this approach is expected to bias the results, since removal of FC1 censored data almost halves the IBD prevalence in the YFCCP dataset from 8.2% to 4.3%. Perhaps more importantly this approach lacks face validity, since in reality censored FC values are not discarded but are treated as clinically meaningful results. This sensitivity analysis is therefore presented for completeness only.

5.3.1.2.2 Parametric sampling method

Under the parametric sampling method, parametric distributions were assigned to the IBD population FC1 values and the IBS population FC1 values. A range of distributions (Normal, Lognormal, Gamma and Weibull) were fitted using the R 'fitdistrplus' package, which uses a maximum likelihood estimation (MLE) process

to fit selected parameterizations to the data provided (277). The optimal fitting distribution for each population was determined via an analysis of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) metrics, with the parameterisation with the lowest AIC and BIC values for each subgroup representing the distribution with the best statistical fit to the data (277).

The MLE process used within the 'fitdistrplus' package enables parameterisations to be applied over both complete and censored data regions. To allow for censored data, the package uses the 'fitdistcens' function, which requires the expected regions for censored data to be specified (277). This function estimates a vector of univariate distribution parameters by maximising the likelihood of censored data using the inputted data on non-censored observations, and left- and right-censored observations. The MLE estimation process in this case is based on the likelihood estimation formula originally derived by Klein and Moeschberger in 2003:

$$L(\theta) = \prod_{i=1}^{N_{nonC}} f(x_i|\theta) \times \prod_{j=1}^{N_{leftC}} F\left(x_j^{upper}\middle|\theta\right) \times \prod_{k=1}^{N_{rightC}} \left(1 - F\left(x_k^{lower}\middle|\theta\right)\right) \times \prod_{m=1}^{N_{intC}} \left(F\left(x_m^{upper}\middle|\theta\right) - F\left(x_m^{lower}|\theta\right)\right) \quad \textbf{(5.2)}$$

with x_i the N_{nonC} non-censored observations; x_j^{upper} upper values defining the N_{leftC} left-censored observations; x_k^{lower} lower values defining the N_{rightC} right-censored observations; $[x_m^{lower}; x_m^{upper}]$ the intervals defining the N_{intC} interval-censored observations; and F the cumulative distribution function of the selected parametric distribution (277, 278).

Whilst left- and right-censored data in this method are specified as falling within specified regions, it should be noted that the MLE distributional parameters produced from this method result in distributions which may extend beyond the extremes of the censored data regions. In other words, the resulting probability density profiles are not truncated as a result of specifying censored data regions.

In this case study, left-censored data were naturally specified as falling within the range 0-10 µg/g, given that negative assay values are not possible. For right-censored data, the lower bound for this censored data region is given as 600 µg/g, however specifying an upper bound is less straightforward. The 600 µg/g upper limit of quantification currently achieved using this version of the Bühlmann

assay is restrictive: alternative assays report measurement ranges spanning as high as 3,000 µg/g, for example (see Table 4-1). Three alternative values were therefore explored for the upper bound of the right-censored data region: 1,000, 2,000 and 3,000 µg/g. Based on an analysis of AIC and BIC criteria for parameterisations produced from each of these specifications, the upper bound of 1,000 µg/g was applied within the base case analysis (i.e. application of this upper bound for right-censored data produced parameterisations with the lowest AIC and BIC results). The two alternative upper bounds for right-censored data were applied within sensitivity analyses. (Note, a full summary of the sensitivity analyses conducted is provided in section 5.3.1.2.5). As previously mentioned, specifying an upper bound of 1,000 µg/g for right-censored data does not mean that the resulting parametric distributions are truncated at this point (see the associated simulated distributional plots provided in Appendix I.2 which show upper distribution tails extending beyond this point).

Table 5-2 shows the AIC and BIC results for each population, using an upper bound for the right-censored data region of 1,000 µg/g. Associated tables using the alternative upper bound values are provided in Appendix I.1; Appendix I.2 also provides a series of density plots showing the simulated probability density distributions for each of the parameterisations provided in Table 5-2 (based on n=10,000 draws from each distribution). Appendix I.3 further provides example R code for fitting parametric distributions to censored data using the 'fitdistrplus' package, based on a hypothetical dataset of test values.

Within Table 5-2, the optimal parameterisation for each population (according to the AIC and BIC results) is highlighted in blue, and the distributional specifications provided by the 'fitdistrplus' package for each parameterisation are reported under the 'Parameterisation' heading (e.g. 'meanlog' and 'sdlog' for the lognormal parameterisation). A face validity metric is also provided, equal to the proportion of 10,000 simulations from each distribution falling equal to or above 50 µg/g (i.e. the NICE FC pathway cut-off threshold). This proportion can be compared to that observed in the YFCCP dataset FC1 data (shown in the final column of the table), to assess how closely the diagnostic accuracy of the NICE FC pathway as estimated from the simulated parametric distributions would be expected to match that based on the empirical YFCCP dataset FC1 values. If the specified

distribution over-estimates the proportion of IBS patients above the cut-off threshold, for example, then the pathway's baseline specificity will be underestimated.

Table 5-2.[34] NICE FC pathway: AIC and BID criteria for FC1 parametric distributions (upper bound for right-censored FC1 data = 1,000 µg/g)

Subgroup	Parameterisation	AIC	BIC	% values ≥ 50 µg/g (simulated data)	% values ≥ 50 µg/g (YFCCP FC1 data)
IBS FC1	**Lognormal** [meanlog = 3.706564; sdlog = 1.22697]	8592.265	8601.809	44.1%	40.3%
	Weibull [shape = 0.7619224; scale = 73.83344]	8728.559	8738.103	48.0%	
	Gamma [shape = 0.6889856; rate = 0.007743206]	8781.936	8791.479	49.2%	
	Normal [mean = 86.97938; SD = 134.9182]	10375.290	10384.830	59.3%	
IBD FC1	**Lognormal** [meanlog = 6.007434; sdlog = 0.8615051]	615.288	620.002	99.5%	96.2%
	Weibull [shape = 1.74992; scale = 589.0402]	588.486	593.199	98.8%	
	Gamma [shape = 2.025055; rate = 0.003819662]	596.323	601.036	94.8%	
	Normal [mean = 526.4776; SD = 291.2407]	588.660	593.373	98.4%	

[34] Note: density plots for these parameterisations are provided in Appendix I.2.

Based on the reported AIC and BIC results, the lognormal parameterisation was selected for the IBS FC1 subgroup and the Weibull parameterisation was selected for the IBD FC1 subgroup. Within the parametric method base case analysis, $FC1_{true}$ values within the error model simulation (n=10,000) were randomly drawn from the lognormal and Weibull FC1 distributions according to the observed YFCCP IBD prevalence (i.e. 8.2% of simulations were drawn from the IBD Weibull distribution [n=9,180]; 91.8% from the IBS lognormal distribution [n=820]). Sensitivity analyses were also conducted applying each of the alternative parametric distributions listed in Table 5-2 (see section 5.3.1.2.5).

With regards to censored data, as well as exploring different values for the upper limit applied to the right-censored data region, a complete case analysis was also conducted. This uses the 'fitdistr' function within the 'fitdistrplus' package, which similarly uses a MLE process to derive data parameterisations – but in this case ignores any censored data (277). As with the bootstrap method, the complete case analysis should be interpreted with caution given the proportion of data discarded.

5.3.1.2.3 Uncertainty

A certain degree of "noise" exists in the simulation results, due to both the process of randomly sampling FC values, and applying a random imprecision factor within the error model. In both the bootstrap and parametric base case analyses, a pragmatic approach was adopted to smooth the simulation output, based on calculating moving averages of the diagnostic accuracy results. For the majority of sensitivity and specificity values, moving average values were based on a moving window of 10 using central positioning; i.e. for any given point (e.g. sensitivity result), the average value was based on an average of 10 points including 5 points either side of the value position. Towards the edges of the data (where a window of 10 was not possible) a moving window of 5 was used. At the very edges of the data (where central positioning was not possible), moving averages were based on a window of 5 using left/right positioning as necessary.

An alternative, more robust (but computationally expensive) method is to run a greater number of simulations. An additional sensitivity analysis was therefore conducted for both sampling methods, based on running 100,000 simulations (compared to 10,000 in the base case). For these analyses, results are reported

based on the “noisy” sensitivity and specificity values (i.e. without applying the smoothing algorithm), to enable assessment of the agreement between these results and the pragmatic smoothing algorithm applied in the base case.

For the parametric method, additional uncertainty stems from the parametric specifications selected. In the base case analysis, distributions were modelled based on the mean MLE parameters provided by the ‘fitdistrplus’ package. $FC1_{true}$ values for the IBS population, for example, were derived by taking random draws from a lognormal distribution with ‘meanlog’ = 3.706564 and ‘sdlog’ = 1.22697 (as per the specifications provided in Table 5-2). The uncertainty inherent in the ‘meanlog’ and ‘sdlog’ parameters, due to the fact they are derived on sample data, is not captured using this approach.

It is possible to derive a series of stochastic values for each of the distribution parameters listed in Table 5-2, using an inbuilt function within the ‘fitdistrplus’ package called ‘bootdistcens’. This function produces a user-defined number of stochastic values for the distributional parameters, by iteratively resampling the provided sample data (using bootstrap sampling) and rerunning the MLE procedure for each bootstrap sample (277). The resulting set of stochastic parameter values can then be iteratively applied within the error model simulation by adding an *outer loop* to the simulation process: that is, the base case error model simulation (n=10,000 simulations for each bias and imprecision pair) is run by drawing from distributions defined according to the first set of stochastic parameter estimates, then again for the second, then the third, and so on. For each bias and imprecision pair assessed, the diagnostic accuracy results are then based on an average value over both the inner and outer simulation loops.

Given the additional computational burden associated with adding an outer simulation loop to the analysis, a sensitivity analysis was run using 1,000 inner simulation loops (i.e. a reduced version of the base case) and 1,000 outer loop simulations. This results in 1 million simulations for each bias and imprecision pair assessed. As before, the results of this analysis are reported based on the “noisy” values of sensitivity and specificity, thus allowing an overall comparison of: (i) applying the pragmatic smoothing algorithm (base case analysis); (ii) increasing the sampling number to 100,000 (sensitivity analysis); and (iii) increasing the sampling number and accounting for parametric uncertainty

(sensitivity analysis). Results for all other sensitivity analyses were based on applying the smoothing algorithm, as in the base case.

A final sensitivity analysis was conducted for the bootstrap sampling method, in which the bootstrap sampling process was removed entirely. That is, the error model was applied directly to the YFCCP dataset FC1 values (n=951) alone. This analysis enables an assessment of whether or not increasing the simulation dataset via sampling alters the results, compared to simply applying the error model directly to the empirical dataset.

5.3.1.2.4 Outputs

For each sampling method, the simulation process was repeated for CV% values ranging from 0-100% in 0.5% increments, and for bias ranging from -100 to +100 µg/g in 1 µg/g increments, resulting in a total of 40,401 (201x201) bias and imprecision pairs. For each CV and bias pair, the base case simulation process produces 10,000 $FC1_{sim}$ values, from which the NICE FC pathway diagnostic accuracy was calculated (resulting in 40,401 diagnostic sensitivity and specificity values).

For the base case analyses, the simulation results were illustrated using contour plots, which show how the sensitivity and specificity values change over the joint space of bias (x-axis) and CV (y-axis) inputs (as previously described in Chapter 3, section 3.4.1, Figure 3-4). These plots were used to provide an initial assessment of the robustness of the NICE FC pathway's diagnostic accuracy to increasing imprecision and bias. In addition the contour plots were used to illustrate a novel concept: "*acceptable regions"* of bias and imprecision (see section 5.3.2.2.1, Figure 5-5 for an example). The acceptable region highlights the area of the contour plot which meets a given diagnostic accuracy requirement: in the first instance this requirement was defined according to the lower 95% CIs achieved within the NICE FC pathway baseline diagnostic accuracy assessment (i.e. sensitivity ≥ 88% and specificity ≥ 56%; as reported in section 5.3.2.1). A lower requirement was also explored, reducing these values arbitrarily by 10% (i.e. sensitivity ≥ 78% and specificity ≥ 46%) – this second criterion provides an illustration of how relaxing the diagnostic accuracy requirement alters the acceptable region. Note that the concept of acceptable regions is further extended to cost-effectiveness outcomes in Chapter 6.

In line with previous studies that have used contour plots to present error model simulation results, TE% bands were also superimposed onto these figures (with TE = 1.96*CV% + bias) (130, 133, 142, 146, 174). Whilst the use of TE remains contentious within the clinical chemistry community (as discussed in section 1.2.3), this metric does provide a useful means of summarising the overall level of error occurring at any given point on the contour plot. Using this approach, a further novel concept is presented: the *maximum allowable TE* (TE_{max}), defined as the highest TE band completely contained within a specified acceptable region. The TE_{max} metric provides a summary statistic with which the acceptable region can be described. A full discussion of the acceptable regions and TE_{max} concepts is provided in section 5.5.2.

For the simulation sensitivity analyses, results are presented in a table (see section 5.3.2.2.2, Table 5-6) including the following information: (i) the diagnostic sensitivity and specificity results at the (0,0) point (i.e. at bias = 0 and CV% = 0); (ii) TE_{max} for the two acceptable regions outlined earlier in this section; (iii) the range of acceptable bias observed at CV = 0%, for each acceptable region (i.e. the width of each acceptable region at zero added CV%); and (iii) the range of acceptable CV% observed at bias = 0, for each acceptable region (i.e. the height of each acceptable region at zero added bias).

5.3.1.2.5 Analysis summary

The simulation processes applied in the NICE FC pathway base case analyses are summarised below.

Bootstrap sampling method:

i. Left- and right-censored FC values in the YFCCP dataset (n=951) are replaced with their associated limit values (10 and 600 µg/g respectively);
ii. An expanded bootstrap dataset (n=10,000) is generated by sampling with replacement from the YFCCP dataset rows;
iii. For each $FC1_{true}$ value in the bootstrap dataset, the error model is applied to generate $FC1_{sim}$ values at a given level of bias and CV%;
iv. Diagnostic accuracy of the NICE FC pathway including additional FC bias and CV is calculated by comparing diagnoses based on the $FC1_{sim}$ values

(using a 50 µg/g cut-off threshold) with patients' clinical diagnoses in the bootstrap dataset; and

v. Steps (iii) and (iv) are repeated for a range of bias (-100 to +100 in 1µg/g increments) and CV% (0 to 100% in 0.5% increments) values.

Parametric sampling method:

i. Using the parameter specifications provided in Table 5-2, a total of 10,000 $FC1_{true}$ values are drawn from (a) the FC1 lognormal parametric distribution for IBS patients (n=9180), and (b) the FC1 Weibull parametric distribution for IBD patients (n=820);

ii. For each $FC1_{true}$ value in (i), the error model is applied to generate $FC1_{sim}$ values at a given level of bias and CV%;

iii. Diagnostic accuracy of the NICE FC pathway including additional FC bias and CV is calculated by comparing diagnoses based on the $FC1_{sim}$ values (using a 50 µg/g cut-off threshold) with patients' clinical diagnoses (according to the population distribution from which simulations were drawn); and

iv. Steps (ii) and (iii) are repeated for a range of bias (-100 to +100 in 1µg/g increments) and CV% (0 to 100% in 0.5% increments) values.

A summary of the sensitivity analyses conducted for each sampling method is provided in Table 5-3.

Table 5-3. NICE FC pathway: sensitivity analyses conducted

Code	Analysis summary	Analysis details
Bootstrap sampling method analyses		
[1.0]	Bootstrap method base case analysis	Left- and right-censored data (recorded as "<10" and ">600" in the YFCCP database, respectively) replaced with limit values (10 and 600 µg/g respectively); error model based on 10,000 bootstrap samples; smoothing algorithm applied to sensitivity and specificity results
[1.1]	Left-censored data = 5 µg/g; right-censored data =750 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 750 µg/g
[1.2]	Left-censored data = 5 µg/g; right-censored data =900 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 900 µg/g
[1.3]	Left-censored data = 5 µg/g; right-censored data = 1200 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 1200 µg/g
[1.4]	Left-censored data = 5 µg/g; right-censored data = 1800 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 1800 µg/g
[1.5]	Complete case analysis	Censored data excluded
[1.6]	Raw data only (n=951; no bootstrap sampling)	No bootstrap sampling conducted: error model applied to YFCCP database (n=951) FC1 values alone
[1.7]*	Noisy results (no smoothing algorithm)	No smoothing algorithm applied to the sensitivity and specificity results
[1.8]*	100,000 samples	Error model based on 100,000 bootstrap samples; no smoothing algorithm applied
Parametric sampling method analyses		

[2.0]	Parametric method base case analyses	R 'fitdistrplus' function used to fit parametric distributions to required patient subgroups (IBS FC1, IBD FC1) over complete and censored data regions via maximum likelihood estimation; lognormal distribution used for IBS subgroup; Weibull distribution used for IBD subgroup; region for left-censored data set as 0-10 µg/g; region for right-censored data set as 600-1000 µg/g; error model based on 10,000 draws across the IBS FC1 and IBD FC1 distributions; smoothing algorithm applied to sensitivity and specificity results
[2.1]	Right-censored data region set to 600-2000 µg/g	Upper bound for right censored data region set to 2,000 µg/g within the R 'fitdistcens' function
[2.2]	Right-censored data region set to 600-3000 µg/g	Upper bound for right censored data region set to 3,000 µg/g within the R 'fitdistcens' function
[2.3]	Complete case analysis	Censored data excluded
[2.4]	Lognormal parameterisation	Lognormal parameterisations used for both patient subgroup distributions
[2.5]	Weibull parameterisation	Weibull parameterisations used for both patient subgroup distributions
[2.6]	Gamma parameterisation	Gamma parameterisations used for both patient subgroup distributions
[2.7]	Normal parameterisation	Normal parameterisations used for both patient subgroup distributions
[2.8]*	Noisy results (no smoothing algorithm)	No smoothing algorithm applied to the sensitivity and specificity results
[2.9]*	100,000 samples	Error model based on 100,000 draws across the IBS FC1 and IBD FC1 distributions; no smoothing algorithm applied
[2.10]*	Sampling accounting for parametric uncertainty: inner simulations = 1,000 x 40,401; outer simulations = 1,000	"Inner" simulation: base case error model process using 1,000 draws across the IBS FC1 and IBD FC1 distributions for each bias and CV% pair (n=40,401). "Outer" simulation: for each outer simulation, a different stochastic value for the lognormal (IBS) and Weibull (IBD) distribution parameter estimates was drawn and applied across the 1,000*40,401 inner simulations. The 1,000 outer simulation stochastic parametric values were derived using the 'bootdistcens' function from the 'fitdistrplus' package (277).
*These analyses are based on "noisy" results i.e. with no smoothing algorithm applied		

5.3.2 Results

5.3.2.1 Baseline diagnostic accuracy

The baseline diagnostic accuracy results for the NICE FC pathway are summarised in Table 5-4. According to this pathway, 75/78 patients with IBD would have been correctly identified, giving a high sensitivity of 96.2% (95% CI: 0.88 to 0.99); whilst 521 out of 873 of patients with IBS would have been correctly identified, giving a low specificity of 59.7% (95% CI: 0.56 to 0.63).

Table 5-4. NICE FC pathway: baseline diagnosis accuracy results

		Clinical diagnosis		
		IBD	**IBS**	
YFCCP diagnosis	**IBD**	True positives: n= 75	False positives: n= 352	**PPV = 75/427 = 17.6%**
	IBS	False negatives: n=3	True negatives: n= 521	**NPV = 521/524 = 99.4%**
		Sensitivity = 75/78 = 96.2%	**Specificity = 521/873 = 59.7%**	

The low specificity of this pathway is further illustrated by looking at the distribution of FC1 results across each population within the YFCCP dataset. Figure 5-3 provides count plots for each population, which clearly demonstrates the fact that using a cut-off threshold of 50 µg/g faeces misses a high proportion of IBS patients (shown in the top panel of the figure) who fall above this threshold line. Note that, for the purpose of this figure, left-censored data have been re-coded as 10 µg/g and right-censored data have been re-coded as 600 µg/g, resulting in corresponding peaks at these points within each count plot.

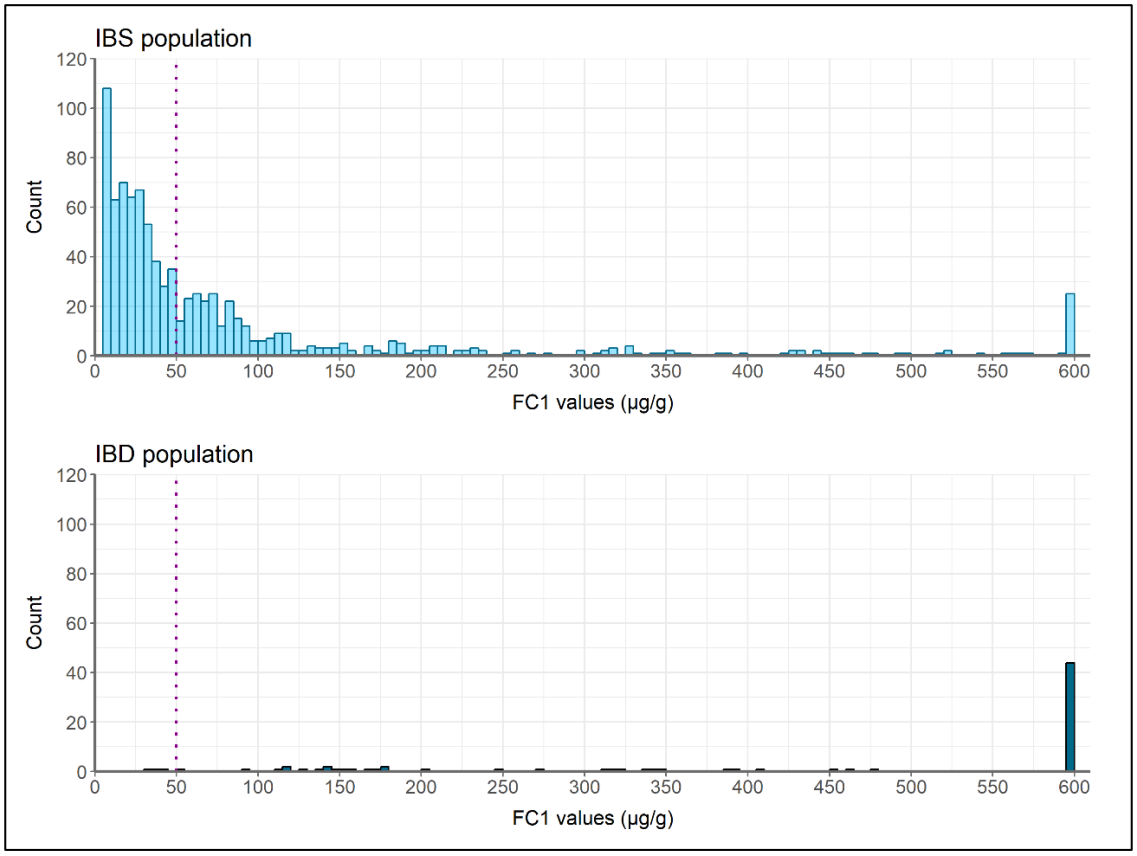


Figure 5-3. NICE FC pathway: count plots showing the distribution of FC1 values for IBS and IBD patients within the YFCCP database

5.3.2.2 Simulated diagnostic accuracy

5.3.2.2.1 Base case analysis

Figure 5-4 provides the base case diagnostic accuracy contour plots for the NICE FC pathway. Note that all plots presented in this section are based on the simulation results with the smoothing algorithm applied (as described in section 5.3.1.2.3). “Noisy” versions of these plots are provided in Appendix J.

The contour plots provide an illustrative tool with which the robustness of the pathway’s diagnostic accuracy to increased FC bias and imprecision may be inspected. For example, consider the specificity contours for the bootstrap method (Figure 5-4, panel A). Starting at the baseline (0,0) point (i.e. zero added bias and imprecision), if we hold added imprecision at 0% and introduce additional positive bias (move horizontally to the right) then we rapidly pass through progressively lower specificity contours (i.e. the pathway specificity is volatile to positive bias), whilst if we introduce negative bias we gradually move through higher specificity contours (specificity is robust to negative bias). Second,

if we hold bias at 0 µg/g and introduce imprecision (move vertically), then we do not pass through any specificity contours (i.e. the pathway specificity is robust to imprecision). Applying the same process to the sensitivity contours, we observe that the sensitivity of the pathway is unaffected by positive bias, whilst negative bias and imprecision result in a gradual decrease in sensitivity.

The same general pattern of results was observed with the parametric method (Figure 5-4, panel B). However, this sampling method produces slightly higher sensitivity (97.7%) and lower specificity (56.5%) at the (0,0) point, compared to the baseline diagnostic accuracy values (sensitivity = 96.2%; specificity = 59.7%). Running the simulation with zero imprecision and bias should produce the same results as the baseline diagnostic accuracy assessment; the fact that this is not the case in this analysis suggests that the parametric sampling method provides a poor fit to the data. The bootstrap method performs better in this respect, reporting 95.9% sensitivity and 60.0% specificity at the (0,0) point.

Figure 5-5 illustrates the same results, this time highlighting the acceptable regions of bias and imprecision relating to an assumed minimum diagnostic accuracy requirement of sensitivity ≥88% and specificity ≥56% (the lower 95% CIs from the baseline diagnostic accuracy evaluation [section 5.3.2.1]). TE% bands are also overlaid onto these plots, to indicate the relationship between the acceptable regions of bias and imprecision and the TE summary metric. Based on these results, TE_{max} (i.e. the maximum TE band completely contained within the acceptable region) is equal to 5% when using the bootstrap method. For the parametric method, due to the fact that this method produces a lower baseline specificity estimate (below the 56% minimum specificity requirement), the acceptable region is offset from (0,0) and there is subsequently no acceptable TE value contained within the acceptable region (Figure 5-5, panel B).

Figure 5-6 highlights the acceptable regions relating to a lower minimum diagnostic accuracy requirement of sensitivity ≥78% and specificity ≥46% (i.e. 10% below the lower 95% CI's from the diagnostic accuracy evaluation). In this case, TE_{max} = 15% with the bootstrap method or 13% with the parametric method. Further discussion of these plots is provided in section 5.5.2.

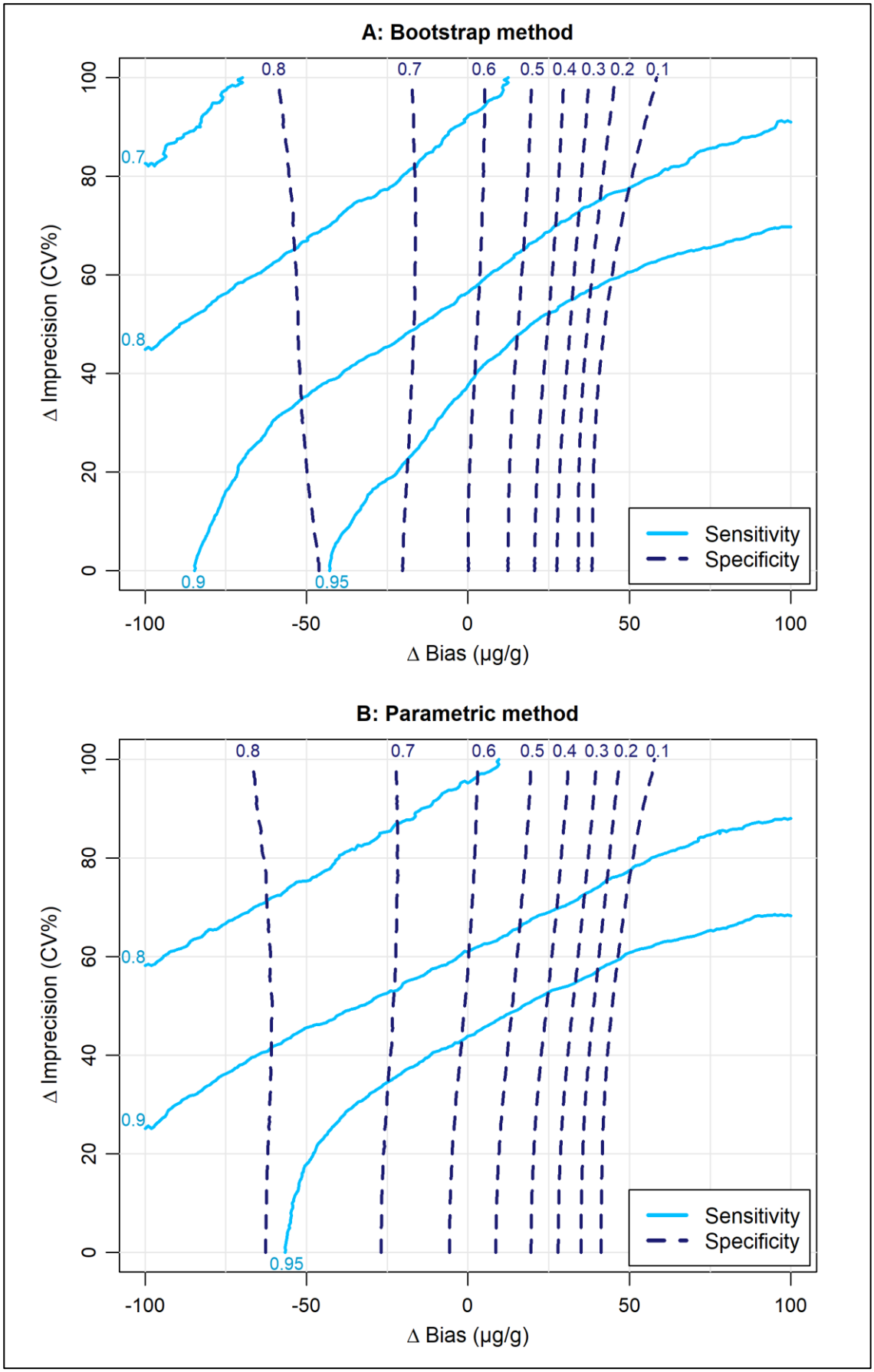


Figure 5-4. NICE FC pathway: base case diagnostic accuracy contour plots

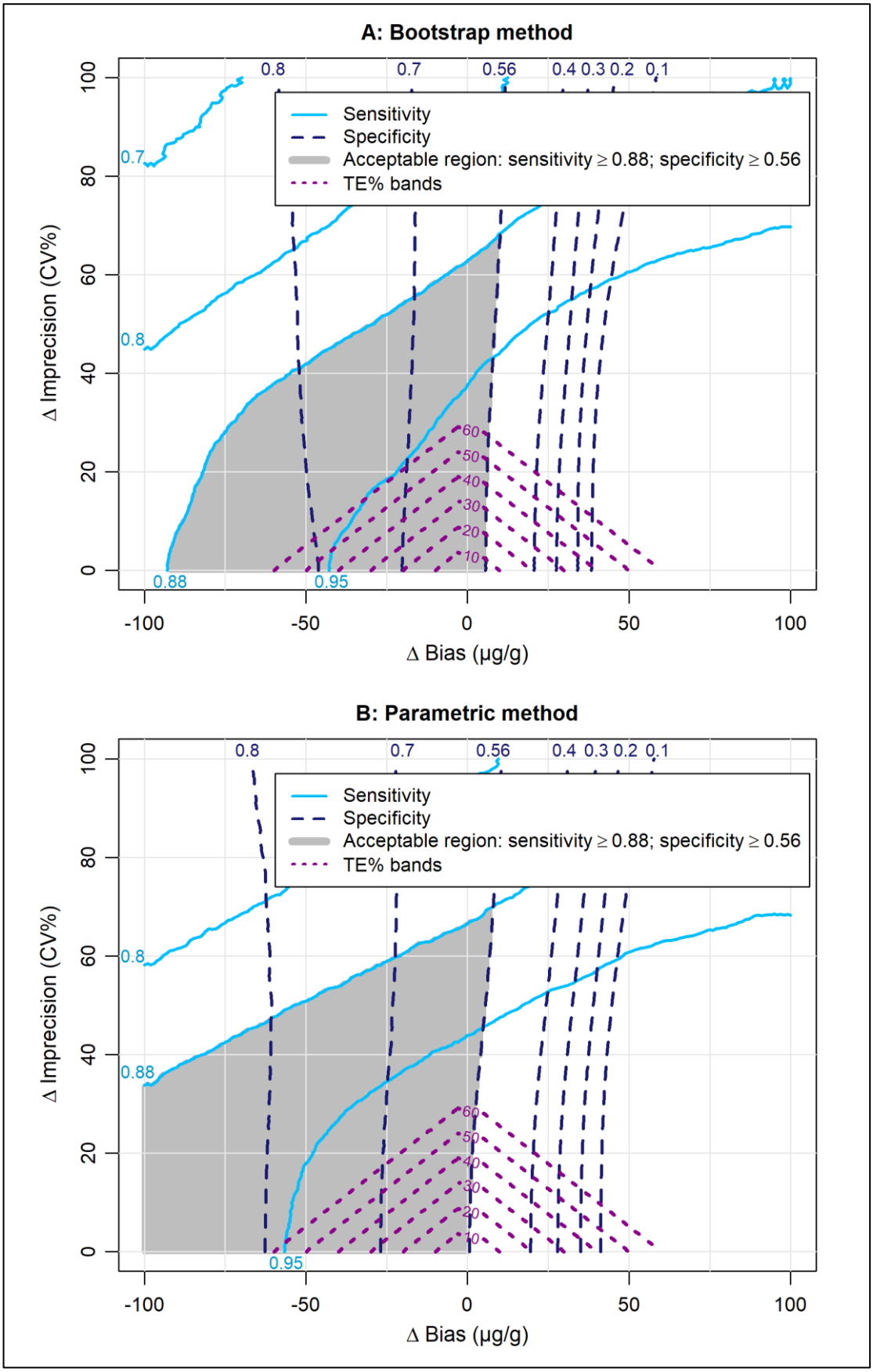


Figure 5-5. NICE FC pathway: base case diagnostic accuracy contour plots showing the acceptable region (maintaining sensitivity ≥0.88 and specificity ≥0.56) and TE% bands

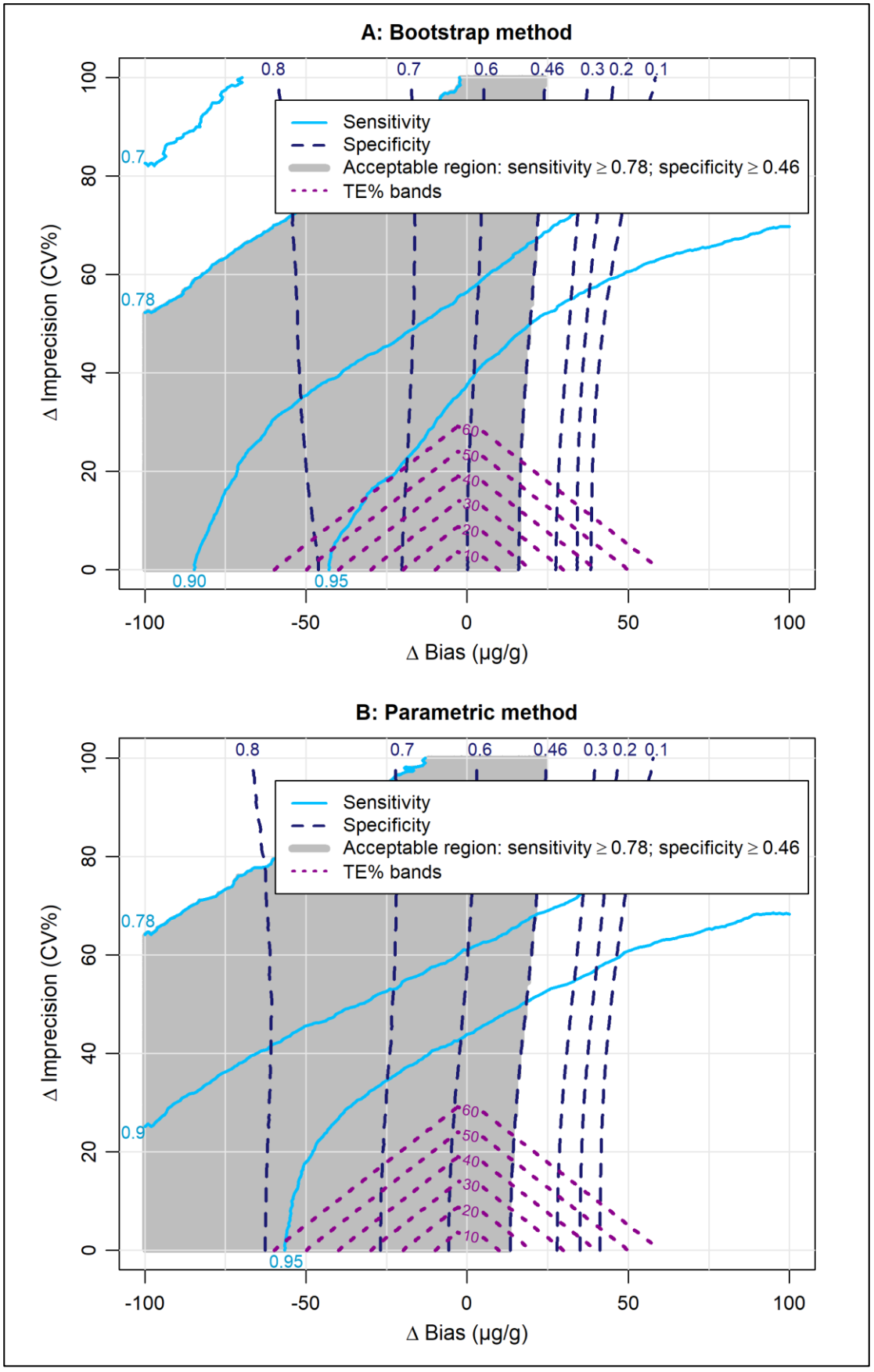


Figure 5-6. NICE FC pathway: base case diagnostic accuracy contour plots showing the acceptable region (maintaining sensitivity ≥0.78 and specificity ≥0.46) and TE% bands

Table 5-5 summarises the key findings from the above contour plots in tabular format. The first set of results provided in this table relate to the simulated diagnostic sensitivity and specificity results when setting added impression and bias to zero. These values can be compared to the empirical baseline diagnostic accuracy results reported in section 5.3.2.1, as a means of assessing the internal validity of the simulated diagnostic accuracy outcomes. That is, the simulated sensitivity and specificity values produced within the error model simulation when setting added imprecision and bias to zero, can be compared to the baseline diagnostic accuracy values based on the YFCCP dataset (sensitivity = 0.962, specificity = 0.597), to assess how closely the simulated data matches the empirical data. As previously discussed, it is evident from these results that the bootstrap method provides a better fit to the empirical data. Further discussion of these results is provided in section 5.5.2.

Table 5-5. NICE FC pathway: simulated diagnostic accuracy base case results

Sampling method	Diagnostic accuracy at bias=0 & CV=0%		Acceptable region 1: sensitivity ≥0.88; specificity ≥0.56			Acceptable region 2: sensitivity ≥0.78; specificity ≥0.46		
	Sensitivity	Specificity	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0
Bootstrap method	0.959	0.600	5%	-92 to 5	0 to 63%	15%	-100 to 15	0 to 100%
Parametric method	0.977	0.565	0%	-100 to 0	0 to 67%	13%	-100 to 13	0 to 100%

5.3.2.2.2 Sensitivity analyses

Table 5-6 reports the results of the sensitivity analyses conducted for the NICE FC pathway assessment. Note that TE_{max} results reported as 'NA' in this table indicate analyses were the acceptable region was offset from the (0,0) point (as opposed to TE_{max} = 0%, which indicates that the acceptable region touched, but did not pass, the baseline point). In addition, any 'Range of CV% at bias=0' results reported as 'NA' similarly indicates cases where the acceptable region did encompass the zero bias line.

Using the bootstrap sampling method, sensitivity analyses exploring alternative specifications for dealing with censored data had no measurable impact on the results (Table 5-6, analyses 1.1-1.5). This is excluding the complete case analysis, which significantly decreases the diagnostic accuracy results and restricts the associated acceptable regions across all of the evaluations. This is due to the fact that this approach discards a high proportion of data (particularly within the IBD population). When using the parametric method, the two analyses exploring the use of higher upper bound values for the right-censored data regions had marginal impact, producing slightly lower baseline sensitivity and specificity values. In the analysis applying an upper bound of 3,000 µg/g limit, the first acceptable region was slightly restricted (Table 5-6, analyses 2.1-2.2).

When using the parametric method, application of the lognormal, Weibull, Gamma and normal distributions produced increasingly lower sensitivity and specificity values and restricted acceptable regions compared to the base case analysis (Table 5-6, analyses 2.4-2.7). The normal distribution was a particularly poor fit to the data, as evident in the distribution plots provided in Appendix I.2. Finally, all of the sensitivity analyses exploring sampling uncertainty (Table 5-6, analyses 1.6-1.8 and 2.8-2.9) and parametric uncertainty (analysis 2.10), had little impact on the base case results.

Full discussion of the NICE FC pathway simulation results is provided in section 5.5.

Table 5-6. NICE FC pathway: simulated diagnostic accuracy sensitivity analysis results

Analysis		Diagnostic accuracy at bias=0 & CV=0%		Acceptable region 1: sensitivity ≥0.88; specificity ≥0.56			Acceptable region 2: sensitivity ≥0.78; specificity ≥0.46		
		Sensitivity	Specificity	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0
Bootstrap sampling method									
Base case analysis	**[1.0] Bootstrap method**	0.959	0.600	5%	-92 to 5	0 to 63%	15%	-100 to 15	0 to 100%
Sensitivity analyses: FC censored data handling	[1.1] Left-censored data = 5 µg/g; right-censored data = 750 µg/g	0.959	0.600	5%	-92 to 5	0 to 64%	15%	-100 to 15	0 to 100%
	[1.2] Left-censored data = 5 µg/g; right-censored data = 900 µg/g	0.959	0.600	5%	-92 to 5	0 to 64%	15%	-100 to 15	0 to 100%
	[1.3] Left-censored data = 5 µg/g; right-censored data = 1200 µg/g	0.959	0.600	5%	-92 to 5	0 to 65%	15%	-100 to 15	0 to 100%
	[1.4] Left-censored data = 5 µg/g; right-censored data = 1800 µg/g	0.959	0.600	5%	-92 to 5	0 to 65%	15%	-100 to 15	0 to 100%
	[1.5] Complete case analysis	0.896	0.564	0%	-6 to 0	0 to 41%	11%	-71 to 11	0 to 75%
Sensitivity analyses: sampling uncertainty	[1.6] Raw data only (n=951; no bootstrap sampling)	0.957	0.592	4%	-86 to 4	0 to 62%	15%	-100 to 15	0 to 100%
	[1.7]* Noisy results (no smoothing algorithm)	0.961	0.605	5%	-94 to 5	0 to 69%	15%	-100 to 15	0 to 99%
	[1.8]* 100,000 samples	0.958	0.599	4%	-88 to 4	0 to 62%	15%	-100 to 15	0 to 99%

Parametric sampling method									
Base case analysis	**Parametric method**	0.977	0.565	0%	-100 to 0	0 to 67%	13%	-100 to 13	0 to 100%
Sensitivity analyses: censored data handling	[2.1] Right-censored data region: 600-2000 µg/g	0.967	0.562	0%	-100 to 0	0 to 65%	13%	-100 to 13	0 to 100%
	[2.2] Right-censored data region: 600-3000 µg/g	0.963	0.555	- NA -	-88 to -1	0 to 65%	12%	-100 to 12	0 to 100%
	[2.3] Complete case analysis	0.950	0.518	- NA -	-29 to -5	- NA -	6%	-62 to 6	0 to 79%
Sensitivity analyses: parameterisation	[2.4] Lognormal parameterisation	0.994	0.564	0%	-100 to 0	0 to 69%	13%	-100 to 13	0 to 100%
	[2.5] Weibull parameterisation	0.989	0.521	- NA -	-100 to -9	51 to 69%	10%	-100 to 10	0 to 100%
	[2.6] Gamma parameterisation	0.986	0.489	- NA -	-100 to -15	51 to 69%	4%	-100 to 4	0 to 100%
	[2.7] Normal parameterisation	0.951	0.391	- NA -	-100 to -58	- NA -	- NA -	-100 to -24	56 to 97%
Sensitivity analyses: sampling and parametric uncertainty	[2.8]* Noisy results (no smoothing algorithm)	0.978	0.568	1%	-100 to 1	0 to 72%	13%	-100 to 13	0 to 100%
	[2.9]* 100,000 samples	0.988	0.563	0%	-100 to 0	0 to 70%	13%	-100 to 13	0 to 100%
	[2.10]* Sampling accounting for parametric uncertainty: inner simulations = 1,000 x 40,401; outer simulations = 1,000	0.985	0.567	1%	-100 to 1	0 to 70%	14%	-100 to 13	0 to 100%

5.4 Part 2: YFCCP evaluation

This section presents the methods and results for the diagnostic accuracy evaluation of the YFCCP. All data analysis and simulation modelling conducted within this assessment were performed using R software (version 3.4.3) (272).

5.4.1 Methods

5.4.1.1 Baseline diagnostic accuracy

The baseline diagnostic accuracy of the YFCCP was determined using both the FC1 and FC2 results within the YFCCP database, as per the YFCCP protocol. That is, patients were diagnosed as having suspected IBS if either their FC1 result was <100 µg/g, or if their FC1 result was ≥100 µg/g and their FC2 result was <100 µg/g; and patients were diagnosed as having suspected IBD if their FC1 result was ≥100 µg/g and their FC2 result was ≥100 µg/g. By comparison with patients' recorded clinical diagnoses (summarised in Table 5-1), each FC diagnosis was classified as true positive, true negative, false positive or false negative; and diagnostic sensitivity and specificity was calculated based on the proportion of results falling into each of these categories (as illustrated in Appendix C). Count plots, showing the distribution of patient FC1 and FC2 values across each of the IBS and IBD populations, were also produced.

5.4.1.2 Simulated diagnostic accuracy

In contrast to single-test strategies, evaluation of repeat-test strategies – such as the YFCCP – requires either multiple applications of the error model, or an alternative simulation approach. As a first step, it is useful to consider the different factors which may introduce longitudinal variation in test results over time. For example, for a given patient, variation in longitudinal FC results (e.g. FC1 vs. FC2 results) may arise from three primary factors:

1) Within-person biological variation (i.e. natural fluctuation in an individual's production of calprotectin);
2) Measurement uncertainty (including imprecision and bias resulting from pre-analytical and analytical processes); and
3) Disease status and/or activity.

If no FC2 data were available for this analysis, or if the aim was to explore alternative repeat-test strategies (e.g. applying different testing frequencies and/or time intervals), then each of the above factors would need to be explicitly modelled in order to estimate the trajectory of serial test values. In this case study however, the aim was to assess the impact of increasing measurement uncertainty on an established clinical pathway (the YFCCP) with a fixed diagnostic protocol, for which sample data is available. The approach taken in this case therefore was to directly apply the error model to the FC1 and FC2 data available. Thus, rather than explicitly modelling each of the factors contributing to longitudinal variation (which would require reliable evidence on each of the factors, not yet available for FC), these factors were instead indirectly captured within the YFCCP database FC1 and FC2 values.

The overall simulation process adopted for the YFCCP analysis is summarised in Figure 5-7 below.

i. A sample of $FC1_{true}$ values is assigned;
ii. For each $FC1_{true}$ value, the addition of bias and imprecision is simulated according to the specified error model to generate $FC1_{sim}$ values e.g.:

$$\mathbf{FC1_{sim} = FC1_{true} + [FC1_{true} \times \mathit{N}(0,1) \times CV] + Bias} \qquad \mathbf{(5.3)}$$

iii. For all $FC1_{sim}$ values ≥100 µg/g, an associated sample of $FC2_{true}$ values is assigned;
iv. For each $FC2_{true}$ value, the addition of measurement uncertainty is simulated according to the specified error model to generate $FC2_{sim}$ values e.g.:

$$\mathbf{FC2_{sim} = FC2_{true} + [FC2_{true} \times \mathit{N}(0,1) \times CV] + Bias} \qquad \mathbf{(5.4)}$$

v. The diagnostic accuracy of the YFCCP including additional imprecision and bias is calculated by comparing diagnoses based on the $FC1_{sim}$ and $FC2_{sim}$ values (using the YFCCP diagnostic protocol) with patients' clinical diagnoses;
vi. Steps (i) to (v) are repeated for a range of CV and bias values.

Figure 5-7. YFCCP: error model simulation approach required for a repeat-test strategy

As in the NICE FC pathway evaluation, two approaches were explored in this assessment for sampling $FC1_{true}$ and $FC2_{true}$ values from the YFCCP dataset – the 'bootstrap method' (discussed in section 5.4.1.2.1 below), and the 'parametric method' (discussed in section 5.4.1.2.2). An additional analysis, based on applying the error model directly to the YFCCP dataset FC1 and FC2 values (i.e. with no sampling process applied), was also considered within a sensitivity analysis under the bootstrap method (see section 5.4.1.2.3).

5.4.1.2.1 Bootstrap sampling method

For the YFCCP evaluation bootstrap method, the same initial process as applied within the NICE FC pathway evaluation was undertaken: a bootstrap simulation dataset was generated by drawing 10,000 bootstrap samples from the YFCCP dataset, with each sampled row now including patients' FC1 *and* FC2 values, as well as their final clinical diagnoses. The FC1 values within each bootstrap were then used as $FC1_{true}$ values within the first error model application (i.e. step (ii) in Figure 5-7). Additional simulation was then required to generate $FC2_{sim}$ values: in this case, for all $FC1_{sim}$ values returned as ≥100 µg/g from the first error model application, the associated FC2 values from the corresponding rows of the bootstrap simulation dataset were then used as the $FC2_{true}$ values within the second error model application (step (iv) in Figure 5-7), thereby maintaining the within-patient correlation between FC1 and FC2 values. The diagnostic accuracy assessment was then based on a comparison of FC diagnoses using the $FC1_{sim}$ and $FC2_{sim}$ values (according to the YFCCP diagnostic protocol), with linked data on patients' clinical diagnoses within the bootstrap simulation dataset.

With regards to censored data, the same method as employed in the NICE FC pathway evaluation (see section 5.3.1.2.1) was again applied for the YFCCP evaluation – i.e. all left-censored data (now including FC1 and FC2 values) were replaced with a value of 10 µg/g and all right-censored data were replaced with 600 µg/g. In addition, the same set of sensitivity analyses (listed in section 5.4.1.2.5) were conducted to explore the impact of the censored data substitution values. Note that the complete case sensitivity analysis is again expected to significantly bias the results, with the removal of FC1 and FC2 censored data in this case more than halving the IBD prevalence in the YFCCP dataset from 8.2% to 3.5%. This analysis should therefore be considered with caution.

A further consideration within the YFCCP evaluation relates to the conduct of FC2 testing. Within the YFCCP dataset, 5 IBD patients and 701 IBS patients did not have an FC2 test conducted due to having an FC1 value <100 µg/g (i.e. no FC2 test was required in these cases according to the YFCCP protocol). A further 10 IBD patients did not have an FC2 test conducted due to being directly referred to secondary care (i.e. non-compliant referrals). Under the bootstrap method therefore, depending on the level of measurement uncertainty applied within the first error model application, a number of $FC1_{sim}$ values returned as ≥100 µg/g could have a missing $FC2_{true}$ value within that row of the bootstrap dataset.

In the base case analysis, required FC2 values not available within the bootstrap dataset were imputed by resampling from the population-specific YFCCP FC2 data – i.e. if the missing $FC2_{true}$ value related to an IBS patient, then the $FC2_{true}$ value was imputed by taking a random draw from the available YFCCP IBS FC2 data (n=172); likewise if the missing value related to an IBD patient, then imputations were drawn from the available YFCCP IBD FC2 data (n=63).

An alternative approach to imputing missing FC2 values was considered within a sensitivity analysis. For this analysis, required $FC2_{true}$ values were imputed by first sampling with replacement from available (population-specific) data on the *between-test proportional differences (FC_{diff}),* defined as FC_{diff} = (FC2 – FC1)/FC1. Within the error-model simulation, for all $FC1_{sim}$ values returned as ≥ 100 µg/g, $FC2_{true}$ values (where unavailable in the bootstrap dataset) were derived by taking the associated $FC1_{true}$ value and applying a randomly drawn FC_{diff} value (i.e. $FC2_{true} = FC1_{true} + FC1_{true}*FC_{diff}$). Note that proportional (rather than absolute) differences were used to avoid generating negative $FC2_{true}$ values. Count plots, illustrating the population-specific distributions of FC_{diff} values derived from the YFCCP dataset, are provided in Appendix K.

5.4.1.2.2 Parametric sampling method

When applying the parametric method in the YFCCP evaluation, the same initial process as applied within the NICE FC pathway evaluation was undertaken: $FC1_{true}$ values (n=10,000) were randomly drawn from the previously specified lognormal and Weibull FC1 distributions for the IBS and IBD populations respectively, according to the observed YFCCP IBD prevalence. These $FC1_{true}$

values were then used within the first error model application (i.e. step (ii) in Figure 5-7). To generate required $FC2_{true}$ values, the same process of deriving optimal parametric distributions as outlined for the FC1 data (section 5.3.1.2.2) was similarly applied to data on two further subgroups: IBS FC2 values, and IBD FC2 values. That is, a range of distributions (Normal, Lognormal, Gamma and Weibull) were fitted to the population-specific YFCCP FC2 data using the R 'fitdistcens' function to derive distributions over both censored and complete data regions (277). The optimal parameterisations used within the base case analysis were then selected based on an analysis of AIC and BIC metrics.

Table 5-7 reports the AIC and BIC criteria for the parametric analysis specifying a left-censored data region of 0-10 µg/g, and a right-censored data region of 600-1000 µg/g (as in the NICE FC pathway base case analysis). Note that the IBS and IBD population FC1 distributions reported in Table 5-7 are the same as previously reported in part 1 (section 5.3.1.2.2); however, an alternative face validity metric is here provided, equal to the proportion of simulated FC values falling above the YFCCP 100 µg/g cut-off threshold. This metric can be used as before to assess the internal validity of the simulated distributions via comparison with the YFCCP empirical dataset values. Associated tables using the two alternative upper bounds for the right-censored data region (2,000 and 3,000 µg/g) are provided in Appendix I.1; Appendix I.2 also provides the simulated probability density distributions for each of the parameterisations listed in Table 5-7 (based on n=10,000 draws from each distribution).

Based on the results reported in Table 5-7, the lognormal distribution was selected for both the IBS FC1 and FC2 distributions, and the Weibull distribution was selected for both the IBD FC1 and FC2 distributions within the parametric method base case analysis. Sensitivity analyses were also conducted to explore the impact of adopting each of the alternative parameterisations listed in Table 5-7 (see section 5.4.1.2.5). Within the error model simulation, the population-specific proportions of $FC1_{sim}$ values returned as ≥100 µg/g within the first error model application informed the number of $FC2_{true}$ simulations drawn from the respective population FC2 distributions, to which the second error model was applied (i.e. step (iii) in Figure 5-7).

Table 5-7. YFCCP: AIC and BIC criteria for FC1 and FC2 distributions (upper bound for right-censored data = 1,000 µg/g)

Subgroup	Parameterisation	AIC	BIC	% values ≥ 100 µg/g (simulated data)	% values ≥ 100 µg/g (YFCCP dataset)
IBS FC1	**Lognormal** [meanlog = 3.70656; sdlog = 1.22697]	8592.265	8601.809	23.2%	19.7%
	Weibull [shape = 0.76192; scale = 73.83344]	8728.559	8738.103	28.3%	
	Gamma [shape = 0.68899; rate = 0.00774]	8781.936	8791.479	29.9%	
	Normal [mean = 86.97938; SD = 134.91820]	10375.29	10384.830	45.4%	
IBS FC2	**Lognormal** [meanlog = 4.41576; sdlog = 1.24684]	1915.081	1921.376	44.3%	40.7%
	Weibull [shape = 0.85262; scale = 152.01680]	1926.026	1932.321	49.8%	
	Gamma [shape = 0.814413; rate = 0.00491]	1929.075	1935.370	51.7%	
	Normal [mean = 160.70540; SD = 185.69180]	2134.028	2140.323	61.8%	
IBD FC1	**Lognormal** [meanlog = 6.00743; sdlog = 0.86151]	615.288	620.002	95.2%	93.6%
	Weibull [shape = 1.74992; scale = 589.04020]	588.486	593.199	95.7%	
	Gamma [shape = 2.02506; rate = 0.00382]	596.323	601.036	94.3%	
	Normal [mean = 526.47760; SD = 291.24007]	588.660	593.373	92.4%	
IBD FC2	**Lognormal** [meanlog = 5.91977; sdlog = 0.70842]	588.898	593.185	96.7%	100%
	Weibull [shape = 1.75989; scale = 520.63000]	583.813	588.099	94.9%	
	Gamma [shape = 2.42639; rate = 0.00524]	585.347	589.633	95.5%	
	Normal [mean = 458.38920; SD = 266.79900]	593.387	597.673	91.2%	

As for the bootstrap method, an alternative approach to sampling the FC2 data was considered within a sensitivity analysis for the parametric method. In this case, all required $FC2_{true}$ values were generated by first sampling from population-specific distributions for the *between-test proportional differences,* defined as FC_{diff} = (FC2 – FC1)/FC1. Within the simulation, for all $FC1_{sim}$ values returned as ≥100 µg/g, $FC2_{true}$ values were derived by taking the associated $FC1_{true}$ value and applying a randomly drawn FC_{diff} value (i.e. $FC2_{true} = FC1_{true} + FC1_{true}*FC_{diff}$). As before, proportional differences were applied so as to avoid generating negative $FC2_{true}$ values. Note that this method differs from the associated bootstrap sensitivity analysis, in that *all* FC2 values were here generated by drawing from the FC_{diff} distribution (rather than only for a subset of 'missing' FC2 values within the bootstrap method).

Parametric distributions for the IBS and IBD FC_{diff} values were derived using a similar process as previously outlined, with four alternative parametrisations (normal, lognormal, gamma and Weibull) applied to the FC_{diff} data, using the 'fitdistcens' function to account for censored data (277). There were two key differences in this case, however. First, numerous censored data regions were possible for the FC_{diff} data, depending on whether the FC1 and/or FC2 values feeding into the FC_{diff} calculation were left- or right-censored.[35] A range of censored data regions for FC_{diff} were therefore assigned within the 'fitdiscens' function, depending on the associated FC1 and FC2 values. The second complication lies in the fact that, in its natural form, the FC_{diff} distribution can span from -1 (when the test drops from a positive value to zero), to infinity (when the test rises from zero to a positive value). In order to enable application of the lognormal, Gamma and Weibull distributions (which cannot be applied to negative data values), a temporary adjustment was applied, adding +1 to each of the FC_{diff} values. This adjustment was removed after sampling the required FC_{diff} values from the given parametric distribution within the error model simulation.

The AIC and BIC criteria for each of the parameterisations applied to the adjusted FC_{diff} data are provided in Table 5-8. Based on these results, the Weibull

[35] For example, if FC1 = "<10" and FC2 = "<10", then both of these values lie somewhere in the left-censored data region (0-10 µg/g). FC_{diff} in this case may lie anywhere between -1 (where FC1=10 and FC2=0) to +infinity (where FC1=0 and FC2=10).

distribution was used for the IBS population FC_{diff} distribution, and the Gamma distribution was used for the IBD population FC_{diff} distribution. Appendix K (section K.2) provides corresponding probability density distributions for the listed parameterisations (based on n=10,000 draws from each distribution). Each of the alternative parameterisations listed in Table 5-8 were also explored in sensitivity analyses (see section 5.4.1.2.5).

5.4.1.2.3 Uncertainty

The same smoothing algorithm previously outlined in section 5.3.1.2.3 was similarly applied to the YFCCP evaluation results, to smooth the sensitivity and specificity outputs. As before, sensitivity analyses were also conducted based on: (i) removing the smoothing algorithm (i.e. “noisy” results); (ii) increasing the simulation number to 100,000; (iii) removing the sampling process altogether within the bootstrap method; and (iv) applying an additional outer simulation loop within the parametric method, to account for parametric uncertainty within this method (in this case, capturing uncertainty within both the FC1 and FC2 distributional parameters). A full list of sensitivity analyses conducted is provided in section 5.4.1.2.5.

5.4.1.2.4 Outputs

For each sampling method, the simulation process was repeated for CV values ranging from 0-100% in 0.5% increments, and for bias ranging from -100 to +100 µg/g in 1 µg/g increments, producing 40,401 diagnostic sensitivity and specificity results within each analysis. The results were illustrated using the same series of contour plots as in the NICE FC evaluation. For the plots illustrating the acceptable regions of bias and imprecision, in the first instance these were specified based on an assumed minimum accuracy requirement set equal to the lower 95% CI’s achieved in the YFCCP baseline diagnostic accuracy assessment (reported in section 5.4.2.1): i.e., sensitivity ≥0.85 and specificity ≥0.90. A lower requirement was also explored, reducing these values by 10% (i.e. sensitivity ≥0.75 and specificity ≥0.80). The results of the sensitivity analyses conducted were reported in tabular format, as for the NICE FC pathway evaluation.

Table 5-8. YFCCP evaluation: AIC and BIC criteria for adjusted FC_{diff} parametric distributions (upper bound for right-censored FC data = 1,000 µg/g)

Subgroup	Parameterisation	AIC	BIC
IBS FC_{diff} (adjusted)	**Lognormal** [meanlog = -1.11652; sdlog = 1.31925]	389.945	396.240
	Weibull [shape = 0.88117; scale = 0.60781]	379.165	385.460
	Gamma [shape = 0.83202; rate = 1.28402]	380.040	386.335
	Normal [mean = 0.64413; SD = 0.72701]	571.098	577.393
IBD FC_{diff} (adjusted)	**Lognormal** [meanlog = -0.15797; sdlog = 0.70691]	171.663	175.949
	Weibull [shape = 1.56416; scale = 1.21244]	174.062	178.349
	Gamma [shape = 2.33821; rate = 2.16255]	171.135	175.422
	Normal [mean = 1.0947054; SD = 0.7227265]	195.799	200.085

5.4.1.2.5 Analysis summary

The simulation processes used for the YFCCP base case analyses is summarised below.

Bootstrap sampling method:

i. Left- and right-censored FC values in the YFCCP dataset (n=951) are replaced with their associated limit values (10 and 600 µg/g respectively);
ii. An expanded bootstrap dataset (n=10,000) is generated by sampling with replacement from the YFCCP dataset rows;
iii. For each $FC1_{true}$ value in the bootstrap dataset, the error model is applied to generate $FC1_{sim}$ values at a given level of bias and CV%;
iv. For $FC1_{sim}$ values ≥100 µg/g with missing $FC2_{true}$ values in the bootstrap dataset, required values are imputed by randomly sampling with replacement from the available population-specific YFCCP FC2 values;
v. For $FC1_{sim}$ values ≥100 µg/g, the error model is applied to the associated $FC2_{true}$ value to generate $FC2_{sim}$ values at a given level of bias and CV%;
vi. Diagnostic accuracy of the YFCCP including additional FC bias and CV% is calculated by comparing diagnoses based on the $FC1_{sim}$ and $FC2_{sim}$ values (as per the YFCCP diagnostic protocol) with patients' clinical diagnoses in the bootstrap dataset; and
vii. Steps (iii) - (vi) are repeated for a range of bias (-100 to +100 in 1µg/g increments) and CV (0 to 100% in 0.5% increments) values.

Parametric sampling method:

i. Using the parameter specifications provided in Table 5-7, a total of 10,000 $FC1_{true}$ values are drawn from (a) the FC1 lognormal parametric distribution for IBS patients (n=9180), and (b) the FC1 Weibull parametric distribution for IBD patients (n=820);
ii. For each $FC1_{true}$ value in (i), the error model is applied to generate $FC1_{sim}$ values at a given level of bias and CV%;
iii. For $FC1_{sim}$ values ≥100 µg/g, population-specific $FC2_{true}$ values are drawn from (a) the FC2 lognormal parametric distribution for IBS patients, and (b) the FC2 Weibull parametric distribution for IBD patients (as per Table 5-7);

iv. For $FC1_{sim}$ values ≥100 µg/g, the error model is applied to the associated $FC2_{true}$ value to generate $FC2_{sim}$ values at a given level of bias and CV%;

v. Diagnostic accuracy of the YFCCP including additional FC bias and CV% is calculated by comparing diagnoses based on the $FC1_{sim}$ and $FC2_{sim}$ values (as per the YFCCP diagnostic protocol) with patients' clinical diagnoses (according to the population distribution from which simulations were drawn); and

vi. Steps (iii) - (v) are repeated for a range of bias (-100 to +100 in 1µg/g increments) and CV (0 to 100% in 0.5% increments) values.

A summary of the sensitivity analyses conducted for each sampling method is provided in Table 5-9.

Table 5-9. YFCCP: sensitivity analyses conducted

Code	Analysis summary	Analysis details
Bootstrap sampling method analyses		
[1.0]	Bootstrap method base case analyses	Left- and right-censored data (recorded as "<10" and ">600" in the YFCCP dataset, respectively) replaced with limit values (10 and 600 µg/g respectively); error model based on 10,000 bootstrap samples; missing FC2 data sampled with replacement from available population-specific FC2 data; smoothing algorithm applied to sensitivity and specificity results
[1.1]	Left-censored data = 5 µg/g; right-censored data = 750 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 750 µg/g
[1.2]	Left-censored data = 5 µg/g; right-censored data = 900 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 900 µg/g
[1.3]	Left-censored data = 5 µg/g; right-censored data = 1200 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 1200 µg/g
[1.4]	Left-censored data = 5 µg/g; right-censored data = 1800 µg/g	Left-censored data replaced with 5 µg/g; right-censored data replaced with 1800 µg/g
[1.5]	Complete case analysis	Censored data excluded
[1.6]	Missing FC2 data imputed using FC_{diff} data	Missing FC2 values required in the simulation generated by sampling from the population-specific FC_{diff} data (FC_{diff} = (FC2 – FC1)/FC1), with FC2 = FC1 + FC1* FC_{diff}.
[1.7]	Raw data only (n=951; no bootstrap sampling)	No bootstrap sampling conducted: initial error model application applied to YFCCP database (n=951) FC1 values alone, followed by the same process of re-sampling for FC2 values as used in the base case
[1.8]*	Noisy results (no smoothing algorithm)	No smoothing algorithm applied to the sensitivity and specificity results
[1.9]*	100,000 samples	Error model based on 100,000 bootstrap samples; no smoothing algorithm applied
Parametric sampling method analyses		

[2.0]	Parametric method base case analyses	R 'fitdistrplus' function used to fit parametric distributions to required patient subgroups (IBS FC1, IBS FC2, IBD FC1, IBD FC2) over complete and censored data regions via maximum likelihood estimation; lognormal distribution used for IBS subgroups; Weibull distribution used for IBD subgroups; region for left-censored data set as 0-10 µg/g; region for right-censored data set as 600-1000 µg/g; error model based on an initial 10,000 draws across the IBS FC1 and IBD FC1 distributions; smoothing algorithm applied to sensitivity and specificity results
[2.1]	Right-censored data region = 600 to 2000 µg/g	Upper bound for right censored data region set to 2,000 µg/g within the R 'fitdistcens' function
[2.2]	Right-censored data region = 600 to 3000 µg/g	Upper bound for right censored data region set to 3,000 µg/g within the R 'fitdistcens' function
[2.3]	Complete case analysis	Censored data excluded
[2.4]	Lognormal parameterisation	Lognormal parameterisations used for all patient subgroup distributions
[2.5]	Weibull parameterisation	Weibull parameterisations used for all patient subgroup distributions
[2.6]	Gamma parameterisation	Gamma parameterisations used for all patient subgroup distributions
[2.7]	Normal parameterisation	Normal parameterisations used for all patient subgroup distributions
[2.8]	$FC2_{true}$ values assigned using FC_{diff} distributions	FC2 values generated by sampling from the population-specific FC_{diff} data (FC_{diff} = (FC2 – FC1)/FC1), drawing on a Weibull distribution for the IBS population and a Gamma distribution for the IBD population. FC2 = FC1 + FC1* FC_{diff}.
[2.9]	FC_{diff} lognormal parameterisation	Lognormal parameterisations use for both IBD and IBS FC_{diff} distributions
[2.10]	FC_{diff} Weibull parameterisation	Weibull parameterisations use for both IBD and IBS FC_{diff} distributions
[2.11]	FC_{diff} Gamma parameterisation	Gamma parameterisations use for both IBD and IBS FC_{diff} distributions
[2.12]	FC_{diff} Normal parameterisation	Normal parameterisations use for both IBD and IBS FC_{diff} distributions
[2.9]*	Noisy results (no smoothing algorithm)	No smoothing algorithm applied to the sensitivity and specificity results
[2.10]*	100,000 samples	Error model based on an initial 100,000 draws across the IBS FC1 and IBD FC1 distributions; no smoothing algorithm applied

[2.11]*	Sampling accounting for parametric uncertainty: inner simulations = 1,000 x 40,401; outer simulations = 1,000	"Inner" simulation: base case error model process using an initial 1,000 draws across the IBS FC1 and IBD FC1 distributions for each bias and CV pair (n=40,401). "Outer" simulation: for each outer simulation, a different stochastic value for the lognormal (IBS) and Weibull (IBD) distribution parameter estimates was drawn and applied across the 1,000*40,401 inner simulations. The 1,000 outer simulation stochastic parametric values were derived using the 'bootdistcens' function from the 'fitdistrplus' package (277).
*These analyses are based on "noisy" results i.e. with no smoothing algorithm applied		

5.4.2 Results

5.4.2.1 Baseline diagnostic accuracy

The diagnostic accuracy of the YFCCP is summarised in Table 5-10. Assigning FC diagnoses as per the YFCCP diagnostic protocol results in: 73/78 patients with IBD being correctly identified, giving a sensitivity of 93.6% (95% CI: 0.85 to 0.97); and 803/873 patients with IBS being correctly identified, giving a specificity of 92.0% (95% CI: 0.90 to 0.94).

Table 5-10. YFCCP: baseline diagnosis accuracy results

		Clinical diagnosis		
		IBD	**IBS**	
YFCCP diagnosis	**IBD**	True positives: n= 73	False positives: n= 70	**PPV = 73/143 = 51.0%**
	IBS	False negatives: n= 5	True negatives: n= 803	**NPV = 803/808 = 99.4%**
		Sensitivity = 73/78 = 93.6%	**Specificity = 803/873 = 92.0%**	

A further summary of FC diagnoses according to the YFCCP diagnostic protocol is provided in Figure 5-8. Note that, whilst the YFCCP includes a safety-net GP review at 6 weeks (in which patients with persisting symptoms may receive a secondary care referral), this element of the pathway is not considered within the diagnostic accuracy calculation (i.e. diagnostic accuracy is based only on the initial FC1 and FC2 results only, as outlined in Figure 5-8).

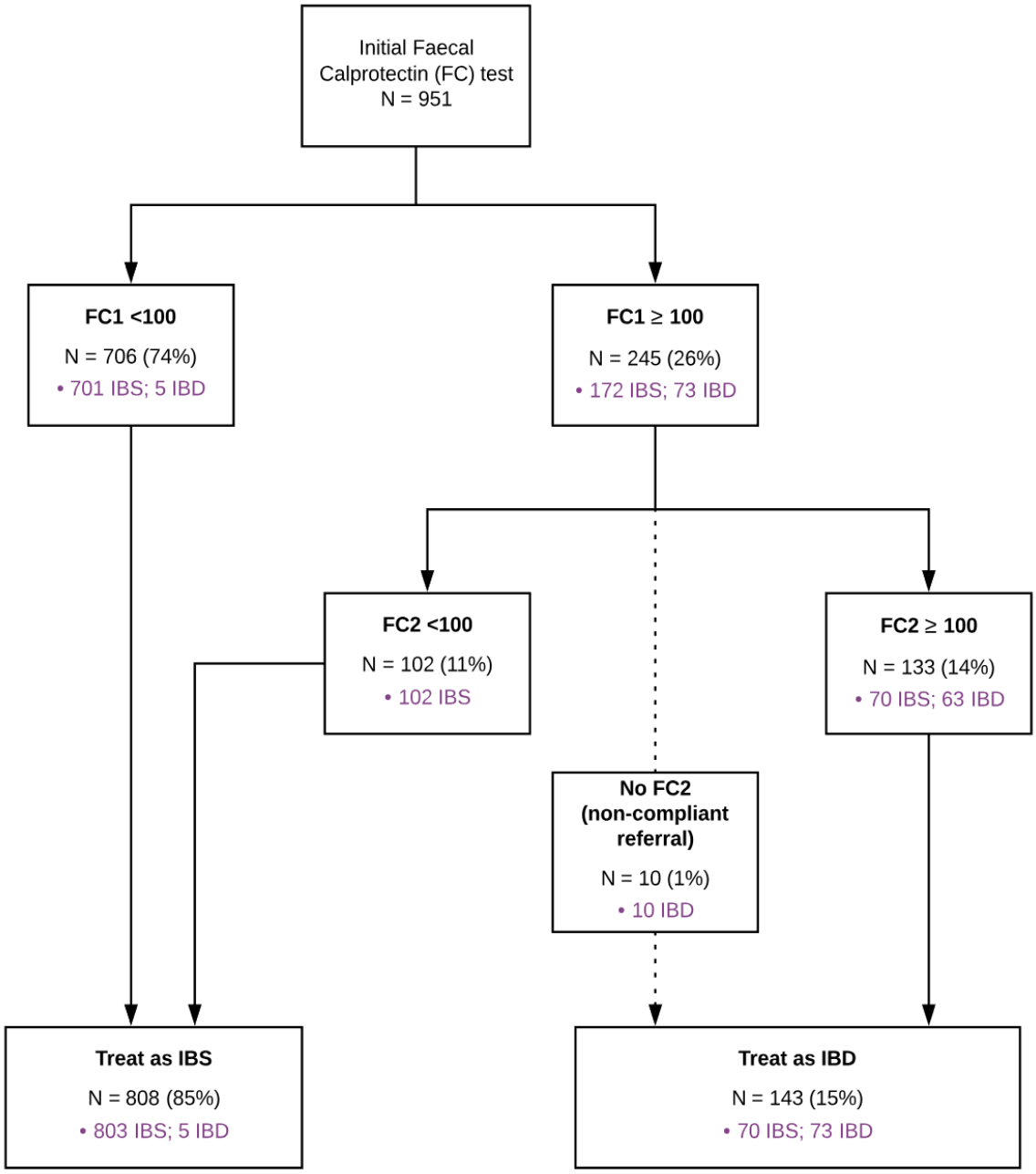


Figure 5-8. YFCCP: flow diagram of FC test results

The mechanism behind the higher diagnostic specificity of the YFCCP (compared to the NICE FC pathway) is illustrated by looking at the distribution of FC1 and FC2 results across each population within the YFCCP dataset, shown in Figure 5-9. Looking first at the top two panels (FC1 results), we can see that using the elevated threshold of 100 µg/g in this population leads to significantly fewer false positive cases (i.e. fewer IBS patients lying above the threshold cut-off value), compared to the standard cut-off threshold of 50 µg/g (Figure 5-3). In addition, it is apparent from the FC2 results (shown in the bottom two panels in Figure 5-9) that a significant proportion of IBS patients with an initially raised result fall back down below the 100 µg/g threshold upon re-testing, whilst all IBD patients remain elevated, allowing the repeat-test to further increase the pathway's specificity without reducing the sensitivity.

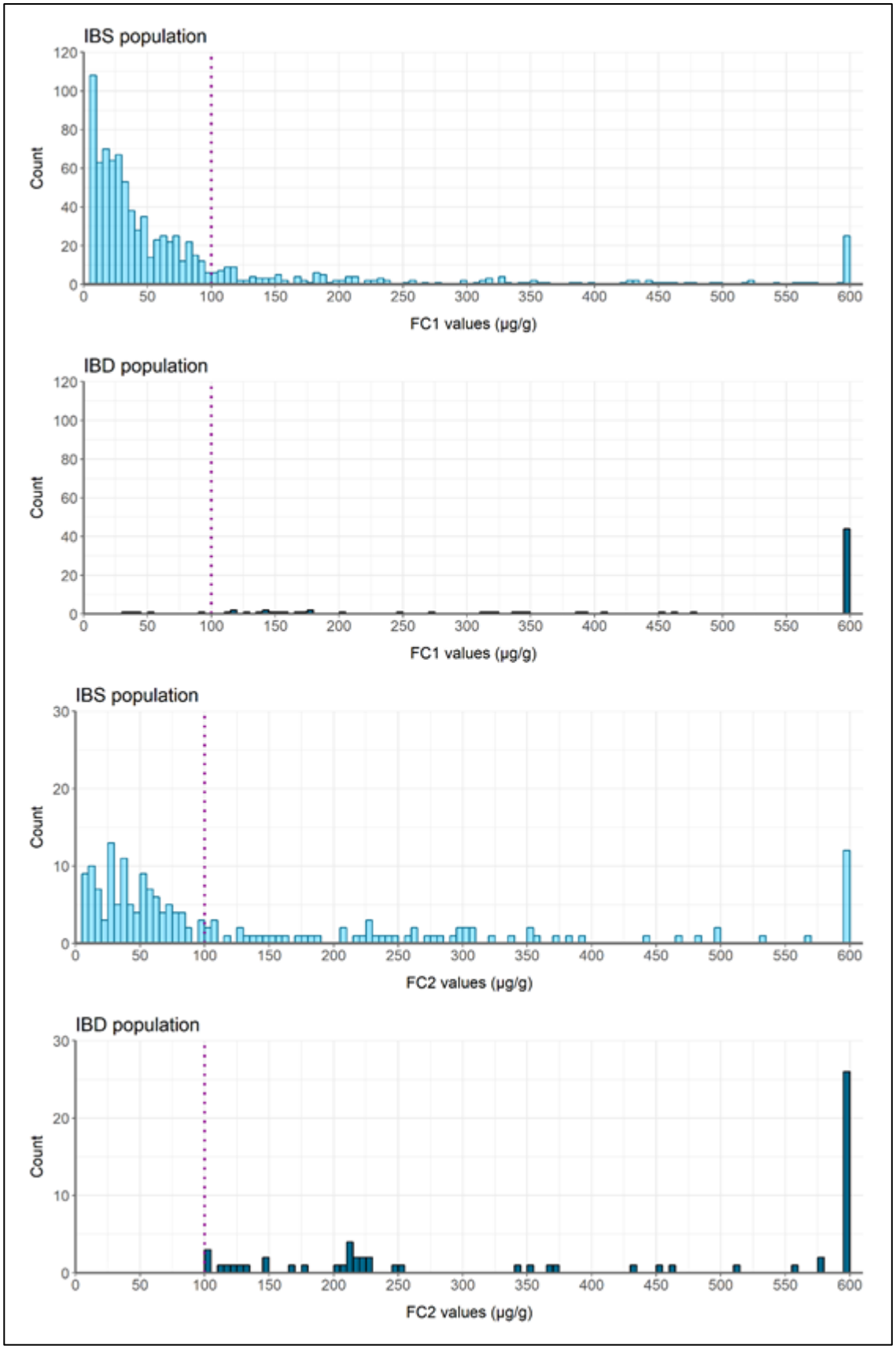


Figure 5-9. YFCCP: count plots showing the distribution of FC1 and FC2 values for IBS and IBD patients within the YFCCP database

5.4.2.2 Simulated diagnostic accuracy

5.4.2.2.1 Base case analyses

Figure 5-10 provides the base case diagnostic accuracy contour plots for the YFCCP. “Noisy” versions of these plots, based on the raw simulation results without applying the base case smoothing algorithm, are provided in Appendix J.

As for the NICE FC pathway, these plots can be used to study the robustness of the YFCCP’s diagnostic accuracy to increased measurement uncertainty. Looking first at the bootstrap method specificity contours, if we hold imprecision at 0% and introduce negative bias (move horizontally to the left from the (0,0) point) then we see that the pathway’s specificity is unaffected by negative bias; whilst if we apply positive bias (move horizontally to the right from the (0,0) point) we gradually pass through lower specificity contours (Figure 5-10, panel A) (note that the rapidity of change here is much less than that previously observed with the NICE FC pathway; section 5.3.2.2, Figure 5-4). Next, if we hold bias at 0 µg/g and introduce imprecision (move vertically from the (0,0) point), then we see that the pathway’s specificity is largely unchanged by increasing imprecision. Applying the same process to the sensitivity contours, we observe that the YFCCP’s sensitivity is unaffected by positive bias, but gradually reduces in response to negative bias and imprecision. In this case, the drop in sensitivity observed in response to negative bias and imprecision is more pronounced for the YFCCP than previously seen for the NICE FC pathway (Figure 5-4).

The same pattern of results was observed with the parametric method (Figure 5-10, panel B). However, this sampling method produces slightly lower sensitivity (90.6%) and lower specificity (89.5%) values at the (0,0) point, compared to the baseline sensitivity (93.6%) and specificity (92.0%) values reported in section 5.3.2.2.1. The bootstrap method meanwhile produces a perfect match to the baseline sensitivity and specificity in this case. A full discussion of these results is provided in section 5.5.2.

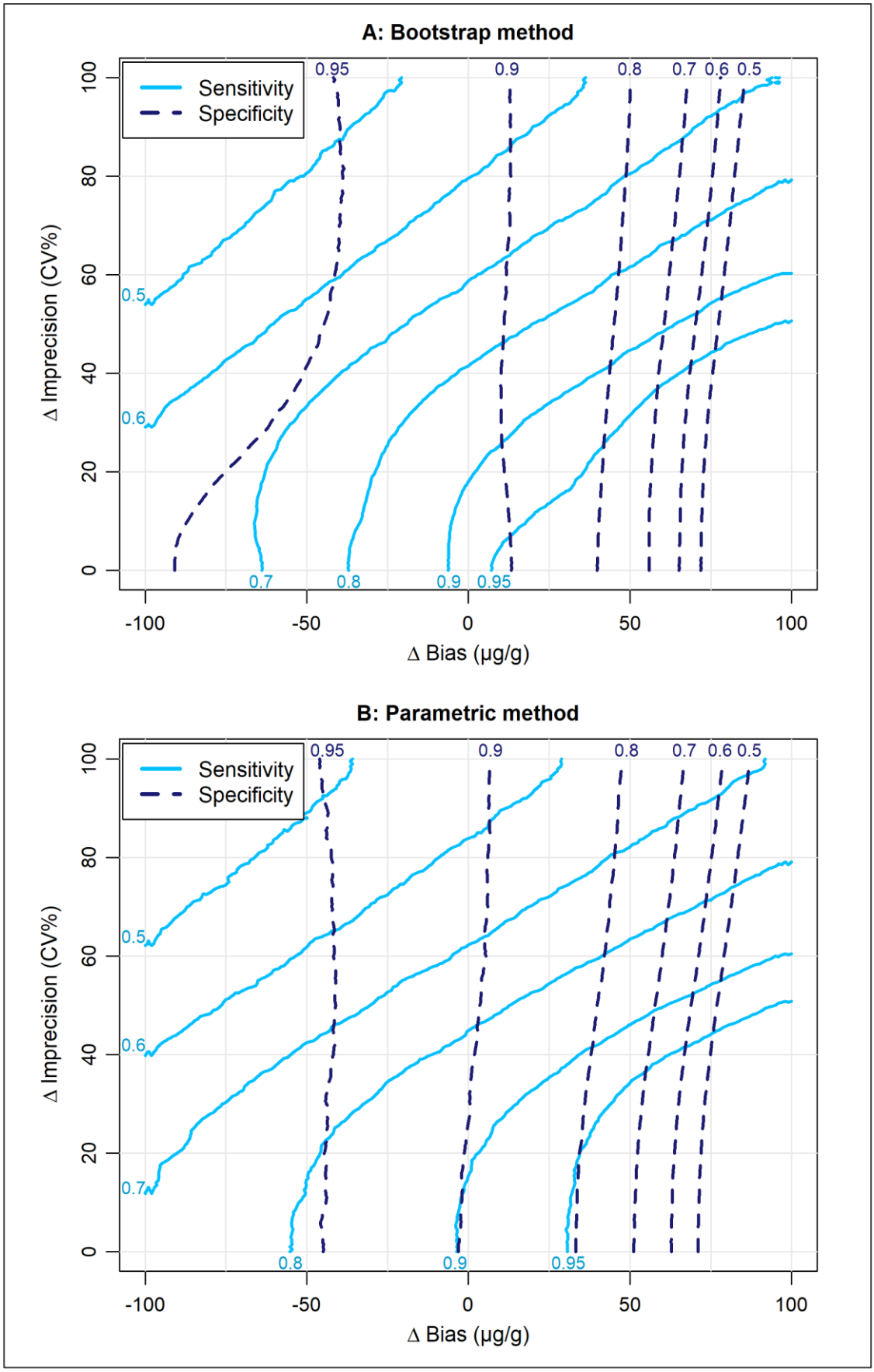


Figure 5-10. YFCCP: base case diagnostic accuracy contour plots

Figure 5-11 shows the same results, this time highlighting the acceptable regions of bias and imprecision relating to an assumed minimum diagnostic accuracy requirement of sensitivity ≥85% and specificity ≥90% (the lower 95% CI's from the YFCCP baseline diagnostic accuracy evaluation [section 5.4.2.1]). TE bands are also overlaid onto these plots. Based on these results, TE_{max} = 13% when using the bootstrap method; whilst for the parametric method, due to the fact that this method produces a lower baseline specificity (below the 90% minimum specificity requirement), the acceptable region is offset from the (0,0) point and there is subsequently no acceptable TE value contained within the acceptable region (Figure 5-5, panel B).

Figure 5-12 further highlights the acceptable regions relating to a lower minimum diagnostic accuracy requirement of sensitivity ≥75% and specificity ≥80% (i.e. 10% below the lower 95% CI's from the diagnostic accuracy evaluation). In this case, TE_{max} = 39% with the bootstrap method or 33% with the parametric method (compared to 15% and 13% for the associated NICE FC pathway analysis respectively, shown in Figure 5-6).

Table 5-11 summarises key findings from the above contour plots in tabular format.

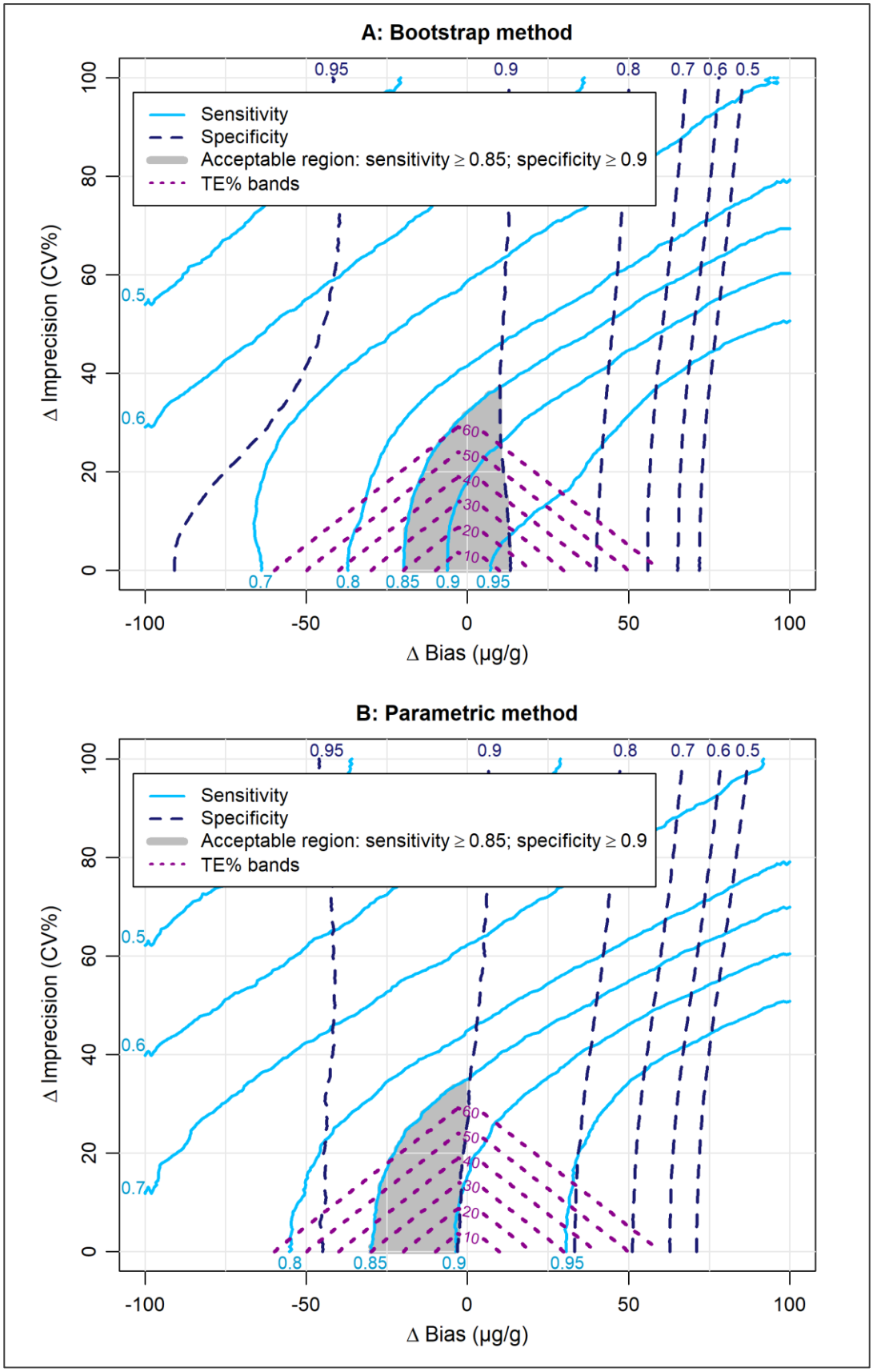


Figure 5-11. YFCCP: base case diagnostic accuracy contour plots showing the acceptable region (maintaining sensitivity ≥0.85 and specificity ≥0.9) and TE% bands

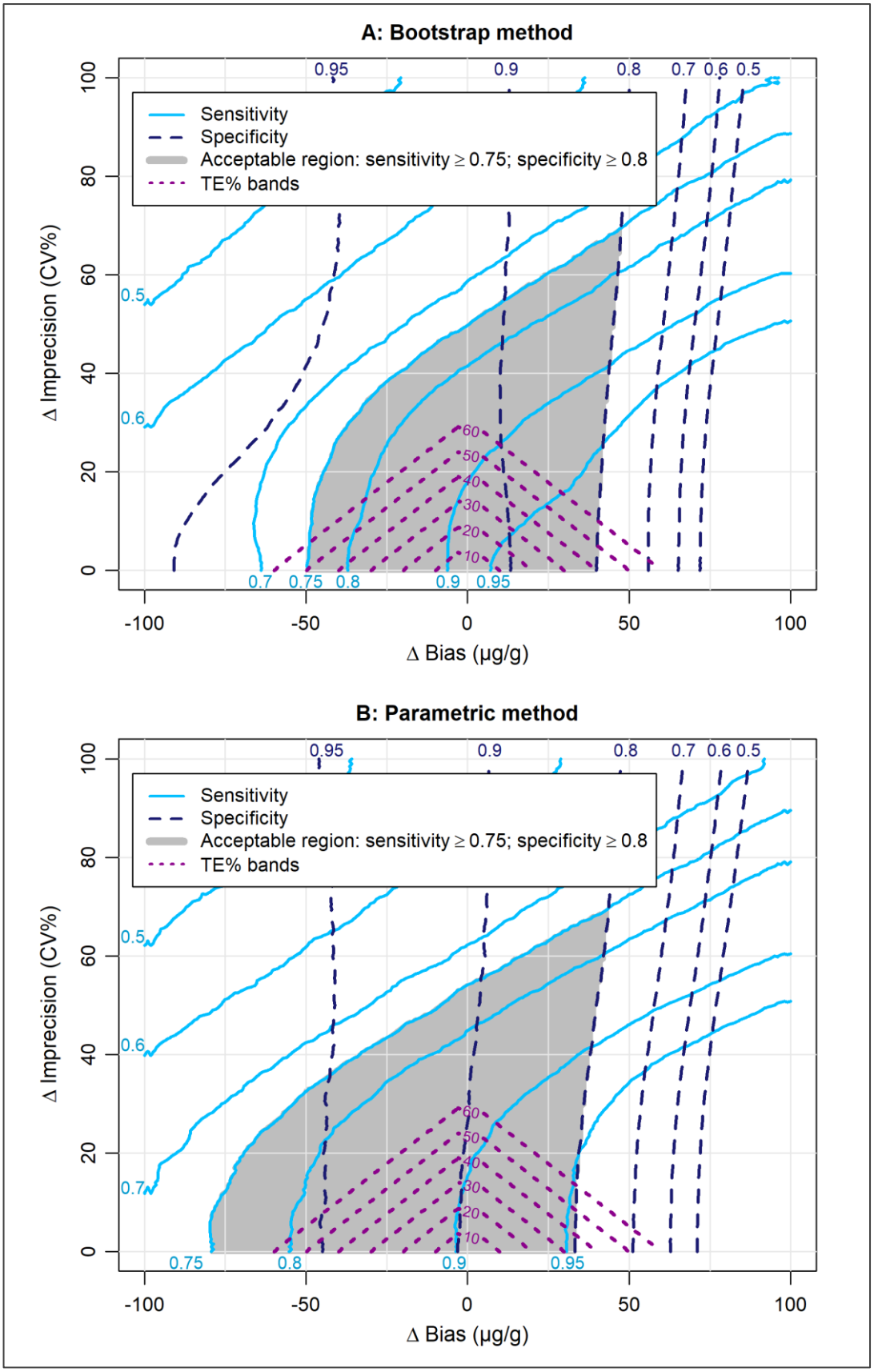


Figure 5-12. YFCCP: base case diagnostic accuracy contour plots showing the acceptable region (maintaining sensitivity ≥0.75 and specificity ≥0.8) and TE% bands

Table 5-11. YFCCP: simulated diagnostic accuracy base case results

Sampling method	Diagnostic accuracy at bias=0 & CV=0%		Acceptable region 1: sensitivity ≥0.85; specificity ≥0.90			Acceptable region 2: sensitivity ≥0.75; specificity ≥0.80		
	Sensitivity	Specificity	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0
Bootstrap method	0.936	0.920	13%	-19 to 13	0 to 32%	39%	-49 to 39	0 to 50%
Parametric method	0.906	0.895	NA (no results in range)	-30 to -4	27 to 35%	33%	-79 to 33	0 to 54%

5.4.2.2.2 Sensitivity analyses

Table 5-12 reports results of the sensitivity analyses conducted for the YFCCP evaluation. The pattern of results is broadly similar to that observed with the NICE FC pathway (section 5.3.2.2.2). Excluding the complete case analysis (which produced significantly lower sensitivity values across both sampling methods), the use of alternative censored data substitution values had no measurable impact on the bootstrap method results (Table 5-12, analyses 1.1-1.5). The two analyses exploring higher upper bound values for the right-censored data region in the parametric method analysis produced similar specificity values and slightly lower sensitivity values, however the drop in sensitivity values in this case resulted in more noticeably diminished acceptable regions (Table 5-12 analyses 2.1-2.2). When using the parametric method, application of the lognormal, Weibull, Gamma and normal distributions produced (generally) increasingly lower sensitivity and specificity values and restricted acceptable regions, with the normal distribution again providing a particularly poor fit to the data (Table 5-12, analyses 2.4-2.7).

Across both sampling methods, sensitivity analyses exploring sampling uncertainty (Table 5-12, analyses 1.7-1.9 and 2.13-2.14) had little impact on the results. The exception to this were the results of the acceptable range of CV% at zero bias for the YFCCP parametric method (assuming a required sensitivity of ≥85% and specificity of ≥90%): this region widened when increasing the sampling number to 100,000, or removing the smoothing algorithm (analysis 2.13-2.14 in Table 5-12). This pattern of results was reflected in the sensitivity analysis applying an outer simulation loop to capture parametric uncertainty within the parametric sampling method (analysis 2.15 in Table 5-12). The instability of this range is due to the placement of the 90% specificity contour in this analysis, which straddles the zero bias line: introducing slight uncertainty in the placement of this contour line can therefore have a significant impact on how much CV% can be tolerated at zero bias, when assuming a specificity requirement of ≥90% (Figure 5-11, panel B). The impact of this uncertainty is further illustrated in the “noisy” contour plots (i.e. with no smoothing algorithm applied), provided in Appendix J.

An additional set of sensitivity analyses were conducted in this assessment, relating to the method used to sample $FC2_{true}$ values within the simulation. Recall

that in the base case analyses, missing $FC2_{true}$ values within the bootstrap method simulation were replaced by randomly sampling with replacement from the available (population-specific) FC2 values; whilst all FC2 values within the parametric method simulation were generated by directly sampling from the (population-specific) FC2 parametric distributions. In the bootstrap method sensitivity analysis (analysis 1.6 in Table 5-12), missing $FC2_{true}$ values were instead generated by sampling from the population-specific empirical FC_{diff} distributions as described in section 5.4.1.2.1; and in a corresponding set of sensitivity analyses under the parametric method (analyses 2.8-12 in Table 5-12) all $FC2_{true}$ values were generated by sampling from the population-specific FC_{diff} parametric distributions, as described in section 5.4.1.2.2.

Using this approach to sampling $FC2_{true}$ values, the parametric method achieved a worsened performance in terms of internal validity, with the diagnostic accuracy values at the (0,0) point moving further away from the baseline diagnostic accuracy results (analyses 2.8-12 in Table 5-12). As such, the parametric method sensitivity analyses results based on drawing from FC_{diff} distributions are not considered further. For the bootstrap method however, this approach to sampling missing $FC2_{true}$ values maintains the same level of internal validity as in the base case analysis (with the sensitivity and specificity values reported at the (0,0) point again matching the baseline diagnostic accuracy results). A key difference resulting from this analysis however, is that the pathway exhibits greater robustness to positive and negative bias, resulting in wider acceptability regions (analysis 1.6 in Table 5-12). Further discussion of these results is provided in section 5.5.2.

Table 5-12. YFCCP: simulated diagnostic accuracy sensitivity analysis results

		Diagnostic accuracy at bias=0 & CV=0%		Acceptable region 1: sensitivity ≥0.85; specificity ≥0.90			Acceptable region 2: sensitivity ≥0.75; specificity ≥0.80		
		Sensitivity	Specificity	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0
Bootstrap sampling method									
Base case	**[1.0] Bootstrap method**	0.936	0.920	13%	-19 to 13	0 to 32%	39%	-49 to 39	0 to 50%
Sensitivity analyses: FC censored data handling	[1.1] Left-censored data = 5; right-censored data = 750 µg/g	0.936	0.920	13%	-19 to 13	0 to 32%	39%	-49 to 39	0 to 51%
	[1.2] Left-censored data = 5; right-censored data = 900 µg/g	0.936	0.920	13%	-19 to 13	0 to 32%	39%	-49 to 39	0 to 52%
	[1.3] Left-censored data = 5; right-censored data = 1200 µg/g	0.936	0.920	13%	-19 to 13	0 to 33%	39%	-49 to 39	0 to 52%
	[1.4] Left-censored data = 5; right-censored data = 1800 µg/g	0.936	0.920	13%	-19 to 13	0 to 33%	39%	-49 to 39	0 to 53%
	[1.5] Complete case analysis	0.810	0.927	- NA -	- NA -	- NA -	6%	-6 to 36	0 to 14%
Sensitivity analyses: FC2 data handling	[1.6] Missing $FC2_{true}$ values sampled using FC_{diff} empirical distribution	0.936	0.921	17%	-24 to 17	0 to 33%	50%	-52 to 52	0 to 50%
Sensitivity analyses:	[1.7] Raw data only (n=951; no bootstrap sampling)	0.934	0.920	11%	-17 to 11	0 to 31%	38%	-48 to 38	0 to 50%

sampling uncertainty	[1.8]* Noisy results (no smoothing algorithm)	0.941	0.921	14%	-19 to 14	0 to 33%	41%	-50 to 41	0 to 50%
	[1.9]* 100,000 samples	0.933	0.919	12%	-19 to 12	0 to 31%	38%	-48 to 38	0 to 49%
Parametric sampling method									
Base case	**[2.0] Parametric method**	0.906	0.895	- NA -	-30 to -4	27 to 35%	33%	-79 to 33	0 to 54%
SA: censored data handling	[2.1] Right-censored data region: 600-2000 µg/g	0.858	0.894	- NA -	-3 to -3	- NA -	33%	-61 to 33	0 to 49%
	[2.2] Right-censored data region: 600-3000 µg/g	0.837	0.894	- NA -	- NA -	- NA -	33%	-48 to 33	0 to 46%
	[2.3] Complete case analysis	0.657	0.915	- NA -	- NA -	- NA -	- NA -	22 to 32	- NA -
SA: parameterisation	[2.4] Lognormal parameterisation	0.910	0.895	- NA -	-28 to -3	31 to 34%	33%	-58 to 33	0 to 51%
	[2.5] Weibull parameterisation	0.912	0.858	- NA -	-31 to -24	- NA -	21%	-86 to 21	0 to 54%
	[2.6] Gamma parameterisation	0.903	0.843	- NA -	- NA -	- NA -	14%	-71 to 14	0 to 51%
	[2.7] Normal parameterisation	0.841	0.706	- NA -	- NA -	- NA -	- NA -	-76 to -37	- NA -
SA: FC2 data handling	[2.8] $FC2_{true}$ values sampled using FC_{diff} distributions (IBS = Weibull; IBD = Gamma)	0.882	0.902	1%	-18 to 1	0 to 30%	48%	-81 to 48	0 to 51%
	[2.9] FC_{diff} lognormal parameterisation	0.850	0.912	1%	-1 to 10	0 to 4%	42%	-42 to 53	0 to 44%
	[2.10] FC_{diff} Weibull parameterisation	0.873	0.882	- NA -	-17 to -14	- NA -	39%	-76 to 39	0 to 50%
	[2.11] FC_{diff} Gamma parameterisation	0.868	0.775	- NA -	- NA -	- NA -	- NA -	-61 to -11	32 to 48%
	[2.12] FC_{diff} Normal parameterisation	0.808	0.753	- NA -	- NA -	- NA -	- NA -	-56 to -23	- NA -

SA: sampling and parametric uncertainty	[2.13]* Noisy results (no smoothing algorithm)	0.901	0.895	- NA -	-32 to -1	3 to 35%	34%	-90 to 34	0 to 56%
	[2.14]* 100,000 samples	0.908	0.898	- NA -	-31 to -2	8 to 36%	34%	-82 to 34	0 to 53%
	[2.15]* Sampling accounting for parametric uncertainty: inner simulations = 1,000 x 40,401; outer simulations = 1,000	0.902	0.899	- NA -	-30 to -1	15 to 35%	34%	-82 to 34	0 to 53%
*These analyses are based on “noisy” results i.e. using the direct simulation results for sensitivity and specificity, with no moving average calculation applied									

5.5 Discussion

5.5.1 Baseline diagnostic accuracy

Two FC strategies were evaluated in this study: the NICE FC pathway (a single-test strategy using the standard 50 µg/g cut-off threshold), and the YFCCP (a repeat-test strategy using a raised 100 µg/g threshold). The first exercise undertaken was to calculate each pathway's baseline diagnostic accuracy using FC values reported in the YFCCP empirical database (see sections 5.3.1.1 and 5.4.1.1).

The results of this analyses indicate that additional diagnostic yield can be obtained by moving from the NICE FC pathway to the YFCCP. Although the YFCCP was associated with a slightly lower sensitivity compared to the NICE FC pathway (93.6% vs 96.2%), this was offset by a significant gain in specificity (92.0% vs. 59.7%). This shift occurs in this example due to particular features of the FC data: first, IBD patients exhibited FC values consistently above the 100 µg/g threshold, both on initial and repeat testing; and second, whilst many IBS patients exhibited raised FC1 values (19.7%; 172/873), the majority of these fell below the cut-off threshold upon re-testing (59.3%; 102/172) (see Figure 5-9). Based on the FC1 results alone, therefore, the raised 100 µg/g cut-off threshold was able to correctly reclassify 180 IBS patients who would have been incorrectly referred under the NICE FC strategy (with an additional 2 IBD patients incorrectly re-classified as having IBS at this point)[36]; and a further 102 IBS patients were able to be correctly reclassified following the repeated test (with no further loss of IBD patients).

The validity of these findings (and the subsequent simulation exercises) depends on the reliability of the clinical diagnoses against which the FC strategies were judged. Clinical diagnoses within the YFCCP database were determined according to the results of endoscopic investigations where available (i.e. only for patients referred to secondary care as per the YFCCP protocol), and assuming IBS classifications were otherwise correct. This data is therefore at risk of *partial*

[36] Applying a single-test strategy using the 100 µg/g threshold produces a sensitivity of 93.5% (95% CI: 0.90 – 0.94) and specificity of 80.3% (95% CI: 0.77 - 0.83).

verification bias: that is, since a non-random subset of patients underwent the gold standard reference test (endoscopy), the number of false negative and true negative cases within the non-verified subgroup (patients with assumed IBS who did not receive endoscopy) could have been under- and overestimated, respectively (279). However, the YFCCP study also included a 6-month follow-up period, intended to capture patients with persisting symptoms referred to secondary care at a later date. Assuming that patients with IBD would indeed experience persisting symptoms and return to their GP within this timeframe, this safeguard should be sufficient to offset the risk of partial verification bias. This was the same assumption as applied in previous publications based on this same data (239, 240).

The applicability of the baseline diagnostic accuracy results to other UK regions, meanwhile, will depend upon the generalisability of the YFCCP data. If regional variation in the distribution of IBS or IBD patients' FC results is expected, then the high diagnostic accuracy of the YFCCP may not be replicated elsewhere. In a recent primary care diagnostic accuracy study conducted in Sheffield, for example, a proportion of IBD patients were found to have FC results within the 50-100 µg/g range (which reduces the sensitivity associated with any strategy employing the raised 100 µg/g threshold) (238).[37] Whilst this finding may be driven by between-assay differences (due to the fact that the Sheffield study used the Immundiagnostik (IDK) Calprotectin ELISA rather than the Bühlmann assay used within the YFCCP), rather than true population FC differences, it nevertheless highlights the fact that application of the YFCCP (or any other pathway) across alternative regions should include validation and/or monitoring of the strategy's diagnostic accuracy.

Alternative FC strategies, incorporating repeated testing within a restricted 'intermediate range' of initial FC results, have also been proposed (227, 240). For

[37] Applying a similar eligibility criteria as the YFCCP, this study evaluated diagnostic accuracy at both the standard (50 µg/g) and raised (100 µg/g) thresholds: at the standard threshold, the authors found a sensitivity and specificity of 72.7% and 64.9% respectively, whilst at the raised threshold these values changed to 54.6% and 80.5%. The sensitivity reduction in this case indicates that a proportion of IBD patients in the Sheffield population had FC values between 50 and 100 µg/g, which contrasts to the FC1 distribution observed in the YFCCP dataset.

example, repeated testing could be restricted to FC1 values falling within the range 100-400 µg/g, with direct referral to secondary care for values >400 µg/g. Based on the YFCCP data, this strategy would maintain the same sensitivity as the YFCCP (since all IBD patients' FC2 values were >100 µg/g), but would incorrectly refer an additional 26 IBS patients who had FC1 values >400 µg/g and FC2 values <100 µg/g. Figure 5-13 illustrates how the YFCCP specificity would alter over various intermediate ranges (using a lower bound of 100 µg/g for the intermediate range). The optimal specificity is achieved for an upper bound of 610 µg/g (equivalent to having no intermediate range at all, since 600 is the upper analytical measurement range for this test); whilst diminishing specificity values are observed as the intermediate range is tightened. All of the ranges explored here exhibited statistically significantly lower specificity compared to the YFCCP strategy (95% CI: 0.90 to 0.94). Overall therefore, based on this diagnostic accuracy assessment, the YFCCP appears to be the preferred strategy over both the NICE FC pathway single-test strategy and intermediate-range strategies.

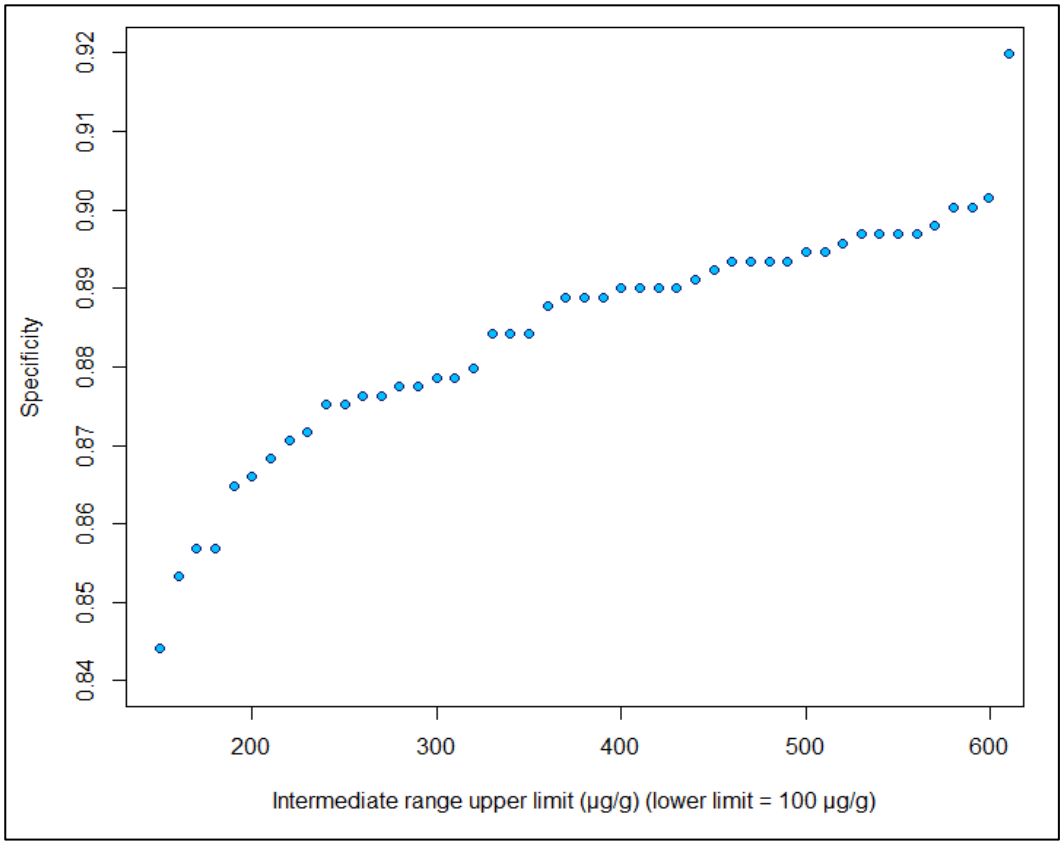


Figure 5-13. YFCCP: specificity results when restricting FC2 testing to an intermediate range of FC1 results

5.5.2 Simulated diagnostic accuracy

The primary objective of this chapter was to explore methodology for assessing the impact of increasing measurement uncertainty on the diagnostic accuracy of testing strategies. To that end, the error model simulation approach introduced in Chapter 3 was used to assess the impact of increasing FC imprecision and bias on the diagnostic accuracy of the NICE FC pathway and the YFCCP.

Both pathways exhibited the same overall pattern of results with respect to the impact of increasing measurement uncertainty. Additional bias has an intuitive impact, with positive bias increasing diagnostic sensitivity and decreasing diagnostic specificity (and vice versa for negative bias), due to patients being pushed above (below) the cut-off threshold. The impact of imprecision meanwhile depends on the probability density of the measurand distribution to which it is applied. For the IBD population, FC1 and FC2 values <100 µg/g were rare, with a cluster of values above 100 µg/g: applying imprecision to this population increases the spread of results, thereby pushing some IBD patients' values below the threshold and reducing the diagnostic sensitivity. The IBS population meanwhile exhibited a more even distribution of results around the cut-off for both FC1 and FC2 results, leading to imprecision having less of an impact on diagnostic specificity within both pathway analyses.

Whilst the pattern of results was similar across both assessments, the YFCCP produced significantly higher specificity results and slightly lower sensitivity results compared to the NICE FC pathway, resulting in a corresponding shift in the position of the contour lines across each evaluation (Figure 5-4 vs. Figure 5-10). Inspection of the NICE FC pathway contour plot also provides further indication that this pathway is under-performing in terms of diagnostic accuracy (Figure 5-4). In particular it can be seen that (with either sampling method), if a negative bias is applied, then significantly higher specificity can be achieved without overtly affecting sensitivity. Applying a negative bias of ~50 µg/g, for example, increases the NICE FC pathway specificity from ~60% to ~80% and maintains a high sensitivity of ~94%. Applying negative bias in this way is equivalent to increasing the FC cut-off threshold (in this example, to 100 µg/g); these results therefore illustrate the fact that, were a higher cut-off to be used

within this population, then a significantly higher diagnostic yield could be achieved (although still at a level inferior to the YFCCP).

In line with the higher sensitivity and lower specificity of the NICE FC pathway, this strategy showed greater robustness to increasing measurement uncertainty with regards to sensitivity, and lower robustness with regards to specificity. For example, over the bias range -50 to 50 µg/g and up to 40% CV, the NICE FC pathway maintained sensitivity above 90% (Figure 5-4); whilst over the same region the YFCCP sensitivity dropped to just below 70% (Figure 5-10). The main driver of reduced sensitivity here was negative bias. For example, applying a -50 µg/g bias (equivalent to raising the NICE FC threshold to 100 µg/g, or likewise raising the YFCCP threshold to 150 µg/g) has a greater detrimental impact on the YFCCP sensitivity due to the fact that several IBD patients had results within the 100-150 µg/g range and would therefore be missed in this scenario. In contrast, the YFCCP was substantially more robust to increasing measurement uncertainty with regards to specificity. Within the added bias range -50 to 50 µg/g, the YFCCP maintained specificity above ~75% across the entire CV range, whilst the NICE FC pathway specificity dropped to 0% within this same region. Positive bias was the primary driver here: for example, applying a 50 µg/g bias to the NICE FC pathway is equivalent to decreasing the threshold to 0 µg/g, which leads to all IBS patients being incorrectly referred, resulting in a 0% specificity.

Two sampling methods were explored in this analysis: parametric and bootstrap sampling. For both pathways, the bootstrap method provided a better fit to the data, producing baseline results which closely matched the diagnostic accuracy results calculated based on the empirical YFCCP dataset (i.e. this method had high internal validity). Within the NICE FC pathway evaluation, the parametric method overestimated the baseline sensitivity and underestimated the specificity, due to the fact that both the IBD Weibull FC1 distribution and the IBS lognormal FC1 distribution overestimated the proportion of values >100 µg/g, respectively (as shown by the face validity metric in Table 5-2). Within the YFCCP evaluation, the parametric method underestimated both sensitivity and specificity, due to the fact that the IBS lognormal FC2 distribution overestimated the proportion of values >100 µg/g (i.e. decreasing specificity), whilst the IBD Weibull FC2

distribution underestimated this same proportion (i.e. decreasing sensitivity) (Table 5-7).

A range of sensitivity analyses were conducted to explore the impact of key methodology assumptions on both sampling methods (Table 5-6 and Table 5-12). The bootstrap method was largely robust to the analyses explored. In particular, ignoring the biased complete-case analysis, censored data handling had no measurable impact within either pathway evaluation for the bootstrap method. This indicates that the pragmatic option of replacing censored data with their associated limit values was, in this example, a reasonable approach. The pragmatic approach to handing sampling uncertainty meanwhile – namely applying the smoothing algorithm – also appeared to be reasonable, producing similar results to both the "noisy" analysis and the extended approach of increasing the sampling number to 100,000, for both sampling methods. Interestingly, removing the sampling process altogether and running the error model directly on the YFCCP dataset produced only slightly different results to the bootstrap base case, suggesting that the dataset in this case study was sufficiently large to avoid sampling altogether.

In contrast to the bootstrap method, the parametric method exhibited large variability to several of the sensitivity analyses conducted. In particular the parametric method was sensitive to the parameterisation selected, with the adoption of alternative parametric distributions leading to a further reduction in the internal validity of this method; and this method was sensitive to the choice of upper bound selected for right-censored data within the 'fitdistcens' function used to elicit parameterisations. Based on these findings it would appear reasonable to discount this approach in favour of the bootstrap method, which is clearly optimal in terms of internal validity and stability.

The one sensitivity analysis which did have a notable impact on the bootstrap method results, was the analysis in which missing $FC2_{true}$ values were sampled by drawing from the population-specific empirical FC_{diff} distributions (as described in section 5.4.1.2.1). This approach maintained the high internal validity of the bootstrap method, but produced wider regions of acceptable bias within the acceptable regions (Table 5-12). Note that the acceptable bias boundaries here are driven by the sensitivity contours in the negative bias region, and the

specificity contours in the positive bias region (see Figure 5-10). The wider region of acceptable bias therefore indicates an increase in the robustness of the pathway's sensitivity and specificity to negative and positive bias, respectively. These two changes result from the fact that: (i) the IBD population FC2 values generated from this approach are higher on average than those produced in the base case analysis – resulting in a greater tolerance to negative bias in terms of correctly classifying IBD patients (i.e. sensitivity); and (ii) the IBS population FC2 values generated from this approach are lower on average than in the base case analysis – resulting in a greater tolerance to positive bias in terms of correctly classifying IBS patients (i.e. specificity). Crucially, without access to data on the expected distribution of FC2 values for patients with FC1 values <100 µg/g, it is not possible to provide any definitive conclusions as to the best approach to sampling missing FC2 values with the bootstrap method. Nevertheless, the results of this sensitivity analysis indicate that the approach taken to sampling missing FC2 values in the base case analysis, is likely to have produced conservative estimates of the acceptable regions.

Overall the simulation results showed that, whilst the NICE FC pathway's high sensitivity is robust to increases in bias and imprecision, the low specificity of this pathway is highly volatile – particularly to positive bias. The YFCCP meanwhile exhibits greater overall robustness: although the high sensitivity of this pathway is slightly less robust compared to the NICE FC pathway, the specificity is substantially more stable; furthermore, the diagnostic performance of this pathway is maximised around the point of zero added measurement uncertainty – as would be expected for an optimised pathway.

5.5.2.1 Acceptable regions

The primary tool used within this study to illustrate the simulation results was the contour plot. Behind each contour plot lies a 201x201 matrix of simulation results for each of the sensitivity and specificity outcomes (i.e. for each of the bias and imprecision pairs simulated); the contour plot function highlights the position of points within each matrix that matches selected levels (i.e. contours) for each outcome. In the first instance, these plots provide a visual summary of how the baseline diagnostic accuracy of each FC pathway changes over the space of increasing bias (in this case shown on the x-axis) and imprecision (y-axis). This

provides a useful tool for assessing the robustness of each pathway's diagnostic accuracy to increasing measurement uncertainty, as previously discussed.

In addition to providing a framework for the assessment of the robustness of each pathway to increased measurement uncertainty, the contour plots were further utilised in this study to present a new concept of *acceptable regions* of bias and imprecision. These regions were derived by specifying a minimum requirement for diagnostic sensitivity and specificity. The purpose was to illustrate how much additional FC measurement uncertainty could be tolerated within each pathway, in order to maintain a given outcome of interest: the acceptable regions, therefore, represent a form of outcome-based APS.

For the NICE FC pathway, both of the acceptable regions explored (including (a) the lower 95% CI values based on the baseline diagnostic accuracy assessment, and (b) 10% below the lower 95% CI values), appeared as off-centre from the baseline (0,0) point, for both sampling methods (Figure 5-5 and Figure 5-6). This is a result of the differential impact of bias on the pathway's sensitivity and specificity: positive bias resulted in a rapid decline of specificity, meaning that a limited amount of positive bias could be tolerated before breaching the minimum specificity level; whilst negative bias resulted in a less pronounced decline in sensitivity, meaning that a greater magnitude of negative bias could be tolerated before breaching the minimum sensitivity level. In particular, due to the fact that the parametric method underestimates the baseline specificity, the affect is most pronounced for the parametric method. The petal-like shape of the regions is due to the differential impact of imprecision: the right-side boundary is an (almost) straight vertical edge, resulting from the fact that increasing imprecision has little impact on the pathway specificity; whilst the left-side boundary is a diagonal sloping edge, resulting from the fact that increasing imprecision reduces the pathway sensitivity, as previously outlined.

For the YFCCP, the overall shape of the acceptable regions was similar to that observed for the NICE FC pathway, however the regions were much smaller (in terms of area under the curve). This was due to the greater impact of negative bias and increasing CV% on reducing the sensitivity of this pathway, resulting in lower levels of negative bias and CV being tolerated compared to the NICE FC pathway (Figure 5-11 and Figure 5-12). However, due to the increased

robustness of this pathways' specificity to positive bias, the resulting acceptable regions tended to lie more symmetrically over the (0,0) point, which resulted in higher TE_{max} values (compared to the NICE FC pathway). The exception to this was with the parametric method for the more stringent diagnostic accuracy requirement (Figure 5-11): in this case, the acceptable region was completely offset from the (0,0) point, due to the fact that this method underestimates the pathways' baseline specificity, to a value below the specified minimum requirement.

An interesting consequence of the off-centred positioning of the acceptable regions for the NICE FC pathway is that, whilst these regions are larger than those presented for the YFCCP in terms of area under the curve, the associated TE_{max} values are much smaller (Table 5-5 and Table 5-11). In this respect, the use of the TE_{max} summary metric represents a loss of information, since it fails to indicate that the NICE FC pathway is robust to high values of negative bias (albeit actually due to the fact that this pathway diagnostic cut-off threshold is not optimised). Nevertheless, the higher TE_{max} statistic achieved with the YFCCP appropriately reflects the fact that this pathway is more evenly robust to positive and negative bias, around an optimised (0,0) point.[38]

The principle question to consider when defining outcome-based APS is how to define a meaningful outcome specification – that is, what should be the minimum requirement for diagnostic sensitivity and specificity? For an intermediate outcome such as diagnostic accuracy, the knock-on impacts of misclassifying patients are not explicitly captured, and instead a judgement must be made about the clinical impact of a reduction in sensitivity and specificity and what is acceptable in this respect. Such a judgement will be context dependant. In this study, the primary utility of FC lies in avoiding unnecessary referrals for IBS patients, with minimal risk of severe outcomes resulting from delayed diagnosis

[38] It should be noted here that the acceptability regions across the two pathway analyses are not comparing like-for-like in absolute terms: if the primary acceptable region for the YFCCP (sensitivity ≥0.85; specificity ≥0.90) was applied directly to the NICE FC pathway, then the acceptability regions in this case would be empty (due to the fact that the NICE FC pathway fails to achieve the required specificity level at any point over the simulation space).

for IBD patients. High diagnostic specificity is therefore critical in this case. In alternative contexts, delayed diagnosis may be associated with greater risks, and high diagnostic sensitivity would instead be priority.

The primary acceptable region considered in this analysis was determined according to specifying minimum diagnostic accuracy at the lower 95% CI level of that achieved in practice. Whilst this was a somewhat arbitrary choice, it nevertheless aligns with the priority of maintaining high specificity, and reflects a reasonable assumption that laboratories and clinicians alike should want to maintain the performance levels of a test achieved within the research used to inform the adoption of that test. Another approach would be to conduct additional consultations with key stakeholders (e.g. clinicians, patients and payers), to identify a consensus on the minimum outcome required. Whilst this would require additional resources, it would ensure that the assumed requirement of clinical performance was acceptable to the relevant end-users.

Another alternative (evidence-based) approach would be to extend the framework presented herein to formally account for the knock-on effects of misclassifying patients, both in terms of costs and clinical consequences. This is the approach explored in Chapter 6, wherein the acceptable regions for FC within each clinical pathway are instead derived based on an analysis of cost-effectiveness – thus introducing an alternative concept of “cost-effective regions” of bias and imprecision.

5.5.3 Limitations

There are several key limitations with this study which relate to the data underpinning the analysis. Potential issues concerning partial verification bias within the diagnostic accuracy assessment, and the need to verify the applicability of the diagnostic accuracy findings beyond the York region, have already been discussed. Two further key limitations are discussed below.

The first issue relates to the incomplete availability of FC2 values within the YFCCP dataset. That is, the fact that FC2 data was only available for patients with an FC1 result above the 100 µg/g threshold (as per the YFCCP protocol), necessitating the assumption that missing FC2 values for FC1 values <100 µg/g (when required in the simulation) would follow the same distribution as

population-specific FC2 values available in the dataset. For the IBD population, this assumption only had to be applied for a maximum of 15 patients (19% of the IBD population): those who had FC1 values < 100 µg/g (n=5) and those who had FC1 values ≥100 µg/g but who were directly referred to secondary care without an FC2 test (n=10) (i.e. non-compliant referrals). For the IBS population, this assumption could, depending on the level of imprecision and bias applied within the initial error model, affect up to 701 patients who had an FC1 value <100 µg/g (80% of the IBS population).

Whilst the assumption of distributional equivalence cannot be directly tested or verified in the absence of complete FC2 data, a sensitivity analysis was conducted to explore an alternative approach: deriving missing FC2 values by sampling from available FC_{diff} values. As previously discussed, this analysis resulted in larger acceptable regions due to an increased tolerance to positive and negative bias (stemming from higher FC2 values being generated in the IBD population and lower FC2 values being generated in the IBS population). As such, the acceptable region estimated in the base case analysis may be considered conservative. No definite conclusions regarding the validity of either approach can be made, however, given per protocol truncated testing within the YFCCP dataset. In particular, if 'missing' FC2 values are expected to have a significantly different distribution than the available FC2 values, then both the base case and sensitivity analysis in this case may be biased. Clearly, if future studies wish to evaluate repeated testing scenarios using a similar simulation approach then attempts should be made to ensure that the data upon which the analysis is based provides complete information on all repeated tests (rather than per-protocol truncated testing, as in the YFCCP dataset).

The second key limitation in this analysis relates to the applicability of the study findings, and concerns the fact that the YFCCP FC data were not in fact "error-free" but rather incorporated a level of baseline uncertainty. The simulation results must therefore be interpreted as indicative of the change in diagnostic accuracy resulting from *additional* bias and imprecision, on top of this baseline uncertainty. Ideally, for the findings to be of use to other laboratories outside of York (i.e. within the context of APS), one would want to know the impact of bias and imprecision starting from an error-free position, such that laboratories could directly relate

their levels of bias and imprecision to the acceptable region presented. As it is, the results provided herein represent levels of bias and imprecision which can be tolerated *on top of* that contained within the YFCCP data itself, which is a clear barrier to wider implementation of the study findings.

This issue can be partially addressed by attempting to quantify the baseline uncertainty contained within the YFCCP dataset FC values. Two pieces of information are available for this task. First, the internal quality control results conducted within the York laboratory over the study period provide evidence of the imprecision resulting from the ELISA platform (with a reported CV of 7% at ~50 µg/g, and 4% at ~150 µg/g) (271). Second, results from a further in-house analysis assessed variability resulting from the sample extraction process. This study found a CV range of 10-15% (271). Using the upper value of 15%, for example, one can calculate a combined baseline imprecision of 16.6% (at ~50 µg/g) and 15.5% (at ~150 µg/g), using the sum of squares rule.[39] The baseline 'zero' CV% point on the contour plots therefore actually represents this baseline imprecision, which would need to be subtracted from the CV associated with any new FC assay being assessed, in order to avoid double counting. Note, however, that there are likely other elements of imprecision, resulting from additional pre-analytical and analytical processes, which would also need to be accounted for in this calculation. Furthermore, reliable estimation of baseline bias is unfortunately not possible in this case study due to a lack of reference measurement procedure for FC.

An alternative approach to this issue, if using the parametric method, is to apply statistical adjustments to baseline distributions to remove known bias and imprecision (2). As previously outlined in Chapter 3 (section 3.4.1), this approach has been used in a handful of previous studies assuming simple normal or lognormal distributions, to remove analytical variability from an associated estimate of total imprecision to isolate variability associated with the "pure biologic distribution" (as illustrated in Appendix H) (119-121, 125, 127, 143). It should be noted however that, like with the above analysis, in order to provide valid results this approach would require complete and reliable information on the baseline

[39] At 50 µg/g: $\sqrt{15^2 + 7^2} = 16.6$. At 150 µg/g: $\sqrt{15^2 + 4^2} = 15.5$.

levels of bias and imprecision. In addition in the context of the current case study, the parametric method exhibited a lack of internal validity – this approach to removing baseline uncertainty was therefore not explored further in this study.

Determining better approaches to the issue of baseline measurement uncertainty is an important aspect for consideration in future research. In particular more sophisticated, prospective approaches could be explored: for example, the value of reference measurement procedures and certified reference materials (not yet available for FC) for deriving $Test_{true}$ values, could be considered for alternative test evaluations. In the current example, in the absence of any clear means to removing baseline measurement uncertainty, the results of this case study should be interpreted as illustrating the impact of *additional* bias and imprecision, on top of baseline measurement uncertainty.

5.6 Summary

- In this chapter, the error model simulation approach was used to assess the impact of increasing FC measurement uncertainty on the diagnostic accuracy of the NICE FC pathway and the YFCCP. These results support hypothesis C of this thesis: that methods from the broader literature (i.e. identified in Chapter 3) may be applied within HTA-style assessments, to evaluate the impact of measurement uncertainty on clinical performance outcomes (note clinical utility and cost-effectiveness outcomes are further evaluated in Chapter 6).
- The simulated diagnostic accuracy results were presented using contour plots, which provided a visual aid to assess the robustness of each pathway's diagnostic accuracy to increased bias and imprecision.
- The contour plots were also used to illustrate a new concept of "acceptable regions" of bias and imprecision, defined according to an assumed minimum diagnostic accuracy requirement. This concept relates to hypothesis D of this thesis: that the application of methods from the broader literature to HTA-style assessments could enable outcome-based APS to be derived.
- The results indicated that the NICE FC pathway is a sub-optimal pathway which is highly volatile to positive bias. In contrast, the YFCCP was found to be to be an optimised and relatively robust strategy.
- Whilst the acceptability regions provided useful information on maximum boundaries for bias and imprecision, a key limitation of this approach is the need to set a minimum diagnostic accuracy requirement.

In **Chapter 6**, the analysis presented in this chapter is extended to clinical utility (QALY) and cost-effectiveness (NMB and INMB) outcomes (described in Appendix D).

Chapter 6
The impact of measurement uncertainty on the cost-effectiveness of FC testing strategies

6.1 Chapter outline

In Chapter 5, the error model simulation approach was used to assess the impact of increasing FC imprecision and bias on the diagnostic accuracy of the NICE FC pathway and the YFCCP. Although these results illustrate the robustness of each pathway's diagnostic accuracy to measurement uncertainty, the question remains as to how these findings translate to "end stage" outcomes, such as patient health outcomes (e.g. QALYs) and cost-effectiveness. The aim of this chapter was to extend the simulation framework presented in Chapter 5, to evaluate the impact of increasing FC measurement uncertainty on end-stage outcomes. As in Chapter 5, the analysis presented in this chapter addresses hypotheses C and D of the thesis (see section 1.5.3).

A linked-evidence economic decision model, previously developed to evaluate the cost-utility of the YFCCP (which also included a NICE FC pathway-equivalent comparator arm), was used as the foundation for this analysis. The impact of increasing FC measurement uncertainty on the modelled cost, QALY and NMB outcomes was explored by embedding the simulation results from Chapter 5 (including the parametric and bootstrap sampling method results) within the FC cost-utility model. The findings are again presented using contour plots, which are here used to illustrate the concept of "cost-effective regions" of bias and imprecision. In addition, the concept of "optimal regions" is introduced, as a means of setting analytical performance to maximise NMB. Section 6.2 below describes the methods of this analysis, followed by the results (section 6.3) and a discussion (section 6.4).

6.2 Methods

6.2.1 YFCCP economic model

The economic model used as the basis for this analysis was previously developed by the York Health Economics Consortium (YHEC) group. An initial version of the model, commissioned by the Yorkshire and Humberside Academic Health

Science Network (YHAHSN), was developed by YHEC to conduct a cost-consequences analysis of the YFCCP (i.e. evaluating cost and effect estimates separately) (239). The cost-consequences model was subsequently updated by YHEC in 2018/19 to capture the impact of delayed diagnosis on patient health-related quality of life, thereby providing cost-utility[40] estimates (i.e. ICERs and INMB) (241). The updated model (henceforth referred to as the 'FC cost-utility model') was used for the basis of this analysis. A final version of the FC cost-utility model was kindly provided for the purpose of this analysis by the model developer, Hayden Holmes (Senior Consultant, YHEC), in June 2019. This section provides an overview of the FC cost-utility model; the model details are also provided in an associated YHEC publication (241). Note that, within the *de novo* analysis presented in this chapter, no alterations were made to the FC cost-utility model structure or parameters, other than those required to capture the impact of measurement uncertainty as outlined in section 6.2.2.

6.2.1.1 Model structure

The FC cost-utility model adopted a 1-year time horizon, which was intended to track patients from initial presentation with lower gastrointestinal symptoms in primary care, through to confirmed diagnosis. The structure of the model is illustrated in Figure 6-1 (YFCCP intervention pathway) and Figure 6-2 (example comparator pathway, using a single FC test). A full list of input parameters used in the model is provided in Appendix L.

In the FC cost-utility model YFCCP intervention arm (referred to as the 'fixed' YFCCP arm within the subsequent *de novo* analysis), patients present at an initial GP visit and are administered their first FC test (FC1) (Figure 6-1). After this, patients return for a follow-up visit, where a confirmatory FC test (FC2) may be administered (as per the YFCCP protocol). Patients diagnosed with suspected IBD are assumed to be referred directly to secondary care, where all patients receive a specialist visit followed by colonoscopy (after which patients receive a definitive diagnosis, including both true positive and false positive cases).

[40] Note: a *cost-utility analysis* (or model) refers to a cost-effectiveness analysis in which QALYs are used as the measure of health benefit. Since cost-utility analysis is a type of cost-effectiveness analysis, these terms are used interchangeably in this study.

Patients diagnosed with IBS are assumed to be managed in primary care, and receive first-line IBS medication. All false negative patients (i.e. patients with IBD incorrectly classified as IBS) are assumed to return to their GP with persisting symptoms, whilst a proportion of true negative patients are also assumed to return. The majority of returning patients are administered second line IBS treatment: all false-negative patients are assumed to return again with persisting symptoms and are subsequently referred to secondary care (again with a specialist visit followed by colonoscopy); whilst second-line IBS treatment is considered to be effective in all IBS patients. Of those patients referred to secondary care without attempting second line IBS treatment, all IBD patients are assumed to receive a specialist visit followed by colonoscopy; whilst a subset of IBS patients are assumed to receive colonoscopy.

Figure 6-1. FC cost-utility model structure: YFCCP intervention arm

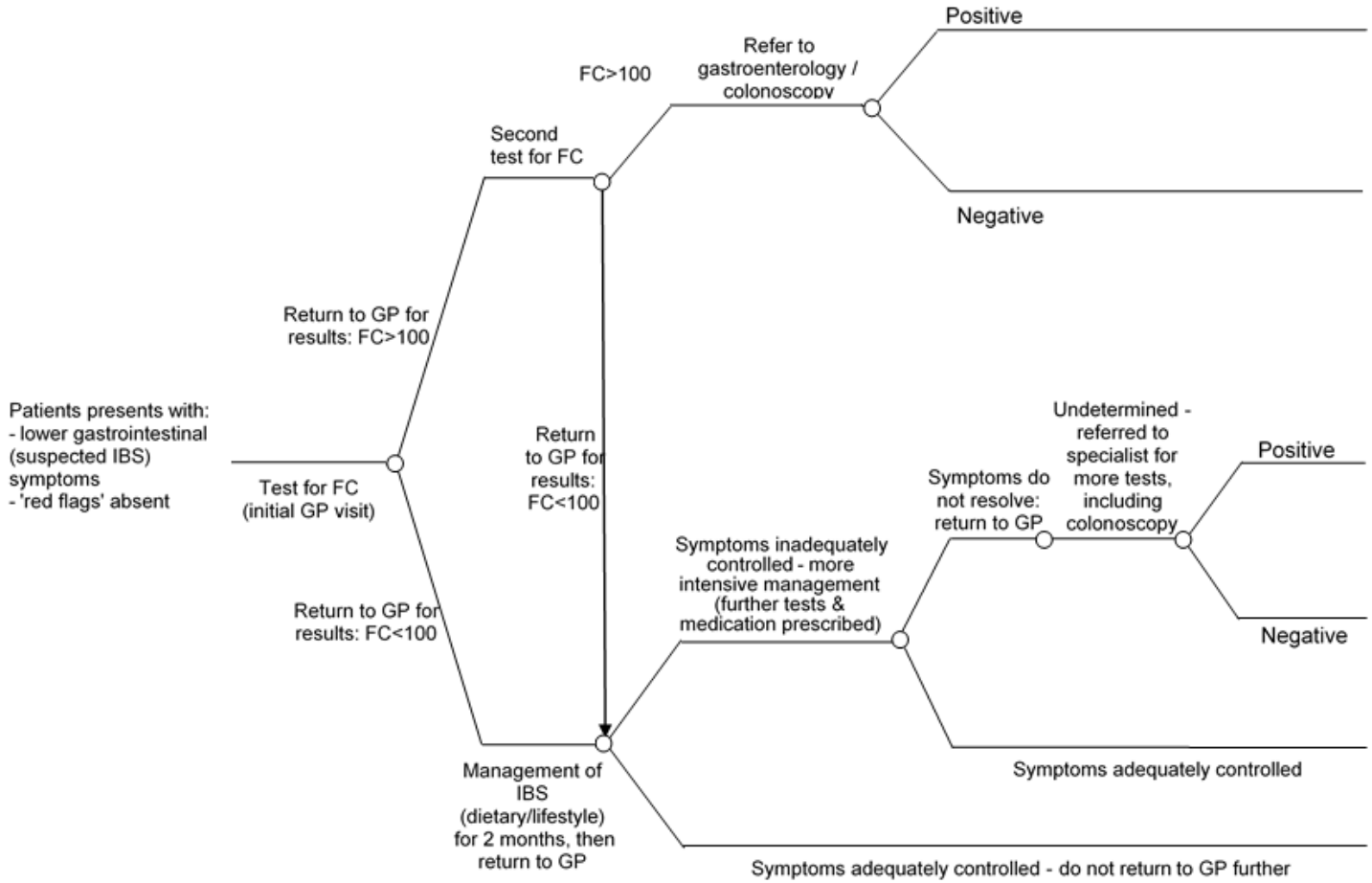
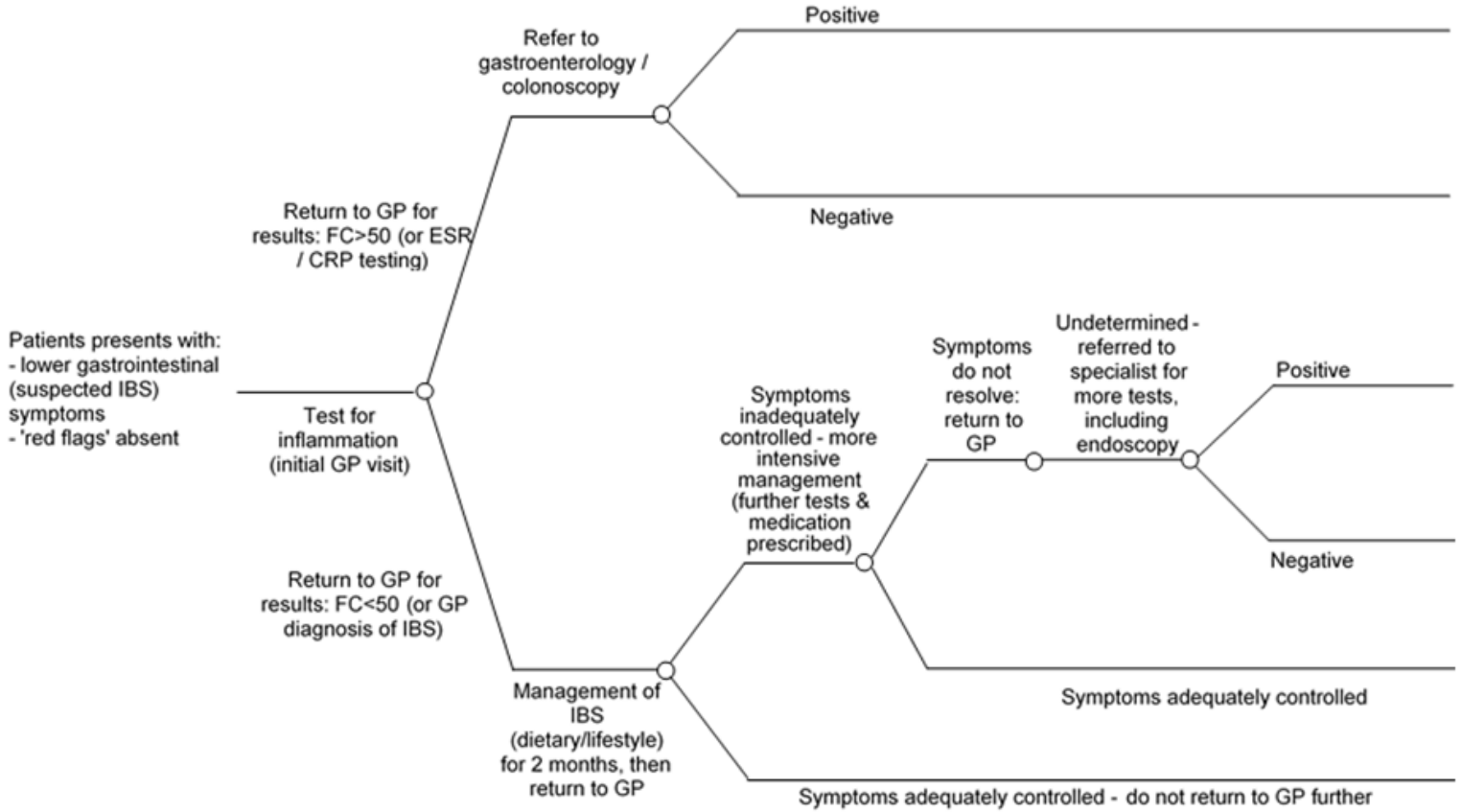


Figure 6-2. FC cost-utility model structure: example comparator arm (single FC test)



Five comparator arms were included in the FC cost-utility model. The general structure of the comparator arms was the same as the intervention arm, however initial diagnoses were based on alternative mechanisms. The comparators included two 'No FC' pathways, and three 'FC testing' pathways (each assuming the standard 50 µg/g cut-off threshold). The different arms in each case relate to alternative data sources used to inform the diagnostic accuracy of the referral strategy, as outlined below:

1. **No FC (Tibble data):** This comparator assumed that secondary care referrals would be based on ESR and CRP tests conducted by the GP. The diagnostic accuracy was based on a published assessment of ESR and CRP conducted within a cohort of patients referred to a secondary care gastroenterology unit (Tibble *et al.* 2002) (218). The accuracy estimates were taken from the study analysis of "low-risk" patients, which excluded patients with red flag cancer symptoms (including anaemia, weight loss or rectal bleeding) (218)).
2. **No FC (NICE data):** This comparator assumed that secondary care referrals would similarly be based on GP decision-making, this time encompassing additional factors considered by the GP as well as ESR and CRP results. The diagnostic accuracy of this pathway was based on data from the systematic review conducted by Waugh *et al*. (2013) as part of the 2013 NICE DAP assessment of FC (220). The majority of studies identified in this review came from secondary care (220).
3. **FC testing (YFCCP data, 50 µg/g cut-off):** This comparator assumed that a single FC test using the standard cut-off threshold would be conducted. Diagnostic accuracy was derived from the YFCCP primary care dataset, by calculating what the accuracy would have been had only the FC1 result been used at the standard cut-off value. Note that this pathway is equivalent to the baseline NICE FC pathway presented in Chapter 5 (but not accounting for the impact of measurement uncertainty).
4. **FC testing (Tibble data):** This comparator also assumed that referrals would be based on the NICE FC pathway, this time basing diagnostic accuracy on an assessment of FC conducted within the Tibble *et al*. (2002) study (again using the "low-risk" patient cohort) (218).

5. **FC testing (NICE data):** This comparator also assumed that referrals would be based on the NICE FC pathway, this time basing diagnostic accuracy on the 2013 NICE DAP systematic review findings (Waugh *et al*. 2013) (220).

6.2.1.2 Model parameters

The FC cost-utility model was deterministic in nature (i.e. all input parameters were fixed at their mean expected value). A full list of the parameters applied in the model (including costs, utilities and event timings, discussed further below), is provided in Appendix L. A summary of the diagnostic sensitivity and specificity values applied within each arm in the model is also provided in Table 6-1 below.

For the YFCCP intervention arm and the FC testing comparator arm (YFCCP data, 50 µg/g cut-off), diagnostic accuracy was derived from the YFCCP dataset – i.e. using the same baseline diagnostic accuracy values as reported in Chapter 5 (sections 5.4.2.1 and 5.3.2.1 respectively). For the remaining comparators, YHEC identified two key data sources to inform the diagnostic accuracy estimates, from a targeted search of the literature: (1) the Waugh et al. (2013) systematic review, which informed the NICE DG11 guidance (hence referred to as the 'NICE data') (220)[41]; and (2) the Tibble et al. (2002) study, which was used as a primary evidence source in a 2010 NHS report on FC (88, 218).

Table 6-1. FC cost-utility model: diagnostic accuracy estimates

Test strategy	Sensitivity	Specificity
Intervention (YFCCP data)	94%	92%
No FC (Tibble data)	35%	73%
No FC (NICE data)	100%	79%
FC testing (YFCCP data, 50 µg/g cut-off)	96%	60%
FC testing (Tibble data)	90%	80%
FC testing (NICE data)	93%	94%

[41] YHEC noted that NICE had conducted a review of the DG11 guidance in May 2017, and subsequently moved this guidance to the static list, indicating that there had been no significant new evidence published in the literature over that time period. As such, YHEC considered the Waugh review to represent a key source for the most up-to-date evidence.

The YHEC cost-effectiveness analysis was conducted from an NHS perspective. Health care costs captured in the model (reported in 2017/ 2018 GBP) included: (i) primary care costs (i.e. GP initial and follow-up visits); (ii) test costs (i.e. FC, ESR, CRP); (iii) secondary care costs (specialist visit and colonoscopy); and (iv) IBS treatment costs (first and second line treatments). Health-related quality of life was captured in the model according to how long patients occupied 'treated' or 'untreated' IBS/IBD health states, which were each associated with different utility values. The time spent in untreated vs. treated health states was derived based on assigned timings of modelled events. For example, the time to achieve a correct diagnosis for IBD true positive cases was composed of the sum of the time assigned to the following events: GP visit and FC1 testing, follow-up GP visit and FC2 testing (where applicable), secondary care specialist visit, and colonoscopy. The 'IBD untreated' utility was then applied in the model for the time to correct diagnosis, and the 'IBD treated' utility was applied for the remainder of the 1-year time horizon. This same process was completed for four diagnostic subgroups: true positives, false positives, true negatives and false negatives.

The impact of the YFCCP intervention in the model depends on the comparator arm selected. For most comparators, the YFFCP was associated with a higher sensitivity (Table 6-1). In those cases, the intervention leads to a higher proportion of true positive cases and a lower proportion of false negative cases, resulting in: more IBD patients receiving a faster diagnosis (and thereby higher QALYs due to spending less time untreated); cost savings resulting from avoiding unnecessary IBS treatment, re-testing and additional GP visits; and a cost increment resulting from the FC testing itself. In most cases the YFCCP was similarly associated with a higher specificity value. In those cases the intervention leads to a higher proportion of true negative cases and a lower proportion of false positive cases, resulting in: more IBS patients receiving a faster diagnosis (and higher QALYs due to spending less time untreated); cost savings resulting from avoiding unnecessary secondary care specialist appointments and colonoscopies; and a cost increment incurred from the FC testing itself. In those cases where the YFCCP sensitivity or specificity was lower than the selected comparator arm, the relationships outlined above were reversed.

Key results of the original FC cost-utility model are presented in Table 6-2[42]. The YFCCP intervention was found to dominate the majority of the comparator strategies evaluated, being associated with lower mean costs and higher mean QALYs. The only comparator that the YFCCP did not dominate was the 'FC testing (NICE data)' strategy, which had a comparable performance to the YFCCP in terms of diagnostic sensitivity and specificity. Nevertheless, the YFCCP was found to be cost-effective compared to this strategy, producing an INMB of £19 per patient. In contrast, the equivalent of the NICE FC pathway, the 'FC testing (YFCCP, 50 µg/g cut-off)' comparator, was associated with the highest costs and lowest QALYs of all the strategies assessed, resulting in the lowest NMB.

Table 6-2. FC cost-utility model: fixed strategy results

Comparator	Model Inputs		Model Results			
	Sensitivity	Specificity	Mean cost	Mean QALY	Mean NMB (£)	INMB (£) YFCCP vs. comparator
YFCCP intervention	94%	92%	£212	0.7896	£15,581	-
No FC (Tibble data)	35%	73%	£259	0.7836	£15,412	£169
No FC (NICE data)	100%	79%	£232	0.7879	£15,526	£55
FC testing (YFCCP, 50 µg/g cut-off)	96%	60%	£314	0.7836	£15,359	£222
FC testing (Tibble data)	90%	80%	£245	0.7860	£15,474	£107
FC testing (NICE data)	93%	94%	£197	0.7880	£15,562	£19

6.2.2 Error model simulation

The FC cost-utility model was based on the "linked evidence" approach – i.e. linking diagnostic accuracy inputs with data on disease prevalence, costs and utilities. This structure enabled the diagnostic accuracy results reported in Chapter 5 to be "bolted-on" or embedded into the FC cost-utility model. That is, a sensitivity analysis was run, in which each of the sensitivity and specificity

[42] For an overview of how NMB and INMB statistics are calculated, please see Appendix D.

results from the base case error model simulation analyses reported in Chapter 5, were iteratively applied within the economic model. The observed cost-utility outputs were then recorded for each iteration, and linked back to the underlying values of bias and imprecision used in the error model simulation.

The FC cost-utility model was built in Excel. As such, the iterative simulation outlined above was implemented using an Excel (2016) macro coded using the visual basics language. Due to the lower computational efficiency of Excel compared to R, a subset of n=10,201[43] diagnostic accuracy outputs from the preceding error model simulation analysis were implemented within the Excel macro (rather than the total set of n=40,401 results). This number was sufficient to produce stable contour graphs when using the base case simulation results (i.e. based on the smoothed simulation results, as described in section 5.3.1.2.3).

For the NICE FC pathway, base case results from the error model simulation (as reported in section 5.3.2.2.1)[44] were iteratively applied to the diagnostic accuracy inputs for the 'FC testing (YFCCP, 50 µg/g cut-off)' comparator arm within the FC cost-utility model. Similarly for the YFCCP, the base case error model simulation results (reported in section 5.4.2.2.1) were applied to the fixed YFCCP intervention arm within the FC cost-utility model. For each iteration of the FC cost-utility model, the following results were recorded: mean costs, QALYs and NMB. Contour plots for each of the modelled outcomes were then constructed (in the same way as for diagnostic accuracy in Chapter 5), to illustrate the impact of increasing FC bias and imprecision on the modelled outcomes for each pathway in isolation (referred to in the following results section as the pathway "absolute" outcomes). Note that in the primary analysis, the calculation of NMB (and INMB, discussed below) assumed a cost-effectiveness threshold of £20,000 per additional QALY, in line with the current threshold adopted by NICE (280).

In addition to absolute outcomes, INMB was derived for each of the simulated pathways. Since the original FC cost-utility model did not explore the impact of

---

[43] Corresponding to an analysis of CV ranging from 0 to 100% (in 1% increments) and bias ranging from -100 to 100 µg/g (in 2 µg/g increments).

[44] Note: this analysis focused on the base case simulations (bootstrap and parametric) for each of the NICE FC pathway and the YFCCP. The sensitivity analyses explored in Chapter 5 were not applied in this analysis.

increasing measurement uncertainty, each of the strategies explored in that model produced fixed NMB results (summarised in Table 6-2). For the simulated YFCCP and NICE FC pathways however (i.e. including the impact of increasing measurement uncertainty), n=10,201 separate NMB results were produced for each of the bias and imprecision values explored. INMB for each of the simulated strategies was therefore calculated by subtracting the fixed intervention and comparator strategy NMB results (n=6, Table 6-2), from each of the two simulated pathway results. Note that, since this process simply subtracts the same fixed value from each of the simulated NMB results (n=10,201), the resulting INMB contour plots exhibit the same shape as the associated absolute NMB contour plots, but with different values attached to the contours (see section 6.3.2 for an example).

An alternative analysis of interest is to compare the two simulated FC strategies against each other. An additional incremental analysis was therefore conducted, comparing the simulated YFCCP with the simulated NICE FC pathway. Note that in this case, since the results of both of these strategies alter over the simulated space of bias and imprecision, the shape of the resulting INMB contour plots is different to that of the associated absolute NMB contours (see section 6.3.2.2).

As in the previous chapter, contour plots were used to highlight acceptable regions of bias and imprecision. For the INMB evaluation, each point on the contour plot (n=10,201) was classified as either cost-effective, where INMB ≥ £0, or non-cost-effective, where INMB<0. The "*cost-effective region*" was then defined as the area of the contour plot in which positive INMB was maintained. The advantage of this approach to specifying a region of acceptable analytical performance, is that it does not require any user-based judgement to be made as to what level of outcome is required. As in Chapter 5, TE bands were overlaid onto the contour plots and TE_{max} values were extracted. In addition, in order to explore the influence of the cost-effectiveness threshold within the evaluation, a further analysis was conducted in which the cost-effective region TE_{max} value was recorded for a range of threshold values (from £0 to £150,000, in £1,000 increments).

In addition to the cost-effective region, a further novel concept is introduced in this analysis: the "*optimal region*" of bias and imprecision. This region is defined

as the area of the INMB (or NMB) contour plot maintaining the top x% of cost-effectiveness results (where 'x' is user defined). In this example, 'x' was arbitrarily set at 10% – for the INMB plot, this is equivalent to maintaining INMB ≥ the INMB 90th percentile. The motivation for presenting the optimal region, and the potential advantages of this approach, are outlined in the following results (section 6.3.2) and discussion sections (section 6.4.2).

6.3 Results

6.3.1 Absolute results: mean costs, QALYs and NMB

6.3.1.1 Simulated NICE FC pathway

Figure 6-3, Figure 6-4 and Figure 6-5 provide contour plots showing the mean cost, QALY and NMB results for the NICE FC pathway respectively.

As in Chapter 5, from these figures we can assess how the specified outcome changes in response to additional FC bias and imprecision. Focusing on costs, for both sampling methods, positive bias increases costs and negative bias decreases costs; while increasing imprecision has marginal impact (Figure 6-3). By reference to Figure 5-4 from Chapter 5 (which showed the base case diagnostic accuracy contour plots for the NICE FC pathway), it is clear that the costs in this case are driven by the pathway specificity values (since the cost and specificity contours follow the same pattern). As such, the pathway costs in this case appear to be similarly volatile to increased positive bias. Observation of the QALY contour plots, meanwhile, illustrates that positive bias decreases QALYs and negative bias increases QALYs; whilst increasing imprecision slightly decreases QALYs (Figure 6-4). Again it appears that the QALY results are driven by the pathway's specificity, with QALYs decreasing rapidly in line with added positive bias. Finally, the NMB contour plots combine the cost and QALY data, with the resulting effect that NMB is volatile to increased positive bias; whilst increasing imprecision has marginal impact (Figure 6-5).

Due to the lower internal validity exhibited with the parametric method (as discussed in Chapter 5), the cost, QALY and NMB estimates for this method at the (0,0) point also displayed discrepancies when compared to the baseline results produced from the FC cost-utility model. Running the FC cost-utility model

for the baseline (fixed) NICE FC pathway (i.e. the 'FC testing (YFCCP, 50 µg/g cut-off)' comparator; sensitivity=96% and specificity=60%) produced a mean cost of £314, QALY of 0.7837, and NMB of £15,359 (Table 6-2). Within the simulation, the bootstrap method produces the closest match to the baseline values (£313, 0.7837 and £15,361 respectively); whilst the parametric method produces slightly diverging values (£325, 0.7833 and £15,342, respectively).[45]

[45] Reported values have been rounded to the nearest pound (for costs and NMB) and to 4 decimal places (for QALYs).

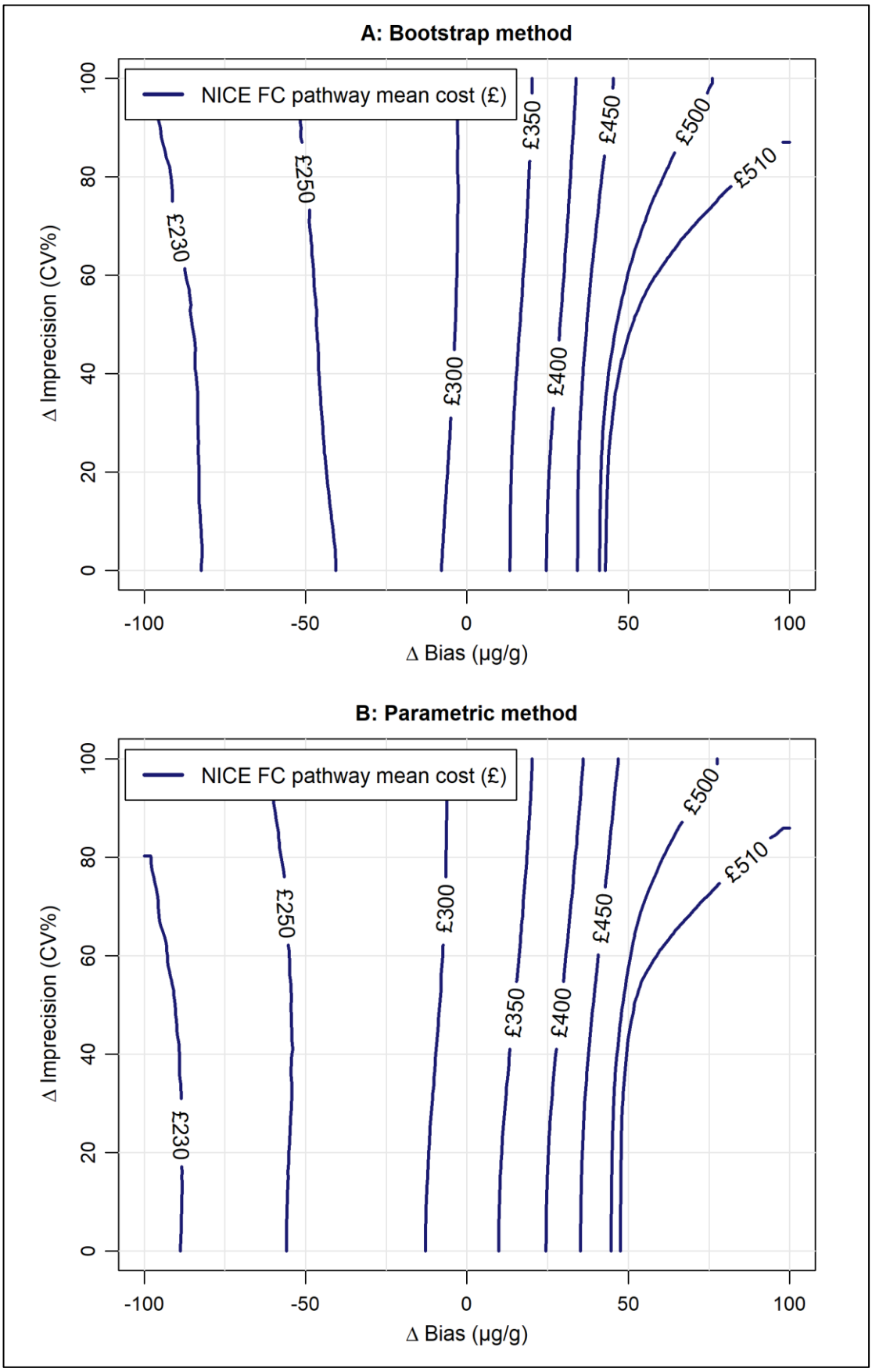


Figure 6-3. NICE FC pathway: contour plot of mean cost (£)

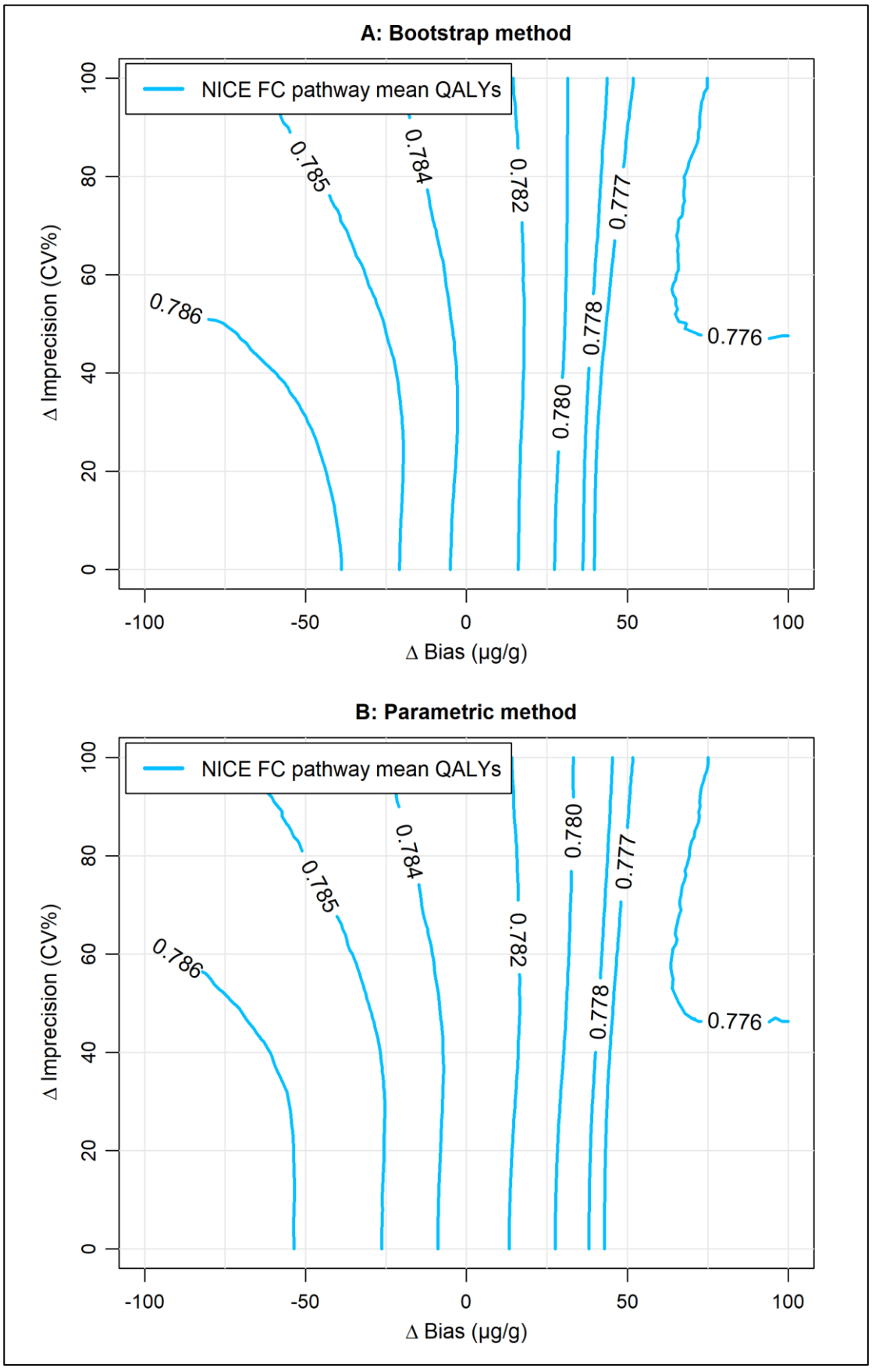


Figure 6-4. NICE FC pathway: contour plot of mean QALYs

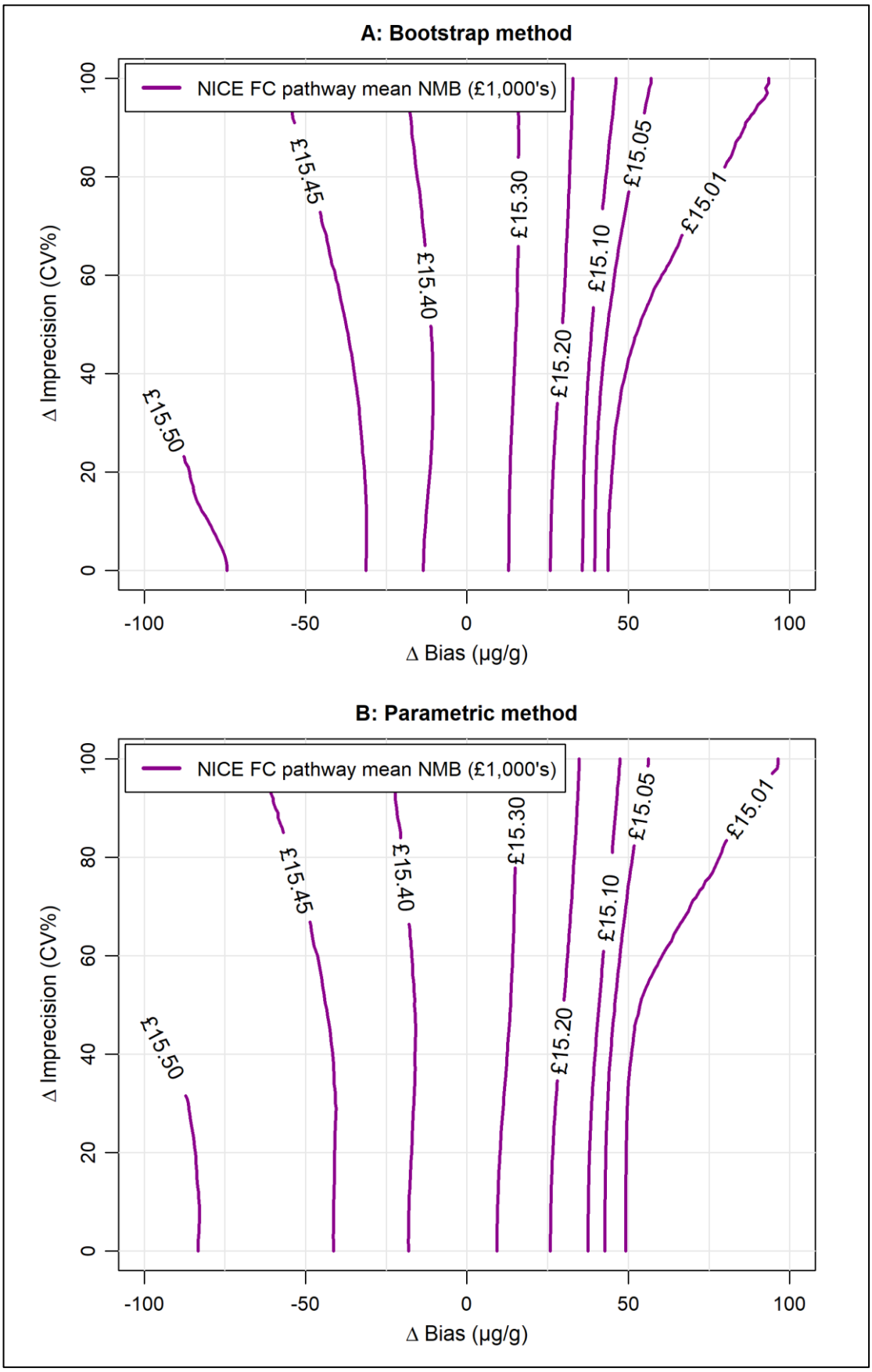


Figure 6-5. NICE FC pathway: contour plot of mean NMB (£1,000's)

6.3.1.2 Simulated YFCCP pathway

Figure 6-6, Figure 6-7 and Figure 6-8 provide contour plots showing the mean costs, QALYs and NMB results for the simulated YFCCP pathway.

Similarly to the NICE FC pathway evaluation, for both sampling methods positive bias increases the YFCCP costs, and negative bias decreases costs; while increasing imprecision has marginal impact on this outcome (Figure 6-6). By reference to Figure 5-10 from Chapter 5 (which showed the base case diagnostic accuracy contour plots for the YFCCP), we can see that the costs in this case are similarly driven by the pathway specificity. Based on the greater robustness of this pathway's specificity to positive bias, the costs in this case similarly exhibit a slower decline in the region of positive bias compared to the NICE FC pathway (Figure 6-6 vs. Figure 6-3). The QALY contour plots meanwhile illustrate that positive bias decreases QALYs and negative bias increases QALYs; whilst increasing imprecision also decreases QALYs, except in the region of high positive bias (Figure 6-7). The shape of the QALY contours is markedly different for the YFCCP compared to the NICE FC pathway. The results indicate that the YFCCP diagnostic sensitivity is a greater driver of QALYs, particularly in the negative bias region, with specificity appearing to be the predominant determinant of QALYs in the higher positive bias region (Figure 6-7). The NMB contour plots reflect a similar shape to the QALYs, with positive bias decreasing NMB and negative bias increasing NMB; and imprecision decreasing NMB up to a moderate range of positive bias (Figure 6-8). Focusing on the bootstrap method results, the overall effect is that the YFCCP's NMB is robust to imprecision and bias up to a moderate region of positive bias.

As with the NICE FC pathway analysis, due to the lower internal validity exhibited with the parametric method, this method produces slightly biased baseline cost, QALY and NMB estimates. Running the original cost-utility model for the YFCCP based on the baseline YFCCP diagnostic accuracy estimates (i.e. 94% sensitivity and 92% specificity) produced a mean cost of £212, a mean QALY of 0.7896, and a mean NMB of £15,581 (Table 6-2). The bootstrap method simulation results produced matching baseline values (£212, 0.7896 and £15,581 respectively); whilst the parametric method produced diverging values (£221, 0.7892 and £15,563, respectively).

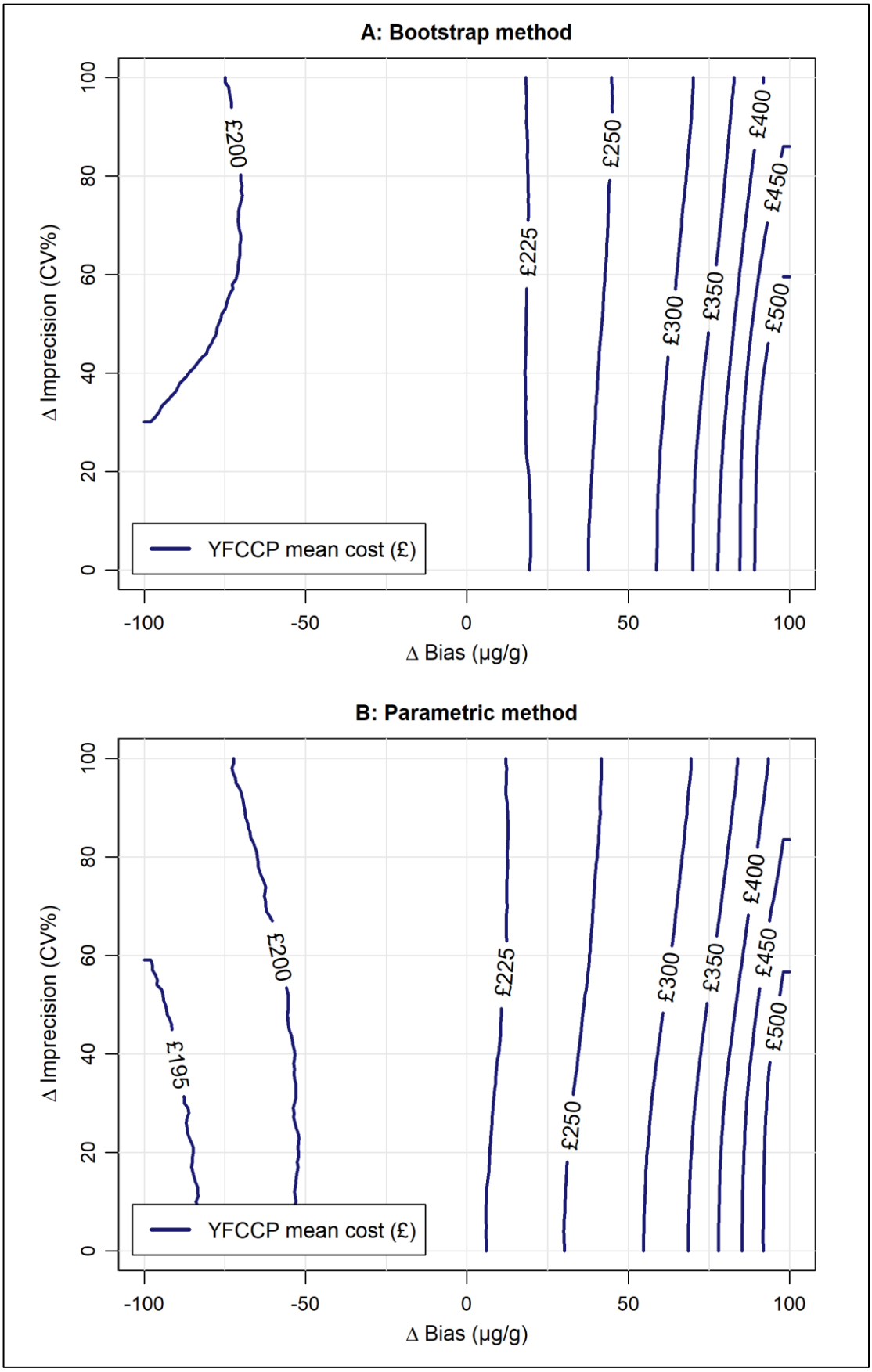


Figure 6-6. YFCCP: contour plot of mean cost (£)

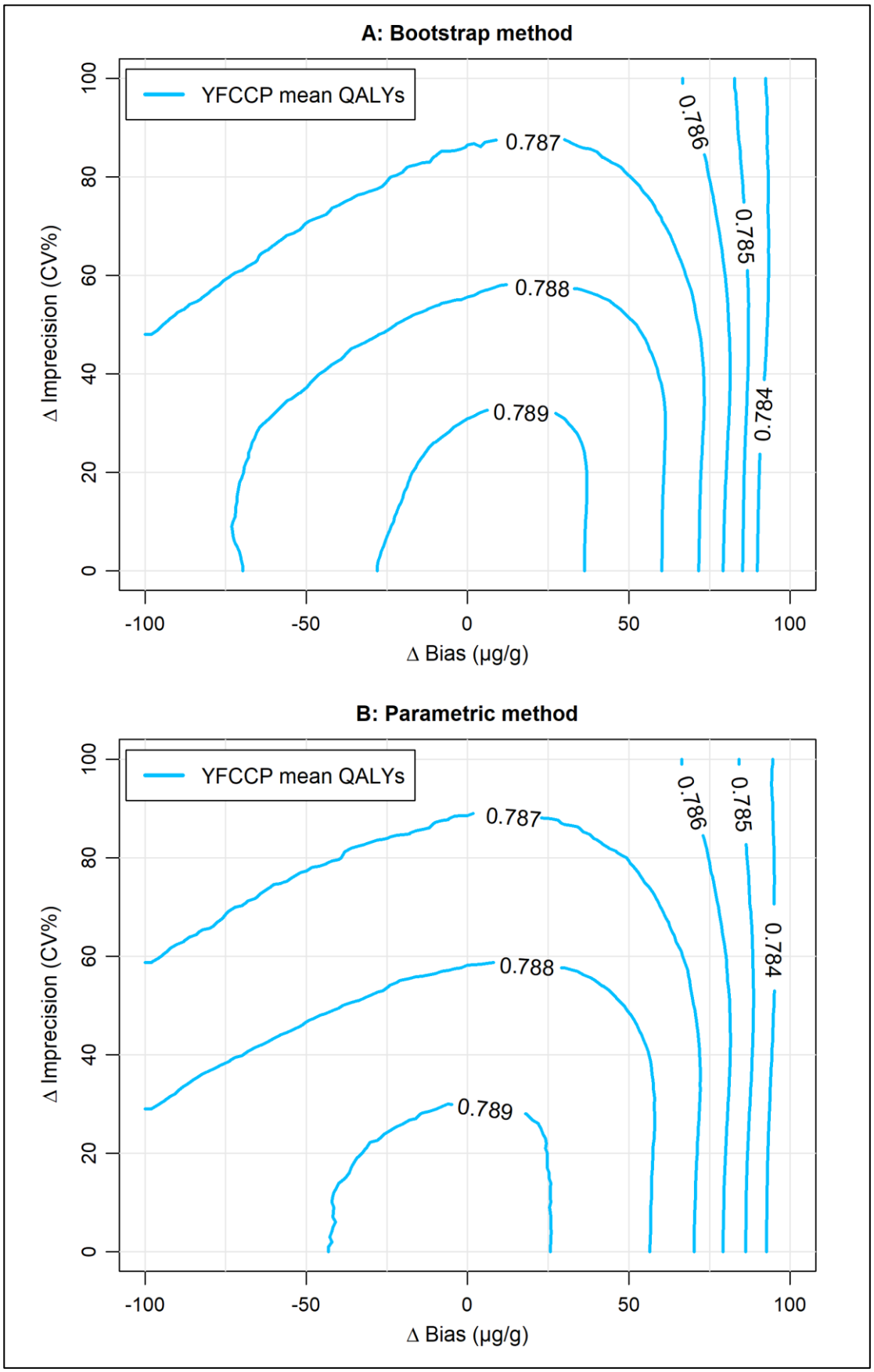


Figure 6-7. YFCCP: contour plot of mean QALYs

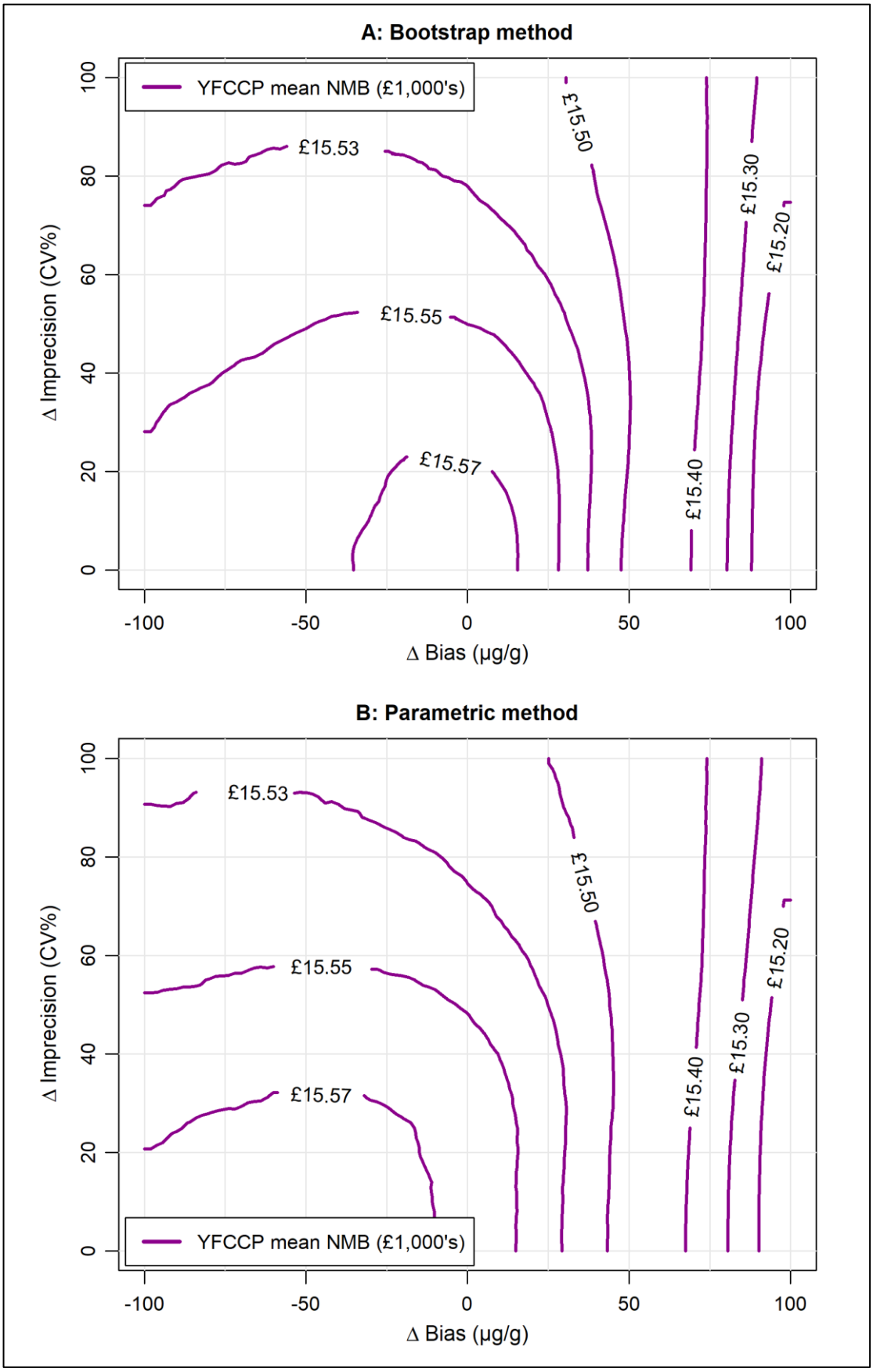


Figure 6-8. YFCCP: contour plot of mean NMB (£1,000's)

6.3.2 Incremental results

6.3.2.1 Simulated FC pathways vs. fixed model comparators

As previously discussed, calculation of INMB for each of the simulated FC pathways (including measurement uncertainty) compared to the six fixed model comparators produces contour plots exhibiting the same shape (but different values) as the NMB plots previously presented. To illustrate this point, an example is provided in Figure 6-9. This shows the INMB contour plots for the simulated YFCCP strategy versus the 'FC testing (NICE data)' comparator strategy, which has a fixed NMB of £15,562 (Table 6-2). This value has therefore been subtracted from each point (n=10,201) on the previously reported YFCCP NMB contour plot (Figure 6-8), to produce the INMB plot reported in Figure 6-9. It is clear by comparison of Figure 6-8 and Figure 6-9, that the shape of the corresponding contour plots is unchanged – only the value attached to the given contours is altered.[46]

Figure 6-9 highlights the position of the zero INMB contour as a dashed line. This contour is of particular interest, as it separates the plot into the cost-effective region (in this case, all points below the zero INMB line, where INMB >£0) and the non-cost-effective region (all points above the zero INMB line, where INMB <£0). The cost-effective region for this example has been highlighted in Figure 6-10, and TE bands have been additionally overlaid. In this case, for the bootstrap method, TE_{max} = 20%; whilst for the parametric method TE_{max} = 0%, due to the fact that the cost-effective region with this method is off-centre from the baseline (0,0) and only just touches the zero bias point.

[46] Note however that due to selecting slightly different relative contour values to present in each figure, the contour lines shown in Figure 6-9 do not appear in exactly the same position as those shown previously in Figure 6-8.

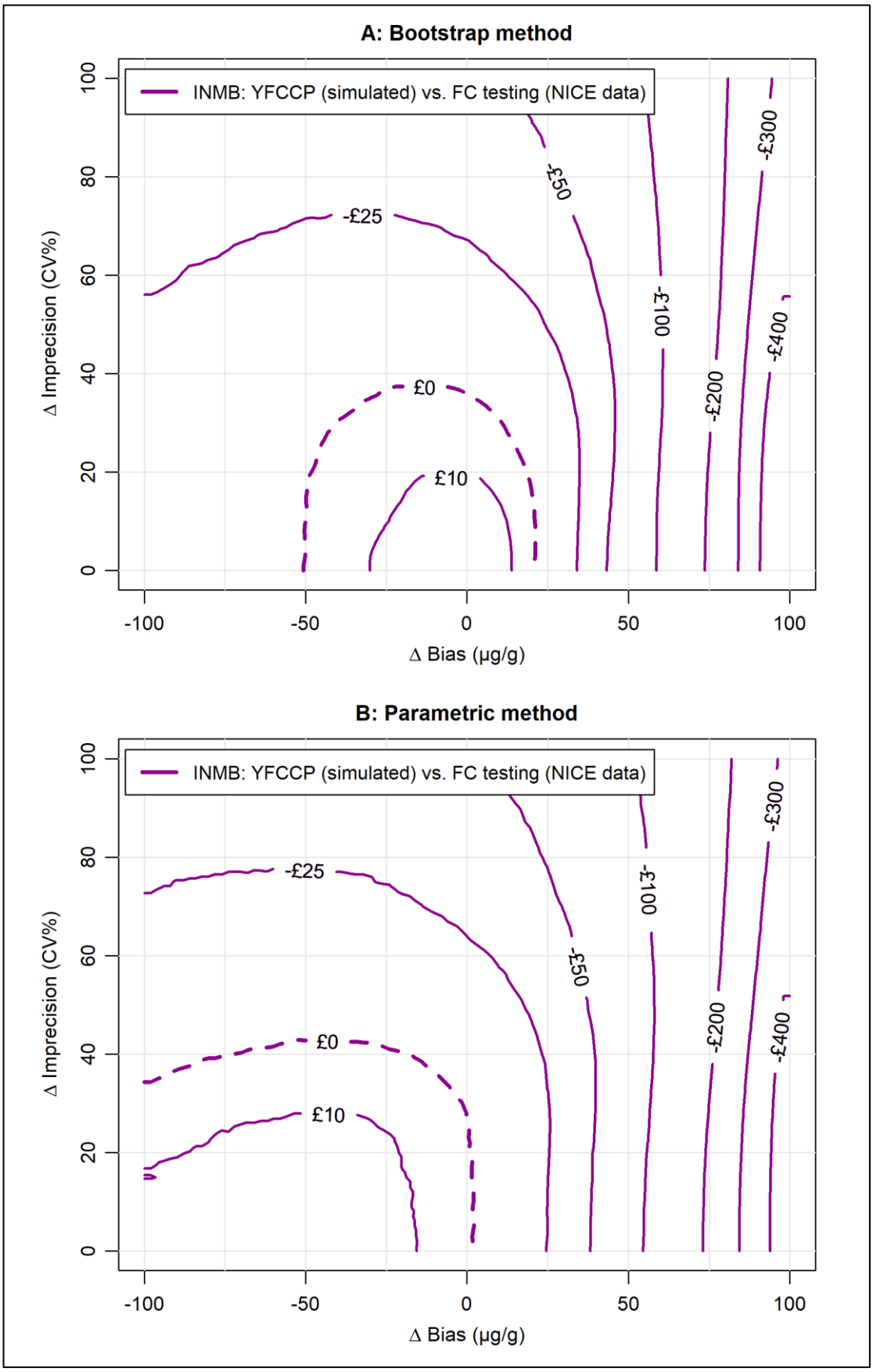


Figure 6-9. YFCCP: contour plot of INMB (£) for simulated YFCCP vs. FC testing (NICE data)

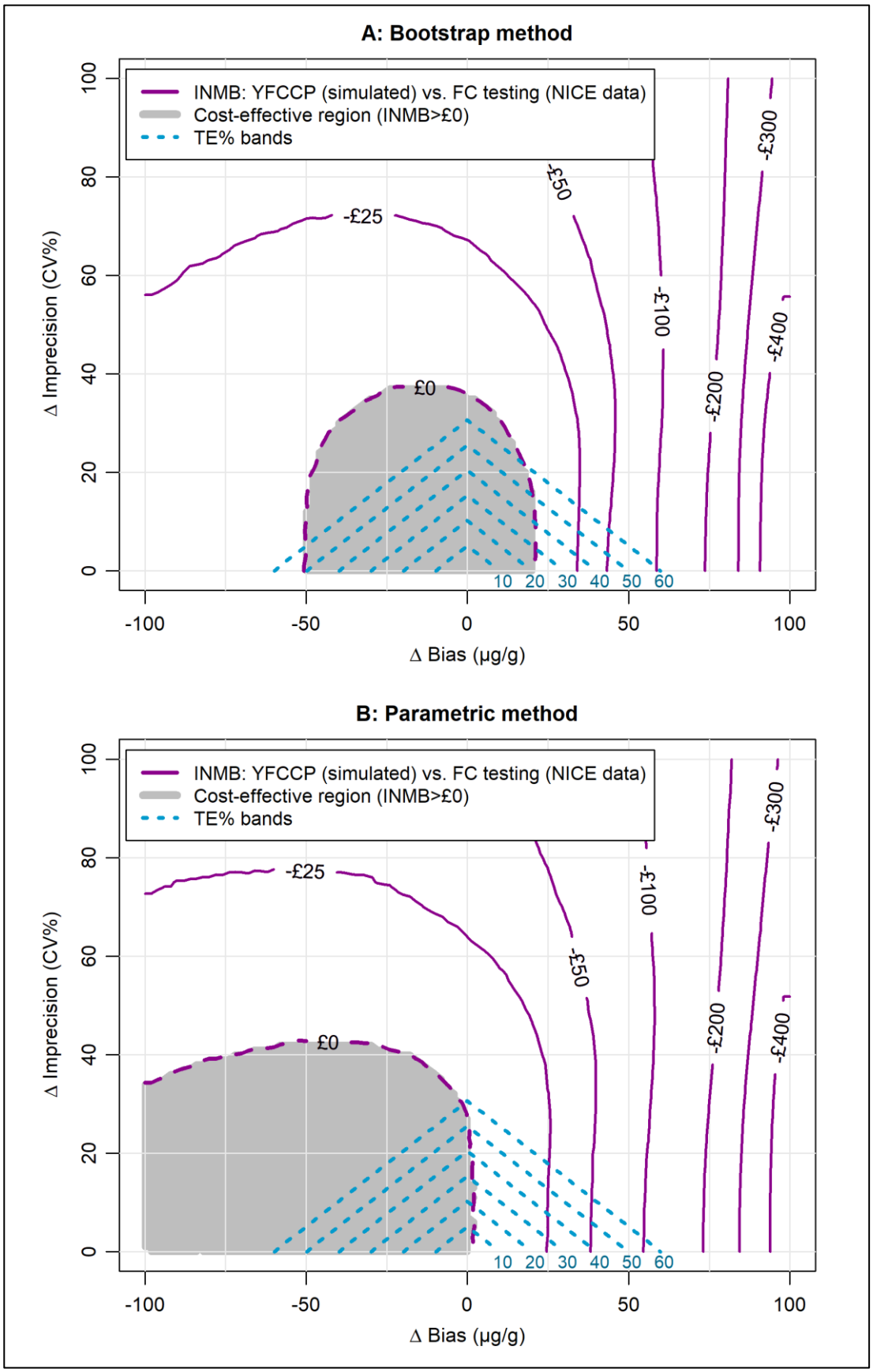


Figure 6-10. YFCCP: contour plot of INMB (£) for simulated YFCCP vs. FC testing (NICE data) showing the cost-effective region (INMB > £0)

In the same way as above, the cost-effective region associated with each of the other fixed comparators can be identified, using both the simulated NICE FC pathway and the simulated YFCCP as the intervention strategies in turn. Rather than presenting each of the individual INMB contour plots, Figure 6-11 and Figure 6-12 below illustrate the position of the zero INMB line (indicative of the cost-effective region) for the simulated NICE FC pathway and the YFCCP respectively, compared to each of the fixed comparators. In addition, Figure 6-13 provides the same plot for the YFCCP analysis, this time with TE bands overlaid. Since TE_{max} was equal to 0 (or NA) for all of the comparators within the NICE FC pathway evaluation, the equivalent plot for that pathway is not shown.

Note that, due to the fact that the simulated NICE FC pathway was dominated by several of the comparators over the simulated space of added bias and imprecision, Figure 6-11 only shows three comparators for which the NICE FC pathway was not universally dominated (i.e. where a non-empty cost-effective region was observed). Similarly for the YFCCP analysis, only the non-dominating comparators are shown: for the case of the bootstrap method, this includes all of the comparators; whilst for the parametric method, this includes all of the comparators except the original fixed YFCCP strategy (Figure 6-12 and Figure 6-13). In each case, the cost-effective regions lie to the left (i.e. south-west) of each of the zero INMB contours presented.

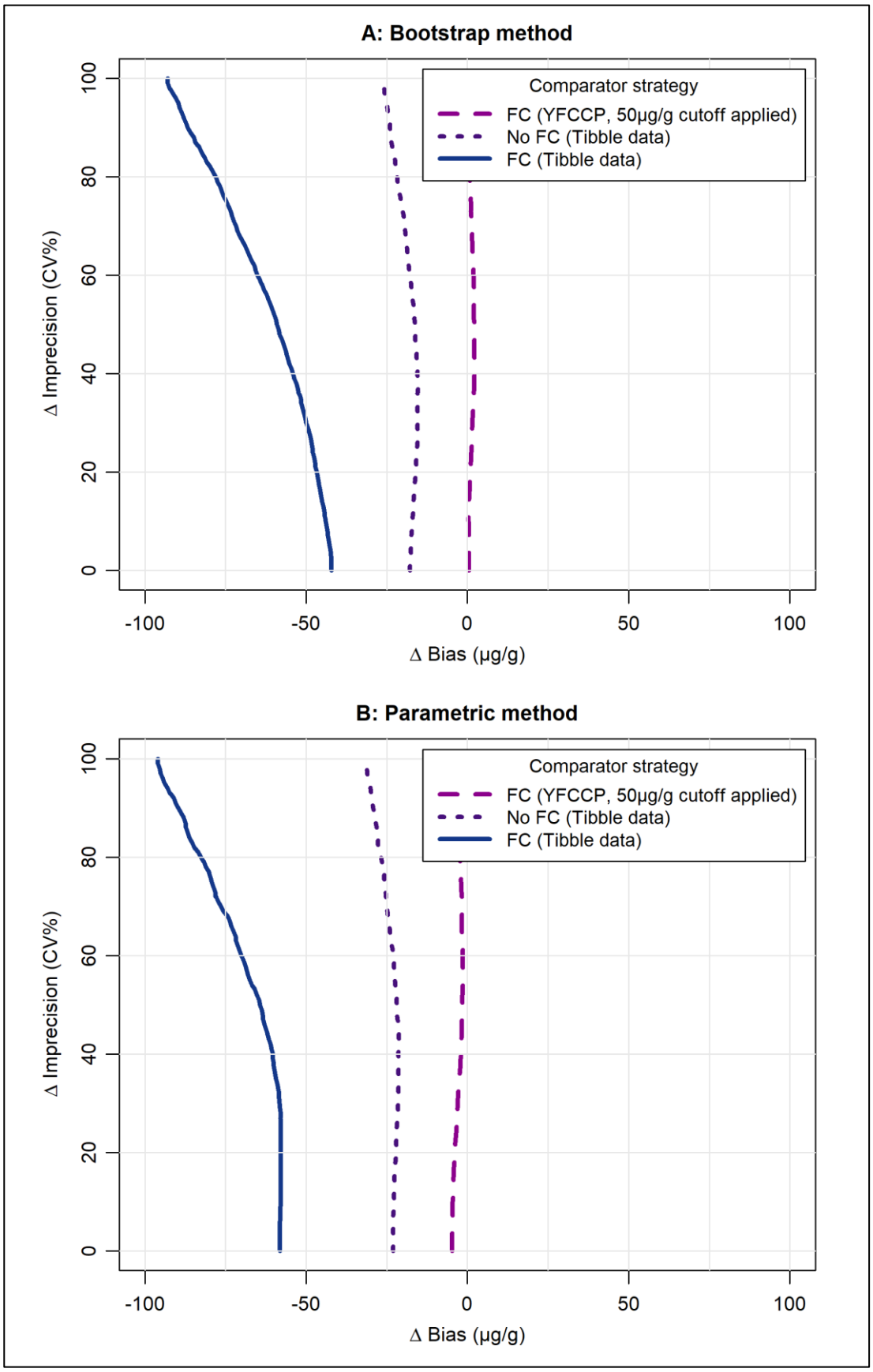


Figure 6-11. NICE FC pathway: contour plot showing the position of the zero INMB contour for the simulated NICE FC pathway vs. fixed comparator strategies

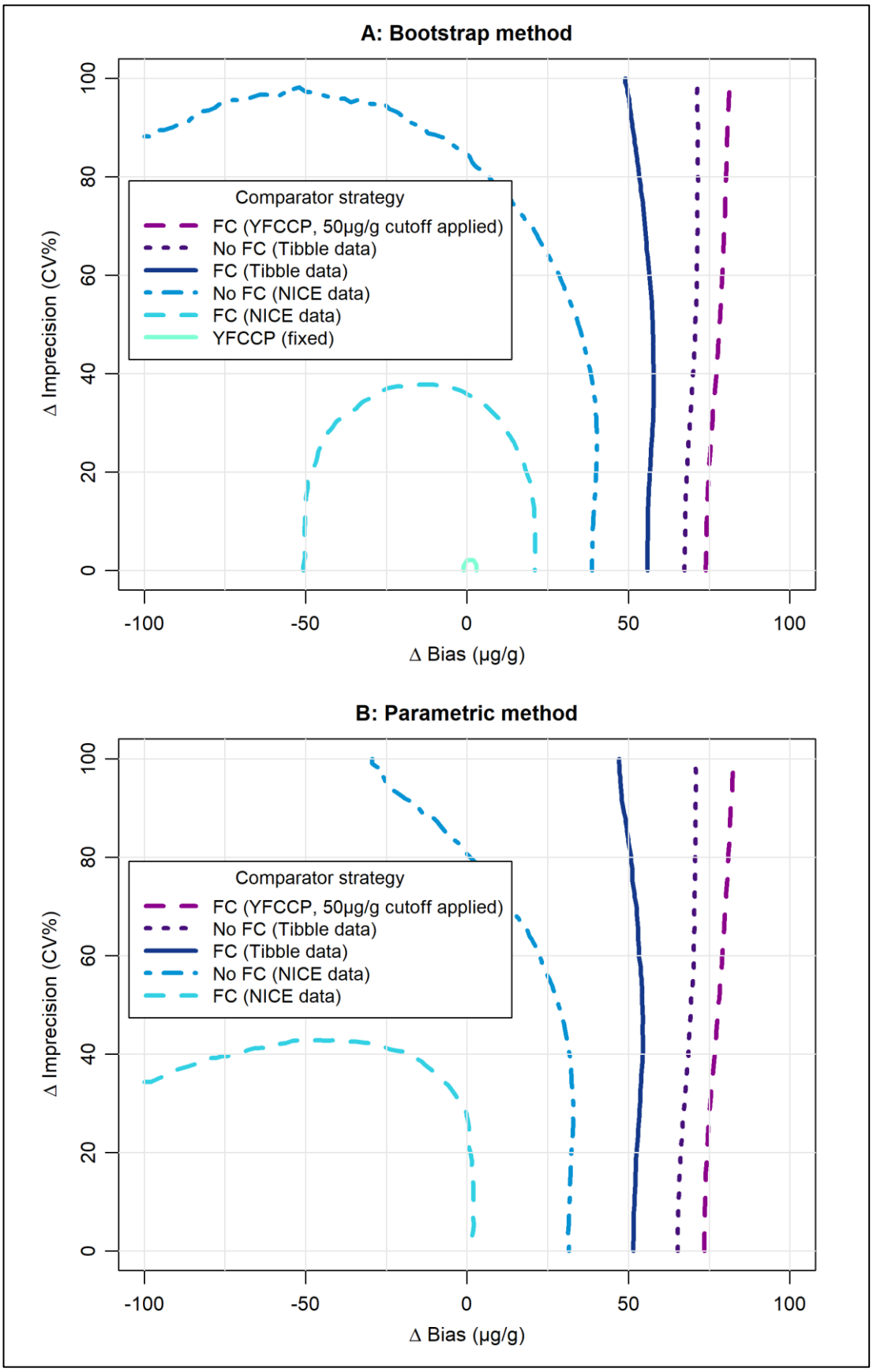


Figure 6-12. YFCCP: contour plot showing the position of the zero INMB contour for the simulated YFCCP vs. fixed comparator strategies

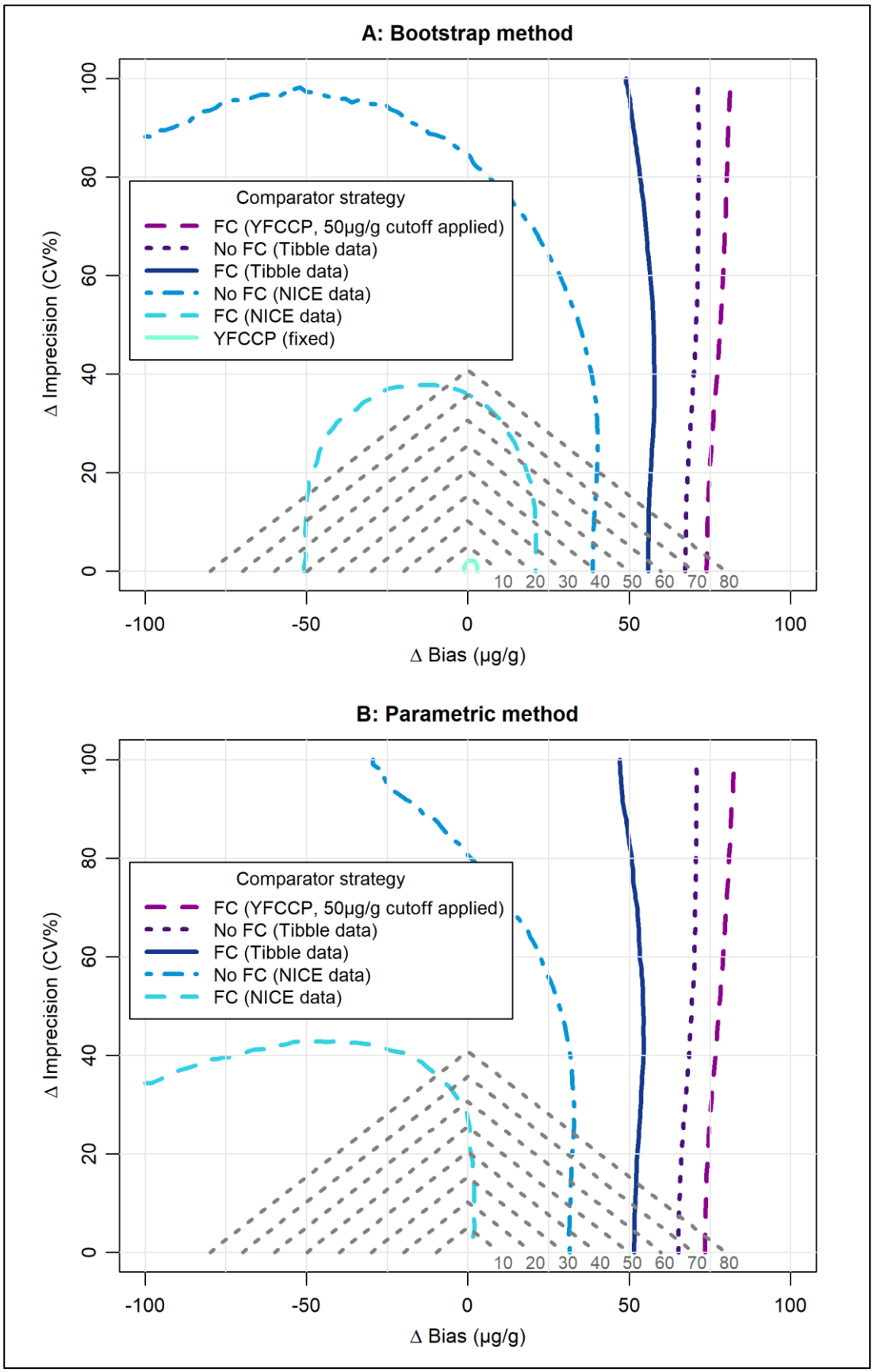


Figure 6-13. YFCCP: contour plot showing the position of the zero INMB contour for the simulated YFCCP vs. fixed comparator strategies, with TE% bands overlaid

The cost-effective regions present areas of acceptable performance in terms of ensuring a positive INMB is maintained. In some cases investigators may wish to explore a more stringent requirement: for example in Figure 6-13, it can be seen that several of the comparators result in large cost-effective regions for the YFCCP, which suggest that considerable magnitudes of imprecision and bias can be tolerated before cost-effectiveness is adversely affected. In the context of setting outcome-based APS, it is unlikely that such low benchmarks of analytical performance would be considered acceptable in the laboratory. Laboratory professionals may therefore be interested in knowing what the added benefit would be, if a more stringent requirement for analytical performance was prescribed. In addition, once the decision to adopt a particular testing strategy has been taken, but where the cost-effective region indicates a wide tolerance to imprecision and bias, it may be of further interest to assess how positive NMB could be *optimised*, rather than simply maintained. If a meaningfully higher NMB[47] could be obtained by demanding an (achievable) higher level of measurement performance, then it would be of interest to determine whether this additional benefit outweighs the expected cost of securing the associated higher region of performance (e.g. the cost associated with increased internal or external quality assurance measures).

Figure 6-14 illustrates an example of an "*optimal region*" for the YFCCP vs. the fixed comparator 'FC testing (NICE data)'. In this case, the optimal region has been arbitrarily defined as the region of bias and imprecision containing the top 10% of INMB achieved across the simulation space (i.e. INMB ≥ the INMB 90th percentile). Due to the fact that the optimal region has here been defined as a *relative* region (i.e. a top percentage of INMB, rather than a fixed INMB requirement), the *position* of this region for both sampling methods is the same regardless of which fixed comparator strategy is applied in the analysis. As a result, whilst there is little difference between the optimal region and the cost-effective region for the example shown in Figure 6-14, the difference is much greater when considering the alternative comparators.

[47] Note that whilst the focus of this analysis is on cost-effectiveness outputs (e.g. NMB), these concepts may equally be applied to other outcomes (e.g. QALYs, life-years, costs etc.).

A key element of interest with the optimal regions, is the value of the minimum INMB that can be achieved within the optimal region (which changes when selecting alterative comparators for the cost-effectiveness calculation). For example, for the comparison illustrated in Figure 6-14, the INMB 90th percentile is £1.5 for the bootstrap method, and £10 for the parametric method, and it is these values which define the minimum INMB achieved within the optimal region. The position of the optimal region boundary for each sampling method does not change when adopting alternative fixed comparators, however the value of the INMB 90th percentile does change (see Table 6-3 and Table 6-4 for the INMB 90th percentile values). The optimal region boundary value is of interest since it reflects the value (in terms of added NMB) which may be obtained by moving from the edge of the cost-effective region (where minimum performance is set such that INMB = £0) to the edge of the optimal region (where minimum performance is set such that INMB ≥ 90th percentile, for example). The possible advantages of being able to quantify this value are discussed in section 6.4.2.

Table 6-3 and Table 6-4 provide a summary of the results for the simulated NICE FC pathway and the YFCCP respectively. This includes the incremental costs, incremental QALYs and INMB for each simulated pathway at the (0,0) point (versus the fixed comparator strategies); alongside key outputs relating to each of the cost-effective and optimal regions – including the TE_{max} values, the range of CV% tolerated at bias = 0 µg/g, and the range of bias tolerated at CV = 0%.

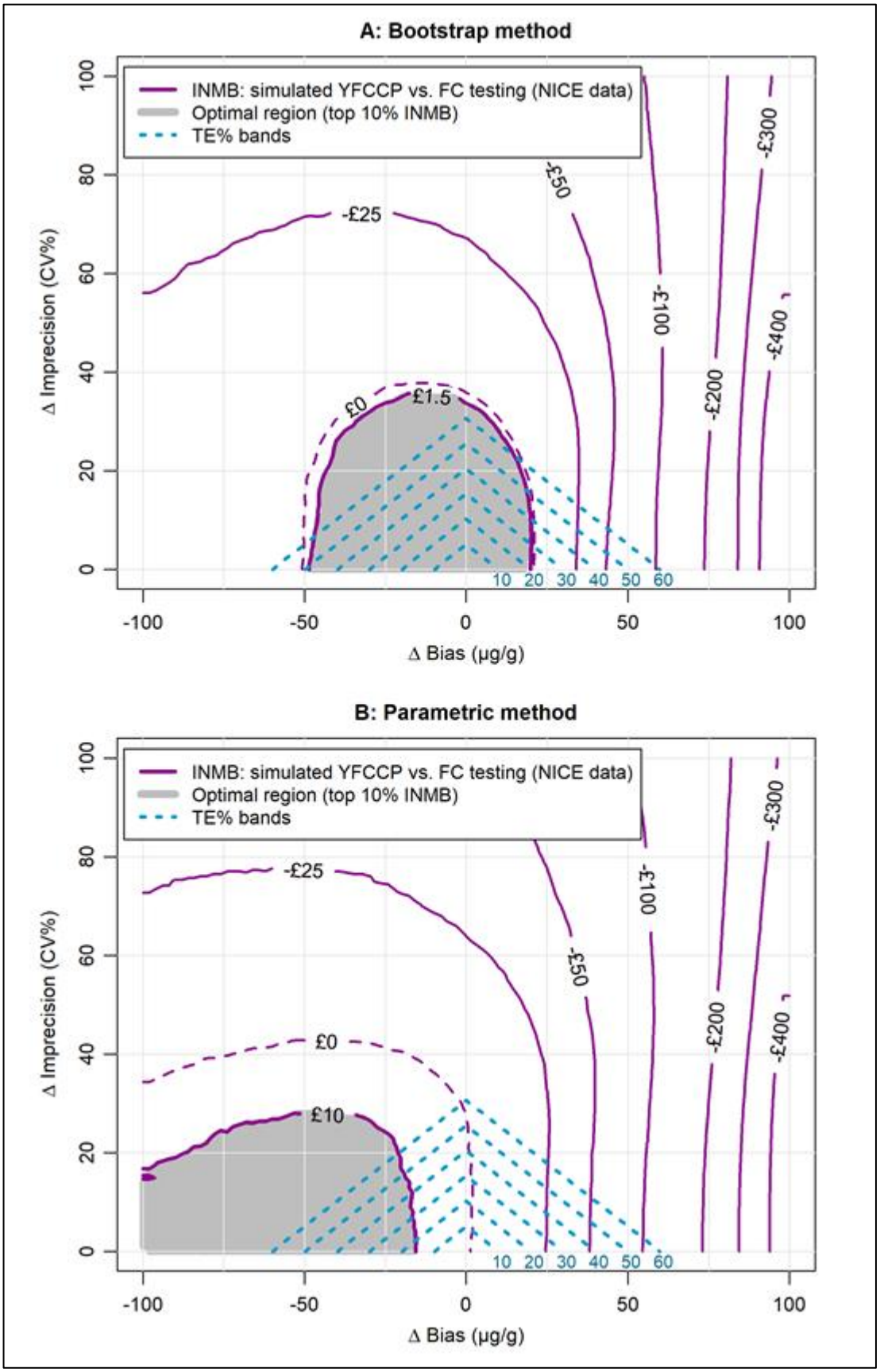


Figure 6-14. YFCCP: contour plot showing the optimal region of INMB (£) for simulated YFCCP vs. FC testing (NICE data) fixed comparator

Table 6-3. NICE FC pathway: incremental results for simulated NICE FC pathway vs. fixed comparators

Fixed comparator strategy	Outcomes at bias=0 & CV=0%			Cost-effective region: INMB > £0			Optimal region: INMB ≥ INMB 90th percentile			
	Δ Cost	Δ QALY	INMB	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	INMB 90th percentile
Bootstrap method										
YFCCP intervention (fixed)	£101	-0.0059	-£219	- NA -	- NA -	- NA -	- NA -	-100 to -58	- NA -	-£92
No FC (Tibble data)	£54	0.0001	-£52	- NA -	-100 to -18	- NA -	- NA -	-100 to -58	- NA -	£76
No FC (NICE data)	£80	-0.0042	-£165	- NA -	- NA -	- NA -	- NA -	-100 to -58	- NA -	-£37
FC testing (YFCCP, 50 µg/g cut-off)	-£1	0.0000	£2	0%	-100 to 0	0 to 88%	- NA -	-100 to -58	- NA -	£129
FC testing (Tibble data)	£67	-0.0022	-£113	- NA -	-100 to -44	- NA -	- NA -	-100 to -58	- NA -	£15
FC testing (NICE data)	£115	-0.0043	-£201	- NA -	- NA -	- NA -	- NA -	-100 to -58	- NA -	-£74
Parametric method										
YFCCP intervention (fixed)	£113	-0.0063	-£239	- NA -	- NA -	- NA -	- NA -	-100 to -68	- NA -	-£96
No FC (Tibble data)	£66	-0.0002	-£71	- NA -	-100 to -24	- NA -	- NA -	-100 to -68	- NA -	£72
No FC (NICE data)	£92	-0.0046	-£184	- NA -	- NA -	- NA -	- NA -	-100 to -68	- NA -	-£41
FC testing (YFCCP, 50 µg/g cut-off)	£11	-0.0003	-£18	- NA -	-100 to -6	- NA -	- NA -	-100 to -68	- NA -	£125
FC testing (Tibble data)	£79	-0.0027	-£133	- NA -	-100 to -60	- NA -	- NA -	-100 to -68	- NA -	£10
FC testing (NICE data)	£127	-0.0047	-£221	- NA -	- NA -	- NA -	- NA -	-100 to -68	- NA -	-£77

Table 6-4. YFCCP: incremental results for simulated YFCCP vs. fixed comparators

Fixed comparator strategy	Outcomes at bias=0 & CV=0%			Cost-effective region: INMB > £0			Optimal region: INMB ≥ INMB 90th percentile			
	Δ Cost	Δ QALY	INMB	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	INMB 90th percentile
Bootstrap method										
YFCCP intervention (fixed)	£0	0	£0	0%	0 to 2	0 to 2%	18%	-48 to 18	0 to 33.0%	-£17
No FC (Tibble data)	-£47	0.0060	£168	66%	-100 to 66	0 to 100%	18%	-48 to 18	0 to 33.0%	£151
No FC (NICE data)	-£21	0.0017	£55	38%	-100 to 38	0 to 84%	18%	-48 to 18	0 to 33.0%	£38
FC testing (YFCCP, 50 µg/g cut-off)	-£102	0.0060	£221	72%	-100 to 72	0 to 100%	18%	-48 to 18	0 to 33.0%	£202
FC testing (Tibble data)	-£34	0.0036	£107	54%	-100 to 54	0 to 100%	18%	-48 to 18	0 to 33.0%	£90
FC testing (NICE data)	£14	0.0016	£18	20%	-50 to 20	0 to 35%	18%	-48 to 18	0 to 33.0%	£1.5
Parametric method										
YFCCP intervention (fixed)	£9	-0.0004	-£17	- NA -	- NA -	- NA -	- NA -	-100 to -16	- NA -	-£8
No FC (Tibble data)	-£38	0.0057	£151	64%	-100 to 64	0 to 100%	- NA -	-100 to -16	- NA -	£160
No FC (NICE data)	-£11	0.0013	£37	30%	-100 to 30	0 to 80%	- NA -	-100 to -16	- NA -	£46
FC testing (YFCCP, 50 µg/g cut-off)	-£93	0.0056	£204	72%	-100 to 72	0 to 100%	- NA -	-100 to -16	- NA -	£213
FC testing (Tibble data)	-£24	0.0032	£89	50%	-100 to 50	0 to 100%	- NA -	-100 to -16	- NA -	£98
FC testing (NICE data)	£24	0.0012	£1	0%	-100 to 0	0 to 27%	- NA -	-100 to -16	- NA -	£10

6.3.2.2 Simulated YFCCP vs. simulated NICE FC pathway

The INMB results for the analysis comparing the simulated YFCCP strategy with the simulated NICE FC pathway (i.e. both accounting for the impact of increasing measurement uncertainty) are provided in Figure 6-15. The YFCCP is associated with lower costs and higher QALYs compared to the NICE FC pathway over (almost) the entire simulated space of bias and imprecision values. As a result, the YFCCP dominates the NICE FC pathway, and is associated with positive INMB values over the whole contour plot (Figure 6-15).

Since the YFCCP in this case is cost-effective compared to the NICE FC pathway over the simulated space, the cost-effective region plot is not provided (i.e. the YFCCP remains cost-effective over the entire contour plot area). Figure 6-16 shows the optimal region of the YFCCP for this analysis, equal to maintaining the YFCCP INMB ≥ the INMB 90th percentile (in this case, equal to £432 for the bootstrap method, and £417 for the parametric method). Tabular results are provided in Table 6-5.

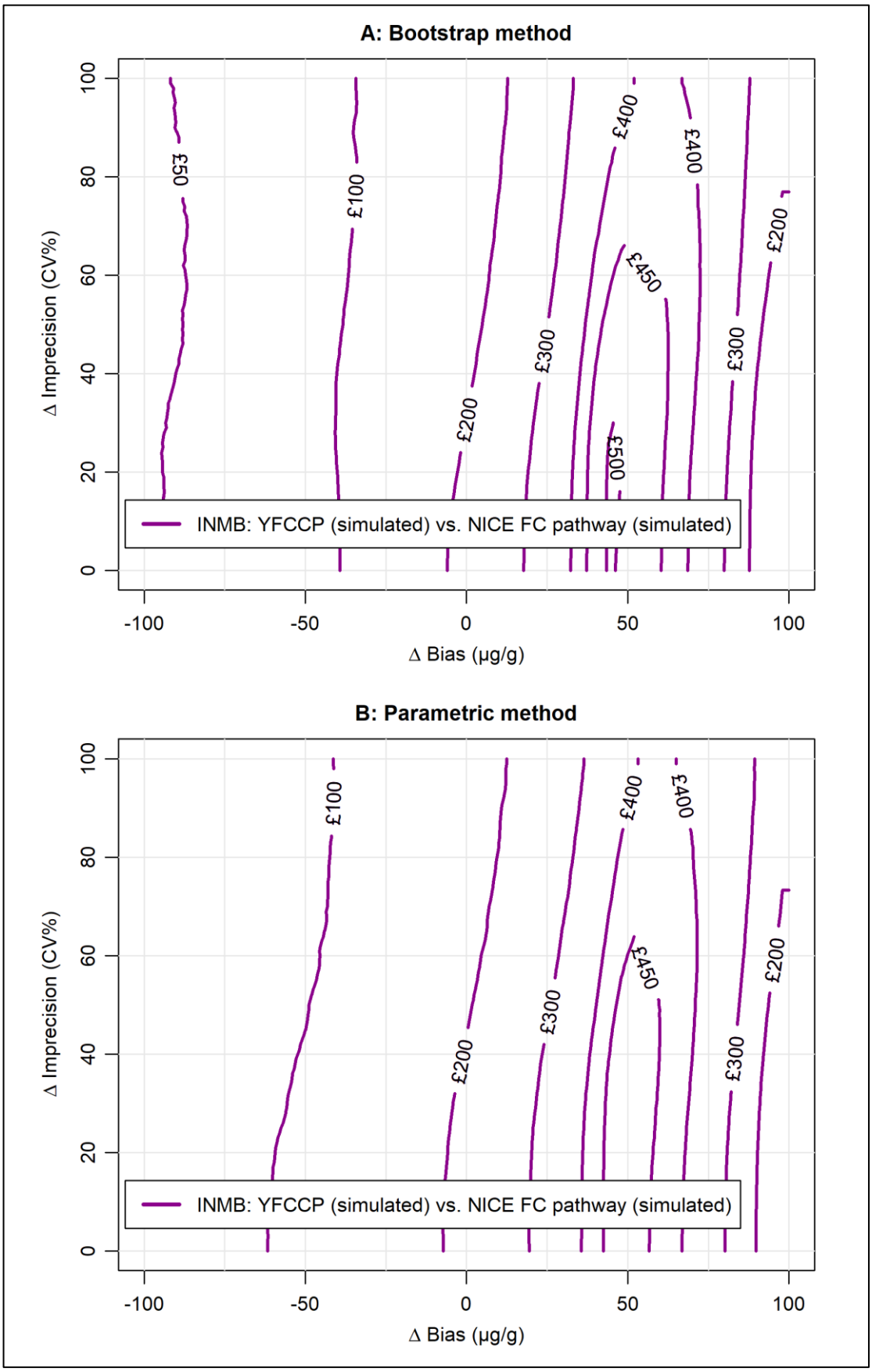


Figure 6-15. YFCCP: contour plot of INMB (£) for simulated YFCCP vs. simulated NICE FC pathway

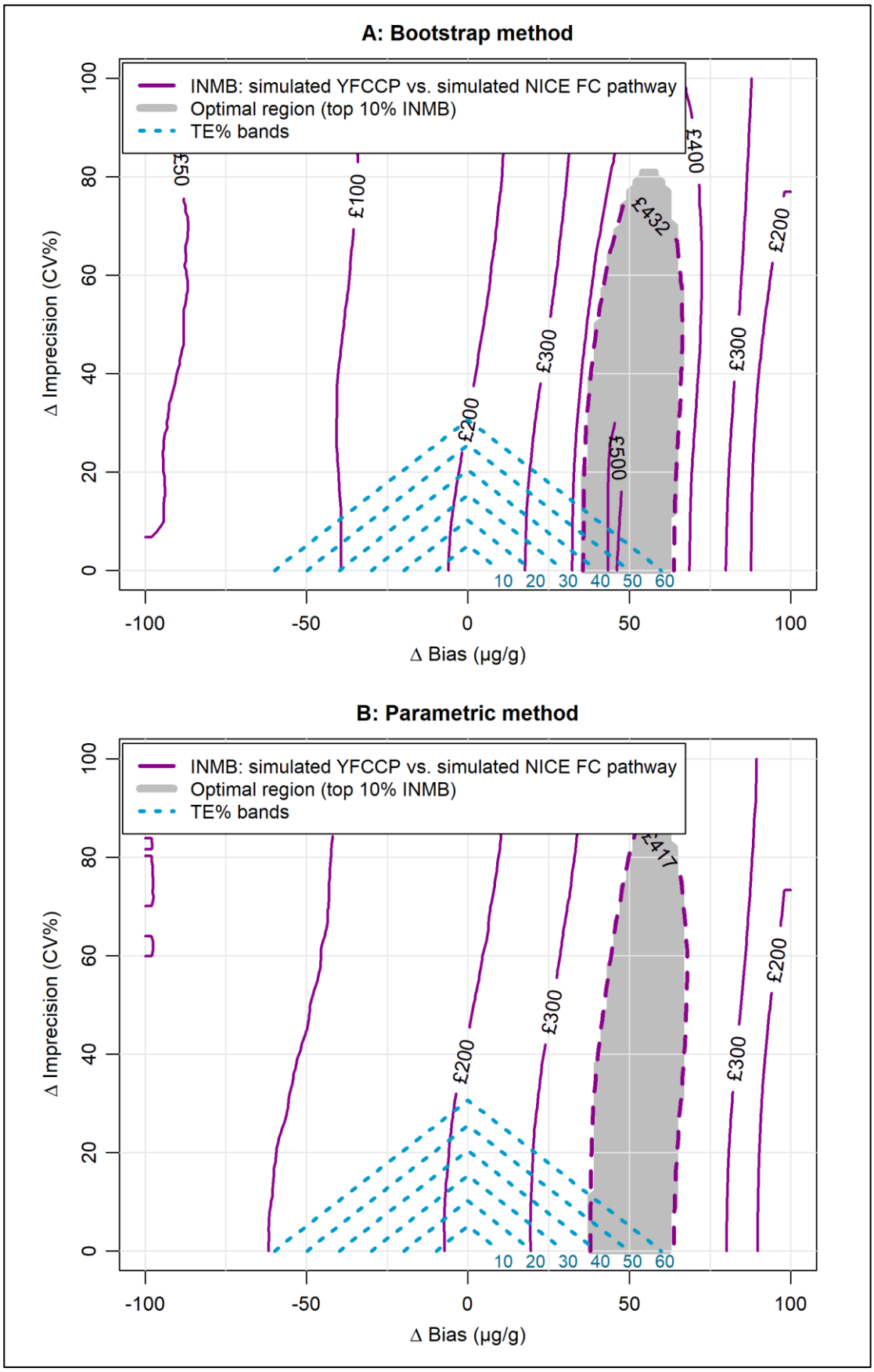


Figure 6-16. YFCCP: contour plot showing the INMB optimal region for the simulated YFCCP vs. the simulated NICE FC pathway

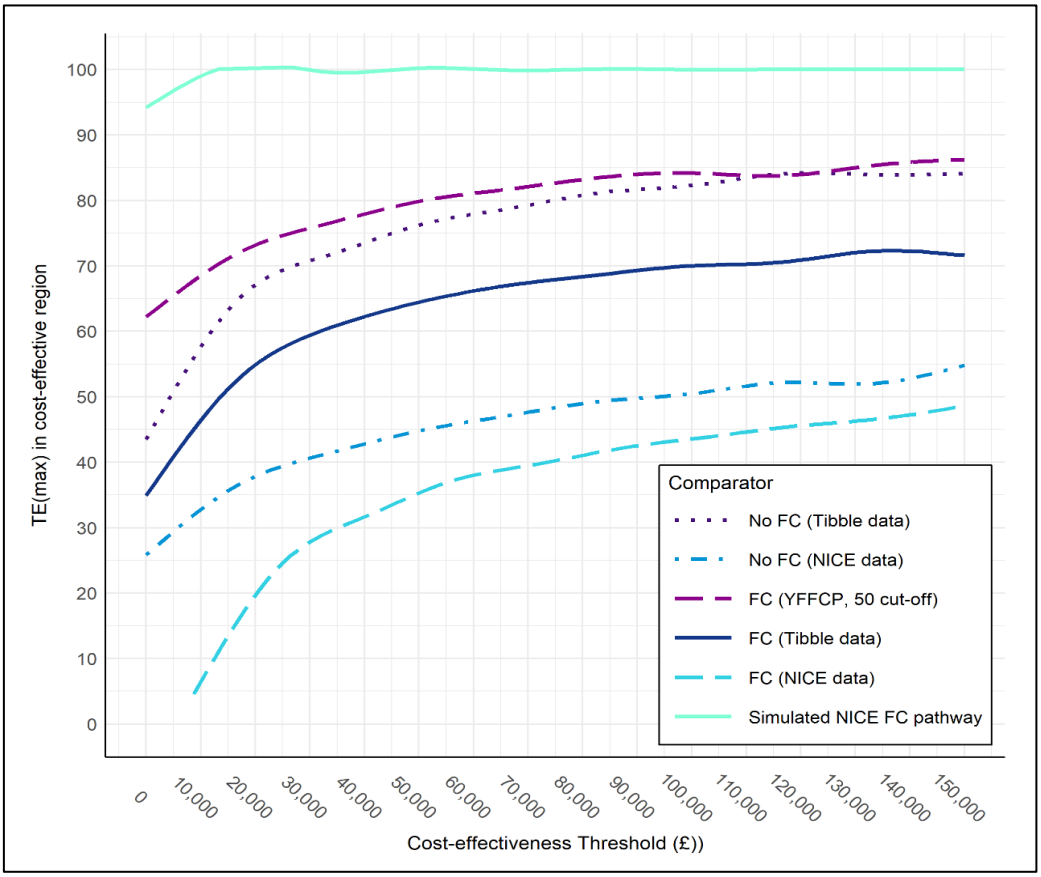
Table 6-5. YFCCP: incremental results for simulated YFCCP strategy vs. simulated NICE FC pathway

Outcomes at bias=0 & CV=0%			Cost-effective region: INMB > £0			Optimal region: INMB ≥ 90th percentile			
Δ Cost	Δ QALY	INMB	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	TE_{max}	Range of bias at CV=0%	Range of CV% at bias=0	INMB 90th percentile
Bootstrap method									
-£101	0.0059	£220	100%	-100 to 100	0 to 100%	- NA -	36 to 62	- NA -	£432
Parametric method									
-£104	0.0059	£222	100%	-100 to 100	0 to 100%	- NA -	38 to 62	- NA -	£417

6.3.2.3 TE_{max} vs. cost-effectiveness threshold

In each of the analyses thus far presented, cost-effectiveness has been calculated assuming a cost-effectiveness threshold of £20,000 per additional QALY. For the simulated YFCCP strategy, Figure 6-17 illustrates how the cost-effective region alters as the cost-effectiveness threshold is varied for the bootstrap method results, using TE_{max} as a summary metric for the size of the cost-effective region.[48] In this case, increasing the cost-effectiveness threshold increases the size of the cost-effective region, more noticeably at the lower threshold values. This result stems from the fact that the YFCCP intervention strategy is mostly cost-saving and QALY-increasing, compared to the comparator strategies considered. As such, increasing the value attached to QALY gains, results in an increase in the associated cost-effective region for the YFCCP.

Figure 6-17. YFCCP: cost-effective region TE_{max}, over a range of cost-effectiveness threshold values (bootstrap method)



[48] Note that this figure excludes the fixed YFCCP comparator, since this strategy dominated the simulated YFCCP over the vast majority of the cost-effectiveness threshold values.

6.4 Discussion

6.4.1 Absolute results: mean cost, QALYs and NMB

The primary objective of this analysis was to extend the methodology presented in Chapter 5, to assess the impact of increasing FC imprecision and bias on end-stage outcomes for the NICE FC pathway and the YFCCP. To this end, the previously reported error model simulation results were applied within an existing FC cost-utility model, to determine the resulting impact on cost, utility and cost-utility outcomes.

As discussed in Chapter 5, the bootstrap method provided a better fit to the data within the error model simulation compared to the parametric method (section 5.5.2). Since the cost-utility results presented herein are directly dependant on the error model simulation results, the bootstrap method similarly provided a close match to the data in this analysis (with cost, QALY and NMB results at the (0,0) point matching those produced from the original FC cost-utility model) (section 6.3.1). In contrast, the parametric method produced diverging results, with the diagnostic accuracy discrepancies outlined in Chapter 5 leading to higher cost and lower QALY estimates for both pathways. This resulted in an underestimation of NMB, leading to the subsequent cost-effective and optimal regions (reported in section 6.3.2) being offset from the (0,0) point for each comparative assessment when using the parametric method. In light of these discrepancies this discussion focuses on the bootstrap method results.

Both pathways exhibited the same overall pattern of results with respect to the impact of increasing measurement uncertainty on costs: additional positive bias increased costs whilst negative bias reduced costs; and imprecision had minimal impact (Figure 6-3 and Figure 6-6). These results reflect the specificity findings from Chapter 5, with increased costs occurring in line with decreased specificity (see section 5.3.2.2, Figure 5-4; and section 5.4.2.2, Figure 5-10). Sensitivity, in contrast, had little impact on costs. The greater influence of specificity in this case was due to the fact that: (a) falsely diagnosed IBD patients in the FC cost-utility model were assumed to *eventually* obtain a secondary care referral (due to persisting symptoms), and incurred minimal costs as a result of delayed diagnosis; whilst (b) falsely referred IBS patients incurred a significant cost in the

model (i.e. colonoscopy), and IBS patients formed a larger percentage of the model population (92%). In alternative clinical contexts, however, sensitivity would be expected to have a greater influence – for example in cases where false negative diagnoses are associated with greater financial implications compared to false positive cases.

With regards to QALYs, markedly different results were observed across the two pathways. For the NICE FC pathway, QALYs were again driven by specificity: additional positive bias decreased QALYs, due to more IBS patients suffering from a delayed diagnosis and spending longer in the 'Untreated IBS' health state (i.e. reduced specificity); whilst negative bias had the opposite effect (Figure 6-4). Recall that, due to the lower FC cut-off threshold applied in this pathway (50 µg/g), a moderate sensitivity was maintained over a wide range of added bias and imprecision, at the expense of a relatively low baseline specificity which was highly volatile to added bias (section 5.3.2.2, Figure 5-4). As such, although sensitivity was a driver of QALYs in the model (since delayed diagnosis for IBD patients is also associated with a utility decrement), specificity had the more dominant impact in this case, with the highest QALYs achieved in the region of maximised specificity (i.e. at -100 µg/g bias).

For the YFCCP, both sensitivity and specificity were key drivers of QALYs. Reductions in QALYs occurred in line with both positive bias (due to reduced specificity), and negative bias and/or imprecision (due to reduced sensitivity), resulting in QALY contours which decreased in distinctive, outward concentric semi-circles from the (0,0) point of highest QALY accrual (Figure 6-7). Recall that, in contrast to the NICE FC pathway, the YFCCP achieved a high baseline diagnostic sensitivity *and* specificity, and exhibited a more even and tempered performance in terms of robustness of these two outcomes to increased measurement uncertainty (section 5.4.2.2.1, Figure 5-10). As a result, QALYs for this pathway are maximised around the (0,0) point (for the bootstrap method), and gradually diminish in response to increasing *or* decreasing bias, and increasing imprecision.

The combined impact of measurement uncertainty on each pathways' costs and QALYs was evaluated using the NMB metric, assuming a cost-effectiveness threshold of £20,000 per additional QALY. For the NICE FC pathway, since

specificity was the key determinant of both costs and QALYs, the NMB was similarly driven by this factor (Figure 6-5). In line with the findings from Chapter 5, the NMB contour plot in this case provides further evidence of the sub-optimal performance of this pathway, with the highest NMB being achieved at -100 µg/g added bias (equivalent to increasing the cut-off threshold to 150 µg/g). For the YFCCP meanwhile, although costs were driven by specificity, QALYs were affected by both sensitivity and specificity; as a result, the NMB for this pathway followed a similar pattern as the corresponding QALY contour plot (Figure 6-8). In addition, the NMB for the YFCCP was optimal close to the (0,0) point (with NMB maximised at CV=0% and bias=+2µg/g for the bootstrap method), indicating that the cut-off threshold used in this pathway is appropriate. Based purely on the individual NMB plots therefore, it would appear that the YFCCP outperforms the NICE FC pathway, even over regions of increased measurement uncertainty – the incremental analysis discussed below further confirms this.

It should be noted that an alternative, more simplistic approach to exploring the robustness of modelled outcomes, would be to run standard one-way sensitivity analysis on the baseline diagnostic sensitivity and specificity values for each pathway in the base case YFCCP cost-utility model. This traditional approach provides information on the sensitivity of the modelled outcomes to changes in the diagnostic sensitivity and specificity inputs. A key limitation with that approach, however, is that it fails to provide any information on how changes in diagnostic accuracy inputs relate to changes in the underlying measurement performance of the test. Instead, by linking changes in the modelled outputs back to changes in the measurement bias and imprecision of the test, the error model simulation approach presented in this thesis provides additional information on what fundamental aspects of the test's performance are driving the observed changes in clinical and health-economic outcomes. This provides useful information for clinical and reimbursement decision makers, and can further help to inform optimal laboratory practices via the derivation of outcome-based APS.

6.4.2 Incremental (INMB) results

6.4.2.1 Cost-effective regions

The cost-effectiveness of the two simulated FC pathways was evaluated using the INMB metric (section 6.3.2). Each pathway was first assessed in relation to the six fixed intervention and comparator arms included in the original FC cost-utility model (section 6.3.2.1). The results of this analysis enable assessment of the robustness of each pathway's cost-effectiveness to added bias and imprecision, and further identification of the region of added bias and imprecision which achieves cost-effectiveness.

For the NICE FC pathway, all of the comparators dominated this pathway at the (0,0) point (i.e. they were cheaper and more effective) (Table 6-2). Figure 6-11 further demonstrates that, for the higher performing comparators (i.e. the two 'NICE data' comparators, and the YFCCP), the NICE FC pathway failed to achieve cost-effectiveness at any value of added bias and imprecision. For the lower performing comparators (i.e. the two 'Tibble data' comparators and the fixed NICE FC pathway), the NICE FC pathway did achieve cost-effectiveness, in the region of negative bias. The offset placement of the cost-effective regions in this case is a clear sign of the sub-optimal design of this pathway, since for an optimal strategy one would expect cost-effectiveness to be maximised around the area of lowest measurement uncertainty. These results confirm that, as well as not being cost-effective at the baseline, the NICE FC pathway is only cost-effective against lower performing comparator strategies when applying high negative bias (equivalent to raising the cut-off threshold).

For the YFCCP, at the (0,0) point this pathway was cost-effective compared to all of the fixed comparators, and dominated all but the FC testing (NICE data) strategy (Table 6-2). As with the NICE FC pathway, the largest cost-effective regions for the YFCCP were obtained via assessment against those comparators with the lowest NMB. However, since the YFCCP's performance was maximised around the baseline (0,0) point and lowest in the region of high positive bias, the zero INMB contour for this pathway when compared to low performing comparators fell within the region of high positive bias. In these cases therefore, the cost-effective region covered the entire domain of imprecision (up to 100%

CV), for a wide range of bias (-100 to >50 µg/g).[49] Meanwhile for higher performing comparators (i.e. associated with higher NMB), the associated zero INMB line for the YFCCP fell closer to the region of optimal performance, resulting in a contraction of the cost-effective region. Nevertheless, even when compared to its highest performing comparator [the 'FC testing (NICE data)' arm][50] the YFCCP maintained a sizable cost-effective region (TE_{max}=20%) (Table 6-4). As such, these results confirm that the cost-effectiveness of the YFCCP is expected to be robust to variations in analytical performance, and further support the case for the adoption of this pathway.

An additional incremental analysis was conducted comparing the two simulated pathways against each other (i.e. assuming that the same magnitude of measurement uncertainty occurs within *both* pathways) (section 6.3.2.2). In the INMB plots thus far presented for the YFCCP, the pathway's optimal INMB was achieved around the (0,0) point, in line with location of the highest NMB achieved with this pathway. In this incremental analysis however, the highest INMB for the YFCCP was achieved just below 50 µg/g bias (0% CV), where the *relative* performance of the YFCCP vs. NICE FC pathway was highest (Figure 6-15) (see section 6.4.2.2 for further discussion of the optimal region of this plot). Due to the consistently higher performance of the YFCCP over the complete space of simulated bias and imprecision, the cost-effective region in this case covered the entire contour plot. This confirms that the YFCCP is consistently more robust to increased measurement uncertainty compared to the NICE FC pathway.

A key advantage of basing APS on cost-effective regions (as opposed to acceptable regions as presented in Chapter 5), is that no user-based, subjective

---

[49] See for example the placement of the zero INMB line for the three worst performing comparators in Figure 6-12: FC testing (YFCCP, 50 µg/g cut-off), No FC (Tibble data), FC testing (Tibble data). In these cases, TE_{max} ranged from 54 to 72%.

[50] Note that the fixed YFCCP comparator in this example represents a special case. For an optimised testing strategy, it is expected that when comparing that strategy (accounting for the impact of added measurement uncertainty), against the same *fixed* strategy (not accounting for added measurement uncertainty), that *any* additional bias or imprecision would reduce the performance of that pathway. Hence why the fixed YFCCP comparator is associated with a near-empty cost-effective region in this analysis. As such, this comparator only really serves as a validity check, to confirm that the pathway is indeed optimised.

judgement is required to identify this region, since cost-effectiveness (within the UK context) can be directly determined based on the NICE cost-effectiveness threshold (280). Thus, once the costs and QALYs associated with an intervention and comparator are known, this threshold can be applied to determine the INMB, with cost-effectiveness defined as INMB > £0. In contrast in Chapter 5, an assumption had to be made regarding the minimum level of diagnostic accuracy which was considered acceptable, in order to define the acceptable region of bias and imprecision. Of course, the validity of the NICE threshold used to define the cost-effective region may be questioned, and alternative values may be explored within sensitivity analyses (as in section 6.3.2.3). Nevertheless, in the UK HTA context, the NICE threshold is typically considered as an externally pre-determined variable. Thus, whilst modelling cost-effectiveness outcomes requires additional work compared to modelling diagnostic accuracy outcomes, the avoidance of subjectivity within setting APS may be considered a useful advantage.

The extraction of clearly defined APS from cost-effectiveness analysis ideally requires a single, clearly defined comparator. In the context of primary care IBD diagnosis, however, selection of an appropriate comparator is challenging due to two issues. First, standard care in the UK is variable, due to an inconsistent uptake of FC across primary care following the NICE 2013 recommendation (i.e. there is no single “standard care” pathway) (239). Second, a paucity of primary care studies in this area means that the validity of secondary-care-based diagnostic accuracy estimates used within the FC cost-utility model can, and have, been questioned (239).

Focusing on the YFCCP (as the only robust cost-effective pathway where APS may therefore be required), it may be assumed that this pathway is most likely to be considered in areas where the single-test NICE FC pathway is already in place. Assuming for arguments sake that this is the case, then the ‘FC testing (YFCCP, 50 µg/g cut-off)’ [i.e. the fixed NICE FC pathway] comparator is likely to be of most relevance, since this is the only FC comparator based on primary care data. Basing APS for FC on this analysis, however, suggests that a substantial magnitude of additional bias and imprecision can be tolerated before breaching cost-effectiveness (TE_{max} = 72%). For the purposes of deciding whether or not to

adopt the YFCCP, such a large cost-effective region is encouraging. For the purposing of setting APS, however, this region sets a very low bar for analytical performance, which would most likely be considered unacceptable as a performance benchmark within the laboratory. A possible alternative route for setting APS in the form of “optimal regions” was therefore presented.

6.4.2.2 Optimal regions

The concept of optimal regions was presented in this analysis as an alternative and/or supplementary means of setting outcome-based APS alongside cost-effective regions. Rather than informing test adoption decisions, the idea behind the optimal region is to assess the appropriate level of analytical performance that should be demanded for a given test in practice. This approach may be useful in two key scenarios. First, a tighter restriction on measurement performance may be desired from the laboratory professionals perspective, in cases where the cost-effective region allows an unacceptably large magnitude of measurement uncertainty (e.g. a level of measurement uncertainty much higher than that typically achieved in the laboratory). Second, if a substantially higher NMB can be achieved within a subsection of the cost-effective region, then this is useful to know. For example, it may be that a meaningfully higher NMB could be obtained by restricting the acceptable bounds of bias and imprecision to a region inside that indicated by the cost-effective region. Assuming that the requirement for heightened measurement performance would be associated with some form of cost (e.g. related to increased internal and/or external quality assurance activities), then the key question in this case concerns whether or not the associated benefit outweighs this cost of raising the analytical performance goals. The use of NMB as the evaluated outcome is useful in this context, since the expected costs associated with achieving a specified optimal region can be directly compared to the expected gain in NMB in monetary terms (i.e. the NMB gain associated with moving from the edge of the cost-effective region, where NMB is zero, to the edge of the specified optimal region). If the cost associated with achieving a higher benchmark of measurement performance outweighs the expected gain in NMB, then the use of the optimal region would not be justified. Conversely, if the expected benefit outweighed the cost, then the use of the specified optimal region would be justified.

In this case study, the optimal region was arbitrarily defined as the region of added bias and imprecision maintaining INMB at or above the INMB 90th percentile (i.e. the top 10% of INMB across the simulated space). Focusing on the YFCCP (as the only robust cost-effective pathway where APS may therefore be required), this optimal region spanned a bias range of -48 to 18 µg/g (at 0% CV), allowed CV up to 33% (at 0 µg/g bias), with an associated TE_{max} of 18% (Table 6-4). Interestingly, when evaluating the YFCCP against its closest performing comparator – the FC testing (NICE data) comparator – there was little to be gained by restricting performance to the optimal region as opposed to the cost-effective region (TE_{max} =20%). In this case, the INMB 90th percentile was equal to £1.5, meaning that an additional benefit of £1.50 is achieved by moving from the zero INMB line (the edge of the cost-effective region), to the 90th percentile line (the edge of the optimal region). In contrast, there is a lot to be gained from imposing the optimal region in cases where cost-effectiveness has been established against a relatively low performing comparator. For example, when comparing the YFCCP to the FC testing (YFCCP, 50 µg/g cut-off) comparator, an average benefit of at least £202 could be obtained by maintaining analytical performance within the optimal region (TE_{max}=18%), as opposed to the cost-effective region (TE_{max}=72%). The comparison of different optimal region boundary values, therefore, provides a mechanism by which to quantify the added value of imposing restricted APS.

For the analyses based on fixed comparator strategies, the presented INMB optimal regions are equivalent (in terms of allowable bias and imprecision) to extracting the top 10% of the absolute NMB for the associated intervention (since in relative terms, the NMB and INMB plots are equivalent). The optimal region in this case can therefore be defined using the NMB plot alone, without incremental assessment against comparator pathways. Of course, assessment against relevant comparators is a key step to ensuring that an intervention does in fact achieve cost-effectiveness. If, however, the cost-effectiveness of a testing pathway has already been confirmed – for example where a test is already in use following an evidence-based adoption decision – then NMB data alone may be used to derive APS.

For the analysis comparing the two simulated pathways against each other, interpretation of the optimal region is more complex. When using the fixed comparator arms, the INMB plots for each pathway exhibited the same shape as their associated absolute NMB plots. The optimal regions in this case therefore reflected the area of measurement performance within which the simulated pathway performed best. This is not so for the analysis comparing the two simulated pathways against each other. The YFCCP optimal region in this case shifts towards +50 µg/g bias, wherein the performance of *both* pathways was reduced, but significantly more so in the NICE FC pathway (due to the FC cut-off threshold in this case dropping to zero, resulting in 0% specificity). It would be inappropriate in this case to conclude that the optimal performance of the YFCCP is at the region of 50 µg/g bias. Rather, this result highlights that the YFCCP is more robust to measurement uncertainty than the NICE FC pathway, and most notably so in the region of positive bias.

Historically, APS have often been set at multiple levels. Performance goals based on biological variation, for example, have typically been reported at three levels – 'minimum, 'desirable', and 'optimal' – based on allowing different proportions of test values to fall beyond test reference limits (35). In a similar way, optimal regions may be presented alongside cost-effective regions, to provide an upper and lower benchmark for APS. Indeed multiple optimal regions may be explored. The optimal region produced according to specifying INMB ≥ 90th percentile above, for example, may instead be considered as a 'desirable' achievement, and an 'optimal' goal may be defined as INMB ≥ INMB 95th percentile. For the YFCCP (bootstrap method), defining the optimal region in this way produces a slightly restricted region spanning a bias range of -34 to 14 µg/g at 0% CV (-48 to 18 in the base case), allowing CV up to 24% at 0 µg/g bias (33% in the base case), and with an associated TE_{max} of 14% (18% in the base case).

As with the acceptable regions presented in Chapter 5, the concept of optimal regions again requires a judgement to be made as to what level of outcome should be considered, in this case, *optimal*. The advantage here, however, is that the full range of INMB percentile values can be compared to provide a quantitative assessment of additional value gained by moving from the cost-effective region (INMB ≥ £0) to the optimal region (INMB ≥ selected INMB

percentile). Various optimal regions may therefore be considered, and the associated added INMB's judged against the expected cost of obtaining and maintaining the higher performance levels. Whilst this would require assessment of the relevant costs, it would allow optimal APS to be set based on a full consideration of both the potential benefits of optimising measurement, and the cost of achieving this optimisation. This concept is akin to that of value of information analysis in health economics, for example – wherein the cost of conducting further research is compared to the potential benefits of reducing decision uncertainty as a result of increasing the precision around key parameters of interest within the health-economic analysis (281). A similar approach could be explored for the design of laboratory quality assurance programs in future studies.

6.4.3 Limitations

The analysis presented in this chapter used the error model simulation results from Chapter 5. As a result, the same limitations apply to this analysis as previously highlighted in Chapter 5 (section 5.5). In particular, the results are similarly affected by the existence of baseline measurement uncertainty within the sampled FC data; and the YFCCP assessment is affected by the need to resample missing FC2 data within the error model simulation, due to per protocol truncated testing applied within the YFCCP dataset (section 5.5.3).

There are further limitations in this case which relate to the FC cost-utility model used to inform the assessment of cost-effectiveness. As previously mentioned, there are limitations with the evidence used to inform several of the comparator strategies in this model, which came from secondary care settings due to a paucity of primary care evidence in the literature. This means that the validity of secondary-care-based diagnostic accuracy estimates used within the FC cost-utility model can, and have, been questioned (239). The primary limitation however, concerns the fact that the model uses a short time horizon (1 year), which only captures patient outcomes up to the point of true diagnosis; and the fact that the model is deterministic, meaning that uncertainty around the fixed model input parameters has not been captured. This means that: first, the base-case cost and QALY estimates produced from the FC cost-utility model may not provide a complete and accurate reflection of the cost-effectiveness of primary care FC testing strategies; and second, that the analysis of the impact of

measurement uncertainty on the base-case outputs has not been able to address questions around how FC measurement uncertainty impacts on the *probability of cost-effectiveness* for the FC strategies (which requires probabilistic sensitivity analysis [PSA] results), or how the impact of measurement uncertainty on the modelled outcomes compares to the overall impact of other sampling (i.e. second-order, parameter) uncertainty.

In their evaluation of the YFCCP, NICE considered the deterministic model results to be sufficient to guide a positive recommendation of this pathway (most likely due to the fact that this pathway was found to be highly cost-effective). In most cases, however, a probabilistic model would be expected to be required in order to accurately guide policy decision making, and in this context the exploration of additional methods to address the impact of measurement uncertainty on probabilistic outputs would be useful. It is expected that, if implementing the methods presented in this chapter within a probabilistic economic model, one would need to apply an inner and outer loop of simulation. That is, for each level of bias and imprecision applied, the associated sensitivity and specificity values (produced from the error-model analysis) would be applied as fixed parameters within the economic model, and the PSA analysis would be run (e.g. n=10,000 Monte-Carlo simulations) to produce a probabilistic estimate of cost-effectiveness at that specific level of bias and imprecision. This process would then be repeated for each of the bias and imprecision values explored, which would allow contour plots to be constructed in the same way as presented in this thesis, but for the alternative outcome of the probability of cost-effectiveness. This would mean, for example, that the cost-effective region as presented in this chapter could be extended to instead present the region of analytical performance maintaining a specified minimum likelihood of cost-effectiveness (e.g. ≥80% probability of being cost-effective).

It should be noted that the above approach – centring on the addition of *hypothetical* bias and imprecision via the error model – does not allow for the assessment of the *relative* importance of measurement uncertainty compared to other parameter uncertainty in the model. In order to address this question, one would first need to establish an *expected distribution* of measurement uncertainty around the baseline test measurements in the model, and apply this as an

uncertain parameter in the same way as for other stochastic parameters feeding into the PSA model. This approach would require additional information on the real world distribution of measurement uncertainty, which may be hard, if not impossible, to identify. Nevertheless, in cases where this is achievable, then a clear advantage lies in the fact that the expected value of perfect parameter information on measurement uncertainty could be evaluated within a value of information analysis, using the PSA outputs[51]. This would mean that the relative importance of measurement uncertainty compared to other uncertain parameters could be addressed, enabling a more comprehensive assessment of whether measurement uncertainty matters for clinical decision making. Further research is required, however, to confirm whether both of the above suggested approaches are feasible and appropriate.[52]

In the context of this case study the primary purpose of this analysis was to explore and illustrate how the error model simulation approach could be embedded within a decision analytic modelling framework to evaluate cost-effectiveness outcomes. Whilst the restriction to deterministic outputs represents a clear limitation, this study nevertheless clearly demonstrates a useful mechanism for linking measurement uncertainty to health-economic outcomes, and presents possible approaches to deriving outcome-based APS and assessing the value of different APS requirements. Future studies could apply this same methodology within long-term, fully probabilistic decision models.

---

[51] Modern approaches to value of information analysis (e.g. efficient regression—based approaches) can be undertaken using PSA outputs alone. For example the Sheffield Accelerated Value of Information (SAVI) online tool can calculate value of information results, including the expected value of perfect parameter results, based on user-inputted information on the probabilistic parameters and the PSA incremental cost and effect results (see http://savi.shef.ac.uk/SAVI/).

[52] It should be noted that with either approach, the inclusion of direct test measurements, measurement uncertainty, and an assigned diagnostic cut-off threshold in the model essentially subverts the need to include user-defined diagnostic sensitivity or specificity values (since diagnostic sensitivity and specificity are products of the true values, measurement uncertainty, and the cut-off threshold).

6.5 Summary

- In this chapter, the error model simulation results from Chapter 5 were embedded within a previously developed economic decision model, to extend the case study analysis to cost, QALY, and cost-effectiveness outcomes. Together with Chapter 5, these results support hypothesis C of this thesis: that methods from the broader literature (i.e. identified in Chapter 3) may be applied within HTA-style assessments, to evaluate the impact of measurement uncertainty on clinical performance, clinical utility and cost-effectiveness outcomes.
- The results were presented using contour plots, which were again used to assess the robustness of each pathway's outcomes to increase bias and imprecision. Based on this analysis, the NICE FC pathway was confirmed to be a sub-optimal pathway which is volatile to positive bias; whilst the YFCCP was confirmed to be to be an optimised and robust strategy.
- The contour plots were further utilised to illustrate two new concepts: "cost-effective regions" of bias and imprecision (where INMB ≥ £0); and "optimal regions" of bias and imprecision (where INMB ≥ a selected INMB percentile). As for the acceptable regions presented in Chapter 5, these concepts similarly relate to hypothesis D of this thesis: that the application of methods from the broader literature to HTA-style assessments could enable outcome-based APS to be derived.
- For the YFCCP, the cost-effective regions were found to allow high levels of imprecision and bias. Optimal regions provided a useful alternative in this scenario, and further enable the added benefit of imposing tighter APS to be quantified.

Overall, the results from **Chapter 5** and **Chapter 6** have illustrated how the impact of measurement uncertainty on pathway outcomes can be assessed, based on applying hypothetical bias and imprecision via error model simulation. In **Chapter 7**, an alternative error model simulation analysis is presented, drawing on RWE: in this example, between-assay performance data derived from a national EQA scheme for FC is used to evaluate the impact of between-assay differences on clinical and health-economic outcomes.

Chapter 7
Real World Evidence (RWE) analysis

7.1 Chapter Outline

In Chapter 5 and Chapter 6, the error model simulation approach was used to explore the impact of *hypothetical* bias and imprecison on FC pathway outcomes. The aim of this chapter was to explore how RWE may be used within the error model simulation approach, to address hypothesis E of this thesis: that methods from the broader literature may be applied or adapted to allow RWE to be utilised within outcome-based assessments. In this case, RWE on FC measurement performance was used as a means of assessing the impact of between-assay differences on clinical and health-economic outcomes. Using data from a national EQA scheme for FC, a RWE analysis was conducted to assess whether or not the YFCCP could achieve the same clinical and economic benefit as demonstrated in the previous chapters, if alternative FC assays were applied.

Two alternative assays were explored in this analysis: the BÜHLMANN fCAL® turbo assay, and the Thermo Fisher EliA™ Calprotectin 2 assay. Between-assay bias and SD profiles for each assay were first derived from the EQA data, using the BÜHLMANN fCAL® ELISA assay as the reference measurement. Using a modified version of the error model presented in Chapter 5, the bias and SD profiles were used to estimate the diagnostic accuracy of the YFCCP when using the alternative assays. These estimates were then applied within the FC cost-utility model, to assess the pathway's cost-effectiveness. The EQA data is first summarised below (section 7.2), followed by the methods (section 7.3), results (section 7.4) and discussion (section 7.5).

7.2 Data

The ongoing UK NEQAS EQA scheme for FC (run by the Birmingham Quality group) was established in February 2012 to monitor the performance of FC assays (230). Under this scheme, three specimens of human faeces are distributed monthly to participating laboratories[53] for analysis using each centre's

[53] Participation in this scheme is open to NHS laboratories, private laboratories, university departments, diagnostic manufacturers and point-of-care users (e.g. GP

chosen FC assay. The three specimens are typically formulated from a mixture of IBD patient samples, and vary in terms of composition from month to month. Each month the participating laboratories analyse the three specimens provided (based on an individual sample analysis) and submit their results electronically to Birmingham Quality.

Based on the submitted FC results, monthly EQA reports are compiled and returned to the participating laboratories (see Appendix M for an example anonymised report provided by Birmingham Quality). Due to a curent lack of reference measurement procedure or CRM for FC (268), rather than presenting a "true" target value for each EQA specimen, the 'all laboratory trimmed mean' (ALTM) is instead reported. The ALTM represents the Healy-trimmed mean (in which specified outlier values are excluded) of all numerical measurements for a given specimen, irrespective of the assay used to obtain the result (282). For each specimen, the EQA reports include data on the ALTM, as well as assay-specific means, CV% and SD values (also based on numerical FC data only). The reports also include an end table (see pages 9-18 in the example report provided in Appendix M) which lists the individual FC results returned across all laboratories, including both numerical values *and* semi-quantitative results (i.e. left- and right-censored vales, such as "<10" or ">600").

The analysis presented in this chapter uses the individual results reported in the end tables of 11 EQA reports, covering the period January 2018 to December 2018.[54] Note that, although individual laboratory results are provided in the report end tables, the laboratory from which each individual result pertains is not reported (i.e. each laboratory is anonymised). No laboratory-identifiable data was thus used in this analysis. The EQA reports were provided by the laboratory at the York Teaching Hospitals NHS Trust in January 2019. Since the reports relate to routine quality assurance data and do not include any patient-identifiable data, formal ethical approval was not required for this case study. Nevertheless, use of this data for publication purposes requires approval from Birmingham Quality

practices, private clinics etc.) (Personal communication with Jane French, Consultant Clinical Scientist at Birmingham Quality, January 2020).

[54] Note that no samples were distributed by NEQAS in February 2018. As a result, there are 11 rather than 12 reports coving the 2018 period.

(who run the scheme), since each EQA report is considered confidential between Birmingham Quality and the participant laboratory. Permission to use this data was obtained from Birmingham Quality in August 2019.

Individual FC values (including numerical and semi-quantitative results) listed in the end tables of the 11 EQA reports, were extracted for three assays: (1) BÜHLMANN fCAL® ELISA – the assay used to measure FC values in the YFCCP dataset (listed as 'Bühlmann [2BU]' in the EQA reports); (2) BÜHLMANN fCAL® turbo (listed as 'Bühlmann fCAL turbo [4BU]'); and (3) (Thermo Fisher) EliA™ Calprotectin 2 (listed as 'Thermo EliA Calpro 2 [2KO2]'). Henceforth, these assays are referred to as 2BU, 4BU and 2KO2 respectively. Although several other FC assays were included in the EQA reports (as reported in Chapter 4, Table 4-1), these three assays represent those most commonly used across the 2018 EQA reports. From the 11 reports extracted (each including 3 FC specimens), a total of 804 results were returned for the 2BU assay, followed by 686 for the 2KO2 assay, and 366 for the 4BU assay. Thus, on average, 24 laboratories returned results for the 2BU test, followed by 21 for the 2KO2 test, and 11 for the 4BU test.

7.2.1 Censored data

As with the YFCCP dataset, the EQA data used in this analysis included left- and right-censored FC values. In this case, a range of lower and upper measurement range limits were observed, due to the fact that different assays and different laboratories report different limit values. A summary of censored data observed within the EQA reports for the three assays evaluated is provided in Table 7-1. For each observed limit, this table reports the frequency of associated censored data for each assay.

Table 7-1 indicates that the 2BU assay had the smallest measurement range, with an upper limit of either 600 µg/g (as in the YFCCP dataset) or 1800 µg/g (achieved by applying different dilution factors to the test samples). The alternative assays meanwhile reported up to ~2,000 µg/g. As a result, the 2BU assay was associated with the highest frequency of right-censored data. The 4BU assay meanwhile had the highest proportion of left-censored data, with lower limits of 20 or 30 µg/g commonly in use; whilst the most common left-censored

data for the 2BU and 2KO2 assays related to lower limits of 30 µg/g and 15 µg/g respectively. Two different methods for handling censored data within the following RWE analysis are discussed in section 7.3 below.

Table 7-1. FC EQA data: censored data

Reported lower/ upper limit	Frequency of censored data N (%)		
	2BU [Total N = 804]	4BU [Total N = 366]	2KO2 [Total N = 686]
Left-censored data			
<4	-	-	3 (<1%)
<10	4 (<1%)	-	-
<15	-	-	16 (2%)
<19	-	2 (<1%)	-
<20	1 (<1%)	36 (10%)	-
<24	-	3 (1%)	2 (<1%)
<25	-	1 (<1%)	-
<26	-	2 (<1%)	-
<30	46 (6%)	23 (6%)	4 (<1%)
<50	-	9 (2%)	-
Right-censored data			
>600	100 (12%)	2 (<1%)	-
>1,799	-	2 (<1%)	-
>1,800	13 (2%)	2 (<1%)	-
>1,932	-	1 (<1%)	-
>2,000	-	2 (<1%)	1 (<1%)
>2,100	-	-	1 (<1%)

7.3 Methods

7.3.1 EQA data analysis: bias and SD profiles

The EQA data summarised in section 7.2 was used to evaluate the comparative performance of two alternative assays, 2KO2 and 4BU, compared to 2BU. The research question framing this analysis, is whether or not the clinical and economic performance of the YFCCP could be maintained when using alternative assays to the 2BU assay. Whilst the 2BU assay is not representative of "true" measurement, it is the assay which has informed the development and optimisation of the YFCCP. For the specific research question considered in this analysis therefore (and in the absence of a reference measurement procedure for FC), 2BU represents the relevant reference assay.

For each specimen included in this analysis (n=33), the 2BU reference measurement was calculated as the mean of the reported 2BU measurements for that specimen[55]. Mean 2BU specimen values were calculated using both numerical and semi-quantitative FC results, with semi-quantitative values set equal to their respective limits in the base case analysis (full details of censored data handling are provided further below). Between-assay differences were calculated by subtracting the 2BU specimen mean reference measurement (n=33) from each 2KO2 (n=686) and 4BU (n=366) individual specimen value. For completeness, this calculation was also conducted for the individual 2BU measurements (n=804) (but note that in the following bias analysis the expected (i.e. mean) bias for the 2BU reference assay is, by default, zero).

Using the calculated between-assay differences, bias profiles were produced for each assay. The EQA data was first plotted on a difference plot, which shows the individual between-assay differences (presented on the y-axis) against the 2BU reference assay measurements (presented on the x-axis). At each reported reference measurement (n=33 specimens), a scatter of between-assay difference values were produced for each assay, depending on the number of laboratories returning results for that assay. Using these difference plots, bias

[55] Note that there was little difference between mean vs. median values across the 2BU measurements for each specimen.

profiles for each assay were derived by fitting a loess regression model to the difference plot.

Loess is a non-parametric regression approach, similar to ordinary least squares (OLS) regression, in which a series of OLS models are fitted to localised subsets of the data (283). This form of regression is useful for fitting a function of expectation over data where the relationship between variables is non-linear or the parametric form is complex or unclear (283). In contrast to the typical approach taken in Bland Altman difference plots – where mean bias is presented as a fixed value over the measurement range (284, 285) – the use of loess regression provides bias *profiles* which show how the value of expected bias (i.e. average between-assay differences) changes over the reference measurement range (see the resulting bias profile plots, reported in section 7.4.1).

In a similar way to above, SD profiles were also derived for each assay, according to the SD observed across the between-assay differences. For each specimen (n=33), the SD of the reported between-assay differences was calculated for each assay. These values were similarly plotted on a scatter plot as described for the between-assay differences above (this time presenting SD on the y-axis). In this case however, only a single SD estimate for each assay specimen was derived, based on the spread of values observed at each reference measurement point for each assay. The SD profiles for each assay were again derived by fitting a loess regression model to the plotted data. The resulting plots (reported in section 7.4.1) illustrate how the expected variability of between-assay differences for each assay changes over the observed FC measurement range.

The bias and SD profiles were constructed using all available FC data provided within the EQA report end tables, including both numerical and semi-quantitative (i.e. left-censored or right-censored) values. Two approaches were implemented for dealing with censored data. In the base case analysis, all censored FC data were replaced with their respective limit values (i.e. values reported as “>600” were replaced with 600, “<10” replaced with 10, etc.). Alternatively in a sensitivity analysis, censored data were instead replaced with associated median estimates derived from the EQA data. For this sensitivity analysis, median estimates were calculated by taking the median of numerical FC values reported across all three assays and all EQA reports, where data within the associated lower and upper

regions were available. For example, for right-censored data reported as “>600” in the EQA reports, an associated quantitative estimate for FC was calculated by taking the median of *all* numerical FC values reported as >600 µg/g, across all the EQA reports and all three assays (in this case providing a median estimate of 915 µg/g). The median estimates and number of values informing these estimates for each of the observed lower and upper limits, are provided in Table 7-2.

Table 7-2. FC EQA data: median estimates for censored data

Reported Limit	Frequency of censored data N (%)			Median estimate	Number of values informing median estimate
	2BU [Total N = 804]	4BU [Total N = 366]	2KO2 [Total N = 686]		
Left-censored data					
<4	-	-	3 (<1%)	NA	0
<10	4 (<1%)	-	-	8	27
<15	-	-	16 (2%)	11	63
<19	-	2 (<1%)	-	13	103
<20	1 (<1%)	36 (10%)	-	14	120
<24	-	3 (1%)	2 (<1%)	17	169
<25	-	1 (<1%)	-	17	179
<26	-	2 (<1%)	-	18	192
<30	46 (6%)	23 (6%)	4 (<1%)	19	239
<50	-	9 (2%)	-	29	459
Right-censored data					
>600	100 (12%)	2 (<1%)	-	915	260
>1,799	-	2 (<1%)	-	2,318	31
>1,800	13 (2%)	2 (<1%)	-	2,318	31
>1,932	-	1 (<1%)	-	2,581	25
>2,000	-	2 (<1%)	1 (<1%)	2,654	22
>2,100	-	-	1 (<1%)	2,797	19

7.3.2 Outcome assessment

7.3.2.1 Error model framework

In the error model formula used in Chapter 5 (Figure 5-2, Equation 5.1), fixed values of bias and CV were applied to a sample of underlying "true" measurement values. Bias and CV were "fixed" in the sense that, within a given simulation run (e.g. n=10,000), the same values of bias and CV were applied across all of the sampled YFCCP FC1 and FC2 values. In the current analysis however, the impact of between-assay differences (i.e. bias), and variability around those differences (i.e. SD), can be explored using the bias and SD *profiles* outlined in section 7.3.1. With this data, individual bias and SD values can instead be applied to each of the sampled FC values within a given simulation run, by drawing the expected bias and SD values for each $Test_{true}$ value from the bias and SD profiles. A slightly modified error model was therefore used in this case, as outlined in Equation 7.1 below:

$$\boldsymbol{Test_{sim_i} = Test_{true_i} + N(Bias_i, SD_i)} \qquad \textbf{(7.1)}$$

where $Bias_i$ represents the mean between-assay difference associated with the $Test_{true}$ value in the bias profile, and SD_i represents the expected variation in between-assay difference at the same $Test_{true}$ point in the SD profile.

Figure 7-1 summarises the modified error model simulation process used to model the impact of FC between-assay differences on the performance of the YFCCP. As in the analysis presented in Chapter 5, "true" FC values within this simulation were again sampled from the YFCCP dataset, which used the 2BU assay. The "true" FC values may again be sampled using the bootstrap method or the parametric method. However, as previously discussed in Chapter 5 (section 5.5.2) and Chapter 6 (section 6.4.1), the bootstrap method provides a close fit to the YFCCP data, in contrast to the parametric method which produces consistently biased results. For the purposes of this anlaysis therefore, only the bootstrap method is used.

i. A sample of $FC1_{true}$ values is assigned;

ii. For each $FC1_{true}$ value, the addition of measurement uncertainty is simulated according to the specified error model to generate $FC1_{sim}$ values:

$$\boldsymbol{FC1_{sim_i} = FC1_{true_i} + N(Bias_i, SD_i)} \qquad \textbf{(7.2)}$$

iii. For all $FC1_{sim}$ values ≥100 µg/g, an associated sample of $FC2_{true}$ values is assigned;

iv. For each $FC2_{true}$ value, the addition of measurement uncertainty is simulated according to the specified error model to generate $FC2_{sim}$ values:

$$\boldsymbol{FC2_{sim_i} = FC2_{true_i} + N(Bias_i, SD_i)} \qquad \textbf{(7.3)}$$

v. The diagnostic accuracy of the YFCCP including additional measurement uncertainty is calculated by comparing diagnoses based on the $FC1_{sim}$ and $FC2_{sim}$ values (using the YFCCP diagnostic protocol) with patients' clinical diagnoses;

vi. Steps (i) to (v) are repeated for each assay's bias and SD profiles

Figure 7-1. Modified error model simulation approach: two-stage FC testing method

The process outlined in Figure 7-1 was run twice: once for each of the alternative FC assays assessed. Each analysis in this case produces one diagnostic sensitivity and specificity result, based on the specified bias and SD profile. Contour plots and acceptability regions are therefore not applicable in this assessment, and no smoothing algorithm – as used in the base case analyses presented in Chapter 5 – could be applied. The simulations in this case were therefore based on drawing a higher number of bootstrap samples (n=100,000).

For each alternative FC assay assessed, the individual values of $Bias_i$ and SD_i in the simulation were drawn from the bias and SD profiles reported in section 7.4.1. These values were simulated using the loess 'predict' function in R – a function which draws values from a fitted loess function, for each point along the observed measurement range. Note that values of bias and associated SD can only be derived in this way over the range of observed data: no predictions of bias or SD can be provided below the lowest reference 2BU measurement, or above the

highest reference 2BU measurement. As such, within the error model simulation, $Bias_i$ and SD_i values relating to $FC1_{true}$ or $FC2_{true}$ values below the lowest defined value, or above the highest defined value within the bias and SD profiles, were set equal to the lowest and highest observed values respectively (i.e. "nearest neighbour" method).

In the base case analysis, all censored FC data – both within the EQA data and the YFCCP dataset – were replaced with their respective limit values. In the sensitivity analysis, all censored FC data were instead replaced with the associated median estimates reported in Table 7-2. In order to be consistent, the same substitutions were applied within the YFCCP bootstrap dataset as in the EQA data in this analysis: that is, left- and right-censored YFCCP FC data ("<10" and ">600") were replaced with the associated EQA median estimates (8 and 915 µg/g, respectively) within the sensitivity analysis. In both the base case and sensitivity anlayses, the simulation results were compared to the baseline diagnostic accuracy of the YFCCP, which was based on 2BU FC values (as reported in Chapter 5, section 5.4.2.1).

7.3.2.2 Cost-effectiveness outcomes

As well as evaluating the diagnostic accuracy of the two alternative FC assays, the simulation results were applied to the YFCCP arm of the FC cost-utility model described in Chapter 6 to determine the expected cost, QALY and NMB results for the YFCCP using the alternative assays. Note that it was assumed within this analysis that the cost of the alternative assays would be the same as the cost of the 2BU assay applied within the FC cost-utility model (£24)[56] – i.e. only the diagnostic accuracy inputs within the FC cost-utility model were updated. As such, any resulting cost differences between the 2BU, 4BU and 2KO2 YFCCP strategies reported in section 7.4.2, are driven purely by changes in the diagnostic accuracy of the assays.

The resulting NMB estimates were compared to the NMB results produced by the FC cost-utility model intervention and comparator strategies (as reported in

[56] YHEC derived the unit cost of FC from the NICE MIB 132 report, based on the cost quoted for standard care laboratory-based FC ELISA testing (£23.30, inflated by YHEC to 2017/18 prices).

section 6.2.1, Table 6-2). In particular, INMB values were calculated based on comparing the YFCCP using each of the alternative assays, against the YFCCP using the 2BU reference method (i.e. the original intervention arm of the FC cost-utility model). These INMB values indicate the closeness of agreement between the YFCCP strategy when using the alternative assays vs. the 2BU assay, with a negative INMB indicating that the alternative assays are associated with worse NMB than the 2BU assay.

7.3.3 Bias correction exercise

It may be that, even if the clinical and economic performance of the YFCCP is negatively affected by adopting the 4BU or 2KO2 assays, this difference could be easily corrected by applying a simple adjustment to the measured FC values to offset the between-assay differences. For example, a simple FC adjustment may adequately restore the YFCCP performance in cases where a relatively constant bias is the main driver of between-assay differences. To explore this possibility, a simulation was run applying a series of alternative correction values to the measured FC1 and FC2 values (i.e. $FC1_{sim}$ and $FC2_{sim}$) within the simulation.

Two exercises were conducted. In the first, the addition of an absolute correction value was applied, ranging from -50 to +100 µg/g in 5 µg/g increments. Note that this analysis is equivalent to adjusting the diagnostic cut-off threshold for the alternative assays: for example, applying a uniform +10 µg/g adjustment to measured test values is equivalent to reducing the cut-off threshold by 10 µg/g (from 100 to 90 µg/g). In the second exercise, a proportional adjustment factor was instead applied. In this case, factors ranging from 0.5 (equivalent to a 50% reduction in the measured FC values) to 2.0 (equivalent to a uniform 100% increase in the measured FC values) were explored, in 0.05 increments.

The results of this analysis were assessed in terms of the YFCCP diagnostic sensitivity and specificity achieved within each scenario. The objective was to determine if the YFCCP's sensitivity and specificity could be made to match that of the baseline levels achieved when using the 2BU assay within the YFCCP dataset (94% sensitivity, 92% specificity). In each case the correction value achieving the highest combined sensitivity and specificity (i.e. highest diagnostic yield) was identified; and the correction value achieving the closest match to the

YFCCP [2BU] sensitivity and specificity values was also identified[57]. The latter values were applied within the FC cost-utility model to further determine the mean cost, QALYs and NMB associated with this particular point. As before, the simulations in this analysis were based on drawing 100,000 bootstrap samples.

7.4 Results

7.4.1 EQA data analysis: bias and SD profiles

Figure 7-2 presents the bias profiles derived in the base case analysis (plot A) and sensitivity analysis (plot B). The vertical scatter of difference results distributed at fixed points along the measurement range illustrates the cluster of values reported for each EQA specimen. The dashed light green line illustrates the bias profile for the 4BU assay (i.e. the loess regression function fitted to the 4BU assay between-assay difference results); and the dot-dashed purple line illustrates the bias profile for the 2KO2 assay. The 2BU reference assay has zero expected bias by definition, and the fitted loess function in this case therefore presents as a solid black line along the zero bias line.

Figure 7-3 illustrates the corresponding SD profiles for each assay. Here one value of SD is calculated for each assay for each specimen. The 2BU reference assay is associated with non-zero SD values in this case, since a spread of individual 2BU specimen results informs each of the mean specimen values used as the reference measurements within the between-assay difference calculations.

Note that, whilst the bias associated with the alternative FC assays appears to be marginal across a large portion of the measurement range, the scale of the plot is heavily skewed by large difference values occurring at both the lower and higher end of the measurement range. At the point of 100 µg/g, for example, in the base case analysis the 4BU assay is associated with an expected bias of -22 µg/g compared to the 2BU assay, and the 2KO2 assay is associated with an

[57] The correction value achieving the "closest match" to the YFCCP [2BU] strategy was defined as the correction value associated with sensitivity (A) and specificity (B) values which had the lowest combined squared difference from the YFCCP [2BU] sensitivity (93.6%) and specificity (92.0%) i.e. those values for which $((93.6\text{-}A)^2 + (92.0\text{-}B)^2)^{1/2}$ was minimised.

expected bias of -39 µg/g. Although this magnitude of bias appears marginal in Figure 7-2, at this region of the measurement range (i.e. close to the FC 100 µg/g cut-off value) these values could nevertheless be influential in terms of patient outcomes.

Appendix N.1 also provides density plots for the $FC1_{sim}$ and $FC2_{sim}$ values produced within the error model simulation from the 4BU and 2KO2 bias and SD profiles. These figures further illustrate the impact of bias and SD on the simulated FC values for each alternative assay in the analysis, compared to the bootstrapped YFCCP data.

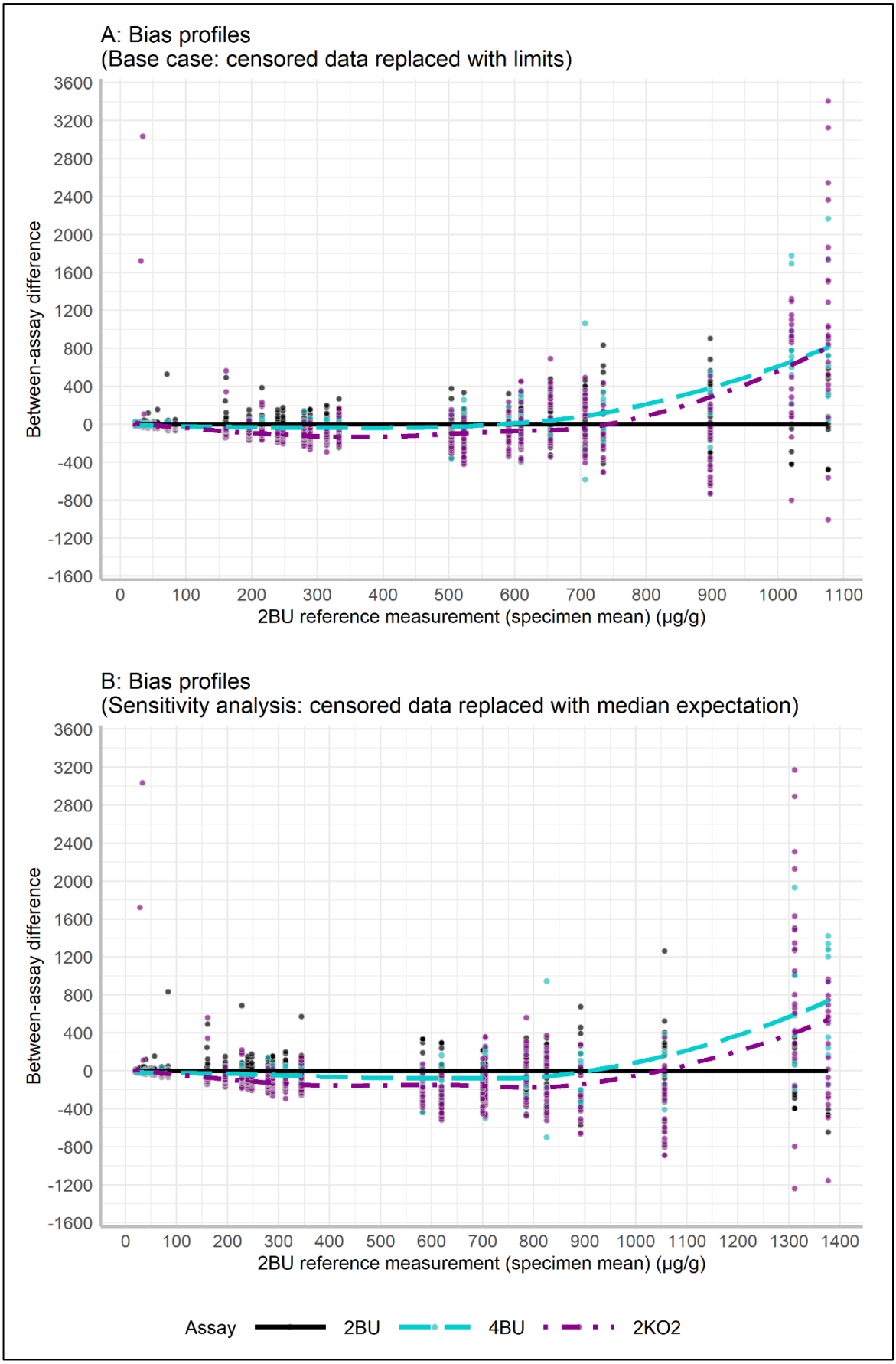


Figure 7-2. FC EQA data: bias profiles

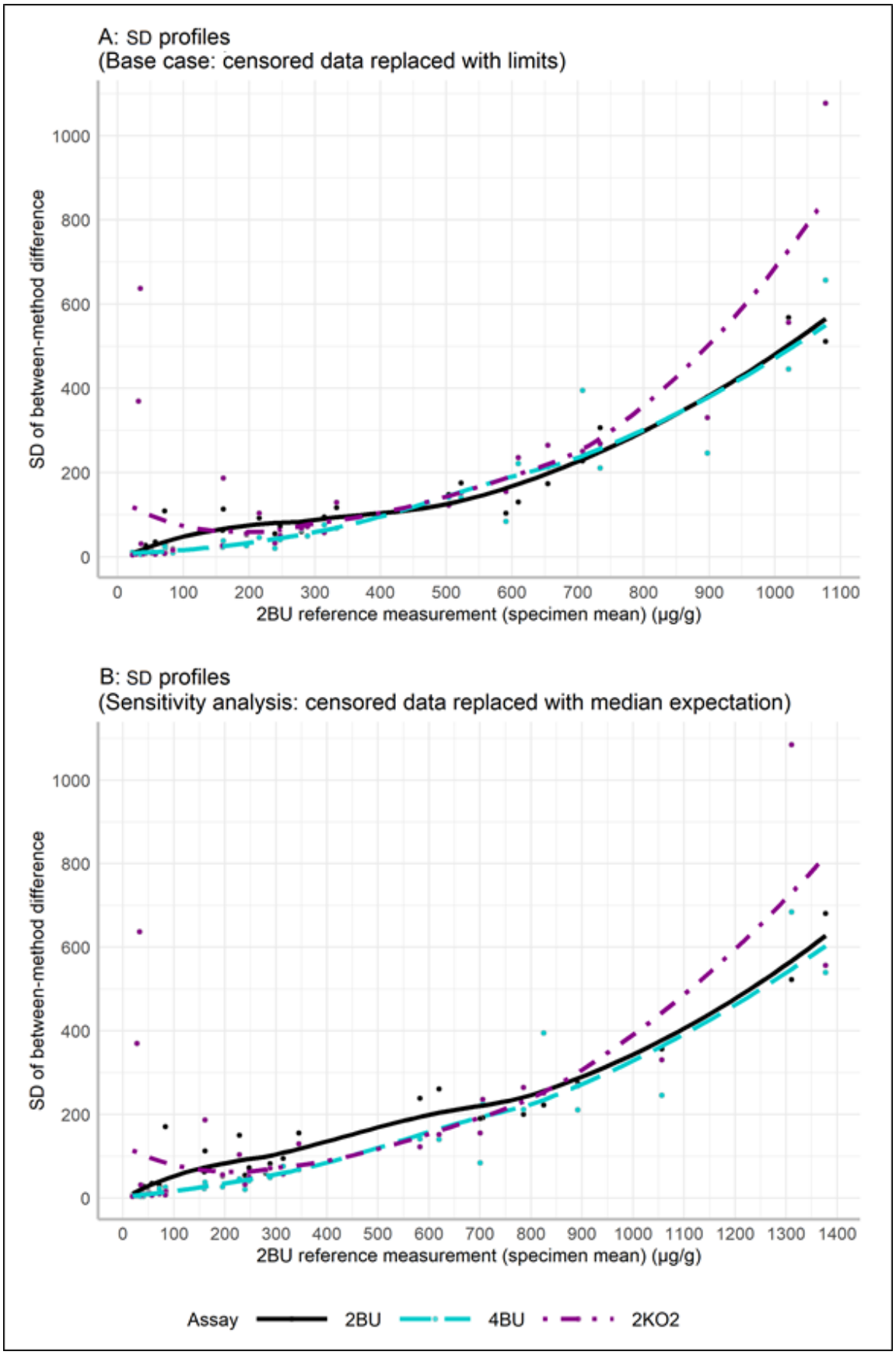


Figure 7-3. FC EQA data: SD profiles

7.4.2 Outcome assessment

Table 7-3 presents the diagnostic accuracy results, and associated cost-effectiveness results, for the 4BU and 2KO2 assays. For reference, the bottom half of this table also presents the diagnostic accuracy inputs and cost-effectiveness outputs for each strategy evaluated in the original FC cost-utility model (as reported in Chapter 6, section 6.2.1). The YFCCP intervention arm of the FC cost-utility model used the baseline diagnostic accuracy results from the YFCCP dataset, and therefore represents the 2BU-relevant comparator in this analysis. The INMB results in this table present the INMB for each alternative assay YFCCP strategy vs. the YFCCP [2BU] comparator strategy. A negative INMB therefore indicates that the strategy in question is not cost-effective compared to the YFCCP [2BU] strategy.

From the diagnostic accuracy results, it can be seen that, when using the 4BU assay the YFCCP is associated with notably lower sensitivity (80.2% vs. 93.8%) and marginally higher specificity (93.5% vs. 92.0%) compared to using the 2BU assay. The 2KO2 assay meanwhile is associated with lower sensitivity (65.5% vs. 93.8%) *and* specificity (82.8% vs. 92.0%) compared to the 2BU assay. For both of the alternative assays, there was marginal difference between the base case and sensitivity analysis results. Further discussion of these results – in particular the key determinants driving the different findings for each assay – is provided in the discussion (section 7.5.1).

In terms of cost-effectiveness, when using the 4BU assay the YFCCP is associated with slightly lower mean costs and QALYs than the 2BU assay; with the associated NMB also being reduced (£15,567 vs. £15,581; INMB = -£14). However, when compared to the comparator strategies from in the original FC cost-utilty model, the YFCCP remained cost-effective when applying the 4BU assay sensitivity and specificity results, producing higher NMB than each of the fixed comparator strategies.

When using the 2KO2 assay meanwhile, the YFCCP was associated with higher costs and lower QALYs than the 2BU assay, resulting in a lower NMB (£15,493 vs. £15,581, INMB = -£88). In this case the YFCCP was no longer cost-effective compared to the two highest performing comparator strategies within the FC cost-utility model [i.e. the ‘FC (NICE data)’ and the ‘No FC (NICE data)’ strategies].

Table 7-3. YFCCP RWE analysis: outcome results

FC assay	Diagnostic accuracy		Cost-effectiveness			
	Sensitivity	Specificity	Cost	QALY	NMB	INMB (£) vs. YFCCP [2BU]
EQA analysis results						
YFCCP [4BU] base case	0.802	0.935	£207	0.7887	£15,567	-£14
YFCCP [4BU] sensitivity analysis	0.812	0.935	£208	0.7889	£15,569	-£12
YFCCP [2KO2] base case	0.655	0.828	£246	0.7869	£15,493	-£88
YFCCP [2KO2] sensitivity analysis	0.650	0.831	£247	0.7870	£15,493	-£88
FC cost-utility model diagnostic accuracy inputs and cost-effectiveness outputs (for reference)						
YFCCP [2BU] intervention	0.936	0.920	£212	0.7896	£15,581	-
No FC (Tibble data)	0.350	0.730	£259	0.7836	£15,412	-£169
No FC (NICE data)	1.000	0.790	£232	0.7879	£15,526	-£55
FC testing (YFCCP, 50 µg/g cut-off)	0.960	0.600	£314	0.7836	£15,359	-£222
FC testing (Tibble data)	0.900	0.800	£245	0.7860	£15,474	-£107
FC testing (NICE data)	0.930	0.940	£197	0.7880	£15,562	-£19

7.4.3 Bias correction exercise

Figure 7-4 illustrates the results of the bias correction exercise when applying an absolute correction value; and Figure 7-5 illustrates the results when applying a proportional correction factor. In both exercises, the results of the base case and sensitivity analyses were very similar: for simplicity therefore, only the base case results have been presented. In Figure 7-4, an absolute correction value of 0 indicates that no adjustment was made to the measured FC values within the simulation; whilst a correction value of -50 µg/g corresponds to increasing the FC cut-off threshold from 100 to 150 µg/g; and a correction value of 100 µg/g corresponds to decreasing the cut-off threshold from 100 to 0 µg/g. In Figure 7-5 meanwhile, a proportional adjustment factor of 1 indicates that no adjustment was applied; whilst an adjustment factor of 0.5 corresponds to halving the measured results, and a factor of 2 corresponds to doubling the measured FC results.

The results of both exercises indicate that with the 4BU assay, significant gains in pathway sensitivity could be achieved if a positive correction value were applied. In particular, the highest diagnostic yield was achieved when: (i) applying an absolute correction value of 35 µg/g (which increased the YFCCP [4BU] sensitivity from 80% to 93%, and decreased specificity from 93% to 89%); or (ii) applying a proportional correction factor of 1.45 (which increased sensitivity to 92%, and reduced specificity to 90%). In the proportional correction factor exercise, the point of highest diagnostic yield (sensitivity 93%, specificity 89%) was also the closest match to that of the YFCCP [2BU] reference strategy (sensitivity 94%, specificity 92%). In the absolute correction value exercise however, a correction value of 30 µg/g achieved the closest match, giving a sensitivity of 91% and specificity of 90%. When applying either a proportional correction factor of 1.45, or an absolute correction value of 30 µg/g, the same cost-effectiveness results were produced: the YFCCP [4BU] strategy costs were increased from £207 (in the base case) to £219 (applying the measurement correction factor), mean QALYs were increased from 0.7887 to 0.7893, and the expected NMB was marginally increased from £15,567 to £15,568.

For the 2KO2 assay, it appears that the performance of the YFCCP when using this test remains substantially below that associated with the 2BU assay, regardless of what correction value is applied. Similar to the 4BU analysis, higher

pathway sensitivity can be achieved with the 2KO2 assay by applying a positive absolute or proportional correction value. In this case, the highest diagnostic yield and closest match to the YFCCP [2BU] strategy results was achieved when applying an absolute correction value of 20 µg/g, which increased sensitivity from 66% to 74% and decreasing specificity from 83% to 77%; or a proportional correction value of 1.35, which increased sensitivity to 75% and reduced specificity to 75%. However, in both cases the sensitivity and specificity remained notably lower than that achieved in the YFCCP when using the 2BU assay. When running these results through the YFCCP arm of the FC cost-utility model, the 2KO2 pathway base case mean cost (£246) was increased to £268 (20 µg/g absolute correction) or £274 (1.35 proportional correction); the pathway mean QALYs (0.7869) was marginally increased to 0.7870 (both correction values); and the pathway NMB (£15,493) was decreased to £15,473 (20 µg/g absolute correction value) or £15,467 (1.35 proportional correction factor). The reduced specificity in this case therefore results in worsened, rather than improved, cost-effectiveness.

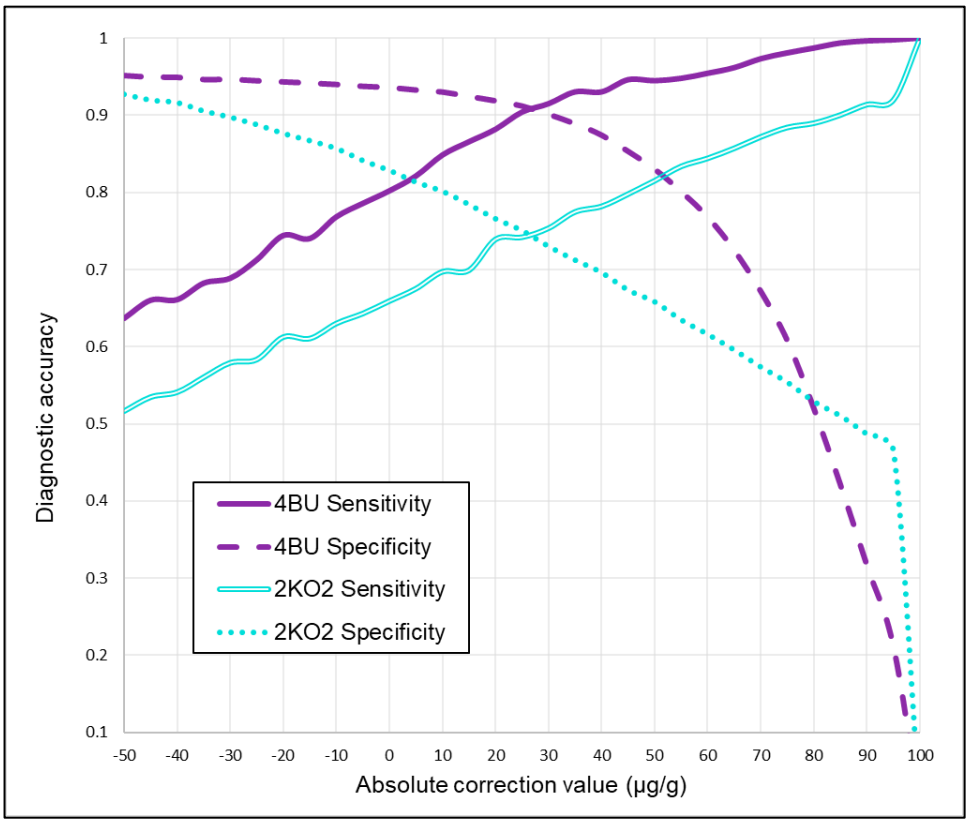


Figure 7-4. RWE analysis: plot of absolute adjustment value vs. diagnostic accuracy for 4BU and 2KO2 FC assays

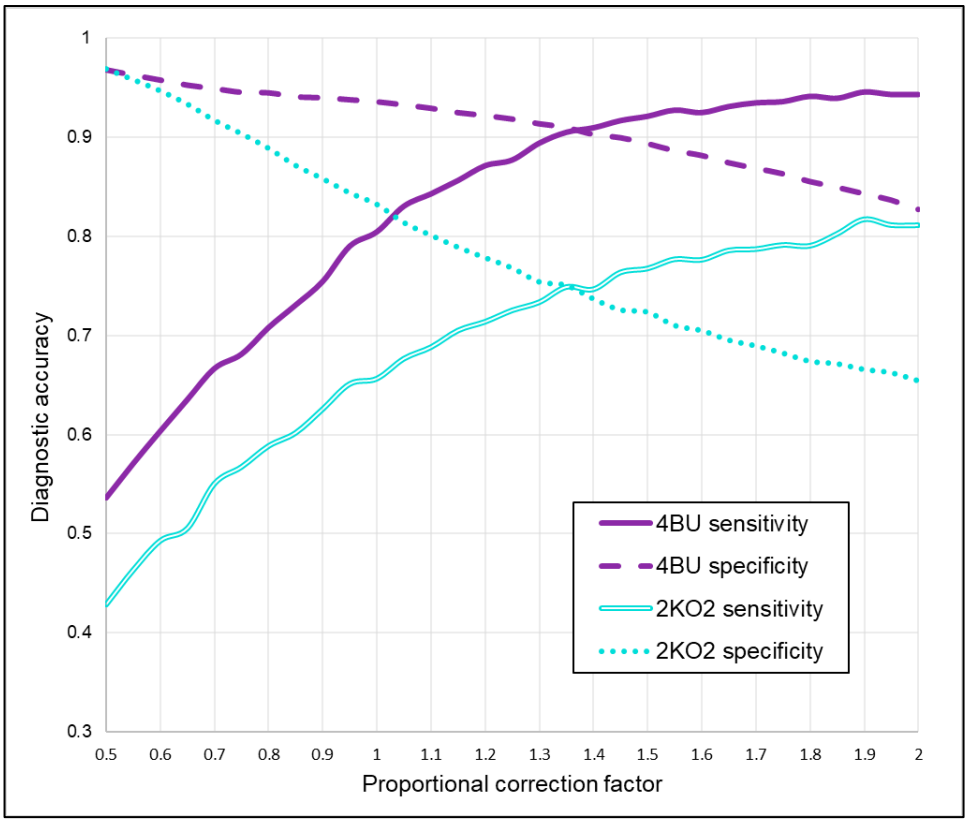


Figure 7-5. RWE analysis: plot of proportional adjustment factor vs. diagnostic accuracy for 4BU and 2KO2 FC assays

7.5 Discussion

7.5.1 Outcome assessment

The objective of this analysis was to explore how RWE may be used within the error model simulation framework presented in the previous case study chapters. For this analysis, RWE was available in the form of EQA reports produced from the UK NEQAS EQA scheme for FC. This data – which included individual test results for a range of assays – enabled an evaluation of the impact of between-assay differences on clinical and health-economic outcomes. In particular the individual-level nature of the EQA data enabled bias and SD *profiles* to be utilised within an error model simulation, to capture the impact of variable between-assay differences across the FC measurement range.

For the 4BU assay, the results indicate that if used in the YFCCP, this assay would produce notably worse sensitivity (80% vs. 94%) and slightly higher specificity (94% vs. 92%) compared to the same pathway using the 2BU reference assay (Table 7-3). In terms of cost-effectiveness, the higher pathway specificity achieved with the 4BU assay resulted in lower mean costs compared to the 2BU assay (£207 vs £212); whilst the lower sensitivity in this case resulted in reduced mean QALYs (0.7887 vs. 0.7896) (Table 7-3). The overall impact on the cost-effectiveness of the YFCCP was marginal: although a slightly lower mean NMB was achieved with the 4BU assay compared to the 2BU assay (£15,567 vs. £15,581), the YFCCP [4BU] strategy nevertheless remained cost-effective when compared to all comparators included in the FC cost-utility model (Table 7-3). In this case therefore, whilst a drop in sensitivity was observed with the 4BU assay, the cost-effectiveness of the YFCCP was maintained thanks to the YFCCP's robustness to increases in measurement uncertainty (as demonstrated in Chapter 6).

For the 2KO2 assay, the results indicate that if used in the YFCCP, this assay would produce substantially worse sensitivity (66% vs. 94%) *and* specificity (83% vs. 92%), compared to the same pathway using the 2BU reference assay (Table 7-3). This results in higher mean costs for the YFCCP compared to using the 2BU assay (£246 vs. £212), as well as lower QALYs (0.7869 vs. 0.7896) and lower NMB (£15,493 vs. £15,581) (Table 7-3). The reduced NMB in this case leads to

the YFCCP no longer being cost-effective when compared to the two highest performing comparator strategies included in the FC cost-utility analysis (Table 7-3). For the 2KO2 assay therefore, the drop in diagnostic accuracy associated with this assay was sufficient to render the YFCCP no longer uniformly cost-effective.

To understand the different results obtained for the 4BU and 2KO2 assays outlined above, a closer inspection of the associated bias and SD profiles is required (Figure 7-2 and Figure 7-3). For the 4BU assay, a negative mean bias was obtained at the lowest reference measurement point of 24 µg/g (bias= -8 µg/g), with a peak in negative bias at 380 µg/g (bias = -38 µg/g). Positive bias did not occur with this assay until a reference measurement of 579 µg/g was reached, after which a steep rise in positive bias occurred. The consistent negative bias observed below 579 µg/g, together with moderate SD over the same region, means that $FC1_{sim}$ and $FC2_{sim}$ values within the error model simulation tended to be under-estimated compared to the 2BU assay, for both the IBD and IBS populations. This results in decreased sensitivity (due to IBD patients being incorrectly pulled under the 100 µg/g cut-off threshold) and increased specificity (due to IBS patients being correctly pulled under the cut-off threshold).

For the 2KO2 assay, the bias profile exhibited a small *positive* bias at the lowest reference measurement point of 24 µg/g (bias= 15 µg/g), which gradually reduced to zero bias at a reference measurement of 44 µg/g (Figure 7-2). Negative bias in this case reached its peak at 359 µg/g (bias = -135 µg/g), before returning to a positive bias at 743 µg/g and rapidly increasing thereafter. The positive bias observed with this assay at the lower end of the measurement range leads to test values within the IBS population, but *not* the IBD population, being pushed above the test cut-off threshold (100 µg/g). This differential impact occurs for two reasons. First, it is almost exclusively IBS patients who occupy the lower region of the measurement range within the YFCCP dataset (see Chapter 5, Figure 5-9). Second, a corresponding peak in SD at the lower measurement range with the 2KO2 assay means that the bias applied within the error model simulation was sufficient to push some low $FC1_{true}$ and $FC2_{true}$ values above the 100 µg/g YFCCP cut-off threshold (Figure 7-3). As such, the 2KO2 bias profile has a negative

impact on both the sensitivity and specificity of the YFCCP, leading to reduced cost-effectiveness.

With regards to the 2KO2 findings, the positive bias and heightened variation observed at the lower measurement range for this assay appears to be driven by two disproportionately high FC measurements in this region. At a reference measurement of 31 µg/g, one laboratory returned a 2KO2 value of 1,751 µg/g (producing a between-assay difference of 1,720 µg/g); and at a reference measurement of 34 µg/g one laboratory returned a 2KO2 value of 3,068 µg/g (producing a between-assay difference of 3,034 µg/g) (Figure 7-2). These two extreme values increased the average bias and SD associated with the given samples (Figure 7-2 and Figure 7-3). It is unknown if these extreme values were a result of between-assay differences (and therefore have appropriately informed the 2KO2 bias and SD profiles), or if the error in these cases relates to pre-analytical and/or post-analytical factors (such as a technical misreporting of the data). If it could be shown that the extreme values were a result of reporting error, then it may be argued that these values should be treated as outliers and removed from the analysis. Whilst this question cannot be answered with the data at hand, it is of interest to explore how influential these extreme values were in the analysis.

Appendix N.2 presents the results of a post-hoc sensitivity analysis in which the two extreme 2KO2 values discussed above were removed from the analysis. In brief, removal of these values results in an important shift in the bias and SD profiles: the positive expected bias previously observed at the lower end of the measurement range for 2KO2 was no longer maintained, with negative bias instead occurring from the lowest reference measurement up to 742 µg/g; and at the same time, the large SD previously observed at the lower measurement range for 2KO2 was notably attenuated. The combined effect of these changes is that the FC measurements are more consistently under-estimated in the simulation analysis, resulting in reduced YFCCP sensitivity (from 66% in the primary 2KO2 analysis to 62% in the post-hoc sensitivity analysis) and increased specificity (from 83% to 96%). This leads to reduced costs and increased QALYs for the YFCCP [2KO2] strategy, resulting in an increase in NMB (from £15,493 to £15,551). In this scenario the YFCCP [2KO2] strategy is cost-effective against all

but the highest performing comparator in the NICE FC cost-utility model (i.e. the 'FC testing (NICE data)' comparator). Based on these results, it is clear that the two extreme values in this case are having a disproportionate effect on the results for this assay. Nevertheless, the reliability of the two extreme values explored in this analysis cannot be ascertained based on the available data. If the two extreme values could be confirmed as being valid, then the base case analysis should be applied.

Two methods for dealing with censored data were explored in this analysis. In the base case, censored data were replaced with their associated limit values (as in the base case analyses presented in Chapter 5 and Chapter 6); whilst in a sensitivity analysis, all censored data (both in the YFCCP bootstrap dataset and in the EQA data) were alternatively replaced with median estimates derived from the EQA data (Table 7-2). A key motivation for running this sensitivity analysis was to assess whether the relatively low upper limit of the 2BU assay used within the YFCCP dataset (600 µg/g) could have introduced bias into the analysis. Such bias may be expected, for example, since higher numerical and semi-quantitative values achieved with the alternative assays (both of which tended to achieve higher upper limits than the 2BU assay) would be penalised with higher bias and SD values in the analysis, due to their between-assay differences being calculated against 2BU values truncated at 600 µg/g. Based on the results however, there was little difference between the base case and sensitivity analyses. This indicates that overestimating between-assay differences within FC values at the higher end of the measurement range had little impact on the simulation results. It should be noted, however, that the median estimates applied within this sensitivity analysis are highly uncertain, since they were based on pooling data across different assays and specimens. Ideally one would need to derive such estimates based on assay- and population-specific data in order to ascertain reliable results.

7.5.2 Bias correction exercise

The results of the outcome assessment discussed above illustrate that switching to an alternative FC assay could have a negative impact on both the diagnostic accuracy and cost-effectiveness of the YFCCP. A bias correction exercise was undertaken to explore if this drop in performance could be offset via a simple

measurement correction. Two exercises were conducted: in the first, a fixed absolute correction value was applied to all measured FC values within the simulation (ranging from -50 to +100 µg/g); and in the second, a proportional correction factor was applied (ranging from 0.5 to 2.0).

For the 4BU assay, the results of the bias correction exercise indicate that the diagnostic accuracy of the YFCCP could be restored to a similar level as achieved when using the 2BU assay. The closest diagnostic accuracy was obtained by applying a fixed absolute correction value of 30 µg/g, or a proportional correction factor of 1.45 (i.e. 45% increase). Both of these correction values caused the YFCCP [4BU] sensitivity to increase from ~80% to ~91%, with a corresponding drop in specificity from ~93% to ~90%. Recall that for the 4BU assay, the bias profile exhibited a consistently negative bias up to a moderate region of the measurement range, with an expected bias of -8 µg/g at the lowest reference measurement point (24 µg/g), increasing to -22 µg/g bias at 100 µg/g, and peaking at -38 µg/g bias at a reference measurement of 380 µg/g. As such, whilst the negative bias was variable in magnitude, it was consistent enough to be adequately offset by a fixed correction value. Although high positive bias values were observed at the upper end of the measurement range, this appears to be of little importance in terms of affecting the pathway diagnostic accuracy.

For the 2KO2 assay, the results of the bias correction exercise indicate that the diagnostic accuracy of the YFCCP cannot be straightforwardly corrected in the same way as the 4BU assay. In this case, the optimal diagnostic yield was achieved when: (i) applying a positive absolute correction value of 20 µg/g, which increased the YFCCP [2KO2] sensitivity from 66% to 74% but also reduced the specificity from 83% to 77%; or (ii) applying a proportional correction value of 1.35, which increased sensitivity to 75% and reduced specificity to 75%. For the 2KO2 assay, the inability to restore diagnostic accuracy lies in the inconsistency of bias over the lower region of the measurement range i.e. the fact that positive bias and high SD occurred at the lower end of the measurement range (below 44 µg/g), followed by negative bias up to 359 µg/g. This inconsistency in direction and magnitude of error in regions close to the test cut-off threshold means that the between-assay differences can no longer be easily adjusted using a fixed absolute or proportional correction factor.

As outlined in section 7.5.1, the results for the 2KO2 assay appear to be driven by two extreme values at the lower end of the measurement range. As well as re-running the outcome assessment removing these two values, the bias correction exercise was also repeated for the post-hoc sensitivity analysis (see Appendix N.2). Recall that in the base case analysis, the YFCCP [2KO2] strategy achieved a sensitivity of 66% and a specificity of 83%, and no correction value was able to achieve sensitivity and specificity values both above 80%. Repeating this assessment in the post-hoc sensitivity analysis (i.e. removing the two extreme values), a maximum diagnostic yield was obtained when applying: (i) a fixed absolute correction value of 60 µg/g, which achieved 83% sensitivity and 87% specificity; or (ii) a proportional correction factor of 2.75, which achieved 85% sensitivity and 86% specificity. The shift to a more consistent bias profile and less volatile SD profile for 2KO2 in this sensitivity analysis therefore resulted in more favourable results within the correction exercise. Nevertheless, in contrast to the 4BU assay, the performance levels for 2KO2 in this case remained sub-optimal compared to that achieved with the 2BU assay (94% sensitivity, 92% specificity). This is due to the fact that, compared to the 4BU assay, the 2KO2 maintained higher magnitudes of negative bias and SD[58] across the majority of the lower measurement region, even after removal of the two suspected outliers.

The bias correction exercise results presented herein indicate that, where between-assay differences have been confirmed to impact on clinical and/or health-economic outcomes, an effective approach to restoring expected outcomes may be to apply a simple measurement adjustment – at least in cases where consistent bias is the key driver of between-assay differences (particularly around the cut-off threshold). An alternative but equivalent approach, in this scenario, is to apply assay-specific diagnostic cut-off thresholds. Indeed, several authors have previously suggested such an approach for FC assays, given documented evidence on between-assay biases in this context (see section 4.3.4.2). In this case study, for example, for the 4BU assay the correction value

[58] For example, at a reference measurement of 100 µg/g, the 4BU assay was associated with an expected bias of -22 µg/g (SD = 16 µg/g), whilst the 2KO2 assay (excluding the two extreme values) was associated with an expected bias of -47 µg/g (SD = 34 µg/g).

of +30 µg/g (which restored diagnostic accuracy) implies an assay-specific cut-off threshold of 70 µg/g (i.e. 100 - 30). This approach is similar to using health economic analysis to optimise diagnostic thresholds; but in this case patient test measurements available for one assay have been combined with data on between-assay measurement differences, to enable threshold optimisation for the additional assays (with available comparative measurement performance data). This approach may be useful, therefore, in scenarios where patient test measurements are not available for all assays, but comparative measurement performance data is available. The question remains as to what could be done in practice to reconcile complex between-assay differences, such as that observed with the 2KO2 assay. If we accept the base case results for the 2KO2 assay, then simply altering the cut-off threshold in this case does not meaningfully improve the YFCCP diagnostic accuracy or cost-effectiveness. It may be that more sophisticated methods of measurement adjustment could be applied to counteract inconsistent and/or higher magnitudes of between-assay bias; however offsetting the impact of high variability in bias (i.e. SD) is likely to be challenging – if not impossible. Where the volatility of an assay's performance has been confirmed, it would clearly be preferable, where possible, to suspend use of such an assay in favour of a higher performing, more consistent procedure.

7.5.3 Limitations

The analysis conducted in this chapter applied a modified version of the error model to the YFCCP dataset, using the bootstrap sampling method. Since this analysis draws on the YFCCP dataset, the same limitations as originally discussed in Chapter 5 are applicable here (see section 5.5.3).

Additional limitations in this case relate to the RWE used to inform the assessment – i.e. assay measurement performance data extracted from 11 NEQAS EQA reports for FC. This is high-order measurement data collected under reproducibility conditions – as such, this data captures variability resulting from assay-specific factors, within-laboratory factors, and between-laboratory factors. An advantage of using this data, therefore, is that the results should reflect how different assays would be expected to perform in practice across the NHS (i.e. incorporating all of the factors of measurement uncertainty listed above). There are two key limitations with this data however. The first relates to

what extent the EQA samples may be considered to be representative of patient samples. *Commutability* of EQA samples – that is, that they should behave in the same way as patient samples do in routine practice – is a desired requirement for any EQA scheme, and current EQA guidelines stipulate that a clear statement of commutability (whether or not this is achieved, and how) should be provided alongside EQA reports (286). No such statements, however, are currently included in the NEQAS reports for FC (see Appendix M). The only sample information provided in the EQA reports used in this analysis was that the FC EQA samples consisted of 'mixed patient samples'. It is unclear if these samples could be reasonably expected to exhibit commutability – as such, it is not known if the assay differences observed when running these samples in the EQA scheme will accurately reflect assay differences that would occur when analysing individual patients' samples (i.e. non-mixed samples). The second limitation concerns the fact that based on this data, one cannot identify the root causes of observed assay differences. For example, it may be that bias observed between the 4BU and 2BU assays was actually driven by variation in pre-analytical or analytical factors, such as the sample extraction method. Without additional data on the specific pre-analytical and analytical processes applied at each testing laboratory, no definitive conclusions can be drawn regarding the root causes of observed measurement differences.

A further limitation with the EQA data in this case relates to the fact that a range of lower and upper measurement limits were used across assays and laboratories. This introduces censored data into the analysis, which may have biased the study findings. A sensitivity analysis was conducted to explore the potential impact of this censoring, which indicated that the censored data were not expected to have biased the study results in this case. Nevertheless, this sensitivity analysis was based on applying quantitative estimates for censored data calculated by pooling values from across different assays and specimens. Future prospective studies in this area could explore more robust methods for dealing with censored data.

A final limitation in this analysis concerns the fact that only one source of RWE – national EQA data – was explored. It is expected that this type of data would not be routinely available for HTA analyses (although it should be noted that the EQA

scheme for FC was already underway when NICE originally assessed the test back in 2013). Future studies could therefore explore the utility of other sources of RWE, such as published laboratory studies, manufacturer data, laboratory databases and laboratory surveys. It may be that measurement uncertainty profiles could also be derived from these alternative RWE sources, enabling a similar analysis to be conducted as presented in this chapter. In addition, the ability of RWE to address other research questions also warrants further investigation.

7.6 Summary

- This chapter has illustrated how RWE in the form of EQA measurement performance data may be used to evaluate the impact of between-assay differences on clinical and health-economic outcomes. These results support hypothesis E of this thesis: that methods from the broader literature may be applied or adapted to allow RWE (relating to test measurement performance data) to be utilised within outcome-based assessments.
- Using a modified version of the error model, data on between-assay bias and variability was captured by drawing on bias and SD profiles derived from the EQA data. The resulting diagnostic accuracy estimates were extended to cost-effectiveness outcomes using the FC cost-utility model.
- The results showed that between-assay differences can negatively impact on downstream outcomes. In cases where between-assay differences are driven by a consistent bias component, a pathway's performance may be effectively restored by shifting the assay cut-off threshold.

The final chapter of this thesis, **Chapter 8**, provides a summary and discussion of the thesis, outlining key findings and limitations of the research as well as possible areas for future research.

Chapter 8
Discussion

8.1 Chapter outline

The aim of this thesis was to develop a framework for assessing the impact of test measurement uncertainty on clinical and health-economic outcomes. To this end, two reviews were conducted to evaluate the methodological landscape in this area: the first review focused on methods applied specifically within HTAs (Chapter 2), while the second review aimed to capture methods used in the wider literature (Chapter 3). A subsequent case study was conducted to develop key methods identified from the reviews (Chapter 4 to Chapter 7), focusing on the error model simulation approach and decision analytic modelling.

This final chapter of this thesis provides a discussion of the research findings. The main findings from each chapter are first summarised (section 8.2 below), followed by a discussion of the implications of the findings and recommendations for future research (section 8.3). The thesis then closes with a final section summarising the key messages of the research (section 8.4).

8.2 Research findings

Chapter 1 of this thesis highlighted an important inconsistency in the research pathway for tests: that is, that although test measurement uncertainty is a key consideration within the test evaluation pathway, it is rarely assessed within downstream test evaluations – such as HTAs – which direct test adoption decisions. In particular, the argument was made that the failure to quantify the impact of test measurement uncertainty on outcomes within the evaluation process may lead to inefficient reimbursement and funding allocation decisions.

The aim of the systematic review presented in **Chapter 2** was to identify if and how test measurement uncertainty has been assessed within HTAs to date (1). The findings of this review verified the introductory hypothesis that measurement uncertainty has not, thus far, been routinely assessed within this context. In the minority of identified HTAs that did include an assessment of measurement uncertainty (19%; 20/107), most consisted of a narrative review of the measurement literature in which the potential influence of measurement

uncertainty on outcomes was not considered. Similarly, of five identified studies which included measurement uncertainty within the HTA economic model, most simply incorporated a baseline level of measurement uncertainty. Only one study – a model-based assessment – attempted to formally quantify the impact of measurement uncertainty on outcomes (in this case, evaluating cost-effectiveness). Overall therefore, this review confirmed that there has been little consideration of measurement uncertainty – in particular the impact of measurement uncertainty on outcomes – within the HTA context.

Based on the limited applications identified in Chapter 2, a methodology review was conducted in **Chapter 3** (2). This review aimed to identify studies from the wider literature which had used an indirect method to assess the impact of measurement uncertainty on clinical or economic outcomes. Based on 82 identified studies, a three-step analytical framework underpinning the various methods identified was apparent: (1) calculation of "true" test values; (2) calculation of measured test values (i.e. incorporating measurement uncertainty); and (3) calculation of the impact of differences between (1) and (2) on the evaluated outcome(s). Within this framework, the error model simulation approach was indicated as an efficient method for exploring the impact of measurement uncertainty on diagnostic accuracy. In addition, decision analytic modelling was highlighted as a flexible tool for linking diagnostic accuracy outputs to downstream clinical and health-economic outcomes; and contour plots were identified as a useful visual aid for presenting and analysing simulation results. Based on these factors, and the ability of these methods to be straightforwardly integrated into existing HTA methodology (e.g. model-based economic evaluation), these methods were selected for further investigation.

In order to explore and develop the methods highlighted above, a case study assessment was undertaken (Chapter 4 to Chapter 7). **Chapter 4** provided a general introduction to the case study. The role of the case study test, FC, as a diagnostic tool for IBD was first outlined, and two primary care pathways – the NICE FC pathway and the YFCCP – were introduced. Motivating factors for choosing FC as the case study test were also highlighted, including: known concerns with the test's measurement performance; a lack of defined APS in this area; and the availability of clinical, economic and EQA data upon which to base

the case study analysis. In the following analysis chapters, the impact of measurement uncertainty on the diagnostic accuracy, clinical utility and cost-effectiveness of the two FC pathways was evaluated using the previously highlighted methods.

Chapter 5 presented the first part of the case study analysis. Using the error model simulation approach identified in Chapter 3, the impact of increasing FC measurement uncertainty on the diagnostic accuracy of the NICE FC pathway, and the YFCCP, was assessed. The simulation results were presented using contour plots, which provided a useful visual aid to examine the robustness of each pathway's diagnostic accuracy to bias and imprecision. These plots were further utilised to illustrate a novel concept of *acceptable regions* of bias and imprecision. The potential for acceptable regions to inform outcome-based APS was discussed, and key challenges were identified: in particular, baseline measurement uncertainty within "true" test values complicates the interpretation of acceptable regions; and the need to set a minimum diagnostic accuracy requirement relies on subjective judgement. Further limitations, relating to the data underpinning the analysis, were also discussed.

In **Chapter 6**, diagnostic accuracy results from Chapter 5 were embedded into an existing economic model to extend the evaluation to clinical utility (QALY) and cost-effectiveness (NMB and INMB) outcomes. Acceptable performance was here alternatively defined as the *cost-effective region* – i.e. the region of the INMB contour plot maintaining INMB > £0. Whilst this approach avoids subjective judgement, cost-effective regions were found to set an inappropriately low benchmark for analytical performance in certain circumstances. The notion of *optimal regions* was therefore introduced, based on selecting a specified top percentile of INMB (or NMB) results. This method was found to be useful for setting hierarchical APS (by specifying a series of percentile values), and for quantifying the added benefit of imposing tighter APS (by comparing the INMB value of different optimal region boundaries).

In the final analysis presented in **Chapter 7**, the framework presented in Chapter 5 and Chapter 6 was extended to incorporate RWE data. In this case, EQA data was used to evaluate the impact of FC between-assay differences on clinical and health-economic outcomes for the YFCCP. Using a modified version of the error

model (drawing on bias and SD profiles), this analysis illustrated how information on the variability of measurement uncertainty over a test's measurement range may be captured within the simulation. The results showed that assay differences can negatively impact on outcomes, but that this impact may be effectively offset by shifting the test cut-off threshold – at least in cases where between-assay differences are driven by consistent bias component. This approach is similar to using health economic analysis to optimise diagnostic thresholds; but in this case patient test measurements available for one assay have been combined with data on between-assay measurement differences, to enable threshold optimisation for the additional assays (with available comparative measurement performance data). This approach may be useful, therefore, in scenarios where patient test measurements are not available for all assays, but comparative measurement performance data is available.

8.3 Implications of findings and future research recommendations

This thesis has outlined a framework for assessing the impact of measurement uncertainty on outcomes. Whilst this framework is of relevance and interest to the laboratory community (given the focus on outcome-based APS), the case study centred on an HTA-style evaluation in which an extended HTA perspective was adopted. This perspective includes: (i) the traditional HTA remit – focused on the assessment of clinical performance (diagnostic accuracy), utility and cost-effectiveness; and (ii) evaluation of measurement uncertainty (historically confined to laboratory studies) – including assessment of the robustness of test outcomes to measurement uncertainty and the derivation of outcome-based APS. The aim of the extended HTA perspective is to bridge the gap between HTA and laboratory fields (as evidenced in Chapter 2), by establishing test measurement as a core component of HTAs.

In this section, implications of the research findings and recommendations for future research are presented. Section 8.3.1 discusses implications relating to the HTA setting, in particular outlining recommendations for future HTA methods guidance on this topic. Section 8.3.2 then focuses on the topic of outcome-based

APS (which has implications for both HTA and laboratory contexts), highlighting key implementation barriers for consideration in future research.

8.3.1 HTA methods guidance

Appropriate consideration of measurement uncertainty within future HTAs requires HTA authorities to formally recognize this component of the test evaluation pathway, and to provide specific guidance on this topic. In particular, the development of HTA guidance in this area requires two key questions to be addressed: (1) *when* should measurement uncertainty be formally assessed within HTAs; and (2) *how* should HTA assessments of measurement uncertainty be conducted? These questions are considered below.

8.3.1.1 When should measurement uncertainty be formally assessed within HTAs?

This thesis has highlighted the crucial dependence between precise and true measurement on the one hand, and clinical and health-economic outcomes on the other. This relationship means that failing to appropriately assess measurement uncertainty, risks failing to appropriately assess outcomes. For the FC case study presented in this thesis, for example, inclusion of measurement uncertainty in the original HTA assessment for the NICE FC pathway would have highlighted the volatility of this pathway to positive bias, and could have thus triggered a recommendation for raising the test cut-off threshold, further research and/or tighter analytical monitoring procedures. For the YFCCP meanwhile, inclusion of measurement uncertainty in the FC cost-utility model would not have been expected to alter the test adoption decision, but rather would have provided further support for the adoption of the YFCCP over the NICE FC pathway. It should be noted that the limitations of this case study analysis (in particular the short and deterministic nature of the decision model) preclude the possibility of drawing definitive clinical conclusions with regards to the impact of measurement uncertainty comparative to other aspects of sampling uncertainty. Nevertheless, these findings clearly highlight that the impact of measurement uncertainty on outcomes needs to be assessed on a on a case-by-case basis – that is, unless a formal evaluation is undertaken, the impact of measurement uncertainty on outcomes cannot be known, and there can be no guarantee that a different

clinical decision wouldn't be made were this additional information available. For this reason, it is herein recommended that assessment of measurement uncertainty should be considered a *best-practice* requirement for *all* HTAs in which a test or measurement device is evaluated. The remainder of this section therefore focuses on when specific *types* of assessment should be undertaken, and key factors influencing when these analyses should be considered a priority of the HTA.

As outlined in Chapter 2, formal assessment of measurement uncertainty in HTAs may include *pre-model assessments* (e.g. systematic review, laboratory survey etc.) and/or *model-based assessments* (1). The primary utility of pre-model assessments lies in quantifying the level of measurement uncertainty associated with a given test, and identifying important pre-analytical and analytical factors expected to influence measurement uncertainty. These assessments may also include a broader analysis of measurement performance – for example considering aspects such as detection limits and selectivity (Appendix B.3). Model-based assessments, meanwhile, provide valuable information on the impact of measurement uncertainty on outcomes: this enables assessment of the robustness of outcomes to measurement uncertainty; estimation of real-world performance levels (taking into account expected increases in measurement uncertainty); and derivation of outcome-based APS.

Whilst pre-model assessments have proved more common in HTAs to date (1), formal assessment of measurement uncertainty in this context should ideally include *both* a pre-model assessment (to review the measurement evidence base) and a model-based assessment (to evaluate the impact of measurement uncertainty on outcomes). Consider, for example, a "best case" scenario, in which a test is found to have minimal and well-controlled measurement uncertainty on the basis of a pre-model assessment (e.g. from the laboratory professional's perspective). It does *not* follow that measurement uncertainty should not be a concern within the subsequent clinical and economic evaluation: depending on the distribution of patient test values around key decision thresholds, and the knock-on impact of test results within the clinical pathway, small deviations in measurement may absolutely have a significant impact on outcomes. Crucially, without a formal assessment, it is extremely difficult to predict this impact. In

addition, excluding the model-based component of the assessment further ignores the potential utility of this analysis to support the implementation and monitoring of tests in the post-adoption phase (i.e. via the identification of outcome-based APS). As such, it is recommended that formal assessment of measurement uncertainty within HTAs should endeavour to include both the pre-model and model-based components.

There are several factors relating to test evaluations which may warrant the formal assessment of measurement uncertainty to be considered a particular priority of the HTA. Three such factors are highlighted below.

(i) The role of the test in the clinical pathway

The role of a test in the clinical pathway dictates what knock-on effects the test is expected to have with respect to patient health outcomes and resource utilisation. As such, consideration of a test's role is crucial to understanding when measurement uncertainty might be expected to have a serious impact on outcomes. If, for example, a slight change in the rate of false negative cases for a diagnostic test is associated with a significant risk of patient harm, then formal assessment of the impact of measurement uncertainty on outcomes (in addition to a pre-model review of the measurement evidence base) will be of particular importance. Equally, when changes to a test's diagnostic accuracy are expected to result in a significant change in costs, then an impact assessment will similarly be of greater importance.

(ii) ***The need for outcome-based APS***

The introduction to this thesis highlighted the fact that, based on current EFLM guidelines, most tests evaluated within the HTA context are expected to fall under Model 1 of the Milan criteria – i.e. requiring APS to be set based on an assessment of the impact of analytical performance on outcomes (30, 33). Given that the analysis of outcomes is the primary directive of HTAs, these studies are perfectly placed to help inform outcome-based APS and could thereby help to improve system quality, efficiency, and ultimately, patient safety. Formal assessment of measurement uncertainty – in particular using model-based approaches as illustrated in this thesis case study – can therefore provide added utility in this respect. Furthermore, there are particular cases where the need for

outcome-based APS may be heightened. For example, when outcomes are expected to be highly sensitive to slight changes in measurement uncertainty (see point (i) above), or when measurement uncertainty is expected to be variable during the post-adoption phase, then identification of outcome-based APS will be of particular importance. If there is potential for the analysis to inform the establishment and/or design of a test EQA scheme, then the assessment will provide additional utility. In these cases, formal assessment of measurement uncertainty within the HTA should be considered a priority.

(iii) Multiple assay assessments

If several assays (for the same measurand) are under assessment, then the evaluation of between-assay differences will be crucial. A pre-model assessment (e.g. literature review) should be conducted in these cases, to assess the equivalence of assays in terms of measurement (ideally via head-to-head comparison studies). If assays are found to be comparable, then equivalent clinical performance (e.g. diagnostic accuracy) may be assumed within the economic model; if not, then the impact of between-assay differences on outcomes should be determined. In particular, in scenarios where clinical performance data is not available for all of the considered assays, but where between-assay measurement performance data is available, then a similar approach to that undertaken in Chapter 7 could be explored to model the impact of between-assay discrepancies on outcomes.

A final point to note here is that, for formal assessments of measurement uncertainty to be meaningfully undertaken in the HTA context, then greater interaction with laboratory professionals (and other relevant testing experts) is required, throughout the HTA process. For example, laboratory professionals can help provide a steer on which elements should be addressed in the pre-model assessment; relevant search terms to include in literature review strategies; primary evidence sources which may be available; and the need for outcome-based APS. As well as informing the HTA research process, laboratory professionals should be established as a key stakeholder at the adoption-decision point, to ensure that any evaluation of measurement uncertainty is appropriately considered at this stage. Only by instigating relevant measurement

experts within the HTA pipeline will appropriate research and test-adoption decisions be achieved.

8.3.1.2 How should assessments of measurement uncertainty be conducted within HTAs?

As previously highlighted, formal assessment of measurement uncertainty within HTAs may consist of: (1) *pre-model assessments* (to review the measurement evidence base); and/or (2) *model-based assessments* (to evaluate the impact of measurement uncertainty on outcomes). This section outlines recommendations for the conduct of each type of analysis – in particular highlighting outstanding issues for consideration in future research.

8.3.1.2.1 Assessing the evidence base for measurement uncertainty

The HTA systematic review reported in Chapter 2 identified several methods for reviewing the measurement evidence base, in the form of pre-model assessments. Across the identified studies, literature reviewing was found to be the most common approach with respect to pre-model assessments. Based on evaluation of the identified studies, however, several key aspects of the reported reviews were found to be lacking in methodological rigor. These included:

(i) the design of search strategies including a measurement outcome filter;
(ii) statistical methods for the quantitative synthesis of measurement uncertainty data; and
(iii) the selection of tool(s) for the quality assessment of measurement literature.

Whilst each of these aspects was highlighted as a potential issue based on the HTA review findings, it may be that relevant literature and/or guidance on these topics is available in the broader literature (i.e. non-HTA studies). Several tools relevant to point (iii) are available, for example, but have not yet been applied within the HTA context (113-115). The development of HTA-specific guidance on each of the above issues, therefore, may simply require a methodology review of the broader literature and/or consultation with relevant methodology experts. If a paucity of relevant guidance were to be identified in the wider literature, then further research on these topics may be required before formal guidance can be issued.

Other forms of pre-model assessment may also be useful for assessing the measurement evidence base. In situations where there is a paucity of published measurement data, for example, laboratory surveys and/or databases may provide useful information on a test's measurement performance. Future qualitative research, such as consultation with laboratory professionals and test manufacturers, could provide further insight into when and how these alternative forms of pre-model assessments are likely to provide meaningful information.

It should be noted that, in all of the HTA pre-model assessments identified in Chapter 2, data on test measurement performance was summarised independently from clinical performance data. An additional issue of relevance, however, concerns the assessment of measurement uncertainty *within* clinical studies used to estimate clinical performance. In the diagnostic context for example, checklists such as STARD and QUADAS-2 aim to assess the quality of diagnostic accuracy studies; however, neither of these tools address potential issues associated with measurement procedures applied within the clinical studies, which could bias or invalidate the clinical performance findings (107, 287). A tool built for this specific purpose could help to inform when formal assessment of the impact of measurement uncertainty on outcomes is required, due to a lack of applicability of the clinical study findings to real-world testing scenarios.

In a recent HTA (Hall *et al.*, 2018; published after the thesis HTA systematic review was conducted), a checklist for assessing the quality of measurement procedures applied within clinical studies was developed, called the 'Quality Assessment of Measurement Procedures (QAMPs) framework' (288). In this study, 'quality' of measurement procedures was defined according to three features: *bias* (i.e. bias in the clinical performance findings resulting from measurement issues), *reproducibility* (i.e. reproducibility of the study with respect to measurement procedures) and *applicability* (i.e. the applicability of the clinical performance findings to real world practice, with respect to measurement procedures). The framework presents an initial list of measurement-related parameters (including pre-analytical and analytical factors and a range of measurement performance metrics), followed by four signalling questions (below) intended to help reviewers determine whether the risk of bias, irreproducibility

and inapplicability within an individual study should each be considered as 'low', 'high' or 'uncertain':

1) Were measurement procedures different between groups?
2) Were measurement procedures described in enough detail to be repeated?
3) Were measurement factors appropriately controlled for?
4) Were measurement procedures applicable to the final clinical setting?

To explore the potential utility of the QAMPs framework, the authors applied it to four studies reporting on the diagnostic accuracy of Nephrocheck® – a test to identify patients with acute kidney injury (AKI) in the critical care setting. These four studies were included within a meta-analysis of the Nephrocheck® test, conducted as part of the broader HTA study (in which several alternative tests for AKI were evaluated). Interestingly, all of the studies in this pilot were classified as having either 'high' or 'unknown' risk across the three quality features. In particular the authors reported that:

> *"Application of this framework within the four Nephrocheck case studies identified several measurement parameters that present a high risk of irreproducibility, including a failure to exclude samples with known interferents, a lack of internal and external quality control and a complete lack of analytical measurement verification in all studies. It also highlighted several issues that might affect the clinical applicability of test results, including freeze–thawing of samples in the absence of validation data and against the recommendations of the manufacturer, potentially biasing clinical cut-off points and overestimating precision; use of a device in an unvalidated patient population (i.e. aged < 18 years); and reporting the median value of three measurements from different laboratories. Furthermore, it identified several issues that made assessment of the risk of bias uncertain." (288)*

Based on their pilot analysis, the authors concluded that there were likely to be serious issues with the validity and applicability of the Nephrocheck® study findings. This study therefore illustrates how issues with measurement procedures applied within clinical studies can significantly affect the legitimacy of clinical performance estimates – a topic further highlighted in additional recent studies (289, 290). The availability of a quality assessment tool such as the QAMPs framework would therefore be useful for future HTAs, as a means of

determining the validity of clinical performance results, and identifying when further assessment of the impact of measurement on outcomes may be of particular importance within the HTA assessment (e.g. when the applicability of clinical study findings is found to be low). As of yet, however, the QAMPs framework has not undergone any form of comprehensive validation procedure, and the authors highlighted the need for input from the wider IVD community before adopting this tool in future research (288). There is clearly a need, therefore, for future studies to build on this work.

8.3.1.2.2 Assessing the impact of measurement uncertainty on outcomes

Error model simulation and the "two-step" linked-evidence approach

In Chapter 3, various indirect methods for assessing the impact of measurement uncertainty on outcomes were identified, and a common three-step analytical framework was presented (see Figure 3-3). Within this framework, error model simulation and decision analytical modelling were identified as particularly useful methods within the context of HTA analysis and the derivation of outcome-based APS. These methods were therefore explored further in the thesis case study (Chapter 4 to Chapter 7).

The error model simulation approach essentially provides a mechanism for linking test measurements to clinical performance outcomes. Decision modelling meanwhile – in particular using the linked-evidence approach – provides a means of linking clinical performance inputs to downstream clinical utility and cost-effectiveness outcomes. Combination of these two methods therefore enables test measurements to be linked to end-stage outcomes. This overall process may be described as a "two-step" linked-evidence approach: the first link establishes the relationship between test measurements and clinically accuracy (via error model simulation), and the second link establishes the impact of test classifications on downstream outcomes (via decision modelling). Whilst in the case study presented in this thesis was limited to deterministic model outputs, this same general mechanism could be applied to probabilistic models also. Most importantly, in contrast to standard decision modelling techniques, this approach appropriately reflects the true course of testing strategies: that is, starting with test measurement as the first, crucial element of any test-directed pathway.

Within HTAs, decision analytic modelling is a well-established tool for assessing cost-effectiveness, and linked-evidence models are commonplace. The novel aspect in this context, therefore, is the error model simulation component, and the embedding of this analysis within decision analytic models. At its core, error model simulation depends on the ability to represent baseline test measurements using either empirical data or parametric distributions. In the diagnostic context, sampled measurements also need to be linked in some way to true clinical diagnoses. Future studies wanting to adopt the two-step linked-evidence approach must therefore first ascertain if and how the necessary data will be obtained. In most cases, this will ultimately depend on the approach taken to evaluating clinical performance within the HTA, as outlined below.

Consider the case of diagnostic test assessments. Depending on data availability, there are four general approaches taken to the estimation of diagnostic accuracy within HTAs, with estimates being based either on: (i) clinical trial/study data, (ii) a single published paper, (iii) an IPD meta-analysis, or (iv) standard (aggregate-level) meta-analysis. The applicability of the error model simulation approach within each of these scenarios is discussed below. Ultimately, future studies intending to apply the error model approach should consider upfront the data requirements for such an assessment, to ensure feasibility of the analysis alongside the planned clinical performance assessment.

(i) Clinical trial/study data

When diagnostic accuracy is based on clinical trial or study data (i.e. an IPD dataset), the error model simulation approach may be applied using the same methods as demonstrated in the thesis case study. This includes both the parametric and bootstrap sampling methods, as well as direct simulation based on the raw empirical dataset, as presented in Chapter 5.

(ii) ***A single published paper***

When diagnostic accuracy is based on findings from a single study, error model simulation may be undertaken if: (i) the underlying IPD can be obtained from the study authors, or (ii) the test distributional parameters (i.e. for the diseased and non-diseased populations) have been reported, or can be obtained from the study authors. In the latter case, baseline “true” test values within the error model

simulation may be sampled from the assigned parametric distribution(s). Note, however, that the validity of this approach depends essentially on the validity of the assumed parametric distribution(s) – ideally therefore, parametric distributions should be used only if appropriate justification for the chosen parameterisations can be obtained.

(iii) Meta-analysis of IPD data

When diagnostic accuracy is based on a meta-analysis of IPD data, the error model can be applied either to the individual study IPD datasets, or using parametric distributions derived from the evidence synthesis. For example, the meta-analysis method outlined by Steinhauser and colleagues (2016) is based on the concept of estimating the underlying parametric distribution functions of the test for the diseased and non-diseased populations, across the IPD datasets (291). In theory, the error model parametric sampling method could be applied in this case, by using the pooled distributional parameters estimated within the meta-analysis to define parametric distributions for the "true" diseased and non-diseased populations within the simulation. Future studies could explore the feasibility of this approach.

(iv) Meta-analysis of aggregate-level data

When diagnostic accuracy is based on a meta-analysis of aggregate-level data, the error model approach cannot be applied, unless there is some mechanism by which the underlying 'pooled' distributions of non-diseased and diseased populations may be estimated. For example, if individual studies report distributional parameters (or this data can be obtained from the study authors), then it may be possible to synthesise this data. In this case, the results of such an analysis would need to be calibrated against the results of the diagnostic accuracy meta-analysis (i.e. to ensure the two sets of results are compatible). As for point (iii) above, future studies in this area are needed to explore the feasibility and validity of this approach.

In addition to the data requirements highlighted above, future studies intending to use the error model simulation approach should address the particular method considerations highlighted in Chapter 5. In particular, two key limitations were emphasised in the case study: first, the issue of baseline measurement

uncertainty was identified as a potential confounding factor (this topic is discussed further in section 8.3.2); and second, missing test data were found to pose a problem in the context of evaluating the YFCCP repeat-test strategy. Ideally, formal assessments of measurement uncertainty should be planned prior to the collection of measurement data, to ensure that missing data issues can be avoided. For example in the thesis case study, missing data could have been avoided by making sure repeated tests were conducted in all patients, or a relevant random subset of patients.

A final point to note here is that, since this thesis focused on the diagnostic setting, further studies are needed to explore the application of these methods to alternative testing scenarios – e.g. monitoring, predictive and prognostic testing pathways. In the monitoring context, for example, it was previously highlighted in Chapter 5 that rather than basing repeated test values on empirical sample data (as in the thesis case study), these values could instead be simulated using data on individual patients' baseline "true" test values, their trajectory of disease (e.g. an annual rate of progression), and biological variation (see section 5.4.1.2). Future studies could explore whether other amendments and/or variations of the presented methods are possible (or required) for each of the alternative testing scenarios.

A note on alternative approaches

This thesis has focused on application of the error model simulation approach within a linked-evidence modelling framework. It is possible that alternative – if not entirely different – approaches, could also be used for the same purposes. For example: risk categorisations produced from error grid analyses (as described in Chapter 3, section 3.3.3) could be applied within the decision modelling framework, in a similar way as illustrated in this thesis using clinical performance estimates; or alternatively, the regression-based approach (also described in section 3.3.3) could be applied to trial-based cost-effectiveness analyses, to explore the impact of hypothetical measurement error on regression-based cost and utility estimates. Future studies could explore the potential utility of these approaches, and others, for evaluating the impact of measurement uncertainty on outcomes.

Overall, this section has highlighted the need for formal assessment of measurement uncertainty within future HTAs. Where possible this assessment should include both pre-model and model-based analyses, to review the measurement evidence base and quantify the impact of measurement uncertainty on outcomes. It is acknowledged that the requirement for formal assessment of measurement uncertainty places an added burden on an already under-resourced system (292). Nevertheless, this is a necessary demand if HTAs are to retain their position as the gold standard method for technology evaluation, and to ensure that estimated benefits to patient health are realised and maintained in real-world clinical practice.

8.3.2 Outcome-based APS

This thesis has highlighted the potential for HTAs to play a greater role in informing laboratory and testing practices in the post-adoption phase – most notably via the derivation of outcome-based APS. Three novel classifications of outcome-based APS were presented within the case study analysis: (i) *acceptable regions* of bias and imprecision, based on an assumed minimum requirement for diagnostic accuracy (Chapter 5); (ii) *cost-effective regions* of bias and imprecision, based on INMB calculated against a chosen comparator strategy (Chapter 6); and (iii) *optimal regions* of bias and imprecision, based on selecting INMB (or NMB) results falling above a specified percentile value (Chapter 6). For each of these classifications, there are several important implementation issues to consider.

The first concern relates to the issue of baseline measurement uncertainty. As previously discussed in Chapter 5 (section 5.5.3), if “true” test measurements applied within the error model simulation are in fact subject to baseline measurement uncertainty, then the simulation results must be interpreted as indicative of the change in diagnostic accuracy resulting from *additional* bias and imprecision, on top of the baseline uncertainty. This means that in the case study analysis, bias and imprecision boundaries relating to each of the presented APS regions actually represent levels of bias and imprecision which can be tolerated *on top of* that contained within the YFCCP data itself. The existence of baseline measurement uncertainty therefore confounds direct interpretation of the error

model simulation results, and presents a potential barrier to the wider implementation of APS derived in this way.

Two possible means of dealing with this issue were previously discussed: (i) attempting to quantify the baseline measurement uncertainty; and (ii) “stripping” baseline measurement uncertainty via statistical adjustment of parametric distributions (see section 5.5.3). Both of these methods, however, require reliable information on the baseline level of measurement uncertainty, and the statistical adjustment approach is further limited to simulations based on parametric sampling. Clearly, the best remedy for this issue is to avoid it altogether: for example by ensuring that “true” test values applied within the error model simulation are based on a gold standard reference measurement procedure considered to be a reliably proxy for the truth. Whilst this was not a possibility for the FC case study (no reference measurement procedure is yet in place for FC), future studies wishing to derive APS should to aim utilise reference measurement procedures where possible. If no such procedure is available, then additional care may be required to ensure that the baseline measurement uncertainty is closely measured and monitored so that it can be meaningfully quantified. Better approaches for dealing with the issue of baseline measurement uncertainty, in particular strategies to enable meaningful APS to be derived in spite of baseline measurement uncertainty, should be explored in future research.

A second concern with each of the presented APS relates to the use of acceptable *regions* of performance, as opposed to fixed bias and imprecision goals. Representing outcome-based APS as regions reflects the fact that, in practice, random and systematic errors can occur concurrently, and each may have a very different effect on outcomes. This means that assessing the impact of bias and imprecision in isolation provides an incomplete picture of the impact of total measurement uncertainty on outcomes. Nevertheless, within the context of APS implementation, the use of two-dimensional regions presents a challenge, since the maximum allowable level of bias depends on the level of CV achieved (and vice versa), leaving laboratory professionals with no clear, fixed goal to target.

Two possible mechanisms for simplifying and distilling the information contained in the APS regions were presented in the case study analysis: first, fixed

specifications of bias and imprecision were presented based on extracting (a) the range of added bias allowed at zero added imprecision, and (b) the range of imprecision allowed at zero added bias; and second, a maximum combined value of imprecision and bias was presented in the form of the TE_{max} summary metric (see results Table 5-5 and Table 5-11). Whilst these approaches are pragmatically appealing, there are clearly inefficiencies with both methods resulting from a loss of information. In the case study analysis for example, setting bias at the maximum absolute value indicated in (a), and imprecision at the maximum value indicated in (b), pushes the analytical performance outside of the APS acceptable regions; whilst in contrast, higher levels of imprecision than indicated by the TE_{max} values are permissible when bias is restricted to a smaller region (e.g. as illustrated in Figure 5-12). Based on these results, the use of TE_{max} may be preferable (as a simple summary metric), since this approach ensures that the acceptable region of bias and imprecision is not breached[59]. Nevertheless, it would clearly be preferable to retain all of the information contained in the APS regions. Future research in this area, therefore, could further explore how outcome-based APS defined as regions could be better presented and/or summarised, in order to maximise the pragmatic usability of these concepts whilst minimising any loss of detail.

Improved implementation of outcome-based APS could be partly addressed by supplying the error-model simulation via a user-friendly web-based application – for example using the R shiny app platform (293). This type of application could allow users to input their own data and/or parametric distributions; specify their own error model function; access the raw simulation results underlying the contour plots; and explore alternative outcome requirements (e.g. adjusting the acceptable level of diagnostic accuracy assumed within the *acceptable region*). This could help to illustrate to wider audiences how the error model simulation works, and provide a greater understanding of the simulation results. Other types of simulation approaches as outlined in Chapter 3 (e.g. simulation around fixed points along the measurement range), could also be accommodated. Again,

[59] It should be noted however, that there is significant resistance to the use of TE metrics in parts of the clinical chemistry field (as discussed in section 1.2.3).

further consultation with relevant end-users of such an application could be conducted, to identify what features the system should include.

A third issue relating to outcome-based APS, concerns to the acceptability of the outcome assumptions applied in the analysis amongst laboratory, clinical and HTA communities. Acceptability of the presented *acceptable regions*, for example, would likely require some form of consultation with clinical experts in order to ensure that the assumed minimum requirement for diagnostic accuracy was appropriate and acceptable to the clinical audience. Acceptability of regions determined according to cost-effectiveness, meanwhile, may be limited in the laboratory setting due to an unfamiliarity with this concept amongst laboratory professionals – for this reason, clinical outcomes such as diagnostic accuracy or utility (i.e. QALYs) may be preferred in that setting, whilst cost-effectiveness may be more acceptable (if not preferable) in the HTA setting. Further consultation work would also be useful here, to determine which outcomes different stakeholders deem acceptable.

A key novel aspect of this research was the derivation of APS based on cost-effectiveness outcomes, in the form of *cost-effective regions* (as presented in Chapter 6). A concern highlighted with the cost-effective regions, however, was that, whilst they are useful in respect of avoiding any user-based subjective judgements, they may result in inappropriately low APS in certain cases. It should be noted that this issue is not only a concern for cost-effective regions – a clinically agreed level of acceptable diagnostic accuracy, for example, could also result in unacceptably wide *acceptable regions*, from the laboratory perspective. The key issue is that acceptable outcome performance may have very little correlation with acceptable analytical performance (i.e. considering currently achieved levels of analytical performance). If there is a wide discrepancy between these two perspectives, this should be considered. For example, the optimal regions presented in Chapter 6 present one possible solution for scenarios where the benchmark set by the cost-effective region is far below that achievable in practice. Alternatively, in situations where the cost-effective region is believed to be unachievable by current working standards, then the INMB analysis can be utilised to quantify the opportunity cost associated with failing to achieve this benchmark, and this information can be used to inform the test-adoption decision.

A final, more general consideration relating to outcome-based APS, concerns the fact that any performance specifications derived from an analysis of clinical utility and/or cost-effectiveness outcomes will necessarily be subject to the same considerations and limitations that apply to outcome analyses conducted in the HTA setting. In particular, this includes the need to consider the specific clinical pathway that an intervention sits in, and the intervention's role in that pathway: the impact of FC on patient health outcomes, for example, depends crucially on whether the test is used in the diagnostic context, or for some other purpose (e.g. in the IBD monitoring context). Additional considerations apply in analyses of cost-effectiveness: for example, cost-effectiveness depends essentially on the comparative strategy selected, as well as the wider health care infrastructure and payment mechanisms in place. Each of these factors means that different APS may be derived for the same test used in different contexts, and APS results may not be applicable beyond the centre, region or county in which they were derived. Unfortunately, this increased complexity is a necessary consequence of any analysis in which clinical and/or health-economic outcomes are evaluated.

Overall, this section has outlined key concerns relating to the implementation of outcome-based APS. Whilst these factors represent potential barriers to straightforward implementation of outcome-based APS, it should be recognised that even with these limitations, outcome-based APS represent the only approach to setting APS that takes into account the impact of analytical performance on patients health and resource utilisation. If the goal of the health system is to improve patient outcomes, then outcome-based APS are required.

8.4 Summary

The following points summarise the key messages of this thesis:

- Test measurement uncertainty can and does have an impact on downstream clinical and health-economic outcomes.
- Methods to assess the impact of measurement uncertainty on outcomes used in the literature follow a three-step analytical framework: (i) calculation of the "true" test values; (ii) calculation of the measured test values (i.e. incorporating measurement uncertainty); and (iii) calculation of the impact of differences between (1) and (2) on the outcome(s) under consideration.
- Within this framework, the error model simulation approach provides a useful mechanism for assessing the impact of measurement uncertainty on diagnostic accuracy. Using error model simulation outputs, outcome-based APS can be derived based on setting a minimum diagnostic accuracy requirement.
- By embedding the error model within an economic decision model, the impact of measurement uncertainty on clinical utility and cost-effectiveness outcomes can be explored. Using model outputs, outcome-based APS may be derived based on analysis of NMB outcomes.
- Between-assay differences can negatively affect clinical and health-economic outcomes. Whilst assay-specific cut-off thresholds may alleviate the impact of consistent bias, variability in bias is less easy to counteract.
- Within HTAs, evaluation of the impact of measurement uncertainty on outcomes can help to inform appropriate test-adoption decisions. Further guidance from HTA authorities is required to ensure that meaningful assessment of measurement uncertainty is undertaken in future studies.
- Within the laboratory, outcome-based APS are vital for ensuring that expected clinical and health-economic benefits associated with testing strategies are obtained and maintained. Further consideration of the appropriate interpretation and implementation of outcome-based APS is required.

References

1. Smith AF, Messenger M, Hall P, Hulme C. The Role of Measurement Uncertainty in Health Technology Assessments (HTAs) of In Vitro Tests. PharmacoEconomics. 2018;36(7):823-35.

2. Smith AF, Shinkins B, Hall PS, Hulme CT, Messenger MP. Toward 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. 2019:clinchem. 2018.300954.

3. Smith A, Shinkins B, Hulme C, Hall P, Michael M. Beyond the laboratory: A review of indirect methods to assess the impact of test measurement uncertainty on downstream clinical and cost outcomes. Clinica Chimica Acta. 2019;493:S353.

4. Smith A, Shinkins B, Hulme C, Hall P, Messenger M. Methods to assess the impact of test measurement uncertainty on downstream clinical outcomes: A case study of faecal calprotectin (FC) for the diagnosis of Inflammatory Bowel Disease (IBD). Clinica Chimica Acta. 2019;493:S353-S4.

5. St John A, Horvath A, Cobbaert C, Jülicher P, Dahm M, Hopstaken R, et al. 57th Annual Scientific Conference.

6. Clinical and Laboratory Standards Institute (CLSI) Harmonized Terminology Database [Internet]. 2019 [cited January 2020]. Available from: http://htd.clsi.org/default.asp.

7. International Organization for Standardization. ISO 5725-3: 1994: Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 3: Intermediate Measures of the Precision of a Standard Measurement Method: International Organization for Standardization; 1994.

8. 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.

9. Linnet K, Boyd JC. Selection and and analytical evaluation of methods-with statistical techniques. Tietz textbook of clinical chemistry and molecular diagnostics 5 ed: Elsevier Sci. Intl. Congress Series 1100; 2012. p. 7-48.

10. International Organization for Standardization. In vitro diagnostic medical devices—measurement of quantities in biological samples—metrological traceability of values assigned to calibrators and control materials. EN ISO. 2003;2003.

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

12. 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.

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

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

15. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, et al. Guide to the Expression of Uncertainty in Measurement. First Edition1993.

16. ISO. 15189: 2012 Medical laboratories–Requirements for quality and competence. Geneva: International Standardisation Organisation. 2012.

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

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

19. Oosterhuis WP, Bayat H, Armbruster D, Coskun A, Freeman KP, Kallner A, et al. The use of error and uncertainty methods in the medical laboratory. Clinical Chemistry and Laboratory Medicine (CCLM). 2018;56(2):209-19.

20. Westgard JO. Error methods are more practical, but uncertainty methods may still be preferred. Clinical chemistry. 2018;64(4):636-8.

21. EUR-Lex. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (Text with EEA relevance.). 2017 [Available from: http://data.europa.eu/eli/reg/2017/746/oj.

22. Roelofsen-de Beer R, Wielders J, Boursier G, Vodnik T, Vanstapel F, Huisman W, et al. Validation and verification of examination procedures in medical laboratories: opinion of the EFLM Working Group Accreditation and ISO/CEN standards (WG-A/ISO) on dealing with ISO 15189: 2012 demands for method verification and validation. Clinical Chemistry and Laboratory Medicine (CCLM). 2020;58(3):361-7.

23. 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.

24. UKAS. Our Role [Available from: https://www.ukas.com/about/our-role/.

25. UKAS. The Route To Accreditation [Available from: https://www.ukas.com/the-route-to-accreditation/.

26. ACB, AACC. Laboratory Accreditation: the basis for confidence 2019 [Available from: https://labtestsonline.org.uk/articles/laboratory-accreditation.

27. Ranson P. Medical devices: recent developments and the potential impact of Brexit. European Pharmaceutical Review. 2019(6).

28. UKAS. Brexit Update 2020 [Available from: https://www.ukas.com/news/brexit-update-2/.

29. Eisenhart S. UK, European Device Industries Urge Continued Regulatory Alignment Post-Brexit. EMERGO2017 [Available from: https://www.emergogroup.com/blog.

30. 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(6):833-5.

31. Kenny D, Fraser C, Petersen PH, Kallner A. Consensus agreement. Scandinavian Journal of Clinical and Laboratory Investigation. 1999;59(7):585-.

32. 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(6):841-8.

33. 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(2):189-94.

34. Fraser CG, Petersen PH, Libeer J-C, Ricos C. Proposals for setting generally applicable quality goals solely based on biology. Annals of clinical biochemistry. 1997;34(1):8-12.

35. Fraser C. General strategies to set quality specifications for reliability performance characteristics. Scandinavian journal of clinical and laboratory investigation. 1999;59(7):487-90.

36. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Medical Decision Making. 2009;29(5):E13-E21.

37. Drain PK, Hyle EP, Noubary F, Freedberg KA, Wilson D, Bishai WR, et al. Diagnostic point-of-care tests in resource-limited settings. The Lancet infectious diseases. 2014;14(3):239-49.

38. 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.

39. Pitini E, D'Andrea E, De Vito C, Rosso A, Unim B, Marzuillo C, et al. A proposal of a new evaluation framework towards implementation of genetic tests. PloS one. 2019;14(8).

40. Lin JS, Thompson M, Goddard KA, Piper MA, Heneghan C, Whitlock EP. Evaluating genomic tests from bench to bedside: a practical framework. BMC medical informatics and decision making. 2012;12(1):117.

41. Pavlou MP, Diamandis EP, Blasutig IM. The long journey of cancer biomarkers from the bench to the clinic. Clinical chemistry. 2013;59(1):147-57.

42. Rousseau F, Lindsay C, Charland M, Labelle Y, Bergeron J, Blancquaert I, et al. Development and description of GETT: a genetic testing evidence tracking tool. Clinical chemistry and laboratory medicine. 2010;48(10):1397-407.

43. Thompson M, Weigl B, Fitzpatrick A, Ide N. More than just accuracy: a novel method to incorporate multiple test attributes in evaluating diagnostic tests including point of care tests. IEEE journal of translational engineering in health and medicine. 2016;4:1-8.

44. Trenti T. An evidence-based laboratory medicine approach to evaluate new laboratory tests. EJIFCC. 2018;29(4):259.

45. Leeflang M, Allerberger F. How to: evaluate a diagnostic test. Clinical Microbiology and Infection. 2019;25(1):54-9.

46. Huddy JR, Ni M, Misra S, Mavroveli S, Barlow J, Hanna GB. Development of the Point-of-Care Key Evidence Tool (POCKET): a checklist for multi-dimensional evidence generation in point-of-care tests. Clinical Chemistry and Laboratory Medicine (CCLM). 2019;57(6):845-55.

47. Walter FM, Thompson MJ, Wellwood I, Abel GA, Hamilton W, Johnson M, et al. Evaluating diagnostic strategies for early detection of cancer: the CanTest framework. BMC cancer. 2019;19(1):586.

48. Monaghan PJ, Robinson S, Rajdl D, Bossuyt PM, Sandberg S, St John A, et al. Practical guide for identifying unmet clinical needs for biomarkers. EJIFCC. 2018;29(2):129.

49. Petrisor B, Keating J, Schemitsch E. Grading the evidence: levels of evidence and grades of recommendation. Injury. 2006;37(4):321-7.

50. Bossuyt PM, McCaffery K. Additional patient outcomes and pathways in evaluations of testing. Medical Decision Making. 2009;29(5):E30-E8.

51. di Ruffano LF, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. Bmj. 2012;344:e686.

52. 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.

53. Verbakel JY, Turner PJ, Thompson MJ, Plüddemann A, Price CP, Shinkins B, et al. Common evidence gaps in point-of-care diagnostic test evaluation: a review of horizon scan reports. BMJ open. 2017;7(9):e015760.

54. 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.

55. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the economic evaluation of health care programmes: Oxford university press; 2015.

56. 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.

57. 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.

58. National Institute of Health and Clinical Excellence (NICE). Diagnostics Assessment Programme manual2011 07/08/2017. Available from: https://www.nice.org.uk/about/what-we-do/our-programmes/nice-guidance/nice-diagnostics-guidance.

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

60. Chang SM, Matchar DB, Smetana GW, Umscheid CA, Gray R, Torchia M. Methods guide for medical test reviews. Rockville: Agency for Healthcare Research and Quality. 2012.

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

62. International Network of Agencies for Health Technology Assessment (INAHTA). [Available from: http://www.inahta.org/.

63. National Institute for Health Research (NIHR). [Available from: https://www.nihr.ac.uk/.

64. National Institute of Health and Care Excellence (NICE). [Available from: www.nice.org.uk.

65. Cowles E, Marsden G, Cole A, Devlin N. A review of NICE methods and processes across health technology assessment programmes: why the differences and what is the impact? Applied health economics and health policy. 2017;15(4):469-77.

66. Government Digital Service (DGS). UK National Screening Committee [Available from: https://www.gov.uk/government/groups/uk-national-screening-committee-uk-nsc.

67. Canadian Agency for Drugs and Technologies in Health (CADTH). Guidelines for the economic evaluation of health technologies: Canada. 4th Edition 2018 [Available from: https://www.cadth.ca/about-cadth/how-we-do-it/methods-and-guidelines.

68. Government Digital Service (DGS). Evidence review criteria: national screening programmes 2015 [Available from: https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes.

69. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013 2013 [Available from: https://www.nice.org.uk/process/pmg9/.

70. National Institute for Health and Care Excellence. Medical technologies evaluation programme methods guide 2017 [Available from: https://www.nice.org.uk/process/pmg33/.

71. Institute for Clinical and Economic Review (ICER). 2020-2023 Value Assessment Framework 2020 [Available from: https://icer-review.org/methodology/icers-methods/icer-value-assessment-framework-2/.

72. Institute for Quality and Efficiency in Health Care (IQWiG). General Methods Version 5.0 2017 [Available from: https://www.iqwig.de/en/methods/methods-paper.3020.html.

73. Thiry N, Neyt M, Van De Sande S, Cleemput I. Belgian guidelines for economic evaluations. International journal of technology assessment in health care. 2014;30(6):601-7.

74. Health Information and Quality Authority. A Guide to Health Technology Assessment at HIQA 2016 [Available from: https://www.hiqa.ie/reports-and-publications/health-technology-assessment/guide-health-technology-assessment-hiqa.

75. Health Quality Ontario. Health Technology Assessments Methods and Process Guide Version 2.0 2018 [Available from: https://www.hqontario.ca/.

76. Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU). Assessment of methods in health care and social services: A handbook 2018 [Available from: https://www.sbu.se/en/method/.

77. Ellison S, Fearn T. Characterising the performance of qualitative analytical methods: Statistics and terminology. TrAC Trends in Analytical Chemistry. 2005;24(6):468-76.

78. López MI, Callao MP, Ruisánchez I. A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. Analytica chimica acta. 2015;891:62-72.

79. CRD. HTA Database 2017 [Available from: https://www.crd.york.ac.uk/CRDWeb/.

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

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

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

83. INAHTA. NIHR-HTA Database 2019 [Available from: http://www.inahta.org/hta-tools-resources/database/.

84. 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).

85. 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.

86. 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.

87. 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.

88. 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.

89. Medical Advisory Secretariat. Gene expression profiling for guiding adjuvant chemotherapy decisions in women with early breast cancer: an evidence-based and economic analysis. Ont Health Technol Assess Ser [Internet]; 2010.

90. 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).

91. 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).

92. 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).

93. 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).

94. 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).

95. 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).

96. 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.

97. 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.

98. 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.

99. 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. Report No.: Assessment Report 1216.

100. 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).

101. 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).

102. 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).

103. 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).

104. 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.

105. 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).

106. 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.

107. 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.

108. Teutsch SM, Bradley LA, Palomaki GE, Haddow JE, Piper M, Calonge N, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP Working Group. Genetics in Medicine. 2009;11(1):3-14.

109. National Health Medical Research Council (NHMRC). A guide to the development, implementation and evaluation of clinical practice guidelines: NHMRC Canberra; 1999.

110. INAHTA. A checklist for health technology assessment reports 2007 [Available from: http://www.inahta.org/wp-content/uploads/2014/04/INAHTA_HTA_Checklist_English.pdf.

111. Altman DG. Systematic reviews of evaluations of prognostic variables. Bmj. 2001;323(7306):224-8.

112. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of internal medicine. 2013;158(4):280-6.

113. 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.

114. Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JP, Kirsch-Volders M, et al. STrengthening the Reporting of OBservational studies in Epidemiology–Molecular Epidemiology (STROBE-ME): an extension of the STROBE Statement. Mutagenesis. 2011;27(1):17-29.

115. 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.

116. 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(sup154):Suppl:S111-24.

117. 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(sup154):Suppl:S95-100.

118. 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(sup154):Suppl:S77-84.

119. 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((sup154)):Suppl:S85-93.

120. 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(8):699-707.

121. 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(6):561-9.

122. 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(2):143-8.

123. 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(3):339-47.

124. 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(11):2175-82.

125. 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(5):553-61.

126. Boyd JC, Bruns DE. Quality specifications for glucose meters: assessment by simulation modeling of errors in insulin dose. Clin Chem. 2001;47(2):209-14.

127. 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(3):191-204.

128. 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(sup240):Suppl:S51-60.

129. 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.

130. Karon BS, Boyd JC, Klee GG. Glucose meter performance criteria for tight glycemic control estimated by simulation modeling. Clin Chem. 2010;56(7):1091-7.

131. 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(4):644-50.

132. 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.

133. 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(12):1510-8.

134. Wilinska ME, Hovorka R. Glucose control in the intensive care unit by use of continuous glucose monitoring: what level of measurement error is acceptable? Clin Chem. 2014;60(12):1500-9.

135. 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(3):562-70.

136. McQueen RB, Breton MD, Craig J, Holmes H, Whittington MD, Ott MA, et al. 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(5):992-1001.

137. 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(2):366-77.

138. Turner MJ, Baker AB, Kam PC. Effects of systematic errors in blood pressure measurements on the diagnosis of hypertension. Blood Press Monit. 2004;9(5):249-53.

139. 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(2-4):358-73.

140. 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(10):1931-8.

141. 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(2):91-9.

142. 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(12):996-1003.

143. 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.

144. 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(9):1459-64.

145. Kroll MH, Garber CC, Bi C, Suffin SC. Assessing the impact of analytical error on perceived disease severity. Arch Pathol Lab Med 2015;139(10):1295-301.

146. 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.

147. 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(5):622-8.

148. 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(2):189-206.

149. 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(8):1143-8.

150. 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(7):939-49.

151. Rouse A, Marshall T. The extent and implications of sphygmomanometer calibration error in primary care. J Hum Hypertens. 2001;15:587.

152. 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.

153. 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(8):1922-8.

154. 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(3):347-57.

155. 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(9):840-5.

156. 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(4):557-62.

157. Elloumi F, Hu Z, Li Y, Parker JS, Gulley ML, Amos KD, et al. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med Genomics. 2011;4:54.

158. Schlauch RS, Carney E. Are false-positive rates leading to an overestimation of noise-induced hearing loss? J Speech Lang Hear Res. 2011;54(2):679-92.

159. 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(3):309-13.

160. Drion I, Cobbaert C, Groenier KH, Weykamp C, Bilo HJ, Wetzels JF, et al. Clinical evaluation of analytical variations in serum creatinine measurements: why laboratories should abandon Jaffe techniques. BMC nephrology. 2012;13(1):133.

161. 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(3):283-94.

162. 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(16-17):1479-84.

163. Langlois MR, Descamps OS, van der Laarse A, Weykamp C, Baum H, Pulkki K, et al. Clinical impact of direct HDLc and LDLc method bias in hypertriglyceridemia. A simulation study of the EAS-EFLM Collaborative Project Group. Atherosclerosis. 2014;233(1):83-90.

164. 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(3):543-50.

165. 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(12):889-98.

166. 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(2):237-45.

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

168. 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(1):380.

169. 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(10):899-905.

170. 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(10):580-8.

171. 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(6):392-9.

172. 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(6):1049-54.

173. Chai JH, Ma S, Heng D, Yoong J, Lim WY, Toh SA, et al. 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.

174. Lyon AW, Kavsak PA, Lyon OA, Worster A, Lyon ME. Simulation models of misclassification error for single thresholds of high-sensitivity cardiac troponin I due to assay bias and imprecision. Clin Chem. 2017;63(2):585-92.

175. 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(2):454-66.

176. Saugel B, Grothe O, Nicklas JY. Error grid analysis for arterial pressure method comparison studies. Anesth Analg. 2018;126(4):1177-85.

177. 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(3):141-7.

178. Piona C, Dovc K, Mutlu GY, Grad K, Gregorc P, Battelino T, et al. Non-adjunctive flash glucose monitoring system use during summer-camp in children with type 1 diabetes: the free-summer study. Pediatr Diabetes. 2018;19(7):1285-93.

179. 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(7):1276-84.

180. 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(8):541-9.

181. 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(S3):S21-4.

182. 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(3):545-52.

183. 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(7):1051-5.

184. Kovatchev BP, Patek SD, Ortiz EA, Breton MD. Assessing sensor accuracy for non-adjunct use of continuous glucose monitoring. Diabetes Technol Ther. 2015;17(3):177-86.

185. Schnell O, Erbach M. Impact of a reduced error range of SMBG in insulin-treated patients in Germany. J Diabetes Sci Technol. 2014;8(3):479-82.

186. Kovatchev BP, Wakeman CA, Breton MD, Kost GJ, Louie RF, Tran NK, et al. Computing the surveillance error grid analysis: procedure and examples. J Diabetes Sci Technol. 2014;8(4):673-84.

187. Klonoff DC, Lias C, Vigersky R, Clarke W, Parkes JL, Sacks DB, et al. The surveillance error grid. J Diabetes Sci Technol. 2014;8(4):658-72.

188. 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(4):904-12.

189. 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(2):365-80.

190. 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(5):365-71.

191. 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(2):303-10.

192. 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(6):812-20.

193. 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(3):974-80.

194. 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(3):360-7.

195. 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.

196. 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(9):911-22.

197. 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(10):1723-31.

198. Marrero F, Qadeer MA, Lashner BA. Severe complications of inflammatory bowel disease. Medical Clinics of North America. 2008;92(3):671-86.

199. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st

century: a systematic review of population-based studies. The Lancet. 2017;390(10114):2769-78.

200. Burisch J, Pedersen N, Čuković-Čavka S, Brinar M, Kaimakliotis I, Duricova D, et al. East–West gradient in the incidence of inflammatory bowel disease in Europe: the ECCO-EpiCom inception cohort. Gut. 2014;63(4):588-97.

201. Steed H, Walsh S, Reynolds N. Crohn's disease incidence in NHS Tayside. Scottish medical journal. 2010;55(3):22-5.

202. Rubin G, Hungin A, Kelly P, Ling J. Inflammatory bowel disease: epidemiology and management in an English general practice population. Alimentary pharmacology & therapeutics. 2000;14(12):1553-9.

203. Kaplan GG, Ng SC. Understanding and preventing the global increase of inflammatory bowel disease. Gastroenterology. 2017;152(2):313-21. e2.

204. Bernstein CN, Shanahan F. Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. Gut. 2008;57(9):1185-91.

205. Manzel A, Muller DN, Hafler DA, Erdman SE, Linker RA, Kleinewietfeld M. Role of “Western diet” in inflammatory autoimmune diseases. Current allergy and asthma reports. 2014;14(1):404.

206. Somineni HK, Kugathasan S. The microbiome in patients with inflammatory diseases. Clinical Gastroenterology and Hepatology. 2019;17(2):243-55.

207. Crohn's & colitis UK. Crohn's Disease 2013 [Available from: https://www.crohnsandcolitis.org.uk/about-inflammatory-bowel-disease/crohns-disease.

208. Crohn's & colitis UK. Ulcerative Colitis 2017 [Available from: https://www.crohnsandcolitis.org.uk/about-inflammatory-bowel-disease/ulcerative-colitis.

209. Johnston RD, Logan RF. What is the peak age for onset of IBD? Inflammatory bowel diseases. 2008;14(suppl_2):S4-S5.

210. Crohn's & colitis UK. What are the symptoms? 2017 [Available from: https://www.crohnsandcolitis.org.uk/about-inflammatory-bowel-disease/what-are-the-symptoms.

211. Canavan C, West J, Card T. The epidemiology of irritable bowel syndrome. Clinical epidemiology. 2014;6:71.

212. Thompson W, Longstreth G, Drossman D, Heaton K, Irvine E, Müller-Lissner S. Functional bowel disorders and functional abdominal pain. Gut. 1999;45(suppl 2):II43-II7.

213. Mykletun A, Jacka F, Williams L, Pasco J, Henry M, Nicholson GC, et al. Prevalence of mood and anxiety disorder in self reported irritable bowel syndrome (IBS). An epidemiological population based study of women. BMC gastroenterology. 2010;10(1):88.

214. National Institute of Health and Care Excellence (NICE). Irritable bowel syndrome in adults: diagnosis and management (CG61) 2008 [Available from: https://www.nice.org.uk/guidance/cg61.

215. National Institute of Health and Care Excellence (NICE). Crohn's Disease: management (NG129) 2019 [Available from: https://www.nice.org.uk/guidance/ng129.

216. National Institute of Health and Care Excellence (NICE). Ulcerative colitis: management (NG130) 2019 [Available from: https://www.nice.org.uk/guidance/ng130.

217. Walsham NE, Sherwood RA. Fecal calprotectin in inflammatory bowel disease. Clinical and experimental gastroenterology. 2016;9:21.

218. Tibble JA, Sigthorsson G, Foster R, Forgacs I, Bjarnason I. Use of surrogate markers of inflammation and Rome criteria to distinguish organic from nonorganic intestinal disease. Gastroenterology. 2002;123(2):450-60.

219. Dolwani S, Metzner M, Wassell J, Yong A, Hawthorne A. Diagnostic accuracy of faecal calprotectin estimation in prediction of abnormal small bowel radiology. Alimentary pharmacology & therapeutics. 2004;20(6):615-21.

220. Waugh N, Cummins E, Royle P, Kandala N, Shyangdan D, Arasaradnam R, et al. Faecal calprotectin testing for differentiating amongst inflammatory and non-inflammatory bowel diseases: systematic review and economic evaluation. 2013.

221. Vavricka SR, Spigaglia SM, Rogler G, Pittet V, Michetti P, Felley C, et al. Systematic evaluation of risk factors for diagnostic delay in inflammatory bowel disease. Inflammatory bowel diseases. 2011;18(3):496-505.

222. National Institute of Health and Care Excellence (NICE). Faecal Calprotectin in Primary Care as a Decision Diagnostic for Inflammatory Bowel Disease and Irritable Bowel Syndrome 2018 [Available from: https://www.nice.org.uk/guidance/dg11/resources/endorsed-resource-consensus-paper-pdf-4595859614.

223. Reumkens A, Rondagh EJ, Bakker CM, Winkens B, Masclee AA, Sanduleanu S. Post-colonoscopy complications: a systematic review, time trends, and meta-analysis of population-based studies. The American journal of gastroenterology. 2016;111(8):1092.

224. National Institute of Health and Care Excellence (NICE). Suspected cancer: recognition and referral [NG12] 2017 [Available from: https://www.nice.org.uk/guidance/ng12.

225. Costa F, Mumolo M, Bellini Ma, Romano M, Ceccarelli L, Arpe P, et al. Role of faecal calprotectin as non-invasive marker of intestinal inflammation. Digestive and Liver Disease. 2003;35(9):642-7.

226. Tibble J, Teahon K, Thjodleifsson B, Roseth A, Sigthorsson G, Bridger S, et al. A simple method for assessing intestinal inflammation in Crohn's disease. Gut. 2000;47(4):506-13.

227. National Institute of Health and Care Excellence (NICE). Faecal calprotectin diagnostic tests for inflammatory diseases of the bowel (DG11) 2013 [Available from: https://www.nice.org.uk/guidance/DG11.

228. BÜHLMANN. BÜHLMANN fCAL® ELISA Calprotectin 2018 [Available from: https://www.buhlmannlabs.ch/wp-content/uploads/2015/01/EK-CAL_-IFU_CE_2018-04-10.pdf.

229. National Institute of Health and Care Excellence (NICE). Review of Diagnostics Guidance DG11 - Faecal calprotectin diagnostic tests for inflammatory diseases of the bowel 2015 [Available from: https://www.nice.org.uk/guidance/dg11/evidence/appendix-a-review-decision-pdf-4474871534.

230. Birmingham Quality. Faecal Markers 2019 [Available from: https://birminghamquality.org.uk/eqa-programmes/jby/.

231. Immundiagnostik. Calprotectin (IDK® Calprotectin) (MRP 8/14) (Stool, 1h) ELISA 2009 [Available from: http://www.immundiagnostik.com/en/home/products/kits-assays/gastroenterology-nutrition.html?tx_mokom01immunprodukte_pi1%5Ban%5D=K%206927&tx_mokom01immunprodukte_pi1%5Bag%5D=403&cHash=264edf0336.

232. Eurospital. CalFast 2019 [Available from: http://www.calprotectintest.com/english/calfast.html.

233. U.S. Food & Drug Adminitstration (FDA). 510(k) K190784 (BÜHLMANN fCAL turbo) 2019 [Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190784.pdf.

234. U.S. Food & Drug Adminitstration (FDA). 510(k) K170993 (Inova QUANTA Flash Calprotectin) 2016 [Available from: https://www.accessdata.fda.gov/cdrh_docs/reviews/K170993.pdf.

235. Orgentic. Calprotectin 2018 [Available from: https://www.orgentec.com/en/products/alegria/Stool+Diagnostics/ORG+280.html.

236. U.S. Food & Drug Adminitstration (FDA). 510(k) K182698 (LIAISON Calprotectin) 2018 [Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf18/K182698.pdf.

237. Otten CM, Kok L, Witteman BJ, Baumgarten R, Kampman E, Moons KG, et al. Diagnostic performance of rapid tests for detection of fecal calprotectin and lactoferrin and their ability to discriminate inflammatory from irritable bowel syndrome. Clinical chemistry and laboratory medicine. 2008;46(9):1275-80.

238. Conroy S, Hale MF, Cross SS, Swallow K, Sidhu RH, Sargur R, et al. Unrestricted faecal calprotectin testing performs poorly in the diagnosis of inflammatory bowel disease in patients in primary care. Journal of clinical pathology. 2018;71(4):316-22.

239. Turvill J, Turnock D, Holmes H, Jones A, Mclaughlan E, Hilton V, et al. Evaluation of the clinical and cost-effectiveness of the york faecal calprotectin care pathway. Frontline gastroenterology. 2018;9(4):285-94.

240. Turvill J, O'Connell S, Brooks A, Bradley-Wood K, Laing J, Thiagarajan S, et al. Evaluation of a faecal calprotectin care pathway for use in primary care. Primary health care research & development. 2016;17(5):428-36.

241. Holmes H, McMaster J, Davies H, Vaines V, Turvill J. Evaluation of the Cost-Utility of the York Faecal Calprotectin Care Pathway. Expert Review of Pharmacoeconomics & Outcomes Research. Submitted December 2019

242. National Institute of Health and Care Excellence (NICE). The New Faecal Calprotectin Care Pathway 2018 [Available from: https://www.nice.org.uk/sharedlearning/the-new-faecal-calprotectin-care-pathway.

243. Lasson A, Stotzer P-O, Öhman L, Isaksson S, Sapnara M, Strid H. The intra-individual variability of faecal calprotectin: a prospective study in patients with active ulcerative colitis. Journal of Crohn's and Colitis. 2015;9(1):26-32.

244. Moum B, Jahnsen J, Bernklev T. Fecal calprotectin variability in Crohn's disease. Inflammatory bowel diseases. 2010;16(7):1091-2.

245. Reenaers C, Bossuyt P, Hindryckx P, Vanpoucke H, Cremer A, Baert F. Expert opinion for use of faecal calprotectin in diagnosis and monitoring of inflammatory bowel disease in daily clinical practice. United European gastroenterology journal. 2018;6(8):1117-25.

246. Calafat M, Cabré E, Mañosa M, Lobatón T, Marín L, Domènech E. High within-day variability of fecal calprotectin levels in patients with active ulcerative colitis: what is the best timing for stool sampling? Inflammatory bowel diseases. 2015;21(5):1072-6.

247. Kristensen V, Malmstrøm GH, Skar V, Røseth A, Moum B. Clinical importance of faecal calprotectin variability in inflammatory bowel disease: intra-individual variability and standardisation of sampling procedure. Scandinavian journal of gastroenterology. 2016;51(5):548-55.

248. Oyaert M, Van den Bremt S, Boel A, Bossuyt X, Van Hoovels L. Do not forget about pre-analytics in faecal calprotectin measurement! Clinica Chimica Acta. 2017;473:124-6.

249. Røseth A, Fagerhol M, Aadland E, Schjønsby H. Assessment of the neutrophil dominating protein calprotectin in feces: a methodologic study. Scandinavian journal of gastroenterology. 1992;27(9):793-8.

250. Tøn H, Brandsnes Ø, Dale S, Holtlund J, Skuibina E, Schjønsby H, et al. Improved assay for fecal calprotectin. Clinica Chimica Acta. 2000;292(1-2):41-54.

251. Konikoff MR, Denson LA. Role of fecal calprotectin as a biomarker of intestinal inflammation in inflammatory bowel disease. Inflammatory bowel diseases. 2006;12(6):524-34.

252. Padoan A, D'Incà R, Scapellato ML, De Bastiani R, Caccaro R, Mescoli C, et al. Improving IBD diagnosis and monitoring by understanding preanalytical, analytical and biological fecal calprotectin variability. Clinical Chemistry and Laboratory Medicine (CCLM). 2018;56(11):1926-35.

253. Alpha Laboratories. Products for Calprotectin Extraction [Available from: https://www.calprotectin.co.uk/calprotectin-products/products-for-calprotectin-extraction/.

254. Inova Diagnostics. Fecal Extraction Device [Available from: https://www.inovadx.com/fecal-extraction-device.

255. Whitehead S, French J, Brookes M, Ford C, Gama R. Between-assay variability of faecal calprotectin enzyme-linked immunosorbent assay kits. Annals of clinical biochemistry. 2013;50(1):53-61.

256. Oyaert M, Trouvé C, Baert F, De Smet D, Langlois M, Vanpoucke H. Comparison of two immunoassays for measurement of faecal calprotectin in detection of inflammatory bowel disease:(pre)-analytical and diagnostic performance characteristics. Clinical chemistry and laboratory medicine. 2014;52(3):391-7.

257. De Sloovere MM, De Smet D, Baert FJ, Debrabandere J, Vanpoucke HJ. Analytical and diagnostic performance of two automated fecal calprotectin immunoassays for detection of inflammatory bowel disease. Clinical Chemistry and Laboratory Medicine (CCLM). 2017;55(9):1435-46.

258. Juricic G, Brencic T, Tesija Kuna A, Njegovan M, Honovic L. Faecal calprotectin determination: impact of preanalytical sample treatment and stool consistency on within-and between-method variability. Biochemia medica: Biochemia medica. 2019;29(1):112-22.

259. Whitehead SJ, Ford C, Gama RM, Ali A, McKaig B, Waldron JL, et al. Effect of faecal calprotectin assay variability on the management of inflammatory bowel disease and potential role of faecal S100A12. Journal of clinical pathology. 2017;70(12):1049-56.

260. Pelkmans LP, de Groot MJ, Curvers J. Analytical Performance and Clinicopathologic Correlation of Four Fecal Calprotectin Methods. American journal of clinical pathology. 2019;152(3):392-8.

261. Ayling RM, Kok K. Fecal calprotectin. Advances in clinical chemistry. 87: Elsevier; 2018. p. 161-90.

262. Ricciuto A, Griffiths AM. Clinical value of fecal calprotectin. Critical reviews in clinical laboratory sciences. 2019;56(5):307-20.

263. Pathirana WGW, Chubb SP, Gillett MJ, Vasikaran SD. Faecal calprotectin. The Clinical Biochemist Reviews. 2018;39(3):77.

264. Lippi G, Betsou F, Cadamuro J, Cornes M, Fleischhacker M, Fruekilde P, et al. Preanalytical challenges–time for solutions. Clinical Chemistry and Laboratory Medicine (CCLM). 2019;57(7):974-81.

265. Miller WG, Tate JR, Barth JH, Jones GR. Harmonization: the sample, the measurement, and the report. Annals of laboratory medicine. 2014;34(3):187-97.

266. Armbruster D, Donnelly J. Harmonization of clinical laboratory test results: the role of the IVD industry. EJIFCC. 2016;27(1):37.

267. Oyaert M, Boel A, Jacobs J, Van den Bremt S, De Sloovere M, Vanpoucke H, et al. Analytical performance and diagnostic accuracy of six different faecal calprotectin assays in inflammatory bowel disease. Clinical Chemistry and Laboratory Medicine (CCLM). 2017;55(10):1564-73.

268. Ayling RM, Kok K. Fecal Calprotectin. Adv Clin Chem. 2018;87:161-90.

269. Horvath AR, Lord SJ, St John A. Outcome-based analytical performance specifications–Mission impossible? Pathology. 2016;48:S17.

270. Panteghini M, Sandberg S. Defining analytical performance specifications 15 years after the Stockholm conference. Clinical Chemistry and Laboratory Medicine (CCLM). 2015;53(6):829-32.

271. Daniel Turnock (Consultant Clinical Biochemist YTHT, UK). Electronic conversation with: AF Smith (University of Leeds, UK). 2019.

272. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria2013.

273. Arrabal C, Da Rocha R, Nonaka R, Meira S. Comparison of resampling method applied to censored data. International Journal of Advanced Statistics and Probability. 2014;2(2):48-55.

274. Chen K, Lo S-H. On bootstrap accuracy with censored data. The Annals of Statistics. 1996;24(2):569-95.

275. Lo AY. A Bayesian bootstrap for censored data. The Annals of Statistics. 1993;21(1):100-23.

276. Efron B. Censored data and the bootstrap. Journal of the American Statistical Association. 1981;76(374):312-9.

277. Delignette-Muller ML, Dutang C. fitdistrplus: An R package for fitting distributions. Journal of Statistical Software. 2015;64(4):1-34.

278. Klein J, Moeschberger M. Survival analysis: techniques for censored and truncated data. ed 2nd Springer Verlag. New York. 2003.

279. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Archives of pathology & laboratory medicine. 2013;137(4):558-65.

280. National Institute of Health and Care Excellence (NICE). Guide to the methods of technology appraisal 2013 (PMG9) 2013 [Available from: http://nice.org.uk/process/pmg9.

281. Claxton KP, Sculpher MJ. Using value of information analysis to prioritise health research. Pharmacoeconomics. 2006;24(11):1055-68.

282. Healy M. Outliers in clinical chemistry quality-control schemes. Clinical chemistry. 1979;25(5):675-7.

283. Jacoby WG. Loess:: a nonparametric, graphical tool for depicting relationships between variables. Electoral Studies. 2000;19(4):577-613.

284. Euser AM, Dekker FW, le Cessie S. A practical approach to Bland-Altman plots and variation coefficients for log transformed variables. Journal of clinical epidemiology. 2008;61(10):978-82.

285. Chhapola V, Kanwal SK, Brar R. Reporting standards for Bland–Altman agreement analysis in laboratory research: a cross-sectional survey of current practice. Annals of clinical biochemistry. 2015;52(3):382-6.

286. Jones GR, Albarede S, Kesseler D, MacKenzie F, Mammen J, Pedersen M, et al. Analytical performance specifications for external quality assessment–definitions and descriptions. Clinical Chemistry and Laboratory Medicine (CCLM). 2017;55(7):949-55.

287. 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.

288. Hall PS, Mitchell ED, Smith AF, Cairns DA, Messenger M, Hutchinson M, 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. 2018;22(32).

289. Bossuyt PM, Olsen M, Hyde C, Cohen JF. An analysis reveals differences between pragmatic and explanatory diagnostic accuracy studies. Journal of Clinical Epidemiology. 2020;117:29-35.

290. Sun Q, Welsh KJ, Bruns DE, Sacks DB, Zhao Z. Inadequate reporting of analytical characteristics of biomarkers used in clinical research: a threat to interpretation and replication of study findings. Clinical chemistry. 2019;65(12):1554-62.

291. Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC medical research methodology. 2016;16(1):97.

292. O'Rourke B, Werkö SS, Merlin T, Huang LY, Schuller T. The 'Top 10'Challenges for Health Technology Assessment: INAHTA Viewpoint. International journal of technology assessment in health care. 2019:1-4.

293. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny. Web application framework for R. R package version [1.00]. 2017. 2018.

294. Dewitte K, Fierens C, Stöckl D, Thienpont LM. Application of the Bland–Altman plot for interpretation of method-comparison studies: a critical investigation of its practice. Clinical chemistry. 2002;48(5):799-801.

295. Ludbrook J. Confidence in Altman–Bland plots: a critical review of the method of differences. Clinical and Experimental Pharmacology and Physiology. 2010;37(2):143-9.

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

297. Braga F, Panteghini M. Generation of data on within-subject biological variation in laboratory medicine: an update. Critical reviews in clinical laboratory sciences. 2016;53(5):313-25.

298. Ricós C, Alvarez V, Cava F, Garcia-Lario J, Hernandez A, Jimenez C, et al. Current databases on biological variation: pros, cons and progress. Scandinavian journal of clinical and laboratory investigation. 1999;59(7):491-500.

299. Aarsand AK, Røraas T, Bartlett WA, Coşkun A, Carobene A, Fernandez-Calle P, et al. Harmonization initiatives in the generation, reporting and application of biological variation data. Clinical Chemistry and Laboratory Medicine (CCLM). 2018;56(10):1629-36.

300. Aarsand AK F-CP, Webster C, Coskun A, Gonzales-Lao E, Diaz-Garzon J, Jonker N, Minchinela J, Simon M, Braga F, Perich C, Boned B, Roraas T, Marques-Garcia F, Carobene A, Aslan B, Barlett WA, Sandberg S. . The EFLM Biological Variation Database 2019 [Available from: https://biologicalvariation.eu/.

301. Setia N, Nichols JH. Preanalytical Variation in Clinical Laboratory Testing. Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation. 2006.

302. Hawkins R. Managing the pre-and post-analytical phases of the total testing process. Annals of laboratory medicine. 2012;32(1):5-16.

303. Tate JR, Johnson R, Barth J, Panteghini M. Harmonization of laboratory testing—current achievements and future strategies. Clinica Chimica Acta. 2014;432:4-7.

304. Theodorsson E. Validation and verification of measurement methods in clinical chemistry. Bioanalysis. 2012;4(3):305-20.

305. Theodorsson E, Magnusson B, Leito I. Bias in clinical chemistry. Bioanalysis. 2014;6(21):2855-75.

306. Jani D, Allinson J, Berisha F, Cowan KJ, Devanarayan V, Gleason C, et al. Recommendations for use and fit-for-purpose validation of biomarker multiplex ligand binding assays in drug development. The AAPS journal. 2016;18(1):1-14.

307. 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.

308. 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.

309. Curtis LA, Burns A. Unit Costs of Health and Social Care 2018, Personal Social Services Research Unit, University of Kent, Canterbury 2018 [Available from: https://doi.org/10.22024/UniKent/01.02.70995.

310. National Institute of Health and Care Excellence (NICE). Point-of-care and home faecal calprotectin tests for monitoring treatment response in inflammatory bowel disease. Medtech innovation briefing [MIB132] 2017 [Available from: https://www.nice.org.uk/advice/MIB132.

311. NHS Improvement. National schedule of reference costs 2018 [Available from: https://improvement.nhs.uk/resources/reference-costs/#rc1718.

312. National Institute of Health and Care Excellence (NICE). Infliximab for acute exacerbations of ulcerative colitis (TA163) 2008 [Available from: https://www.nice.org.uk/Guidance/ta163.

Appendix A

Glossary table[60]

	Synonyms	General definition	Definition from CLSI Harmonized Terminology Database
Analytical phase	Examination procedure	All processes in the total testing process occurring at the point of sample analysis.	Set of operations, described specifically, used in the performance of examinations according to a given method
Analytical sensitivity	-	The rate of change in the measured test value, in relation to a given increase in the measurand concentration	Quotient of the change in an indication and the corresponding change in the value of a quantity being measured
Analytical variation	-	The component of imprecision attributable to variation in analytical factors (factors occurring during the analytical phase of the total testing process).	-
Batch	Lot	-	One or more components or finished devices that consist of a single type, model, class, size, composition, or software version that are manufactured under essentially the same conditions and that are intended to have uniform characteristics and quality within specified limits.
Between-subject biological variation	Inter-individual biological variation; group biological variation	Variation observed across individuals in terms of their homeostatic set points.	-
Bias	**-**	Systematic error in measurement.	Estimate of a systematic measurement error.
Biological variation	**-**	Within-subject and between-subject variation in measurand concentrations over time.	Consists of within-subject (CV_I, intra-individual) and between-subject (CV_G, inter-individual, group) variation.

[60] This table provides two definitions for each term: first, the general definitions used in this thesis; 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.

	Synonyms	General definition	Definition from CLSI Harmonized Terminology Database
Bland-Altman plot	Difference plot	A scatter plot of the difference between index and reference test measurements vs. the mean of the paired results, allowing estimation of mean difference, limits of agreement, outliers and constant and proportional bias.	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.
Calibration	**-**	The process of testing and adjusting a test instrument or system, to establish a correlation between the measurand and measurement response	Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication
Certified reference material (CRM)	**-**	Materials that have been characterised via an unbroken chain of measurement processes, each with a defined measurement uncertainty, linking back to a reference measurement procedure	Reference material, accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes traceability to an accurate realization of the unit in which the property values are expressed, and for which each certified value is accompanied by an uncertainty at a stated level of confidence
Clinical performance	Clinical validity; test accuracy; test efficacy	The ability of a test to detect patients with a particular clinical condition or in a physiological state	The sum of all attributes that may be important for clinical use of results from a measurement procedure when applied to a specific intended use
Clinical utility	Clinical effectiveness; clinical usefulness	The clinical value that can be derived from a test, which may be quantified in terms of intermediate clinical utility (relating to the impact of test results on patient management decisions e.g. the decision to treat or not treat), or end-stage clinical utility (relating to the impact of test results on patient health outcomes e.g. patient mortality and morbidity).	Value or benefit assigned to a particular outcome or state; diagnostic information that contributes to the identification of a particular condition or disease.
Cost-effectiveness	Efficiency	The ability of an intervention to produce an efficient impact on patient health outcomes in relation to costs.	-
Cross-reactivity	**-**	The existence of obstruction from substances in the test sample which are mistaken for the target analyte 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 measurement procedure.
Diagnostic accuracy	-	The ability of a test to discriminate between diseased and non-diseased subjects, or between two or more clinical states.	The ability of a diagnostic test to discriminate between diseased and non-diseased subjects, or between two or more clinical states.
Diagnostic sensitivity	Clinical sensitivity	The proportion of diseased patients which the test correctly identifies as having the disease.	The proportion of patients with a well-defined clinical disorder (or condition of interest) whose test values are positive or exceed a

	Synonyms	General definition	Definition from CLSI Harmonized Terminology Database
			defined decision limit (i.e. a positive result and identification of the patients who have a disease).
Diagnostic specificity	Clinical specificity	The proportion of healthy patients which the test correctly identifies as not having the disease.	The proportion of patients who do not have a specified clinical disorder whose test results are negative or within the defined decision limit.
False negative case	**-**	A person with the disease or clinical condition of interest, who is incorrectly classified as not having the disease/condition based on a negative test result	-
False positive case	**-**	A person who does not have the disease or clinical condition of interest, who is incorrectly classified as having the disease/condition based on a positive test result	-
Harmonisation	**-**	The comparability of test results, irrespective of the measurement procedure used, and where or when a measurement was made.	The process of recognizing, understanding, and explaining differences while taking steps to achieve uniformity of results, or at a minimum, a means of conversion of results such that different groups can use the data obtained from assays interchangeably.
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.
Interference	**-**	The existence of obstruction from substances in the test sample which either inhibit the process of binding with the target analyte.	Artificial increase or decrease in apparent concentration or intensity of an analyte (measurand) due to the presence of a substance that reacts non-specifically with either the detecting reagent or the signal itself.
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.
Limit of Blank (LOB)	**-**	The highest (apparent) concentration of analyte expected to be identified when processing blank samples (i.e. samples containing zero quantity of analyte).	The highest measurement result that is likely to be observed (with a stated probability [alpha]) for a blank sample.

	Synonyms	General definition	Definition from CLSI Harmonized Terminology Database
Limit of Detection (LOD)	-	The lowest analyte 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.
Lower limit of Quantification (LOQ_{lower})	-	The lowest concentration of analyte in a sample that a test can measure with a specified level of imprecision and trueness.	The lowest concentration of measurand that can be detected with acceptable precision and trueness, under routine clinical laboratory conditions, in a defined type of sample.
Lower limit of Quantification (LOQ_{upper})	-	The highest concentration of analyte in a sample that a test can measure with a specified level of imprecision and trueness.	The highest concentration of measurand that can be detected with acceptable precision and trueness, under routine clinical laboratory conditions, in a defined type of sample.
Linear range	-	The region of measurand values over which linearity is maintained.	The range over which the testing systems results are acceptably linear; that is, where nonlinear error is less than the error criterion.
Linearity	-	Linearity relates to how well the slope of the calibration curve follows a straight line.	The ability (within a given range) to provide results that are directly proportional to the concentration (amount) of the analyte in the test sample.
Measurand	Analyte	The substance intended to be measured by a given test.	Quantity intended to be measured.
Measurement performance	Analytical validity; analytical performance; technical performance; technical efficacy	Refers to the overall technical performance of a test, including the central components of measurement uncertainty (precision and trueness) as well as additional performance parameters including test selectivity, detection and quantitation limits, analytical sensitivity, linearity and measurement range.	-
Measurement range	Reportable range; measuring interval; working interval; working range	The range of measurand concentrations over which a test is demonstrated to perform adequately.	Range of analyte concentrations over which meaningful results can be acquired
Measurement uncertainty	-	Uncertainty around the underlying 'true' measurand value associated with an observed test measurement, resulting from systematic and/or random error.	Non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used.
Medical in-vitro test	-	Tests conducted on patient samples taken from the human body.	-
Metrological chain of traceability	-	The sequence of measurement processes linking a CRM to the reference measurement procedure.	Sequence of measurement standards and calibrations that is used to relate a measurement result to a reference.

	Synonyms	General definition	Definition from CLSI Harmonized Terminology Database
Negative predictive value (NPV)	-	The likelihood that a patient has the disease given that the test result is positive.	The likelihood that an individual with a negative test does not have the disease, or other characteristic, which the test is designed to detect.
Positive predictive value (PPV)	-	The likelihood that a patient is healthy given that the test result is negative.	The likelihood that an individual with a positive test result has a particular disease or characteristic that the test is designed to detect.
Pre-analytical phase	**Pre-examination processes**	All processes in the total testing process occurring prior to the point of sample analysis.	[Defined as "pre-examination 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
Pre-analytical variation	**-**	The component of imprecision attributable to variation in analytical factors (factors occurring during the analytical phase of the total testing process).	-
Precision	-	The closeness of agreement between repeated test results.	Closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions.
Reference change value	-	The change that must occur in an individual's serial results before that change may be considered significant.	Represents the statistically significant difference between consecutive results based on the combined inherent variation of both results. The total variation of a result is a combination of pre-examination, examination, post-examination, and within-subject biological variation
Reference measurement procedure	-	A thoroughly investigated measurement procedure shown to yield values having an uncertainty of measurement commensurate with its intended use, especially in assessing the trueness of other measurement procedures	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.
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.
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

	Synonyms	General definition	Definition from CLSI Harmonized Terminology Database
			locations, operators, measuring systems, and replicate measurements on the same or similar objects.
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
Scientific validity	-	The association between a measurand and a clinical condition or disease state.	-
Test evaluation pathway	-	The trajectory of research required to take a new technology from the biomarker discovery phase, to the test adoption phase	-
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.
Total testing process	**-**	The complete process of events occurring from the point at which a test is initially ordered, through to the point at which the test result is made available to the clinician.	-
Traceability	**-**	-	Property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty
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.
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 analyte.	Non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used.
Within-subject biological variation	**-**	The fluctuation of measurand concentrations in the body over time.	-

Appendix B
Measurement uncertainty and measurement performance: supplementary material

This appendix presents further details relating to test measurement uncertainty and measurement performance, supplemental to the introduction provided in Chapter 1 (section 1.2). Appendix B.1 provides an example of a Bland-Altman plot, relevant to the discussion of bias in section 1.2.2.2. Section B.2 provides a discussion of pre-analytical and analytical factors (including a generic illustration of a 'feather map'), which is relevant to the discussion of bias and imprecision in section 1.2.2. Finally section B.3 provides details of additional metrics of measurement performance (in addition to the central components of measurement uncertainty reviewed in Chapter 1), which are relevant to the discussion of method validation and verification provided in section 1.2.4.

B.1 Example Bland-Altman plot

According to the original Bland-Altman plot (proposed by the authors JM Bland and DG Altman in 1986), the mean of each paired measurement [(A+B)/2] from a method-comparison study is plotted against the absolute difference [A-B] (11). An example of such a plot is provided in Figure B-1. In this case, there does not appear to be a relationship between the mean and differences, and the bias can therefore be summarised as the average difference between the two sets of measurements – shown as the solid blue line in Figure B-1. In addition, based on the fact that the differences are normally distributed, '95% limits of agreement' (equivalent to +/- 1.96*SD of the differences) are also presented (the dashed upper and lower lines in Figure B-1). Based on this example, the average bias is indicated as 15.95, and 95% of differences lie within the lower limit of -3.05 and the upper limit of 34.85.

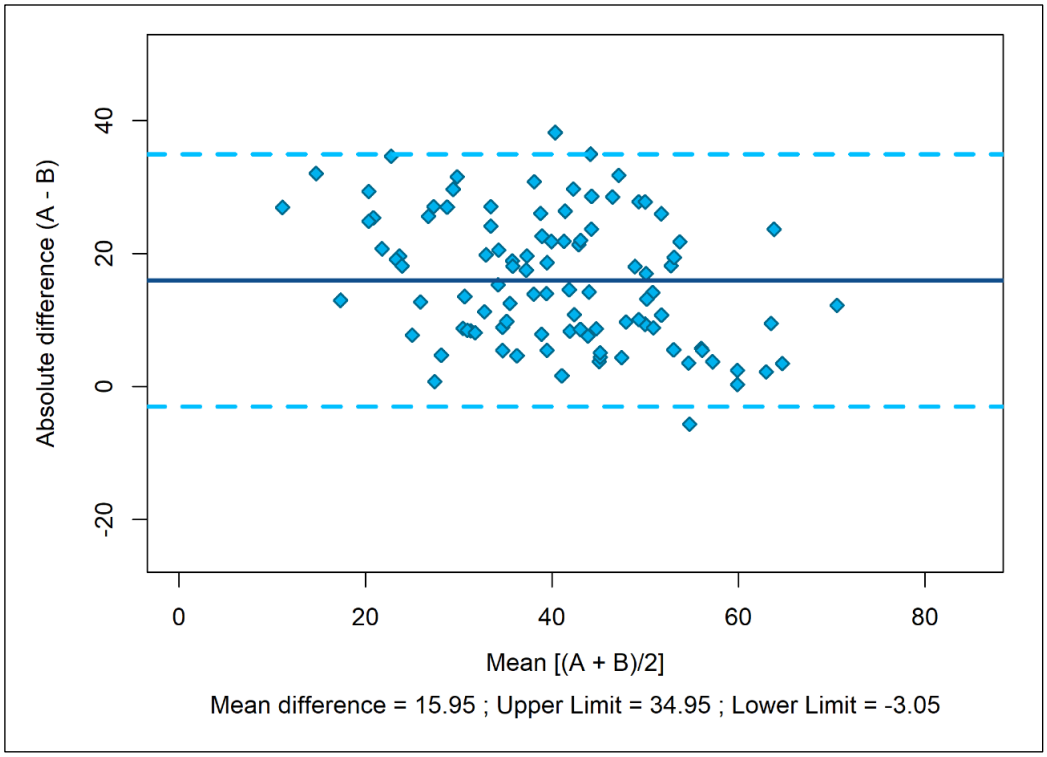


Figure B-1. Example of a Bland Altman plot

In alternative scenarios, different variations of the Bland-Altman plot may be applied. For example, if the spread of differences had been observed to increase in relation to the mean, then a logarithmic or percentage y scale could have been used (11, 294); and if a *proportional bias* had been observed (i.e. if bias increased or decreased in line with the measurement value), then linear regression could have been applied to fit a line of best fit describing the average bias (295). Furthermore, Figure B-1 presents a Bland-Altman plot with the mean of paired measurements [(A+B)/2] presented on the x-axis, as originally proposed by Bland and Altman in their seminal 1986 paper. Taking the mean of paired measurements in this way is the required approach when comparing two uncertain testing procedures. If, however, one set of measurements may be considered to be without measurement uncertainty (e.g. when using a reference measurement procedure or CRMs), then differences may instead be plotted against the reference value directly (rather than the mean) (9).

B.2 Pre-analytical, analytical and post-analytical factors affecting measurement uncertainty

Multiple factors may introduce bias and imprecision into test measurements, occurring at different points along the *total testing process* (i.e. the complete process of events occurring from the point at which a test is initially ordered, through to the point at which the test result is made available to the clinician) (296). The total testing process can be divided into two core phases: (i) the *pre-analytical phase* – which includes all processes occurring prior to the point of sample analysis; and (ii) the *analytical phase* – which includes all processes occurring at the point of sample analysis.[61] For a given test, pre-analytical and analytical factors (i.e. relevant factors occurring in each phase of the testing process) can be summarised using a 'feather diagram'. Figure B-2 provides a generalised example, in which relevant factors are presented in a (roughly) chronological order along the testing pathway.

The first important factor is *within-subject biological variation* – defined as the fluctuation of measurand concentrations in the body over time (6). Technically biological variation occurs during the pre-analytical phase of the testing pathway; however, this component of variation is conceptually different to variation caused by other pre-analytical factors, since it reflects a natural phenomenon occurring in the body. Other pre-analytical factors, meanwhile, result from human processes which can be altered and largely controlled via standardisation of pre-analytical procedures. As such, biological variation is typically considered as a separate component of imprecision, distinct from pre-analytical and analytical variation (discussed further below).

In its simplest form, biological variation relates to the natural fluctuation of measurand concentrations in the body around a 'homeostatic set point' (i.e. the

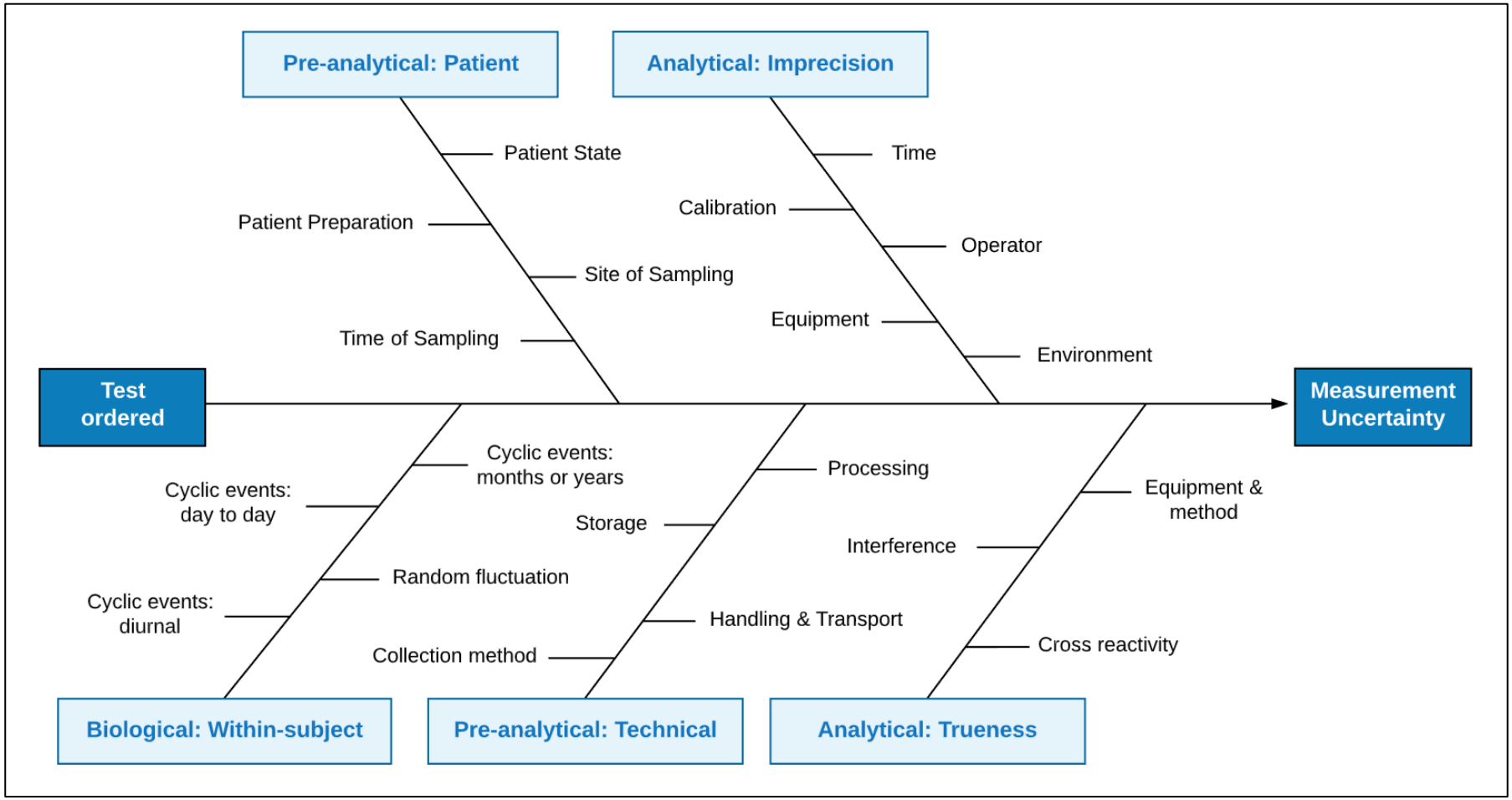
[61] Two further phases of the total testing process may be described: (i) the *pre-pre-analytical phase*, which relates to whether or not the appropriate test was ordered by the clinician and subsequently undertaken in the laboratory; and (ii) the *post-analytical phase*, which relates to how the test result is recorded, and subsequently presented to and interpreted by the clinician. Errors introduced at these phases are of critical importance, since they impact on whether or not the right test is conducted and how that test is acted on. However, these errors do not influence measurement uncertainty *per se*, and are therefore not addressed in this thesis.

average of repeated test values for an individual) (297). For example, many liver function tests (such as serum bilirubin and alanine transaminase) display randomly fluctuating values within monitored individuals (13). Biological variation also describes variation resulting from cyclic events related to circadian rhythms (e.g. time of day, month, or year etc.); female hormones associated with reproductive health, for example, exhibit peaks and troughs relating to the female menstrual cycle (13). For many measurands however, cyclical variations are considered to have limited clinical importance, and biological variation is evaluated as a component of random fluctuation (expressed as an SD or CV) acting on an individual's "true" measurand value (13). At the group level, variation is also observed across individuals in terms of their homeostatic set points – this is referred to as *between-subject biological variation* (13, 297).

The evaluation of biological variation is of particular importance in scenarios where repeated tests are used to inform patient management. In this case, data on biological variation is required to determine the change that must occur in an individual's serial results before that change may be considered significant (i.e. the *reference change value*) (13). Studies of biological variation rely on serial test results from individuals over time, with the influence of other sources of variation minimised via careful control of testing processes (13, 297). The total variation observed in serial test results can then be analysed using statistical techniques (i.e. analysis of variance [ANOVA]) to quantify the individual components of within-subject and between-subject biological variation, and analytical variation[62] (13, 297). Given the importance of biological variation in the interpretation of test results, significant international efforts have been applied to collate and generate evidence on biological variation for a range of measurands (298-300).

[62] The component of analytical variation (discussed later in this section) can be isolated within biological variation studies by including replicate measurements in the experiment, to provide an estimate of repeatability.

Figure B-2. Generalised feather diagram depicting factors contributing to measurement uncertainty



The remaining factors occurring in the pre-analytical phase (henceforth referred to as *pre-analytical factors*) include: (i) pre-analytical patient factors – related to patient preparation prior to taking the test sample, such as their posture, food intake and exercise; and (ii) pre-analytical technical factors – related to how the test sample is collected, transported to and stored in the laboratory. Exposure to sunlight, for example, can cause a breakdown of certain measurands (e.g. bilirubin); care is therefore required in the handling of these samples, to ensure that the measurand is not unknowingly degraded (301). Since sudden changes in pre-analytical processes may lead to bias in test measurements, standardisation of this phase of the total testing process is crucial (302, 303). Once a pre-analytical protocol is in place, then insofar as subsequent deviations in pre-analytical factors would be expected to occur randomly over time, the resulting impact on measurement is increased variability in the form of *pre-analytical variation.* Together with *analytical variation* (below), these components of variation are expressed as an SD or CV, which feed into the overall imprecision of a test.

Analytical factors are those factors occurring during the analytical phase of the total testing process (i.e. at the point of sample analysis). In Figure B-2 these parameters have been divided into those predominantly associated with imprecision, versus those predominantly associated with bias. The influence of key analytical factors on imprecision (i.e. time, operator, calibration, environment and equipment) is discussed in section 1.2.2.1. In the same way as for pre-analytical factors, the key with controlling the influence of these factors is to ensure standardisation of the analytical phase, as far as possible. Subsequent deviation in analytical factors over time may then be considered to act mainly on imprecision, in the form of *analytical variation*.

Analytical factors associated with bias, meanwhile, include bias resulting from the test method and equipment, and bias resulting from other components in the test sample which may interrupt the measurement process (e.g. via *interference* or *cross-reactivity*). Bias resulting from the test method and equipment is the primary focus of bias studies, as discussed in section 1.2.2.2. The concepts of interference and cross-reactivity, meanwhile, are typically examined separately from primary bias assessments, and are thus included under the wider banner of

measurement performance. These concepts are therefore discussed further in Appendix B.3.

The analysis of pre-analytical and analytical factors is relevant at each stage of the test evaluation pathway – from the initial test development phase through to the routine implementation phase. During the initial development and optimisation phases of test development, some exploration of key pre-analytical and analytical factors is routinely conducted, however testing procedures are often developed around tightly controlled pre-analytical and analytical procedures. As such, validation, verification and ongoing quality assurance procedures in the laboratory are crucial (304).

B.3 Measurement performance: additional parameters

Further to the central components of measurement uncertainty (i.e. bias and imprecision), the concept of *measurement performance* includes additional parameters which contribute to and describe the overall technical performance of a test. This includes: *selectivity*, *detection limits*, *analytical sensitivity*, *linearity* and *measurement range*. Whilst the focus of this thesis is on the central components of measurement uncertainty (bias and imprecision), a brief summary of these additional measurement performance parameters is provided here for completeness.

B.3.1 Selectivity

Selectivity is defined as the ability of a test to measure a specified measurand in the presence of interferences that may be expected to be present in the sample *matrix* (i.e. all the components of a sample, excluding the measurand). Naturally occurring substances in the test matrix may interfere with the test measurement in one of two ways: first, substances may inhibit the process of binding with the target measurand (known as *interference*); and second, substances may be mistaken for the target measurand, leading to 'unintentional' binding (known as *cross-reactivity*) (see Figure B-2) (305, 306). For example, a heightened level of bilirubin (a natural by-product of liver functioning), is known to produce bias in certain tests for creatinine (used to monitor kidney functioning) (307).

One way of evaluating test selectivity, is to take test samples with a known quantity of measurand (e.g. CRM samples) and deliberate 'spike' them with a

known quantity of suspected interferent (308). If the additional component leads to a significant change in the test value, then the substance can be confirmed as an interferent. In general, interferents should be identified by the test manufacturer during the test development phase, and appropriately accounted for in test standard operating procedures. Further exploration of interferents in the clinical laboratory may be undertaken if, for example, interference is a suspected cause of an identified bias of unknown origin.

B.3.2 Detection limits

Various limits can be specified which describe the smallest concentration of a measurand that can be reliably measured by the test. 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 the measurand); (ii) the *limit of detection* (LOD), defined as the lowest measurand concentration which a test can reliably distinguish from the LOB; and (iii) the *lower limit of quantification* (LOQ_{lower}), defined as the lowest concentration of measurand which a test can detect with a specified level of precision and trueness (typically set an order of magnitude higher than the LOD) (6). These concepts are of key importance for tests in which trace measurements (i.e. low concentrations of the measurand) are expected; if only high concentrations of the measurand are expected then evaluation of the detection limits is of little consequence. Where trace measurements are expected, then the ability of a test to measure low concentrations will be a key determinant of the test's clinical performance.

A further related concept is that of the upper limit of quantification (LOQ_{upper}). In line with the LOQ_{lower}, the LOQ_{upper} is defined as the highest concentration of measurand which a test can detect with a specified level of precision and trueness (6). Together, the concepts of LOQ_{lower} and LOQ_{upper} are important in determining the measurement range of a test, as discussed in section B.3.3.

B.3.3 Analytical sensitivity, linearity and measuring range

Figure B-3 shows a standard calibration curve, which may be derived by running a series of samples of known concentrations (or known relative concentrations). The *analytical sensitivity* of a test refers to the rate of change in the measured

test value, in relation to a given increase in the measurand concentration – this is equal to the slope of the calibration curve shown in Figure B-3 (6). *Linearity*, meanwhile, relates to how well the slope of this line follows a straight line; and the *linear range* refers to the region of measurand values over which linearity is maintained (6). Beyond the linear range, samples become saturated and the analytical sensitivity of the method begins to drop – shown by a "tailing off" of the calibration curve. Finally, the *measurement range* of a test describes the range of concentrations over which a test is demonstrated to perform adequately. Typically this is set equal to the region in-between the LOQ_{lower} and LOQ_{upper}, as shown in Figure B-3.

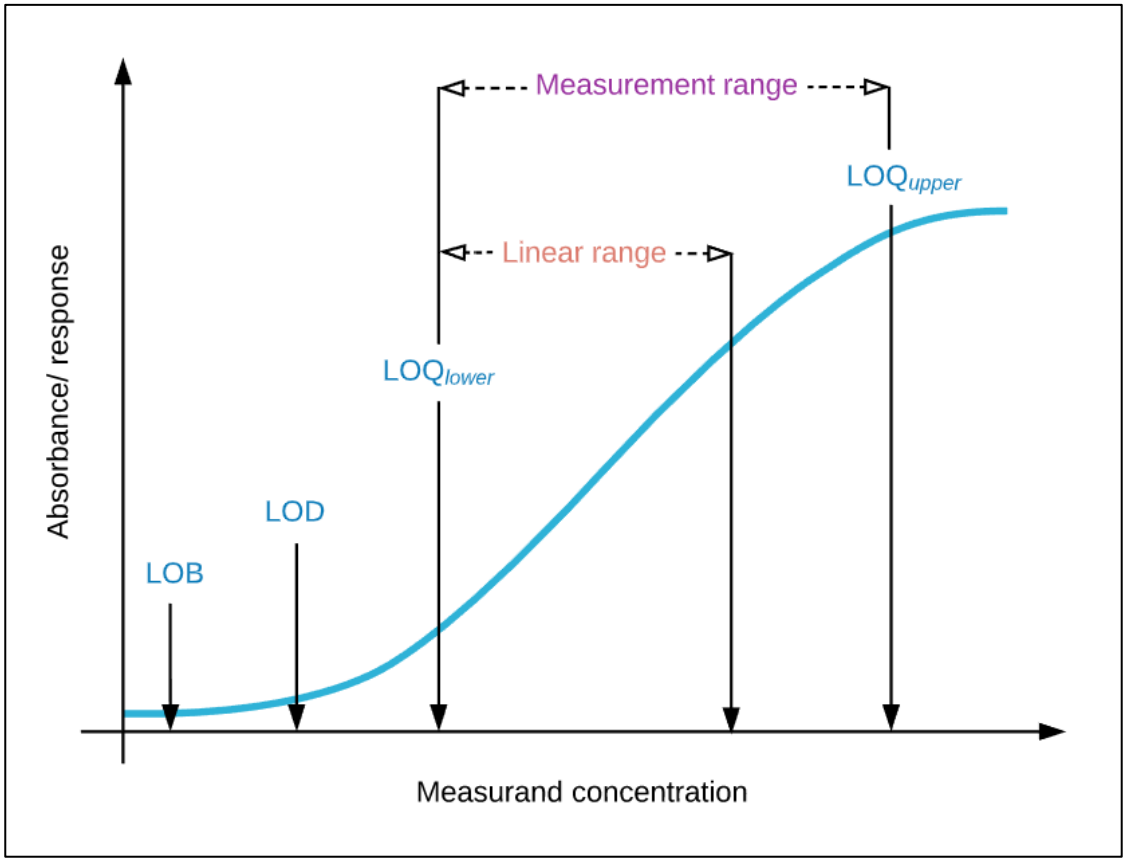


Figure B-3. Calibration curve illustrating limits of detection and measurement range

Appendix C

Diagnostic accuracy calculation

Diagnostic accuracy is defined as the ability of a test to discriminate between diseased and non-diseased subjects, or between two or more clinical states. It is evaluated by comparing test-directed diagnoses against "true" diagnoses (based, for example, on an established gold standard reference test). This comparison enables test-directed diagnoses to be classified as true positive, false positive, true negative or false negative, as illustrated in Table C-1.

Diagnostic sensitivity is calculated as the proportion of diseased patients which the test correctly identifies as having the disease; whilst *diagnostic specificity* is calculated as the proportion of healthy patients which the test correctly identifies as not having the disease. Alternatively, diagnostic accuracy may be summarized in terms of predictive measures: the *positive predictive value* (PPV) of a test is defined as the likelihood that a patient has the disease given that the test result is positive; whilst the negative predictive value (NPV) is defined as the likelihood that a patient is healthy given that the test result is negative.

Table C-1. Confusion matrix showing diagnostic accuracy measures

		True diagnosis		
		Diseased	**Healthy**	
Test diagnosis	**Diseased**	True positives (TP)	False positives: (FP)	**PPV = TP/(TP+FP)**
	Healthy	False negatives: (FN)	True negatives: (TN)	**NPV = TN/(TN+FN)**
		Sensitivity = TP/(TP+FN)	**Specificity = TN/(TN+FP)**	

Appendix D
Cost-effectiveness metrics

Cost-effectiveness may be measured in terms of the *Incremental cost-effectiveness ratio (ICER)*, or *Incremental Net Monetary Benefit (INMB)*. Both of these metrics depend on having estimates for the total costs associated with the intervention test arm (C_T) and the standard care comparator arm (C_{SC}); and estimates for the total health effect for the intervention test arm (E_T) and the standard care comparator arm (E_{SC}). These estimates depend on the perspective taken in the economic evaluation, and may be derived from either a trial-based or model-based study (as discussed in section 1.3.2). In the UK context costs are measured in GBP (£) and a common unit of health effect is the QALY.

The ICER is calculated by dividing the difference in costs between two arms (ΔC), by the difference in health effects (ΔE), as illustrated in equation D.1 below.

$$ICER = \frac{C_T - C_{SC}}{E_T - E_{SC}} = \frac{\Delta C}{\Delta E} \qquad \text{(D.1)}$$

Assuming that the test intervention is more costly and more effective than standard care, the ICER represents the additional cost required to be spent on the intervention to gain an additional unit of health.[63] The cost-effectiveness of an intervention is determined by whether or not this rate of gain in terms of the health effect, outweighs the *opportunity cost* of the additional spending – that is, the amount of health that will be lost to other patients elsewhere in the healthcare system, as a result of redirecting funding to the more costly intervention. Assuming that required additional funding will be taken from the least effective healthcare services currently provided by the NHS, then the opportunity cost is captured by the productivity of the health service at the margin. The threshold capturing the marginal productivity of the NHS is referred to as the *supply side cost-effectiveness threshold*.[64]

---

[63] Note: if the intervention produces lower costs and higher effects than the comparator, then the intervention is said to *dominate* standard care, and the ICER is not required.

[64] The cost-effectiveness threshold may alternatively be conceptualised as the decision maker's *willingness to pay* per additional unit of health (i.e. the *demand side cost-effectiveness threshold*).

In the UK, NICE stipulates a cost-effectiveness threshold (λ) of £20,000 per additional QALY (58). This means that if a new intervention has an ICER of < £20,000 per additional QALY then it is likely to be considered a cost-effective use of NHS resources; whilst an ICER of > £20,000 indicates that the intervention is not expected to be a cost-effective use of resources or is required to meet additional criteria. This decision rule is expressed as per equation D.2 below.

$$ICER = \frac{\Delta C}{\Delta E} < \lambda \qquad \text{(D.2)}$$

When the threshold (λ) is defined, one can multiply the QALYs by the threshold value to express QALYs on the monetary scale (or, conversely, one may divide the incremental cost by the threshold value to convert costs onto the QALY scale). For example, assuming the NICE adopted threshold of £20,000, we can convert 0.5 QALYs into an equivalent cost of £10,000. This enables the total benefit associated with each strategy to be expressed in terms of net monetary benefit (NMB) as illustrated below:

$$NMB_T = (E_T \times \lambda) - C_T \qquad \text{(D.3)}$$

$$NMB_{SC} = (E_{SC} \times \lambda) - C_{SC} \qquad \text{(D.4)}$$

The INMB is then calculated either as outlined in equation D.5 or D.6 below.

$$INMB = NMB_T - NMB_{SC} \qquad \text{(D.5)}$$

$$INMB = (\Delta E \times \lambda) - \Delta C \qquad \text{(D.6)}$$

Unlike the ICER, for which the exact interpretation of cost-effectiveness depends on whether or not the incremental cost and QALYs are positive or negative, the interpretation of the NMB statistics is straightforward: for any given set of strategies, the strategy with the greatest NMB is the most cost-effective alternative; and for a given pair-wise comparison, an intervention is cost-effective if it is associated with a positive INMB.

Appendix E

HTA systematic review: listed authorities on the CRD HTA database

Table E-1. INAHTA members and additional organisations listed on the CRD HTA database (as of March 2017)

N	Organisation abbreviation	Organisation name, country
INAHTA Members		
1	AETS	Agencia de Evaluación de Tecnologias Sanitarias, SPAIN
2	ACE	Agency for Care Effectiveness, SINGAPORE
3	AETSA	Andalusian Agency for Health Technology Assessment, SPAIN
4	Age.Na.S	The Agency for Regional Healthcare, ITALY
5	AHRQ	Agency for Healthcare Research and Quality, USA
6	AHTA	Adelaide Health Technology Assessment, AUSTRALIA
7	AHTAPol	Agency for Health Technology Assessment in Poland, POLAND
8	AQuAS	Agència de Qualitat i Avaluació Sanitàries de Catalunya, SPAIN
9	ASERNIP-S	Australian Safety and Efficacy Register of New Interventional Procedures -Surgical, AUSTRALIA
10	ASSR	Agenzia Sanitaria e Sociale Regionale (Regional Agency for Health and Social Care), ITALY
11	AVALIA-T	Galician Agency for Health Technology Assessment, SPAIN
12	CADTH	Canadian Agency for Drugs and Technologies in Health, CANADA
13	CDE	Center for Drug Evaluation, Taiwan, REPUBLIC OF CHINA
14	CEDIT	Comité d'Evaluation et de Diffusion des Innovations Technologiques, FRANCE
15	CEM	Inspection générale de la sécurité sociale (IGSS), Cellule d'expertise médicale, LUXEMBOURG
16	CENETEC	Centro Nacional de Excelencia Tecnológica en Salud, MEXICO
17	CMeRC HTA Unit	Charlotte Maxeke Research Consortium, SOUTH AFRICA
18	CONITEC	National Committee for Technology Incorporation, BRAZIL
19	DAHTA @ DIMDI	Deutsche Agentur für Health Technology Assessment, GERMANY
20	DECIT-CGATS	Coordenação Geral de Avaliação de Tecnologias em Saúde; Departamento de Ciência e Tecnologia, BRAZIL
21	G-BA	The Federal Joint Committee (Gemeinsamer Bundesausschuss), GERMANY
22	GÖeG	Gesunheit Österreich GmbH, AUSTRIA
23	HAS	Haute Autorité de Santé, FRANCE

24	HealthPACT	The Health Policy Advisory Committee on Technology, AUSTRALIA & NEW ZEALAND
25	HIQA	Health Information and Quality Authority, IRELAND
26	HIS	Healthcare Improvement Scotland, UNITED KINGDOM
27	HQO	Evidence Development and Standards Branch, CANADA
28	DEFACTUM	Social & Health Services and Labour Market, DENMARK
29	IACS	Health Sciences Institute in Aragon, SPAIN
30	IECS	Institute for Clinical Effectiveness and Health Policy, ARGENTINA
31	IETS	Instituto de Evaluación Tecnológica en Salud, COLOMBIA
32	IHE	Institute of Health Economics, CANADA
33	INASante	National Authority for Assessment and Accreditation in Healthcare, TUNISIA
34	INESSS	Institut national d'excellence en santé et en services, CANADA
35	IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, GERMANY
36	KCE	Belgian Health Care Knowledge Centre, BELGIUM
37	LBI-HTA	Ludwig Boltzmann Institute for Health Technology Assessment, AUSTRIA
38	MaHTAS	Health Technology Assessment Section, Ministry of Health Malaysia, MALAYSIA
39	MSP-Uruguay	Ministerio-Salud-Publica, URUGUAY
40	MTU-SFOPH	Swiss Federal Office of Public Health, SWITZERLAND
41	NECA	National Evidence-based healthcare Collaborating Agency, KOREA
42	NIHR	National Institute for Health Research, UNITED KINGDOM
43	NIPH	Norwegian Institute of Public Health, NORWAY
44	OSTEBA	Basque Office for Health Technology Assessment, SPAIN
45	RCHD-CS	Ministry of Public Health of the Republic of Kazakhstan, Republican Centre for Health Development, KAZAKHSTAN
46	SBU	Swedish Agency for Health Technology Assessment and Assessment of Social Services, SWEDEN
47	UVT	HTA Unit in A. Gemelli Teaching Hospital, ITALY
48	ZIN	Zorginstituut Nederland, THE NETHERLANDS
49	ZonMw	The Netherlands Organisation for Health Research and Development, THE NETHERLANDS
Additional organisations		
1	-	Scottish Health Purchasing Information Centre
2	SCTIE	Secretaria de Cencia, Tecnologia e Insumos Estrategicos, Departmento de Ciencia e Tecnologia, BRAZIL
3	VASPVT	State Health Care Accreditation Agency under the Ministry of Health of the republic of Lithuania, LITHUANIA
4	TASK	Technology Assessment at SickKids, CANADA
5	-	Technology Assessment Unit of the McGill University Health Centre
6	-	Stockholm County Council Support for evidence-based medicine Method Council, HTA

7	TNO	Netherlands Organisation for Applied Scientific Research,THE NETHERLANDS
8	OTA	US Congress Office of Technology Assessment
9	UETS	Unidad de Evaluacion de Tecnologias Sanitarias, SPAIN
10	-	Unidad de Tecnologias de Salud, MEXICO
11	HEHTA	Unit of Health Economics and Technology Assessment in Health Care, UNITED KINGDOM
12	UHC	University HealthSystem Consortium, USA
13	-	University of York, UNITED KINGDOM
14	VATAP	Veteran Affairs Technology Assessment Program, USA
15	L&I	Washington State Department of Labor and Industries, USA
16	-	Wessex Institute for Health Research and Development, UNITED KINGDOM
17	WMHTAC	West Midlands Health Technology Assessment Collaboration, UNITED KINGDOM
18	WorkSafeBC	Worksafe British Columbia, CANADA

Appendix F

HTA systematic review: CRD HTA database search strategy

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

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]

Appendix G

Methodology review: search strategies

G.1 EMBASE

Database: Embase Classic+Embase <1947 to 2019 March 27>

1	exp diagnostic test/
2	exp assay/
3	exp laboratory diagnosis/
4	exp molecular diagnosis/
5	exp in vitro study/
6	(assay* or biomarker* or predictor*).tw.
7	test*.ti.
8	1 or 2 or 3 or 4 or 5 or 6 or 7
9	((total or measur* or systematic or random or analytic* or preanalytic* or pre-analytic*) adj3 (error* or uncertain*)).tw.
10	misclassif*.tw.
11	(trueness or imprecision).tw.
12	(bias or precision).ti.
13	(biological adj2 (variation or variability)).tw.
14	((analytic* or preanalytic* or pre-analytic* or technical*) adj3 (goal* or perform* or valid* or verif*)).tw.
15	((limit* adj2 (detect* or blank or quantification or quantitation)) or LOD or LOB or LOQ or LLoQ or ULoQ).tw.
16	9 or 10 or 11 or 12 or 13 or 14 or 15
17	simulation*.tw.
18	exp *Computer Simulation/
19	exp *Monte Carlo Method/
20	exp *Models, Statistical/
21	17 or 18 or 19 or 20
22	methodol*.ti.
23	8 and 16 and 21
24	8 and 16 and 22
25	23 or 24
26	limit 25 to (human and yr="2008 -Current")

G.2 Ovid Medline(R)

Database: **Ovid MEDLINE(R) Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) <1946 to 2019 March 27>**

1	exp Clinical Laboratory Techniques/
2	exp In Vitro Techniques/
3	exp Biomarkers/
4	exp diagnostic techniques, cardiovascular/ or exp diagnostic techniques, digestive system/ or exp diagnostic techniques, endocrine/ or exp diagnostic techniques, neurological/ or exp "diagnostic techniques, obstetrical and gynecological"/ or exp diagnostic techniques, ophthalmological/ or exp diagnostic techniques, otological/ or exp diagnostic techniques, radioisotope/ or exp diagnostic techniques, respiratory system/ or exp diagnostic techniques, surgical/ or exp diagnostic techniques, urological/ or exp diagnostic tests, routine/
5	(assay* or biomarker* or predictor*).tw.
6	test*.ti.
7	1 or 2 or 3 or 4 or 5 or 6
8	((total or measur* or systematic or random or analytic* or preanalytic* or pre-analytic*) adj2 (error* or uncertain*)).tw.
9	misclassif*.tw.
10	(trueness or imprecision).tw.
11	(bias or precision).ti.
12	(biological adj2 (variation or variability)).tw.
13	((analytic* or preanalytic* or pre-analytic* or technical*) adj2 (goal* or perform* or valid* or verif*)).tw.
14	((limit* adj2 (detect* or blank or quantification or quantitation)) or LOD or LOB or LOQ or LLoQ or ULoQ).tw.
15	8 or 9 or 10 or 11 or 12 or 13 or 14
16	simulation*.tw.
17	exp *Computer Simulation/
18	exp *Monte Carlo Method/
19	exp *Models, Statistical/
20	16 or 17 or 18 or 19
21	methodol*.ti.
22	7 and 15 and 20
23	7 and 15 and 21
24	22 or 23
25	limit 24 to (humans and yr="2008 -Current")

G.3 Web of Science (Core Collection)

Set	Search terms
# 13	#12 OR #11 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI Timespan=2008-2019*
# 12	#10 AND #9 AND #1
# 11	#10 AND #8 AND #1
# 10	#6 OR #5 OR #4 OR #3 OR #2
# 9	TI = methodol*
# 8	TS = (simulation* OR "Monte Carlo simulation*" OR "computer simulation*" OR "statistical model*")
# 7	TS = ((limit* NEAR/2 (detect* OR blank OR quantification OR quantitation)) OR LOD OR LOB OR LOQ OR LLoQ OR ULoQ).
# 6	TS = ((analytic* OR preanalytic* OR pre-analytic* OR technical*) NEAR/3 (goal* OR perform* OR valid* OR verif*))
# 5	TS = (biological NEAR/2 (variation OR variability))
# 4	TS = (trueness OR imprecision)
# 3	TI = (bias OR precision)
# 2	TS = (((total OR measur* OR systematic OR random OR analytic* OR preanalytic* OR pre-analytic*) NEAR/3 (error* OR uncertain*)) OR misclassif*)
# 1	TS = ((laboratory NEAR/2 (test* OR diagnosis)) OR (("in vitro" OR in-vitro) NEAR/2 (technique* OR test*)) OR "biomarker*" OR "assay*" OR "predictor*")

G.4 BIOSIS (Citation Index)

Set	Search terms
# 13	#12 OR #11 *Indexes=BCI Timespan=2008-2019*
# 12	#10 AND #9 AND #1 *Indexes=BCI Timespan=2008-2019*
# 11	#10 AND #8 AND #1 *Indexes=BCI Timespan=2008-2019*
# 10	#6 OR #5 OR #4 OR #3 OR #2 *Indexes=BCI Timespan=2008-2019*
# 9	TI = methodol* *Indexes=BCI Timespan=2008-2019*
# 8	TS = (simulation* OR "Monte Carlo simulation*" OR "computer simulation*" OR "statistical model*") *Indexes=BCI Timespan=2008-2019*
# 7	TS = ((limit* NEAR/2 (detect* OR blank OR quantification OR quantitation)) OR LOD OR LOB OR LOQ OR LLoQ OR ULoQ). *Indexes=BCI Timespan=2008-2019*
# 6	TS = ((analytic* OR preanalytic* OR pre-analytic* OR technical*) NEAR/3 (goal* OR perform* OR valid* OR verif*)) *Indexes=BCI Timespan=2008-2019*
# 5	TS = (biological NEAR/2 (variation OR variability)) *Indexes=BCI Timespan=2008-2019*
# 4	TS = (trueness OR imprecision) *Indexes=BCI Timespan=2008-2019*
# 3	TI = (bias OR precision) *Indexes=BCI Timespan=2008-2019*
# 2	TS = ((total OR measur* OR systematic OR random OR analytic* OR preanalytic* OR pre-analytic*) NEAR/3 (error* OR uncertain*)) *Indexes=BCI Timespan=2008-2019*
# 1	TS = ((laboratory NEAR/2 (test* OR diagnosis)) OR (("in vitro" OR in-vitro) NEAR/2 (technique* OR test*)) OR "biomarker*" OR "assay*" OR "predictor*") *Indexes=BCI Timespan=2008-2019*

Appendix H
Error “stripping” example

The total imprecision of a test (CV_T) may be described as comprising of three core components: within-individual biological variation (CV_I), pre-analytical variation (CV_{Pre-A}), and analytical variation (CV_A) (127). In this case, CV_T can be described as per Equation H.1 below:

$$CV_T = \sqrt{{CV_I}^2 + {CV_{Pre-A}}^2 + {CV_A}^2} \qquad \text{(H.1)}$$

To isolate the “pure biologic distribution”, free of pre-analytical and analytical variation, these components of imprecision can be removed or “stripped” from the estimate of total imprecision, to leave the within-individual biological variation (127). Suppose, for example, that quality assurance data indicates that total imprecision for a given test is equal to 12.3%, and the individual components of pre-analytical and analytical variation are equal to 5.9% and 2.9% respectively. CV_I can then be calculated as per Equation H.2:

$$CV_I = \sqrt{{CV_T}^2 - {CV_{Pre-A}}^2 - {CV_A}^2} \qquad \text{(H.2)}$$

Substituting our estimated values we have:

$$CV_I = \sqrt{12.3^2 - 5.9^2 - 2.9^2} = 10.39 \qquad \text{(H.3)}$$

This value can then be used to inform the distribution of “true” measurand values.

Appendix I
Parametric sampling method

I.1 AIC and BIC results for alternative right-censored data regions

For the NICE FC pathway evaluation, Table I-1 and Table I-2 show the AIC and BIC results for the simulated parametric FC1 distributions for each population, using the alternative upper bounds for the right-censored data region of 2,000 and 3,000 µg/g respectively.

Similarly for the YFCCP evaluation, Table I-3 and Table I-4 show the AIC and BIC results for the simulated parametric FC1 and FC2 distributions for each population, using the alternative upper bounds for the right-censored data region of 2,000 and 3,000 µg/g respectively.

Table I-1. NICE FC pathway: AIC and BIC criteria for FC1 parametric distributions (upper bound for right-censored FC data region = 2,000 µg/g)

Subgroup	Parameterisation	AIC	BIC	% values ≥ 50 µg/g (simulated data)	% values ≥ 50 µg/g (YFCCP FC1 data)
IBS FC1	**Lognormal**	8575.469	8585.013	45.6%	40.3%
	Weibull	8723.379	8732.923	49.1%	
	Gamma	8779.891	8789.435	48.1%	
	Normal	10375.29	10384.83	61.9%	
IBD FC1	**Lognormal**	565.9061	570.6195	98.6%	96.2%
	Weibull	551.4574	556.1708	97.5%	
	Gamma	553.3666	558.08	97.4%	
	Normal	568.471	573.1845	92.5%	

Table I-2. NICE FC pathway: AIC and BIC criteria for FC1 parametric distributions (upper bound for right-censored FC data region = 3,000 µg/g)

Subgroup	Parameterisation	AIC	BIC	% values ≥ 50 µg/g (simulated data)	% values ≥ 50 µg/g (YFCCP FC1 data)
IBS FC1	**Lognormal**	8573.435	8582.979	45.7%	40.3%
	Weibull	8723.341	8732.885	49.1%	
	Gamma	8779.891	8789.434	48.1%	
	Normal	10375.29	10384.83	61.9%	
IBD FC1	**Lognormal**	554.2693	558.9827	97.9%	96.2%
	Weibull	545.3042	550.0176	96.8%	
	Gamma*	-	-	-	
	Normal	568.4071	573.1205	92.2%	

*For the IBD FC1 population, the MLE estimation would not converge for the Gamma specification in this analysis, so no results are provided for this parameterisation in this population.

Table I-3. YFCCP: AIC and BIC criteria for FC1 and FC2 parametric distributions (upper bound for right-censored FC data region = 2,000 µg/g)

Subgroup		Parameterisation	AIC	BIC	% values ≥ 100 µg/g (simulated data)	% values ≥ 100 µg/g (YFCCP data)
IBS	FC1	Lognormal	8575.469	8585.013	23.8%	19.7%
		Weibull	8723.379	8732.923	28.4%	
		Gamma	8779.891	8789.435	30.5%	
		Normal	10375.29	10384.83	47.0%	
	FC2	Lognormal	1904.728	1911.023	44.4%	40.7%
		Weibull	1921.202	1927.497	50.1%	
		Gamma	1925.586	1931.881	52.4%	
		Normal	2134.02	2140.315	63.7%	
IBD	FC1	Lognormal	565.9061	570.6195	94.8%	93.6%
		Weibull	551.4574	556.1708	92.8%	
		Gamma	553.3666	558.08	93.5%	
		Normal	568.471	573.1845	90.0%	
	FC2	Lognormal	568.1996	572.4859	95.5%	100%
		Weibull	570.6714	574.9577	92.7%	
		Gamma	569.4821	573.7684	93.9%	
		Normal	588.505	592.7913	90.4%	
		Normal	588.505	592.7912	89.4%	

Table I-4. YFCCP: AIC and BIC criteria for FC1 and FC2 parametric distributions (upper bound for right-censored FC data = 3,000 µg/g)

Subgroup		Parameterisation	AIC	BIC	% values ≥ 100 µg/g (simulated data)	% values ≥ 100 µg/g (YFCCP data)
IBS	FC1	Lognormal	8573.435	8582.979	23.9%	19.7%
		Weibull	8723.341	8732.885	28.5%	
		Gamma	8779.891	8789.434	30.5%	
		Normal	10375.29	10384.83	45.1%	
	FC2	Lognormal	1903.045	1909.34	45.5%	40.7%
		Weibull	1921.075	1927.37	49.2%	
		Gamma	1925.557	1931.852	53.0%	
		Normal	2134.02	2140.315	62.7%	
IBD	FC1	Lognormal	554.2693	558.9827	93.9%	93.6%
		Weibull	545.3042	550.0176	92.6%	
		Gamma	-	-	-	
		Normal	568.4071	573.1205	90.0%	
	FC2	Lognormal	564.8118	569.0981	95.1%	100%
		Weibull	570.0538	574.3401	92.0%	
		Gamma	568.4333	572.7196	93.3%	
		Normal	588.505	592.7912	89.4%	

*For the IBD FC1 population, the MLE estimation would not converge for the Gamma specification in this analysis, so no results are provided for this parameterisation in this population.

I.2 Parametric sampling method – density plots

Figure I-1 and Figure I-2 show the density plots for 10,000 simulations from each of the parametric distributions explored within the parametric method base case, across four patient-test subgroups (IBD FC1, IBS FC1, IBD FC2 and IBS FC2). The probability density based on the underlying YFCCP data is shown by the solid black line in each figure, with the alternative parametric distributions illustrated by the various coloured lines. Note that for the sake of illustration left- and right-censored data values in the YFCCP dataset have again been set equal to 10µg/g and 600 µg/g respectively. This results in secondary 'peaks' at the 600 mark, most noticeable in the FC2 distributions, as a result of the large proportion of right-censored data within the YFCCP dataset: these peaks are artificial however, since in reality these censored values would be spread across the upper measurement range, as in the simulated parametric distributions accounting for censoring.

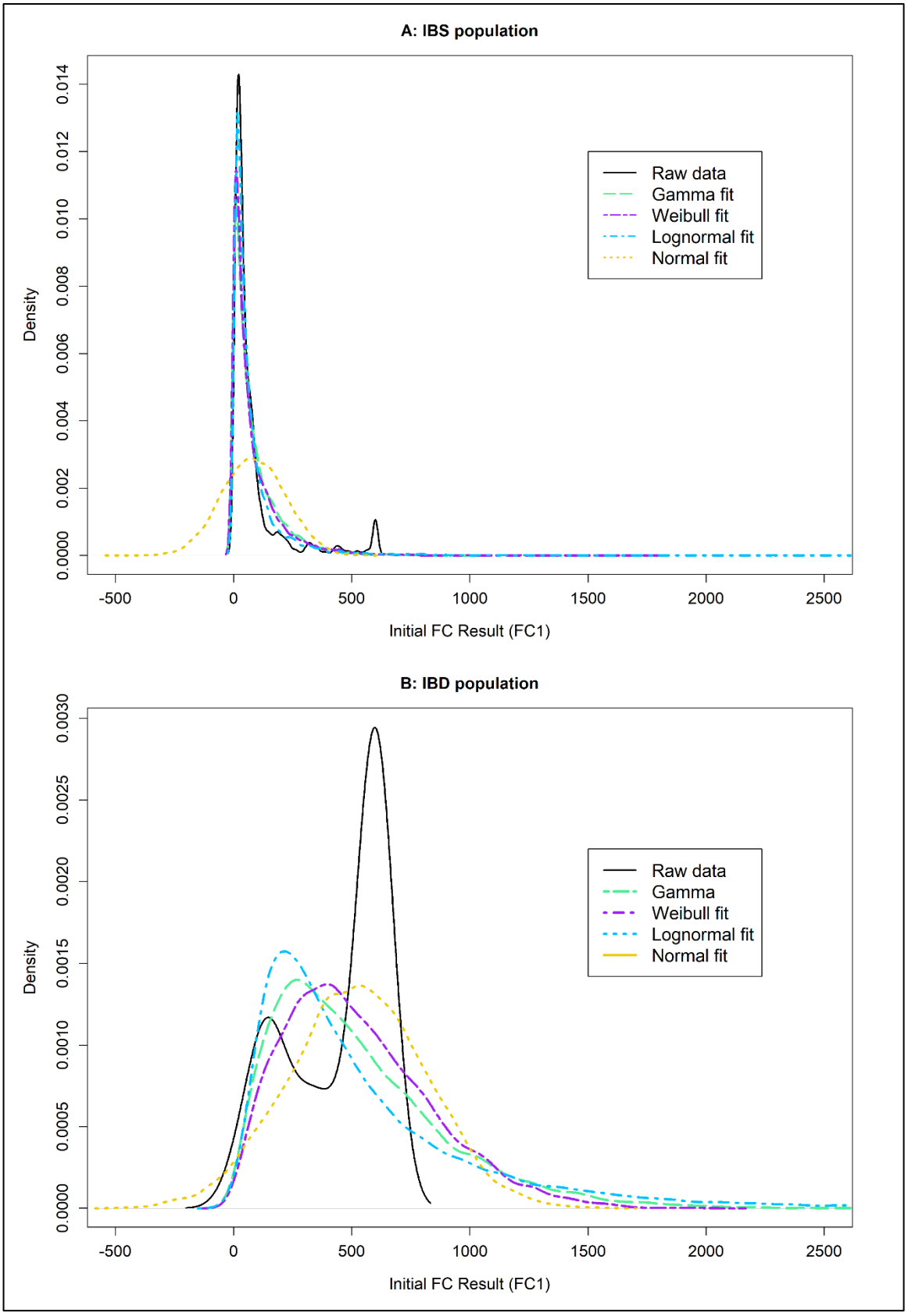


Figure I-1. FC1 values: density plot of parametric distribution fits for IBS and IBD populations

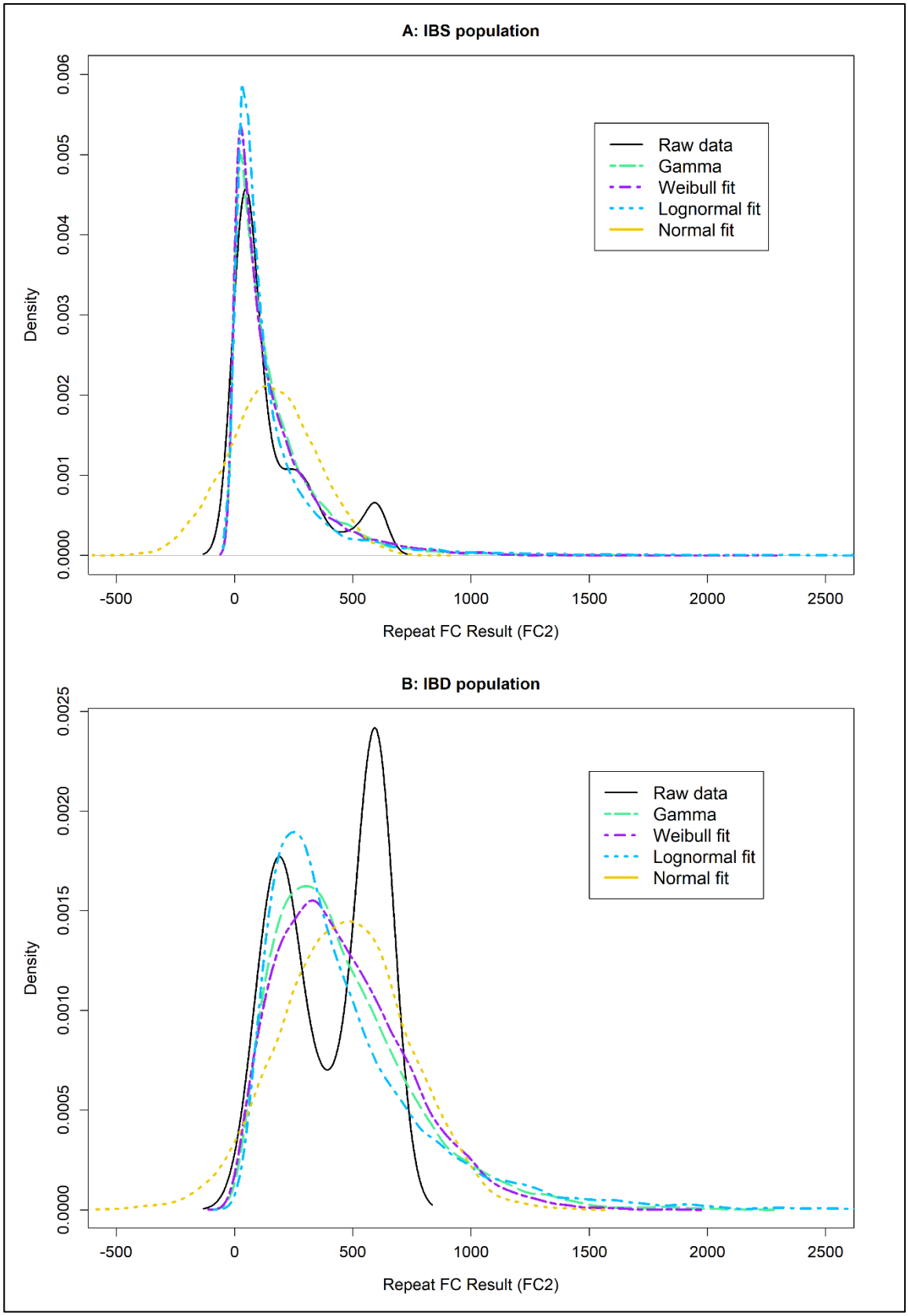


Figure I-2. FC2 values: density plot for parametric distribution fits for IBS and IBD populations

I.3 Example R code for parametric sampling method

#Example code for fitting distributions to censored data using the R 'fitdistrplus' package, using a simulated (hypothetical) dataset.

#---------- Set up code ----------

```
#Set working directory for storing plots:
setwd("insert\\file\\pathname")

#Load required packages:
if (!require("fitdistrplus"))   install.packages("fitdistrplus");    library("fitdistrplus")

#Set random number seed:
set.seed(10)
```

#---------- Generate a hypothetical dataset of lognormally distributed data ----------

```
#Calculate lnorm parameters based on natural mean = 80 and SD = 100:
mean <- 80
sd <- 100
meanlog <- log(mean^2/ sqrt(sd^2 + mean^2))
sdlog <- sqrt(log(1 + (sd^2/ mean^2)))
print(meanlog); print(sdlog)

#Simulate dataset:
data <-  rlnorm(n=1000, meanlog, sdlog)
data <- as.data.frame(data)

#Inspect data:
head(data)
hist(data[,1], breaks=100)
min(data[,1]); max(data[,1])

#Generate a new data column, with values <10 reset to character value "<10", and values >600 rest to ">600":
data[,2] <- data[,1]
colnames(data) <- c("complete", "cens_character")
data$cens_character[data$complete < 10]  <-  "<10"
data$cens_character[data$complete > 600] <-  ">600"

#Generate a final data column, with any censored values set to a temporary numerical value
```

```
#In this case, left-censored values are set to 5, and right-censored values are set to 650
#These temporary numerical values are used as placeholders only.
data[,3] <- data$complete
colnames(data) <- c("complete", "cens_character", "cens_numerical")
data$cens_numerical[data$complete < 10]  <-  5
data$cens_numerical[data$complete > 600] <-  650

#Check the replacement worked:
data$cens_numerical[data$complete < 10]
data$cens_numerical[data$complete > 600]

#---------- Parametric fitting code ----------
#To use fitdistr for censored data the data needs to be in the appropriate format:
#Two columns are required - 'left' and 'right'.
#Left defines the left (i.e. lower) boundary of the censored data region.
#Right defines the right (i.e. upper) boundary of the censored data region.
#For complete data, left and right are set equal to the numerical value observed.

data$left <- data$cens_numerical
data$right  <- data$cens_numerical

#Left-censored data lie somewhere between 0 and 10, so replace left with 0, right with 10:
data$left[data$cens_numerical==5]     <- 0
data$right[data$cens_numerical==5]   <- 10

#Right-censored data lie somewhere above 600. So left = 600, right =NA, or a defined upper
boundary.
#In this case use an upper bound for the right-censored data region of 1000:
data$left[data$cens_numerical==650]   <- 600
data$right[data$cens_numerical==650]  <- 1000

#---- Run the fitdistcens code
temp  <- data.frame(data$left, data$right)
colnames(temp) <- c("left", "right")

#Plot CDF of censored data (turnbull plot)
#plotdistcens(temp)

fit_norm    <- fitdistcens(temp, "norm")     #;summary(fit_norm)
fit_ln      <- fitdistcens(temp, "lnorm")    #;summary(fit_ln)
```

```
fit_gamma   <- fitdistcens(temp, "gamma")    #;summary(fit_gamma)
fit_weib    <- fitdistcens(temp, "weibull")  #;summary(fit_weib)
#Note: summary code provides the AIC and BIC values for each model

#---- Simulate parametric distributions using the estimated distributional parameters
Nsim <- 10000
sim_norm    <- rnorm   (Nsim,  mean=    fit_norm$estimate[1],  sd=    fit_norm$estimate[2])
sim_lnorm   <- rlnorm  (Nsim,  meanlog= fit_ln$estimate[1],    sdlog= fit_ln$estimate[2])
sim_gamma   <- rgamma  (Nsim,  shape=   fit_gamma$estimate[1], rate=  fit_gamma$estimate[2])
sim_weib    <- rweibull(Nsim,  shape=   fit_weib$estimate[1],  scale= fit_weib$estimate[2])

#---- Plot frequency density distributions
tiff(filename="temp.tiff", units="in", width=10, height=7, res=300)
plot (density(data$complete), lwd=2, xlim=c(-200,1500), main="", xlab="Test values",
ylab="Density")
lines(density(sim_gamma), col="seagreen2",   lty="longdash", lwd=3)
lines(density(sim_weib),  col="purple",      lty="twodash",  lwd=3)
lines(density(sim_lnorm), col="deepskyblue", lty="dotdash",  lwd=3)
lines(density(sim_norm),  col="gold2"      , lty="dotted",   lwd=3)
legend(600,0.012, legend=c("Raw data", "Gamma fit", "Weibull fit", "Lognormal fit", "Normal fit"),
col=c("black", "seagreen2", "purple", "deepskyblue", "gold2"), lty=c("solid", "longdash",
"twodash","dotdash","dotted"), lwd=c(2,2,2,2,2), cex=1.2)
dev.off()
```

Appendix J
Results: “noisy” contour plots

The plots below provide “noisy” versions of each of the base case contour plots reported in the simulated diagnostic accuracy results sections (sections 5.3.2.2.1 and 5.4.2.2.1). These relate to the raw simulation results, in which the smoothing algorithm applied within the base case contour plots has been removed (i.e. sensitivity analyses 1.7 and 2.8 as described in Table 5-3; and sensitivity analyses 1.8 and 2.9 as described in Table 5-9).

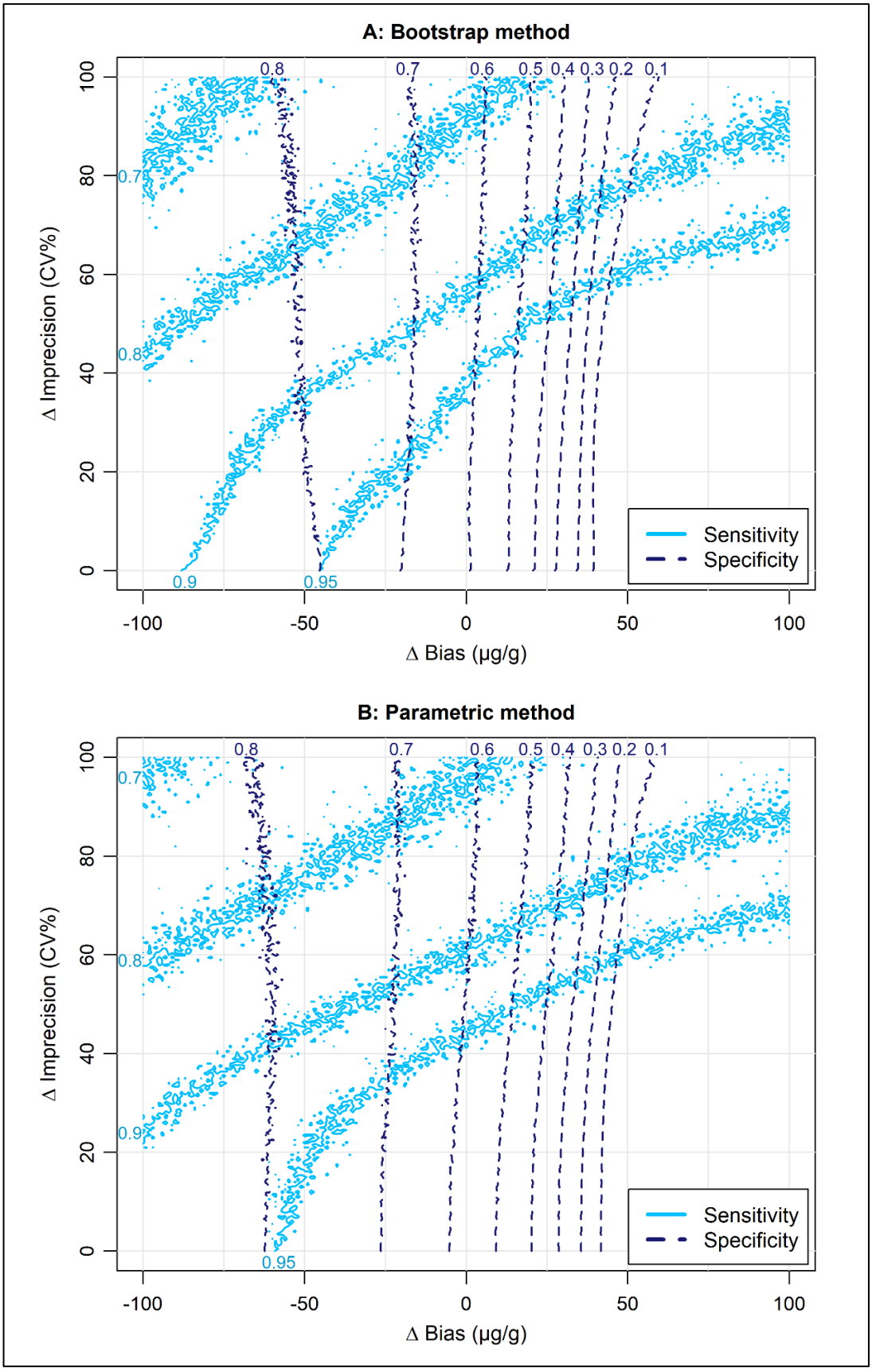


Figure J-1. NICE FC pathway: diagnostic accuracy contour plots (no smoothing algorithm applied)

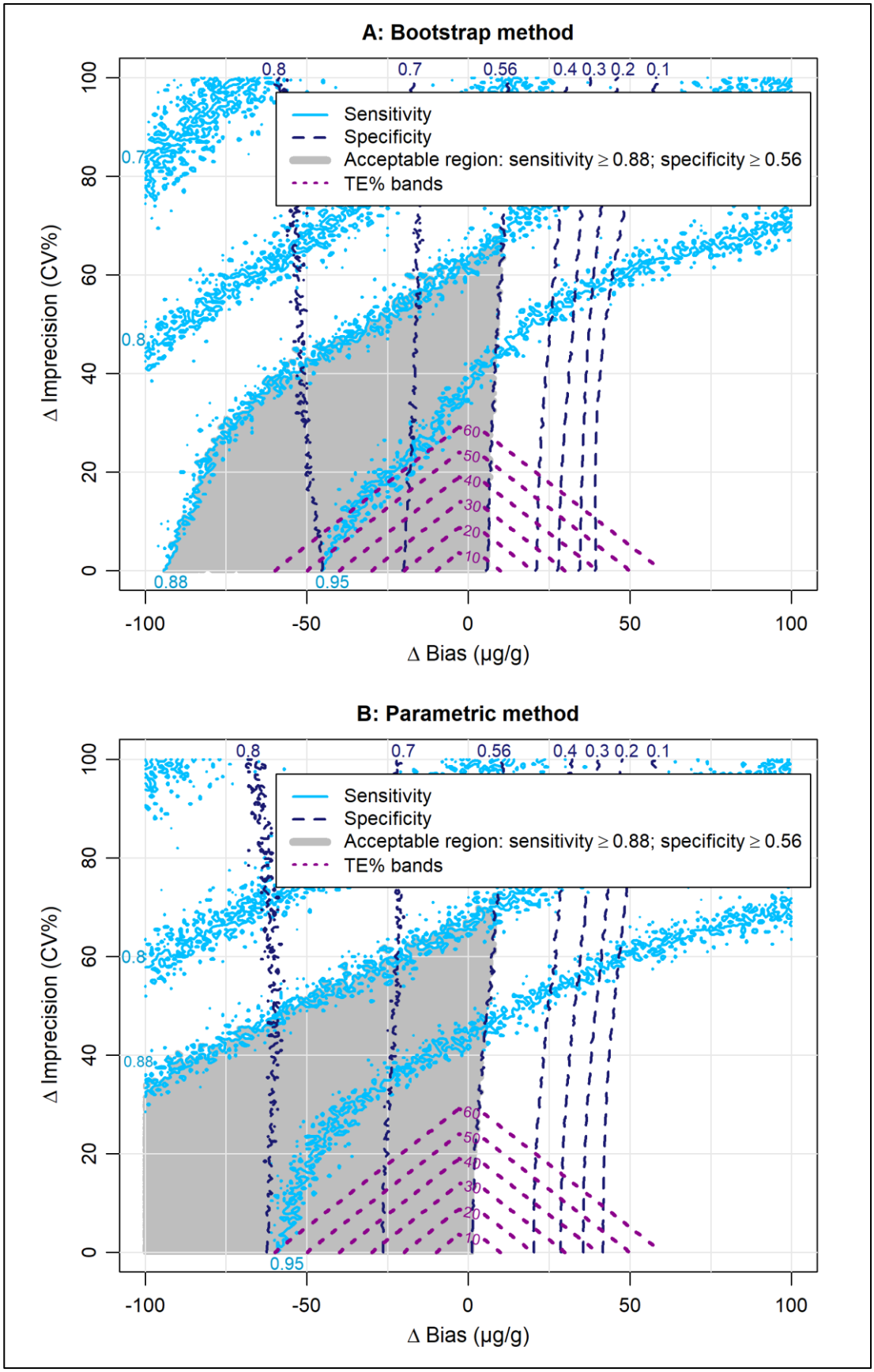


Figure J-2. NICE FC pathway: contour plots showing acceptable region (for diagnostic accuracy requirement: sensitivity ≥0.88, specificity ≥0.56) and TE bands (no smoothing algorithm applied)

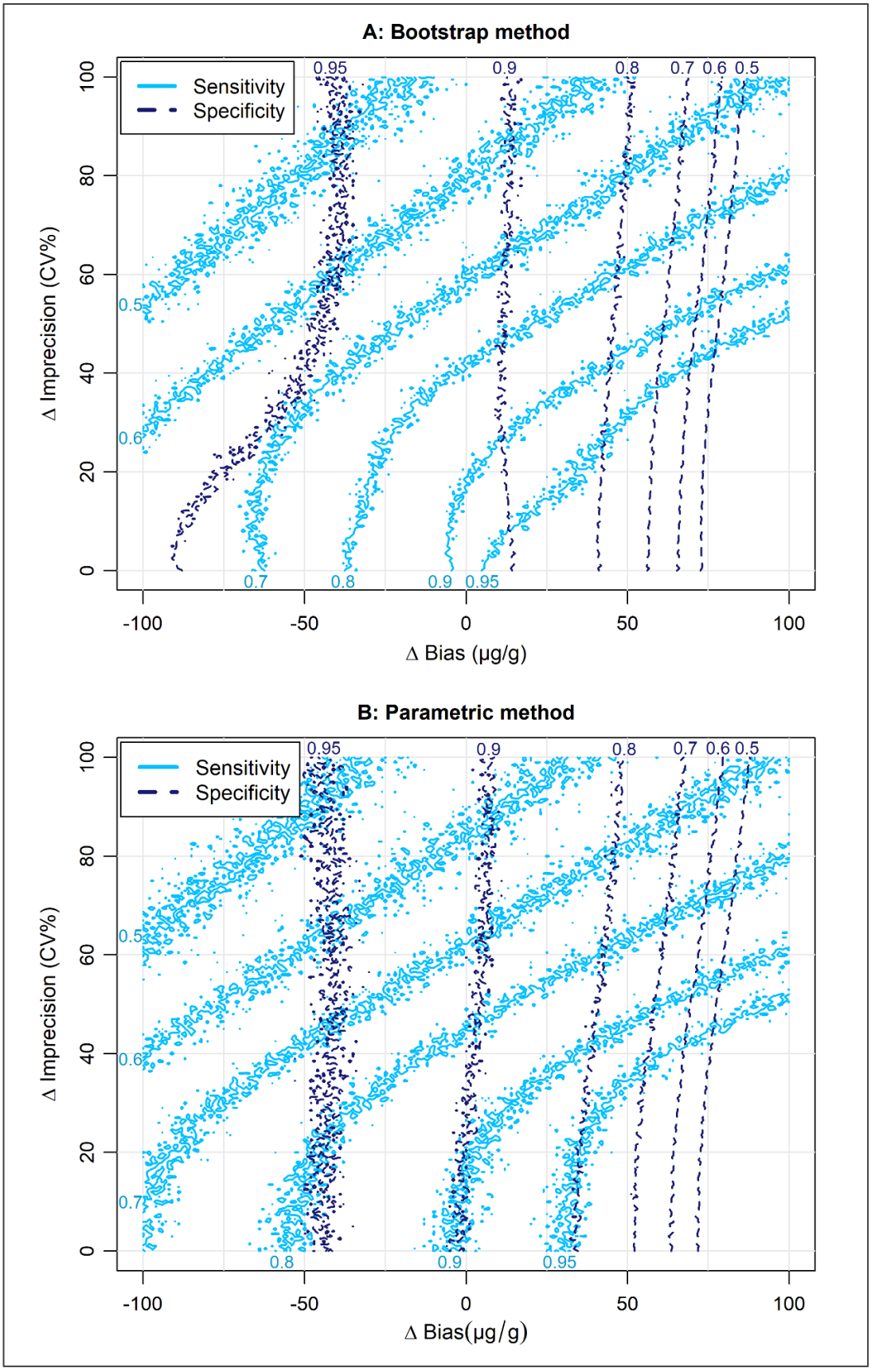


Figure J-3. YFCCP: diagnostic accuracy contour plots (no smoothing algorithm applied)

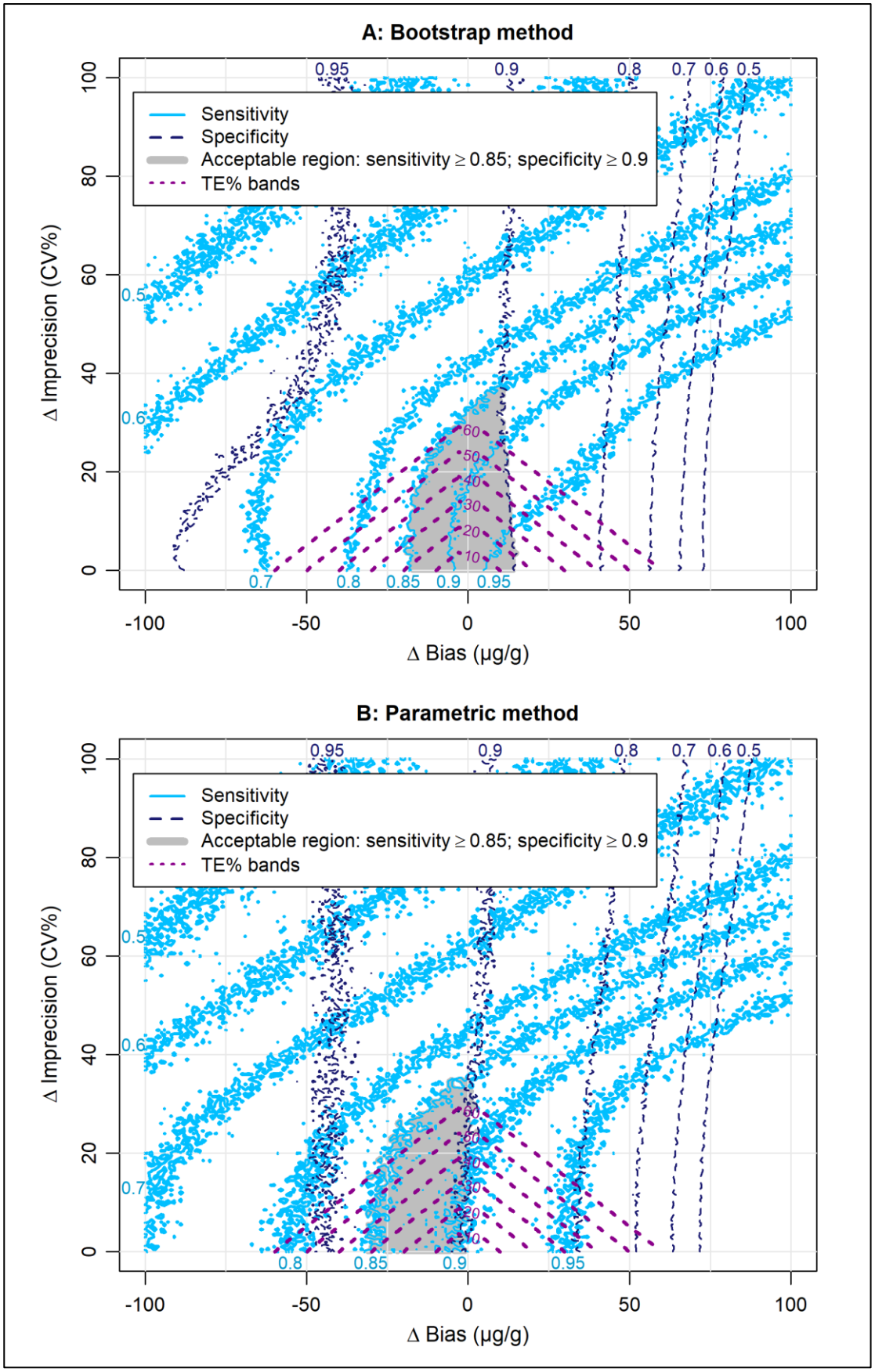


Figure J-4. YFCCP: contour plots showing acceptable region (for diagnostic accuracy requirement: sensitivity ≥0.85, specificity ≥0.9) and TE bands (no smoothing algorithm applied)

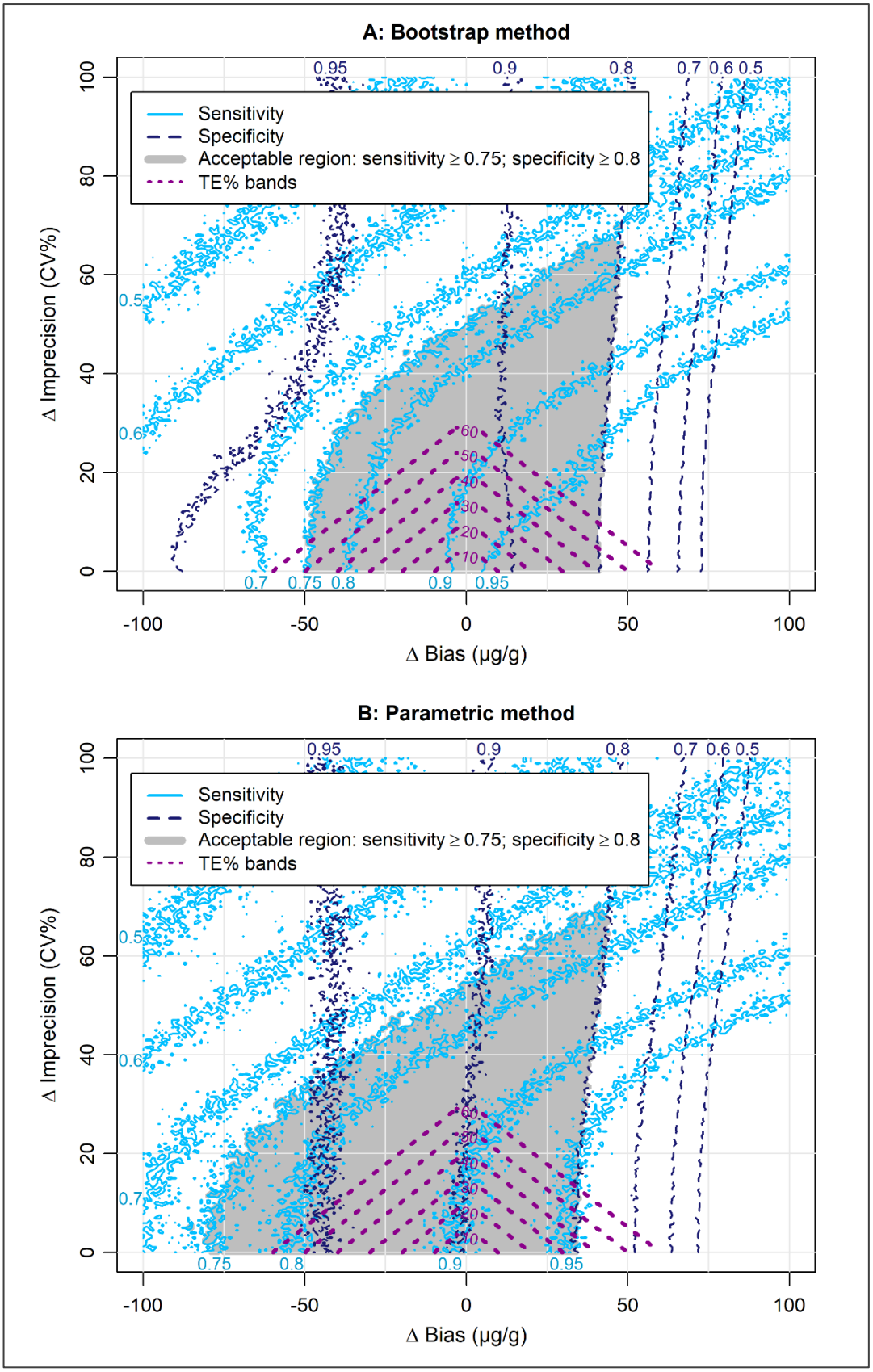


Figure J-5. YFCCP: contour plots showing acceptable region (for diagnostic accuracy requirement: sensitivity ≥0.75, specificity ≥0.8) and TE bands (no smoothing algorithm applied)

Appendix K

FC_{diff} distributions

K.1 Bootstrap sampling method

Figure K-1 presents the FC_{diff} distributions for the IBS and IBD populations, based on FC_{diff} values calculated from the YFCCP dataset FC1 and FC2 values [FC_{diff} = (FC2-FC1)/FC1)]. These empirical distributions were used in the bootstrap method FC_{diff} sensitivity analysis described in section 5.4.1.2.1.

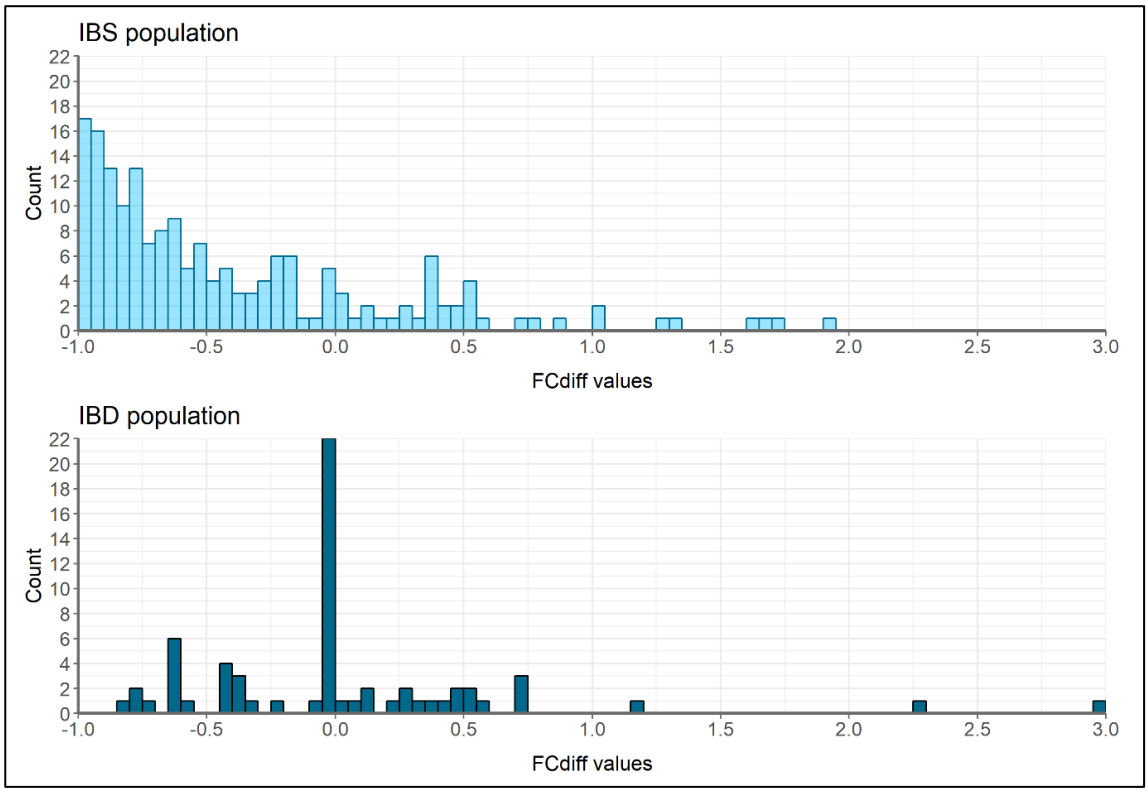


Figure K-1. FC_{diff} count plots

K.2 Parametric sampling method

Figure K-2 and Figure K-3 show the FC_{diff} density plots based on 10,000 draws from each of the parametric distributions explored within the FC_{diff} sensitivity analysis described in section 5.4.1.2.2. Note that, whereas Figure K-1 above reports the FC_{diff} empirical values on their natural scale (spanning from -1 to infinity), Figure K-2 and Figure K-3 report the parametric distributions fitted to the adjusted FC_{diff} values (i.e. in which +1 has been applied to all of the FC_{diff} values).

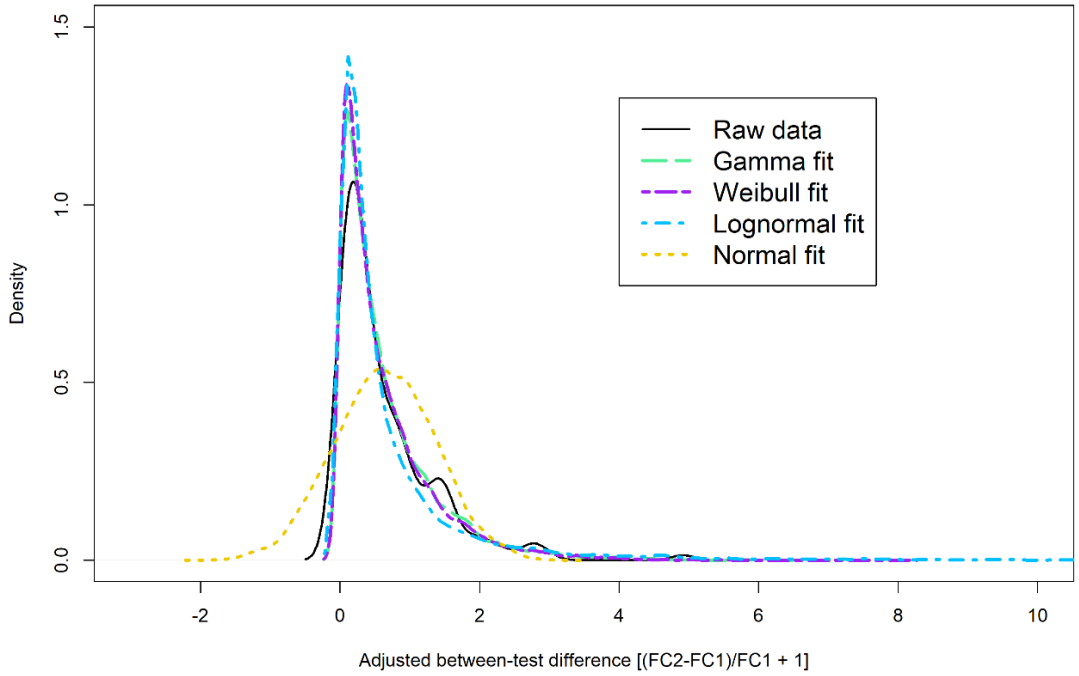


Figure K-2. IBS FC_{diff} values: density plot for parametric distribution fits

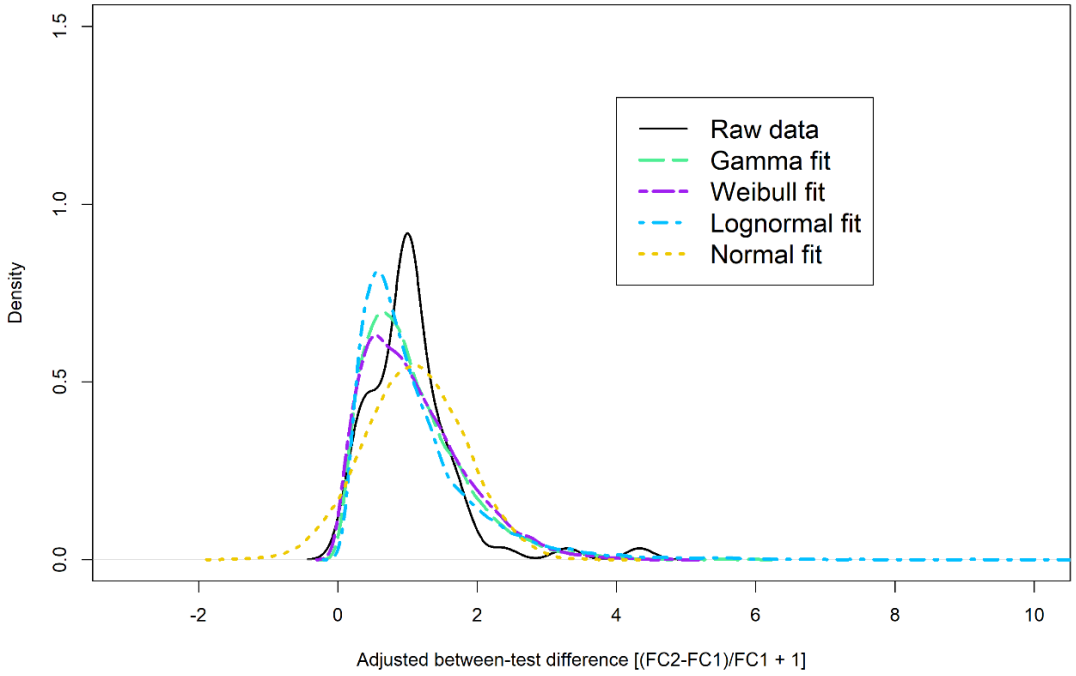


Figure K-3. IBD FC_{diff} values: density plot for parametric distribution fits

Appendix L

FC cost-utility model parameters

Table L-1. FC cost-utility model: model parameters

Parameter	Value in model	Source
Global parameters		
Time horizon	1 year	-
Discount rate	NA	-
IBD prevalence	8%	YFCCP dataset
Diagnostic accuracy		
YFCCP sensitivity	94%	YFCCP dataset
YFCCP specificity	92%	
No FC (Tibble data) sensitivity	35%	Tibble *et al*. (2002) (218)
No FC (Tibble data) specificity	73%	
No FC (NICE data) sensitivity	100%	Waugh *et al.* (2013) (220)
No FC (NICE data) specificity	79%	
YFCCP (50 µg/g cut-off applied) sensitivity [NICE FC Pathway]	96%	YFCCP dataset
YFCCP (50 µg/g cut-off applied) specificity [NICE FC Pathway]	60%	YFCCP dataset
FC testing (Tibble data) sensitivity	90%	Tibble *et al.* (2002) (218)
FC testing (Tibble data) specificity	80%	
FC testing (NICE data) sensitivity	93%	Waugh *et al.* (2013) (220)
FC testing (NICE data) specificity	94%	
Unit Costs		
GP visit	£37	PSSRU, unit costs of health and social care 2018 (309)
Calprotectin test	£24	NICE MIB 132 (costs inflated to 2017/18 using PSSRU inflation index) (310)
Specialist visit	£155	NHS reference costs 2017-18, consultant led gastroenterology outpatient attendance (311)

Colonoscopy	£264	NHS Reference costs 2017-18, Outpatient colonoscopy without biopsy (311)
IBS first line medication	£22	Drug tariff, July 2017
IBS second line medication	£77	Calculated from Drug tariff, July 2017
ESR + CRP test (positive test)	£5.85	NICE MIB 132 (Costs inflated to 2017/18 using PSSRU inflation index) (310)
ESR + CRP test (negative test)	£9.28	NICE costing template for Faecal calprotectin, https://www.nice.org.uk/guidance/dg11/resources
Utilities		
Untreated IBS	0.68	NICE CG61, Appendix G (214)
Treated IBS	0.81	NICE CG61, Appendix G (214)
Untreated Crohn's Disease	0.61	NICE CG152, Appendix H (since updated by NICE NG129) (215)
Treated Crohn's Disease	0.88	NICE CG152, Appendix H (since updated by NICE NG129) (215)
Untreated Ulcerative Colitis	0.32	NICE TA163 (2008) (312)
Treated Ulcerative Colitis	0.79	NICE TA163 (2008) (312)
Proportion IBD Crohn's Disease	39%	Turvill et al. 2018 (239)
Timings (days)		
YFCCP intervention arm		
GP visit (initial)	14	Clinical advice
Calprotectin test (initial)	2	Clinical advice
GP follow-up (for positive screening)	7	Clinical advice
GP follow-up (for negative screening)	18	Clinical advice
Calprotectin re-test (for positive screening)	18	Turvill et al. (2018) (239)
Calprotectin re-test (for negative screening)	0	Turvill et al. (2018) (239)
Specialist visit (for positive screening)	0	Clinical advice
Colonoscopy (for positive screening)	25	Turvill et al. (2018) (239)
GP visit (for negative screening, unresolved symptoms)	30	Clinical advice
Specialist visit (for negative screening, following unresolved symptoms)	42	Clinical advice

Colonoscopy (for negative screening, following unresolved symptoms)	42	Clinical advice
No FC comparator arms (where different to intervention)		
ESR + CRP test (for No FC testing comparators)	1	Clinical advice
Specialist visit (for positive screening, following GP referral)	21	Clinical advice
Colonoscopy (for positive screening, following GP referral)	39	Clinical advice
GP follow-up (following negative GP assessment)	14	
Proportions		
FC arms only: calprotectin re-test (for negative screening)	20%	Clinical advice
IBS medication (for negative screening)	50%	Clinical advice
Returning to GP with unresolved symptoms (for negative screening, true negatives)	20%	YFCCP dataset
Returning to GP with unresolved symptoms (for negative screening, false negatives)	100%	YFCCP dataset
Second line IBS medication (for negative screening, true negatives)	13%	Clinical advice, 65% of patients with unresolved symptoms after first line will be prescribed second line IBS medication
Second line IBS medication (for negative screening, false negatives)	33%	Clinical advice
Additional GP visit (for negative screening, true negatives)	0%	Assumption
Additional GP visit (for negative screening, false negatives)	100%	Clinical advice
Additional specialist visit (for negative screening, true negatives)	7.5%	YFCCP dataset
Additional specialist visit (for negative screening, false negatives)	100%	Clinical advice
Colonoscopy (for negative screening, true negatives)	2.9%	Clinical advice - 38% of patients referred with a negative screening for IBD will have colonoscopy
Colonoscopy (for negative screening, false negatives)	100%	Clinical advice

Appendix M

Example EQA report from the UK NEQAS EQA scheme for FC

An example EQA report from the UK NEQAS EQA scheme for FC (run by the Birmingham Quality group) is provided below. This is an anonymised version of the EQA report (i.e. all laboratory-identifiable data has been removed), which was produced for distribution #167 (distributed in May 2018) and was downloaded from the Birmingham Quality website in December 2019 (230).

Quality Manager
Pathology Laboratory
Biochemistry Department
Town
County
Country

This Scheme is essentially web-based. We can alert you to information regarding the Scheme via email. The e-mail address (or addresses) we are currently using to contact your laboratory is shown below in red . If no e-mail address is displayed or the information shown is incorrect, please email us with an appropriate contact e-mail address as soon as possible, using the word 'feedback' in the title line.

Based on the date information you have provided, the transit time from specimen dispatch [ ] to receipt [ ] was day(s), and the subsequent time to analysis [ ] in your laboratory was day(s). (Missing values indicate dates not provided. "0 days" represents same day).

Any comments you made to us are shown below and have been acted upon where necessary

Any specific comments applicable only to laboratory are shown below

Any general comments applicable to all laboratories are shown below

We are very grateful to Professor Tariq Iqbal and Dr Amanda Rossiter of the University Hospitals NHS Foundation Trust in Birmingham and Professor Matthew Brookes of the Royal Wolverhampton Hospitals NHS Trust for their assistance with this EQA programme.

We must also thank our donors for providing the invaluable specimens distributed through this programme.

Report authorised on Thursday 10 May 2018 by:

Jane French
Programme Director, UK NEQAS for Faecal Markers of Inflammation

Birmingham Quality is part of the University Hospitals Birmingham NHS Foundation Trust and provides this UK NEQAS service from PO Box 3909, Birmingham B15 2UE, UK
To contact us, email birminghamquality@uhb.nhs.uk or phone us on +44 (0)121 414 7300

www.birminghamquality.org.uk

Published at 15:51 on Friday 11 May 2018

Analytical Performance over the last 6 months (rolling time window of 6 distributions)

All our time periods are 'rolling' to give you current information.
You may wish to keep you own log of Calendar Year or Financial Year time points if you require 'year-end' statements for your own internal use.
Any analytes with out of consensus performance will be highlighted in red and can be clicked for further details.

You have out of consensus performance for:	*None*
You have in consensus performance for:	Calprotectin
You have no performance data for:	*None*

Participation and Return Rates

This scheme cycle is notionally every four weeks.
Analytically, we assess you over a six month time window (6 Distributions).
For return rates, late and amended results we assess you over a twelve month period (12 distributions).

	Distributions	Rating	Affected Distributions
Participation	12 distributions out of a possible 12	Satisfactory	
Late Returns	0 distributions from the last 12	Satisfactory	
Amendments	0 distributions accepted from the last 12	Satisfactory	

Analytical Performance for specimens from distribution 167 only

You can judge, in association with your IQC and other QA measures, if your current performance is a blip or part of a trend.

Out of consensus for at least one specimen for:	*None*
In consensus for all specimens for:	Calprotectin
You have no specimen %bias etc. for:	*None*
You are not registered for:	Lactoferrin (quantitative)

If your laboratory is outside of the acceptable limits of performance for any its rolling time-window scores (A, B or C scores), this will be indicated by a red traffic light symbol. It is the responsibility of the laboratory to undertake an internal investigation to establish the underlying cause and put in place corrective and preventive action. Please do not wait to receive a formal notification of performance from the Scheme Organiser or the National Quality Assurance Advisory Panel (NQAAP) before logging the non-conformity and, where necessary, acting upon the data contained in your report. A green traffic light merely reflects that your laboratory is performing as well as the state-of-the-art allows; it does not necessarily mean that your assay / laboratory performance is good enough clinically.

No rolling-time window scores or trend data will be calculated or appear on your report unless you have returned numerical results for at least 7 specimens during the 6-distribution time window.
As many labs often report less than (<) or greater than (>) results, it may not always be possible to calculate rolling time-window scores for them.
We are currently working on a way to accomodate all Participants, but the 'Concentration versus Interpretation Plots' highlight our dilemma!

	Specimen	Pool	Result	Target	Specimen %bias	B score	C score	B	C
Calprotectin (ug/g)	167A	370	<19.5			-1.2	22.1	● ↗	● ↔
	167B	396	118.3	102	+15.7 ◆				
	167C	394	42.8	37.1	+15.4 ◆				
Calprotectin interp (N, E or P)	167A	370	N	N					
	167B	396	E	E					
	167C	394	N	N					

For the time being, the All-laboratory Trimmed Mean (ALTM) will be quoted as the target value for all quantitative analytes in the programme. As we gather more data from all of the available kits and methods, the choice of target value will be re-assessed.

Please do not be too disappointed if your results appear to be a long way from the ALTM. For Calprotectin (or indeed any antibody-based assay), differences in the numerical results obtained by different manufacturer's kits are to be expected. Please do check though, that your laboratory: 1) produces results in keeping with other users of your method and 2) consistently displays a similar Specimen %bias at similar concentrations of calprotectin. The 6-distribution summary table contained in your report should help with 2.

All laboratories (in collaboration with their users/clinicians) should choose (and independently validate) a kit that is fit for their purpose and not worry about being 'different' on UK NEQAS reports.

Distribution 168 will be dispatched on 14/05/2018 . Results are due back in Birmingham by, notionally, 23:59 on 10/06/2018.

Where your results appear as "XPL", it is because you did not report a numerical value for that analyte, but you did provide an explanation as to why a result was not reported.

You have until the close of Distribution 168 (10/06/2018) to submit late results / request amendments to results for Distribution 167.
Please enter any late results or requests to amend non-analytical errors for Distribution 167 on the web under the usual Results button ensuring the correct Distribution number has been selected. You should include your name and a valid reason.
Amending results is at the discretion of the Director and is not an automatic entitlement.
We are not accepting any further requests to amend results or to change methods for Distribution 166.

Spec.	Pool	Pool description / Treatments / Additions
167A	370	Patient with IBD*
167B	396	Patient with IBD*
167C	394	Patient with IBD*

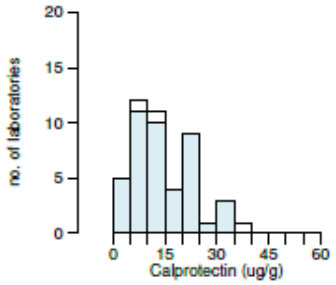
☐ All methods
☐ ELISA
☐ Calpro (ALP) [2CP]

Your B score is -1.2
Your C score is 22.1

The B limit is +/- 75.0
The C limit is 75.0

Specimen : 167A

	n	Mean	SD	CV(%)
All methods [ALTM]	46	14.7	9.8	66.4
ELISA	43	14.4	9.4	65.2
Buhlmann [2BU]	12	23.0	7.9	34.5
Diasorin [2IN]	4	6.3		
Immundiagnostik (K6927)	5	20.3	4.0	19.6
Thermo EliA Calpro 2 [2KO2]	13	10.3	3.7	35.8
Thermo EliA [2KO]	2	18.5		
Chemiluminescence	1	15.0		
non-numeric results	47			



Your result <19.5
Target value
()

Your Interpretation N
Target N

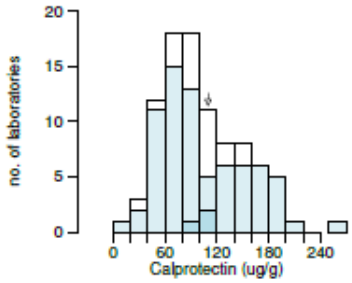
Standard Uncertainty

Your specimen:
%bias
Accuracy Index

Method mean
(Calpro (ALP) [2CP])

Specimen : 167B

	n	Mean	SD	CV(%)
All methods [ALTM]	92	102	47	46.4
ELISA	72	104	53	50.6
Buhlmann [2BU]	24	153	39	25.2
Calpro (ALP) [2CP]	3	110		
Diasorin [2IN]	6	80.8	5.2	6.4
Immundiagnostik (K6927)	7	135	40	29.8
Thermo EliA Calpro 2 [2KO2]	17	70.4	26.0	36.9
Thermo EliA [2KO]	6	75.0	12.8	17.0
Immuno turbidimetric	10	106	17	16.2
Buhlmann fCAL turbo [4BU]	10	106	17	16.2
Chemiluminescence	5	74.3	29.9	40.3



Your result 118.3
Target value 102
(ALTM)

Your Interpretation E
Target E

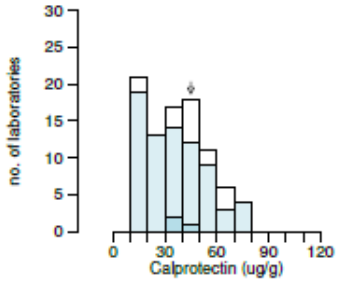
Standard Uncertainty 6.6

Your specimen:
%bias +15.7
Accuracy Index

Method mean 110
(Calpro (ALP) [2CP])

Specimen : 167C

	n	Mean	SD	CV(%)
All methods [ALTM]	90	37.1	19.2	51.8
ELISA	74	35.3	18.9	53.6
Buhlmann [2BU]	24	52.2	12.1	23.1
Calpro (ALP) [2CP]	3	39.6		
Diasorin [2IN]	7	23.9	12.6	52.6
Immundiagnostik (K6927)	7	51.0	19.7	38.7
Thermo EliA Calpro 2 [2KO2]	19	25.9	10.4	40.2
Thermo EliA [2KO]	6	19.0	4.3	22.4
Immuno turbidimetric	9	43.5	5.6	12.8
Buhlmann fCAL turbo [4BU]	9	43.5	5.6	12.8
Chemiluminescence	2	26.5		
non-numeric results	5			



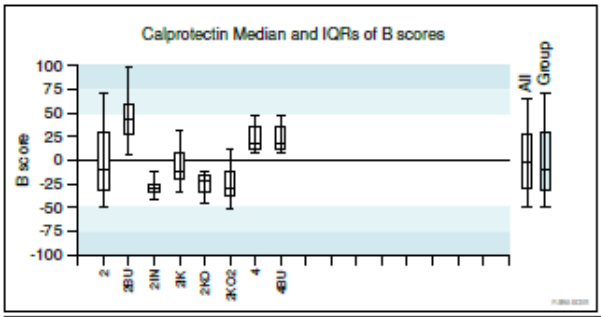
Your result 42.8
Target value 37.1
(ALTM)

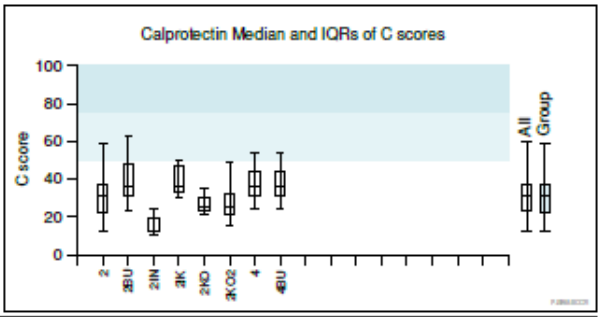
Your Interpretation N
Target N

Standard Uncertainty 2.7

Your specimen:
%bias +15.4
Accuracy Index

Method mean 39.6
(Calpro (ALP) [2CP])



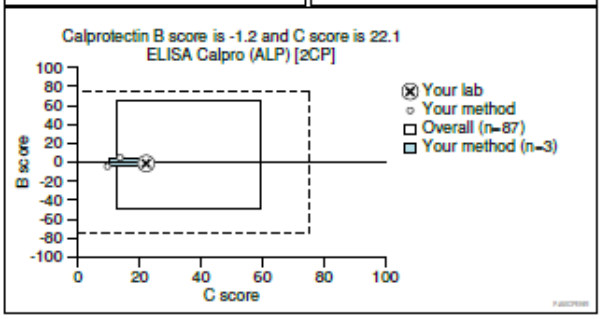


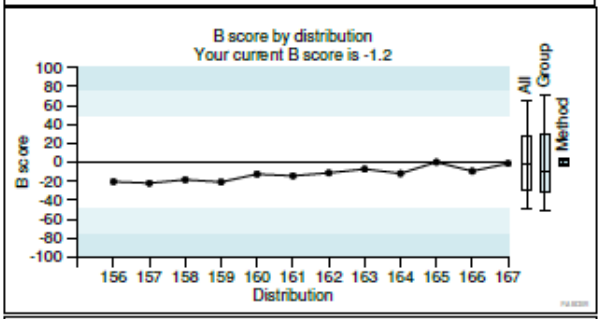
*It is not known if patients were in remission or having a flare-up and/or on medication at the time the specimen was produced.
We have excluded the Calprotectin results for Specimen 167A from the calculation of the rolling-time window scores (low analyte concentration).

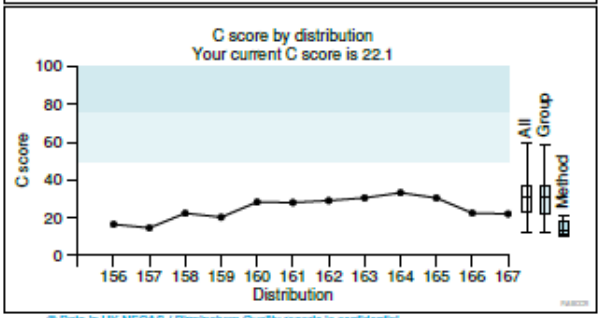
Birmingham Quality is part of the University Hospitals Birmingham NHS Foundation Trust and provides this UK NEQAS service from PO Box 3909, Birmingham B15 2UE, UK
To contact us, email birminghamquality@uhb.nhs.uk or phone us on +44 (0)121 414 7300

Published at 15:51 on Friday 11 May 2018

Pool (exclusion) [Type]	Distribution 162 12-Nov-2017 result	target	%bias	Distribution 163 10-Dec-2017 result	target	%bias	Distribution 164 28-Jan-2018 result	target	%bias	Distribution 165 04-Mar-2018 result	target	%bias	Distribution 166 08-Apr-2018 result	target	%bias	Distribution 167 06-May-2018 result	target	%bias
(370)																(<19.5)		
(404)				(<19.5)														
(401)	(<19.5)	**17.7**																
400							<19.5	**28.1**										
390										37	**33.7**	+9.7						
394																42.8	**37.1**	+15.4
379							23.5	**38.8**	-39.5									
395													83.4	**96.8**	-13.8			
396																118.3	**102**	+15.7
(405)				(144.4)	**127**	(+13.3)												
(402)	(139.9)	**143**	(-2.4)															
388										200	**160**	+25.4						
380							157.6	**207**	-23.8									
381													195.9	**229**	-14.4			
(406)				(273.5)	**255**	(+7.1)												
(403)	(252.8)	**270**	(-6.3)															
392													302.5	**315**	-4.0			
399										568	**538**	+5.6						
Method mean	2CP			2CP			2CP		-31.6	2CP		+13.6	2CP		-10.7	2CP		+15.6
B score	-11.2			-7.2			-11.9			+0.2			-9.4			-1.2		
C score	29.1			30.6			33.3			30.6			22.6			22.1		





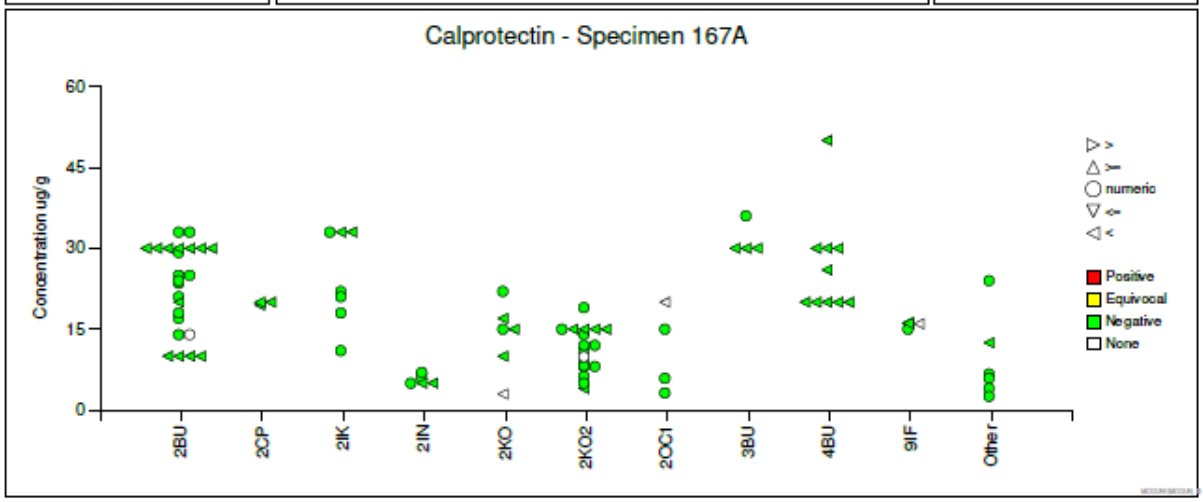


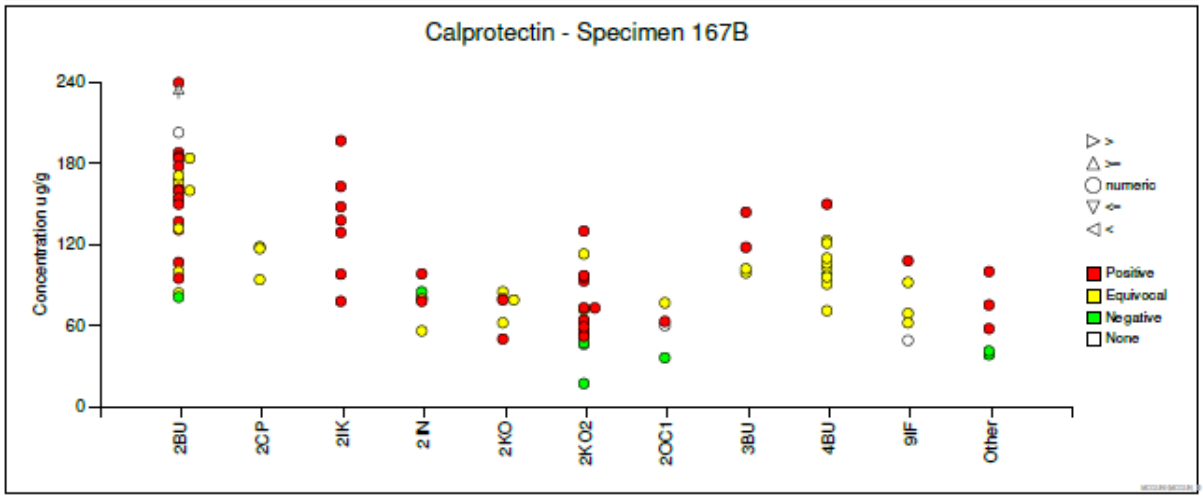
Published at 15:51 on Friday 11 May 2018

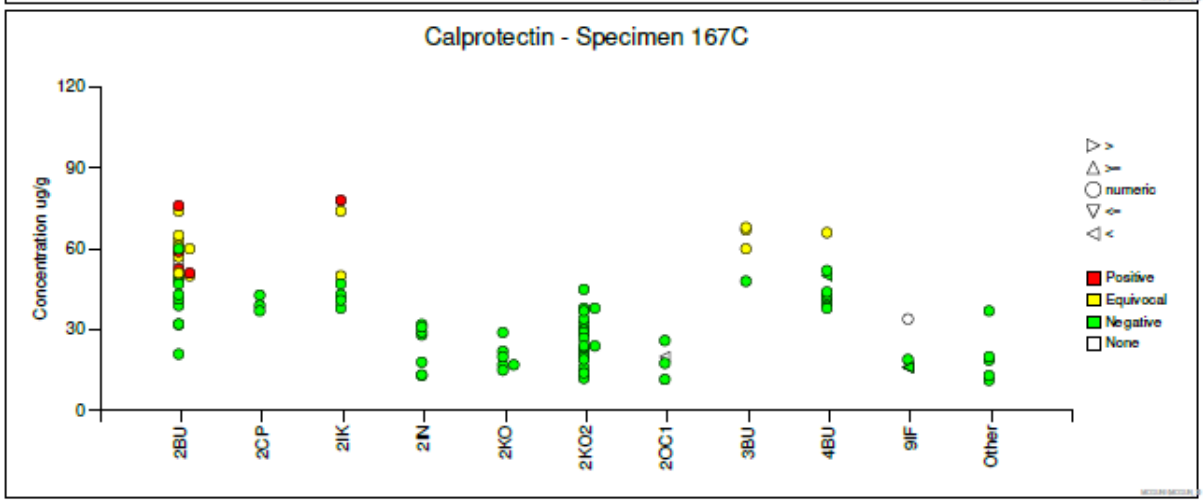
		167A			167B			167C		
	n	Mean	SD	CV(%)	Mean	SD	CV(%)	Mean	SD	CV(%)
All methods [ALTM]	92	14.7	9.8	66.4	102	47	46.4	37.1	19.2	51.8
ELISA	74	14.4	9.4	65.2	104	53	50.6	35.3	18.9	53.6
Accusay [2AY]	4	4.2			49.3			18.9		
Buhlmann [2BU]	24	23.0	7.9	34.5	153	39	25.2	52.2	12.1	23.1
Calpro (ALP) [2CP]	3				110			39.6		
Diasorin [2IN]	7	6.3			80.8	5.2	6.4	23.9	12.6	52.6
Immundiagnostik (K6927)	7	20.3	4.0	19.6	135	40	29.8	51.0	19.7	38.7
Inova [2IF]	1	24.0			100			37.0		
Orgentec Alegria [2OC1]	4	8.0			61.5			18.4		
Thermo EliA Calpro 2 [2KO2]	19	10.3	3.7	35.8	70.4	26.0	36.9	25.9	10.4	40.2
Thermo EliA [2KO]	6	18.5			75.0	12.8	17.0	19.0	4.3	22.4
Lateral flow	4	36.0			110			63.5		
Quantum Blue [3BU]	4	36.0			110			63.5		
Immuno turbidimetric	10				106	17	16.2	43.5	5.6	12.8
Buhlmann fCAL turbo [4BU]	10				106	17	16.2	43.5	5.6	12.8
Chemiluminescence	5	15.0			74.3	29.9	40.3	26.5		
Inova QUANTA Flash [9IF]	5	15.0			74.3	29.9	40.3	26.5		
Not stated, please specify	1	6.7			38.3			13.0		
non-numeric results	47									

		B score			C score		
	n	Median	Interquartile range		Median	Interquartile range	
All methods	87	-2.0	-29.9	+28.5	31.3	23.2	37.4
Chemiluminescence	3	-32.4	-37.9	-21.6	28.3	24.8	31.4
Inova QUANTA Flash [9IF] 9IF	3	-32.4	-37.9	-21.6	28.3	24.8	31.4
ELISA	72	-10.0	-30.9	+30.4	31.2	22.1	37.3
Accusay [2AY] 2AY	4	-50.7	-58.8	-43.9	17.0	15.7	17.3
Buhlmann [2BU] 2BU	25	+43.6	+27.2	+59.6	36.0	31.5	47.9
Calpro (ALP) [2CP] 2CP	3	-1.2	-2.7	+1.8	13.7	11.6	17.9
Diasorin [2IN] 2IN	6	-29.1	-33.7	-25.1	12.2	12.0	19.4
Immundiagnostik (K6927) 2IK	7	-12.3	-20.1	+7.9	36.3	33.4	46.8
Inova [2IF] 2IF	1	-11.3	-11.3	-11.3	36.2	36.2	36.2
Orgentec Alegria [2OC1] 2OC1	4	-35.6	-44.0	-22.7	39.3	26.8	53.3
Thermo EliA Calpro 2 [2KO2] 2KO2	15	-28.9	-37.0	-11.3	25.6	21.0	31.9
Thermo EliA [2KO] 2KO	7	-21.3	-34.0	-16.0	25.4	23.5	30.0
Immuno turbidimetric	9	+17.9	+11.2	+35.1	36.1	31.0	44.0
Buhlmann fCAL turbo [4BU] 4BU	9	+17.9	+11.2	+35.1	36.1	31.0	44.0
Lateral flow	3	+18.3	+16.7	+20.2	32.7	30.4	54.4
Quantum Blue [3BU] 3BU	3	+18.3	+16.7	+20.2	32.7	30.4	54.4

Published at 15:51 on Friday 11 May 2018







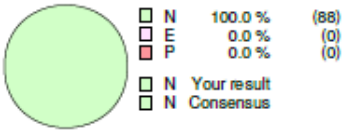
Birmingham Quality is part of the University Hospitals Birmingham NHS Foundation Trust
and provides this UK NEQAS service from PO Box 3909, Birmingham B15 2UE, UK
To contact us, email birminghamquality@uhb.nhs.uk or phone us on +44 (0)121 414 7300

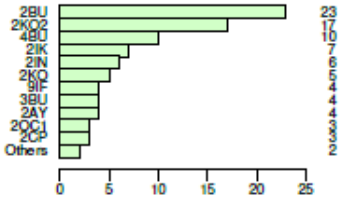
www.birminghamquality.org.uk
Published at 15:51 on Friday 11 May 2018

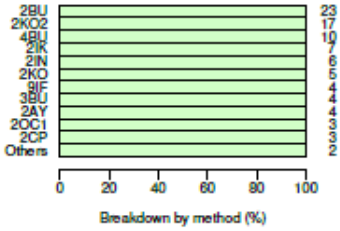
Spec.	Pool	Pool description / Treatments / Additions
167A	370	Patient with IBD*
167B	396	Patient with IBD*
167C	394	Patient with IBD*

Based on the patient's clinical details and numerical calprotectin results:
Sample 167A was designated 'Negative'; the consensus was also Negative.
Sample 167B was designated 'Equivocal'; there was no overall consensus.
Sample 167C was designated 'Negative'; the consensus was also Negative.

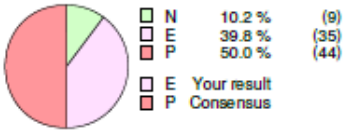
Specimen : 167A

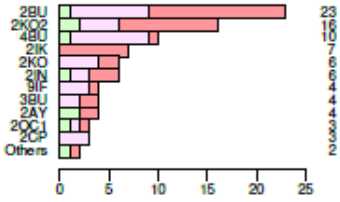


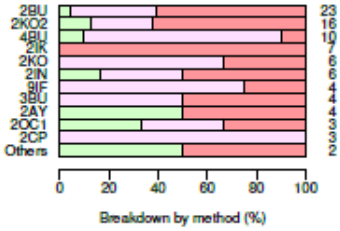




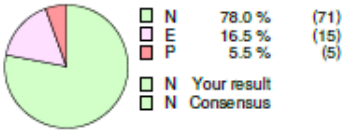
Specimen : 167B

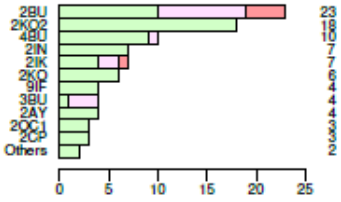


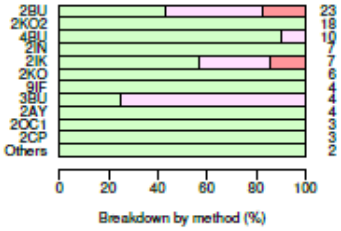




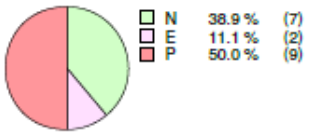
Specimen : 167C





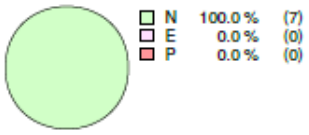


Specimens distributed in each category

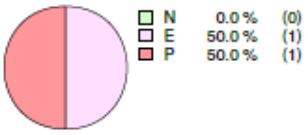


Your interpretation for each category

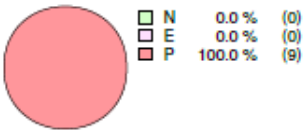
Negative



Equivocal



Positive



Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167A

Kit	Result	Interpretation	Comment
Accusay [2AY]	<12.5	N	
Accusay [2AY]	XPL		
Accusay [2AY]	2.575	N	Negative
Accusay [2AY]	4.028	N	
Accusay [2AY]	5.9	N	
Buhlmann [2BU]	<10	N	Inflammatory bowel disease unlikely.
Buhlmann [2BU]	<10	N	No evidence of GI inflammation - suggestive of IBS.
Buhlmann [2BU]	<10	N	
Buhlmann [2BU]	<10	N	
Buhlmann [2BU]	<20	N	Faecal calprotectin results less than 50 ug/g, inflammatory Bowel Disease is unlikely if patient was symptomatic.
Buhlmann [2BU]	<30	N	Faecal calprotectin within reference range. Unlikely IBD, treat as IBS. Watery stool sample; interpret with caution. If result is not consistent with the clinical picture, suggest repeat on a formed stool sample.
Buhlmann [2BU]	<30	N	A calprotectin result of less than 50 ug/g effectively excludes active inflammatory bowel disease as a cause of symptoms. False negatives can occur (diagnostic sensitivity = 93% [85 - 97]. Normal results can also be seen in circumstances such as lymphocytic colitis, coeliac and stable diverticulosis. If symptoms persist, suggest investigate further or refer as appropriate.
Buhlmann [2BU]	<30	N	
Buhlmann [2BU]	<30	N	Reference Range < 50 ug/g Values <50 ug/g not indicative of inflammation in the gastrointestinal tract. Patient samples with low levels are likely not to be in need of further invasive procedures to determine the inflammation cause.
Buhlmann [2BU]	<30	N	Faecal calprotectin within reference range ?IBS
Buhlmann [2BU]	<30	N	Within reference range. ?IBS. Results <200ug/g are rarely associated with significant pathology. Please see GG&C guidelines for advice and secondary care referral. www.nhsggc.org.uk/media/236675/ggc_fc_guidelines_dec_2015.doc
Buhlmann [2BU]	<30	N	Inflammation of the gastrointestinal tract is unlikely when calprotectin values are <50 ug/g.
Buhlmann [2BU]	14		IBD is unlikely. In patients with symptoms suggestive of IBD a faecal calprotectin level <50ug/g has a negative predictive value of 98%.
Buhlmann [2BU]	14	N	Calprotectin <100: Likely IBS
Buhlmann [2BU]	17	N	No intestinal inflammation detected
Buhlmann [2BU]	18	N	
Buhlmann [2BU]	21	N	
Buhlmann [2BU]	23.5	N	
Buhlmann [2BU]	24	N	IBD is unlikely. In patients with symptoms suggestive of IBD, a faecal calprotectin of <50 ug/g has a negative predictive value of 98%
Buhlmann [2BU]	25	N	No evidence of GI inflammation - suggestive of IBS
Buhlmann [2BU]	25	N	Result of FC <50 ug/g: This result effectively rules out inflammatory bowel disease (IBD). As patient has already met the criteria for irritable bowel syndrome (IBS) and have been refractory to standard treatments please refer to GHNHSFT Complex IBS Service.
Buhlmann [2BU]	29.2	N	<50ug/g Low risk of IBD
Buhlmann [2BU]	33	N	No intestinal inflammation detected. NOTE: No NSAIDs 4 weeks prior to testing.Not to be used for cases of suspected bowel cancer.(NICE DG11 and NICE CG12
Buhlmann [2BU]	33	N	
Buhlmann fCAL turbo [4BU]	<20	N	<50 ug/g :Normal
Buhlmann fCAL turbo [4BU]	<20	N	Not indicative of GI inflammation
Buhlmann fCAL turbo [4BU]	<20	N	Assuming request from a GP: Not indicative of GI inflammation. Please refer to primary care pathway for investigation of chronic diarrhoea.
Buhlmann fCAL turbo [4BU]	<20	N	FCP <100 ug/g suggests GI symptoms are non-inflammatory (e.g. IBS) Suggest manage symptomatically. Please note above are coded comments and not free text comments.

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167A

Buhlmann fCAL turbo [4BU]	<20	N	Initial calprotectin <100ug/g: Likely IBS. Repeat calprotectin <100ug/g: Likely IBS.
Buhlmann fCAL turbo [4BU]	<30	N	FC <100ug/g-IBD unlikely, primary care management.
Buhlmann fCAL turbo [4BU]	<30	N	Normal calprotectin - no evidence of GI inflammation. ?IBS
Buhlmann fCAL turbo [4BU]	<30	N	Negative/Within ref. range Calprotectin....?IBS
Buhlmann fCAL turbo [4BU]	<50	N	Not indicative of inflammation in the GI tract. Patients are likely not to be in need of invasive procedures to determine the inflammation cause.
Calpro (ALP) [2CP]	<19.5	N	
Calpro (ALP) [2CP]	<20	N	
Calpro (ALP) [2CP]	<20	N	
Diasorin [2IN]		XPL	The sample was not tested as it was unformed. This sample would be sent to a referral lab for weighing method.
Diasorin [2IN]	<5	N	<50ug/g faecal calprotectin within reference range ? IBS
Diasorin [2IN]	<5	N	
Diasorin [2IN]	5	N	Negative
Diasorin [2IN]	5.9	N	Faecal Calprotectin Reference Range <50ug/g.
Diasorin [2IN]	6.7	N	<50 ug/g IBD unlikely
Diasorin [2IN]	7	N	IBD is unlikely in this group of patients and should be treated as IBS with a 6 week review. Patients should be monitored and referred routinely to Gastroenterology if second line IBS treatment is unsuccessful.
Immundiagnostik (K6927)	<33	N	Faecal calprotectin concentration <50ug/g wet weight excludes active bowel inflammation with a high degree of confidence. Note that conditions other than IBD (e.g., infection, neoplasia, NSAID treatment) may raise calprotectin. This test should not be used in cases of suspected colorectal cancer.
Immundiagnostik (K6927)	<33	N	
Immundiagnostik (K6927)	11	N	Results >50 regarded as positive 100-150 indicate bowel inflammation >150 consistent with active IBD In children calprotectin values are higher than in adults. For ages 2-9 expect normal values up to 166mg/kg
Immundiagnostik (K6927)	18	N	Faecal Calprotectin within reference range, not suggestive of gastro-intestinal inflammation. If previously diagnosed with IBD, the level would indicate a period of remission
Immundiagnostik (K6927)	21	N	Faecal Calprotectin <50mg/kg does not indicate bowel inflammation. This test should not be used in cases of suspected bowel cancer. Suggest refer to NICE Clinical Guidance 61 (CG61: Irritable bowel syndrome in adults) and NICE Diagnostic Guidance 11 (DG11: Faecal calprotectin diagnostic tests for inflammatory diseases
Immundiagnostik (K6927)	22	N	Calprotectin within reference range.
Immundiagnostik (K6927)	33	N	
Inova QUANTA Flash [9IF]	<16		Result not suggestive of intestinal inflammation. If indicated suggest manage patient as per IBS pathway. Results must be interpreted within the clinical context. Do not use different calprotectin assay results interchangeably.
Inova QUANTA Flash [9IF]	<16	N	
Inova QUANTA Flash [9IF]	<16.1	N	
Inova QUANTA Flash [9IF]	<16.2	N	
Inova QUANTA Flash [9IF]	15	N	Negative (<50 ug/g)
Inova [2IF]	24	N	Negative (<50 ug/g)
Method development [MDV]	<26	N	Faecal Calprotectin Less than or equal to 50ug/g. Not indicative of intestinal inflammation. Levels >50ug/g may be associated with organic intestinal disease. Results should be interpreted in line with clinical assessment.
Not stated, please specify	6.7	N	
Orgentec Alegria [2OC1]	<20		
Orgentec Alegria [2OC1]	3.2	N	
Orgentec Alegria [2OC1]	5.9	N	
Orgentec Alegria [2OC1]	15	N	
Quantum Blue [3BU]	<30	N	
Quantum Blue [3BU]	<30	N	Calprotectin values <50 ug/g do not indicate GI tract inflammation.

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167A

Quantum Blue [3BU]	<30	N	<50ug/g – negative. Manage according to established practice. If IBS suspected, please refer to NICE IBS guidelines at http://publications.nice.org.uk/cg61
Quantum Blue [3BU]	XPL		
Quantum Blue [3BU]	36	N	<60 Faecal Calprotectin within ref range - ?Irritable Bowel Syndrome
Thermo EliA Calpro 2 [2KO2]			sample too liquid for analysis
Thermo EliA Calpro 2 [2KO2]	<4	N	
Thermo EliA Calpro 2 [2KO2]	<15	N	Faecal calprotectin within reference interval.
Thermo EliA Calpro 2 [2KO2]	<15	N	Not indicative of IBD. Levels of >50 ug/g are associated with organic intestinal disease, but should be interpreted in line with clinical assessment.
Thermo EliA Calpro 2 [2KO2]	<15	N	n/a
Thermo EliA Calpro 2 [2KO2]	<15	N	Faecal Calprotectin within reference range, not suggestive of gastro-intestinal inflammation. If previously diagnosed with IBD, the level would indicate a period of remission.
Thermo EliA Calpro 2 [2KO2]	5	N	Negative calprotectin level. Please refer to NICE guidance CG61 for further management.
Thermo EliA Calpro 2 [2KO2]	6.3	N	Negative
Thermo EliA Calpro 2 [2KO2]	8.1	N	Calprotectin within reference range. No evidence of active bowel inflammation. Refer to NICE Clinical Guideline 61.
Thermo EliA Calpro 2 [2KO2]	8.1	N	Faecal Calprotectin levels within normal range. This finding is useful in excluding bowel inflammation due to infection or autoimmune disease though this finding must be interpreted in the clinical context
Thermo EliA Calpro 2 [2KO2]	8.2	N	calprotectin within normal limits
Thermo EliA Calpro 2 [2KO2]	9.2	N	Result <75ug/g: Negative result. Symptoms likely to be due to IBS. Please refer to guideline: http://www.bhrhospitals.nhs.uk/pathology?smbfolder=193 Please note change of cut off as form 24.5.2017
Thermo EliA Calpro 2 [2KO2]	9.9		
Thermo EliA Calpro 2 [2KO2]	11	N	
Thermo EliA Calpro 2 [2KO2]	12	N	No calprotectin detected in this sample (? irritable bowel syndrome)
Thermo EliA Calpro 2 [2KO2]	12	N	A calprotectin result of less than 50mg/kg effectively excludes active inflammatory bowel disease. Normal values can occur in lymphocytic colitis , coeliac disease and stable diverticulosis. If symptoms persist suggest referral to gastroenterology.
Thermo EliA Calpro 2 [2KO2]	14	N	Calprotectin within reference range not suggestive of intestinal inflammation.
Thermo EliA Calpro 2 [2KO2]	15	N	Faecal calprotectin levels within normal range. This is useful in excluding bowel inflammation due to infection or autoimmune disease, although must be interpreted in the clinical context.
Thermo EliA Calpro 2 [2KO2]	19	N	Calprotectin less than or equal to 50ug/g. Not suggestive of intestinal inflammation. Results should be interpreted in line with clinical assessment.
Thermo EliA [2KO]	<3		
Thermo EliA [2KO]	<10	N	Result not suggestive of IBD. Probable IBS follow management pathway for IBS. If symptoms don't improve suggest refer to a gastroenterologist.
Thermo EliA [2KO]	<15	N	<50 ug/g : Negative. Calprotectin level not suggestive of organic pathology.
Thermo EliA [2KO]	<17	N	
Thermo EliA [2KO]	15	N	
Thermo EliA [2KO]	22	N	Negative result. Symptoms likely to be due to IBS. Manage in primary care. Only refer if severe watery diarrhoea or clinical concerns.

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167B

Kit	Result	Interpretation	Comment
Accusay [2AY]	XPL		
Accusay [2AY]	38.6	N	
Accusay [2AY]	41	N	
Accusay [2AY]	57.627	P	
Accusay [2AY]	75.221	P	Positive
Buhlmann [2BU]	81	N	Inflammatory bowel disease unlikely.
Buhlmann [2BU]	84	E	Evidence of moderate GI inflammation. IBD not excluded, but consider other causes e.g. NSAIDS, diverticular disease etc.
Buhlmann [2BU]	95	P	
Buhlmann [2BU]	100.13	E	
Buhlmann [2BU]	106.8	P	>50ug/g Refer to Gastroenterology
Buhlmann [2BU]	131	P	
Buhlmann [2BU]	132	E	Calprotectin values between 50 and 200 ug/g may represent mild organic disease or IBD in remission. Repeat measurement and further investigation may by warranted.
Buhlmann [2BU]	137	P	Moderate intestinal inflammation detected.NOTE: No NSAIDs 4 weeks prior to testing.Not to be used for cases of suspected bowel cancer.(NICE DG11 and NICE CG12
Buhlmann [2BU]	150	P	Moderate intestinal inflammation detected. If over 150ug/g on one occasion or over 50ug/g on two occasions six weeks apart and not on NSAID, then suggest referral to gastroenterology
Buhlmann [2BU]	154	E	Borderline faecal calprotectin. In patients without alarm symptoms or a pre-existing diagnosis of IBD, repeat sample. Ensure NSAIDS and PPIs have been withheld for 4-6 weeks. Exclude alternative causes of mildly elevated calprotectin such as coeliac disease, diverticulitis and gut infections. If Calprotectin is persistently raised a gastroenterology referral will be indicated.
Buhlmann [2BU]	154.4	P	
Buhlmann [2BU]	160	P	
Buhlmann [2BU]	160	E	Calprotectin >100: Please repeat within 2 weeks
Buhlmann [2BU]	161	P	Result of FC >150 ug/g : please refer to Gastroenterology for further investigation of the cause of this raised result.
Buhlmann [2BU]	166	E	Reference range < 50 ug/g Elevated levels 50 - 200 ug/g. Values can represent mild organic disease such as inflammation caused by NSAIDs, mild diverticulitis and IBD in remission phase. The low inflammatory response within this range may suggest repeating the measurement and performing further investigations may be useful.
Buhlmann [2BU]	167	P	Faecal calprotectin suggests gasto-intestinal inflammation. Possible causes include inflammatory bowel disease, infection, polyps, neoplasia and NSAID use. If the patient has previously been diagnosed with IBD, then result may be consistent with active disease. If not previously diagnosed with IBD, further investigation should be considered to establish the aetiology.
Buhlmann [2BU]	171	E	Faecal calprotectin indeterminate. Consider repeat after 4 weeks unless symptoms (e.g. unexplained iron deficiency anaemia) suggest earlier referral to Gastroenterology. Remove NSAIDS for 4 weeks prior to retest.
Buhlmann [2BU]	178	P	Evidence of active GI inflammation. Consistent with IBD or other forms of colitis. Referral advised.
Buhlmann [2BU]	184	E	Calprotectin at a concentration less than 200 ug/g is unlikely to represent an acute inflammatory process if no worrying signs or symptoms (eg weight loss, rectal bleeding, raised CRP) are present.
Buhlmann [2BU]	184	P	Results <200ug/g are rarely associated with significant pathology. Please see GG&C guidelines for advice and secondary care referral. www.nhsggc.org.uk/media/236675/ggc_fc_guidelines_dec_2015.doc
Buhlmann [2BU]	185	P	Faecal calprotectin results greater than 150 ug/g, suggest referral to a Consultant Gastroenterologist.
Buhlmann [2BU]	188	P	
Buhlmann [2BU]	203		Faecal calprotectin suggests organic pathology. Refer urgently to gastroenterology.
Buhlmann [2BU]	406	P	

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167B

Buhlmann fCAL turbo [4BU]	71	E	Moderately elevated levels are associated with organic intestinal disease but should be interpreted in line with clinical assessment. Repeat calprotectin 4 weeks from first test. Stop NSAIDs/aspirin. If repeat test, suggest refer to 'new IBD' clinic.
Buhlmann fCAL turbo [4BU]	90.7	E	50-200ug/g:Mild Organic Disease
Buhlmann fCAL turbo [4BU]	96	E	Initial calprotectin <100ug/g: Likely IBS. Repeat calprotectin <100ug/g: Likely IBS.
Buhlmann fCAL turbo [4BU]	97	N	FCP <100 ug/g suggests GI symptoms are non-inflammatory (e.g. IBS) Suggest manage symptomatically Please note above are coded comments and not free text comments.
Buhlmann fCAL turbo [4BU]	103	E	Calprotectin values 50 - 200 ug/g can represent mild organic disease such as inflammation caused by NSAIDs, mild diverticulitis & IBD in remission phase. The low inflammatory response shown within this range may suggest repeating the test and performing further investigations.
Buhlmann fCAL turbo [4BU]	110	E	FC 101-200ug/g - Indeterminate, rpt in 4-6 wks if symptoms persist. Re-test: FC >100ug/g - Consider referral to Gastroenterology.
Buhlmann fCAL turbo [4BU]	121	E	Indeterminate Calprotectin...see local guideline
Buhlmann fCAL turbo [4BU]	123	E	Borderline raised calprotectin indicating mild inflammation. Stop any NSAIDs and repeat in 4 weeks.
Buhlmann fCAL turbo [4BU]	150	P	Assuming request from a GP: Not indicative of GI inflammation. Please refer to primary care pathway for investigation of chronic diarrhoea.
Calpro (ALP) [2CP]	94	E	
Calpro (ALP) [2CP]	117	E	
Calpro (ALP) [2CP]	118.3	E	
Diasorin [2IN]		XPL	The sample was not tested as it was unformed. This sample would be sent to a referral lab for weighing method.
Diasorin [2IN]	56	E	
Diasorin [2IN]	78	P	Positive
Diasorin [2IN]	79.6	E	5--150ug/g FAECAL CALPRO, indeterminate RESULT. Suggest repeat after 4-6 weeks. Mildly raised figures may still be normal, consider other causes. e.g. infection, polyps, malignancy and NSAID use
Diasorin [2IN]	80.6	P	Faecal Calprotectin Reference Range <50ug/g.
Diasorin [2IN]	85	N	IBD is unlikely in this group of patients and should be treated as IBS with a 6 week review. If still symptomatic at review, suggest routine referral to Gastroenterology.
Diasorin [2IN]	98.3	P	50-150 ug/g - Repeat in 4 wks.', with patient off NSAID or aspirin, if still 50-150 ug/g - refer to GI OPD
Immundiagnostik (K6927)	78	P	
Immundiagnostik (K6927)	98	P	Results >50 regarded as positive 100-150 indicate bowel inflammation >150 consistent with active IBD In children calprotectin values are higher than in adults. For ages 2-9 expect normal values up to 166mg/kg
Immundiagnostik (K6927)	129	P	Faecal Calprotectin result >50mg/kg indicates inflammatory bowel disease. Suggest gastroenterology referral if indicated. Note that inflammatory and non-inflammatory diseases other than IBD and IBS may cause elevated levels of faecal calprotectin. This test should not be used in cases of suspected bowel cancer. Please see NICE Diagnostic Guidance 11 (DG11: Faecal calprotectin diagnostic tests for inflammatory diseases of the bowel) for further information.
Immundiagnostik (K6927)	138	P	Faecal Calprotectin suggests gastro-intestinal inflammation; possible causes include IBD, infection, polyps, neoplasia and NSAID's. If previously diagnosed with IBD, consistent with active disease.
Immundiagnostik (K6927)	148	P	Raised calprotectin. Consider referral if symptoms consistent with IBD.
Immundiagnostik (K6927)	163	P	
Immundiagnostik (K6927)	197	P	Faecal calprotectin concentration >70ug/g indicates active bowel inflammation. Refer to Gastroenterology. Note that conditions other than IBD (e.g., infection, neoplasia, NSAID treatment) may raise calprotectin. This test should not be used in cases of suspected colorectal cancer.

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167B

Inova QUANTA Flash [9IF]	49		Result not suggestive of intestinal inflammation. If indicated suggest manage patient as per IBS pathway. Results must be interpreted within the clinical context. Do not use different calprotectin assay results interchangeably.
Inova QUANTA Flash [9IF]	62	E	
Inova QUANTA Flash [9IF]	69	E	
Inova QUANTA Flash [9IF]	92	E	
Inova QUANTA Flash [9IF]	108	P	Positive
Inova [2IF]	100	P	Positive
Method development [MDV]	106	E	Faecal Calprotectin 51-200ug/g. Borderline Faecal Calprotectin result. Bowel inflammation not excluded. Results should be interpreted in line with clinical assessment.
Not stated, please specify	38.3	N	
Orgentec Alegria [2OC1]	36.1	N	
Orgentec Alegria [2OC1]	60		
Orgentec Alegria [2OC1]	63	P	
Orgentec Alegria [2OC1]	76.8	E	<50ug/g is negative . 50-200ug/g is intermediate and a repeat is suggested after 6 weeks
Quantum Blue [3BU]	XPL		
Quantum Blue [3BU]	99	E	Values of 50-200 ug/g may indicate mild organic disease.
Quantum Blue [3BU]	102	E	50-250ug/g – equivocal. Advise repeat after 6 weeks, ensuring no NSAID use for at least 4-6 weeks. If repeat >50ug/g, advise formal referral by letter to gastroenterology
Quantum Blue [3BU]	118	P	
Quantum Blue [3BU]	144	P	100 - 300 Raised Faecal Calprotectin suggests Inflammatory Bowel Disease
Thermo EliA Calpro 2 [2KO2]		XPL	Bristol stool type 7,unsuitable for analysis
Thermo EliA Calpro 2 [2KO2]			sample too liquid for analysis
Thermo EliA Calpro 2 [2KO2]	17	N	
Thermo EliA Calpro 2 [2KO2]	46		
Thermo EliA Calpro 2 [2KO2]	47	N	Result <75ug/g: Negative result. Symptoms likely to be due to IBS. Please refer to guideline: http://www.bhrhospitals.nhs.uk/pathology?smbfolder=193 Please note change of cut off as form 24.5.2017
Thermo EliA Calpro 2 [2KO2]	52	P	Indicates active intestinal disease
Thermo EliA Calpro 2 [2KO2]	53	E	Borderline range is 50-150 ug/g. We recommend that this test is repeated in 4-6 weeks time if clinically indicated. Raised levels of faecal calprotectin is an indicator of bowel inflammation due to infection or autoimmune disease such as ulcerative colitis or Crohns disease.
Thermo EliA Calpro 2 [2KO2]	55	P	
Thermo EliA Calpro 2 [2KO2]	59	P	An elevated faecal calprotectin suggests gastrointestinal inflammation. Possible causes include IBD, infection,polyps, neoplasia and NSAID's. If previously diagnosed with IBD, consistent with active disease.
Thermo EliA Calpro 2 [2KO2]	63	E	borderline range 50-150 mg/Kg. Suggest repeat in 4-6 weeks if clinically indicated. Raised levels of faecal calprotectin are an indicator of bowel inflammation due to infection or autoimmune disease (such as ulcerative colitis or Crohn's disease).
Thermo EliA Calpro 2 [2KO2]	64	P	Raised calprotectin level. This indicates gastro-intestinal inflammation, suggest referral to gastroenterology.
Thermo EliA Calpro 2 [2KO2]	72	E	Borderline faecal calprotectin. May be due to IBS or IBD, neoplasia, NSAIDS or recent upper respiratory tract infection. Suggest send a repeat stool sample in >1 month, ideally first morning void, at least a 2 pence sized peice.
Thermo EliA Calpro 2 [2KO2]	73	P	Marginally raised faecal calprotectin possibly suggestive of inflammatory bowel disease. Also consider other causes such as infection, polyps, and the use of NSAID's.
Thermo EliA Calpro 2 [2KO2]	73	P	Faecal calprotectin elevated above reference interval. Raised faecal calprotectin may indicate organic gastrointestinal disease. Suggest clinical correlation.
Thermo EliA Calpro 2 [2KO2]	93	P	Positive
Thermo EliA Calpro 2 [2KO2]	96	P	Calprotectin >50ug/g. Elevated level indicates intestinal inflammation. Suggest further investigation for inflammatory bowel disease. (Note calprotectin may also be raised in colorectal neoplasia, GI infections, GI bleeding and NSAID use) Results should be interpreted in line with clinical assessment.
Thermo EliA Calpro 2 [2KO2]	97	P	Faecal calprotectin raised, consistent with GI inflammation. Also consider other causes e.g. infection, polyps, malignancy and NSAID use. If previously diagnosed with IBD, consistent with active disease.

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167B

Thermo EliA Calpro 2 [2KO2]	113	E	Borderline faecal calprotectin result, bowel inflammation not excluded. Results should be interpreted in line with clinical assessment. Suggest consider giving out dietary advice or rule out infection and consider a repeat sample in 4 weeks if symptoms persist.
Thermo EliA Calpro 2 [2KO2]	130	P	Elevated calprotectin suggestive, but not diagnostic of inflammatory bowel disease. Refer to gastroenterology.
Thermo EliA [2KO]	50	P	
Thermo EliA [2KO]	62	E	50-200 ug/g : Gray Zone. Organic pathology cannot be excluded. A repeat sample in 4 to 6 weeks
Thermo EliA [2KO]	79	E	
Thermo EliA [2KO]	79	P	Result suggestive of bowel inflammation. Suggest repeat in 3-4 weeks. Medicines including PPI and NSAIDs can cause elevated calprotectin levels, consider stopping for 3 weeks before testing. Do not carry out faecal calprotectin testing within 1 week of gastrointestinal infection (level will be raised). If repeat elevated suggest referral to a gastroenterologist.
Thermo EliA [2KO]	80	E	
Thermo EliA [2KO]	85	E	Indeterminate result. Exclude infection with a stool culture. If stool culture negative repeat Calprotectin test. If rpt test still >50 or clinical concerns refer patient to gastroenterology using IDB proforma.

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167C

Kit	Result	Interpretation	Comment
Accusay [2AY]	XPL		
Accusay [2AY]	11.3	N	
Accusay [2AY]	18.759	N	Negative
Accusay [2AY]	19.089	N	
Accusay [2AY]	20	N	
Buhlmann [2BU]	21	N	IBD is unlikely. In patients with symptoms suggestive of IBD, a faecal calprotectin of <50 ug/g has a negative predictive value of 98%
Buhlmann [2BU]	32	N	Inflammatory bowel disease unlikely.
Buhlmann [2BU]	32.1	N	No evidence of GI inflammation - suggestive of IBS.
Buhlmann [2BU]	39	N	Faecal calprotectin within reference range. Unlikely IBD, treat as IBS.
Buhlmann [2BU]	41.35	N	
Buhlmann [2BU]	43	N	No intestinal inflammation detected
Buhlmann [2BU]	47	N	No evidence of GI inflammation - suggestive of IBS
Buhlmann [2BU]	48	N	Faecal calprotectin results less than 50 ug/g, inflammatory Bowel Disease is unlikely if patient was symptomatic.
Buhlmann [2BU]	50	E	Minimal intestinal inflammation detected.NOTE: No NSAIDs 4 weeks prior to testing.Not to be used for cases of suspected bowel cancer.(NICE DG11 and NICE CG12
Buhlmann [2BU]	50	N	
Buhlmann [2BU]	51	E	
Buhlmann [2BU]	51	P	
Buhlmann [2BU]	52.5	P	>50ug/g Refer to Gastroenterology
Buhlmann [2BU]	55		Borderline faecal calprotectin. In patients without alarm symptoms or a pre-existing diagnosis of IBD repeat sample. Ensure NSAID's and PPI's have been withheld for 4-6 weeks. Exclude alternative causes of mildly elevated calprotectin such as coeliac disease, diverticulitis and gut infections. If repeat calprotectin is persistently raised a gastroenterology referral will be indicated.
Buhlmann [2BU]	57	E	Calprotectin at a concentration less than 200 ug/g is unlikely to represent an acute inflammatory process if no worrying signs or symptoms (eg weight loss, rectal bleeding, raised CRP) are present.
Buhlmann [2BU]	58.6	E	
Buhlmann [2BU]	59	P	Results <200ug/g are rarely associated with significant pathology. Please see GG&C guidelines for advice and secondary care referral. www.nhsggc.org.uk/media/236675/ggc_fc_guidelines_dec_2015.doc
Buhlmann [2BU]	60	E	Results in this range are clinically indeterminate. Suggest repeat or investigate further as appropriate.
Buhlmann [2BU]	60	N	Calprotectin <100: Likely IBS
Buhlmann [2BU]	61	E	Borderline Raised Result of FC 50 - 150 ug/g : Refer to the GHNHSFT Complex IBS Service and repeat the Calprotectin test in 3 months (avoiding NSAIDs for at least 4 weeks prior to the test).
Buhlmann [2BU]	63	E	Reference range < 50 ug/g Elevated levels 50 - 200 ug/g. Values can represent mild organic disease such as inflammation caused by NSAIDs, mild diverticulitis and IBD in remission phase. The low inflammatory response within this range may suggest repeating the measurement and performing further investigations may be useful.
Buhlmann [2BU]	65	E	Calprotectin values between 50 and 200 ug/g may represent mild organic disease or IBD in remission. Repeat measurement and further investigation may by warranted.
Buhlmann [2BU]	74	E	
Buhlmann [2BU]	76	P	
Buhlmann fCAL turbo [4BU]	<50	N	Not indicative of inflammation in the GI tract. Patients are likely not to be in need of invasive procedures to determine the inflammation cause.

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167C

Buhlmann fCAL turbo [4BU]	38	N	Negative/Within ref. range Calprotectin....?IBS
Buhlmann fCAL turbo [4BU]	39	N	Normal calprotectin - no evidence of GI inflammation. ?IBS
Buhlmann fCAL turbo [4BU]	41	N	Not indicative of GI inflammation
Buhlmann fCAL turbo [4BU]	42	N	Assuming request from a GP: Indicative of active GI inflammation. Refer to Gastroenterology as per primary care pathway for the investigation of chronic diarrhoea
Buhlmann fCAL turbo [4BU]	43	N	FC <100ug/g-IBD unlikely, primary care management.
Buhlmann fCAL turbo [4BU]	43.4	N	<50 ug/g :Normal
Buhlmann fCAL turbo [4BU]	52	N	FCP <100 ug/g suggests GI symptoms are non-inflammatory (e.g. IBS) Suggest manage symptomatically Please note above are coded comments and not free text comments.
Buhlmann fCAL turbo [4BU]	66	E	Initial calprotectin <100ug/g: Likely IBS. Repeat calprotectin <100ug/g: Likely IBS.
Calpro (ALP) [2CP]	37	N	
Calpro (ALP) [2CP]	39	N	
Calpro (ALP) [2CP]	42.8	N	
Diasorin [2IN]	13	N	Negative
Diasorin [2IN]	13.3	N	<50ug/g faecal calprotectin within reference range ? IBS
Diasorin [2IN]	18	N	
Diasorin [2IN]	28.1	N	Faecal Calprotectin Reference Range <50ug/g.
Diasorin [2IN]	29.2	N	Faecal calprotectin within the reference range. ?IBS/quiescent IBD
Diasorin [2IN]	31	N	IBD is unlikely in this group of patients and should be treated as IBS with a 6 week review. Patients should be monitored and referred routinely to Gastroenterology if second line IBS treatment is unsuccessful.
Diasorin [2IN]	32	N	<50 ug/g IBD unlikely
Immundiagnostik (K6927)	38	N	
Immundiagnostik (K6927)	41	N	Calprotectin within reference range.
Immundiagnostik (K6927)	43	N	Results >50 regarded as positive 100-150 indicate bowel inflammation >150 consistent with active IBD In children calprotectin values are higher than in adults. For ages 2-9 expect normal values up to 166mg/kg
Immundiagnostik (K6927)	47	N	Faecal Calprotectin within reference range, not suggestive of gastro-intestinal inflammation. If previously diagnosed with IBD, the level would indicate a period of remission
Immundiagnostik (K6927)	50	E	Faecal Calprotectin <50mg/kg does not indicate bowel inflammation. This test should not be used in cases of suspected bowel cancer. Suggest refer to NICE Clinical Guidance 61 (CG61: Irritable bowel syndrome in adults) and NICE Diagnostic Guidance 11 (DG11: Faecal calprotectin diagnostic tests for inflammatory diseases
Immundiagnostik (K6927)	74	E	
Immundiagnostik (K6927)	78	P	Faecal calprotectin concentration >70ug/g indicates active bowel inflammation. Refer to Gastroenterology. Note that conditions other than IBD (e.g., infection, neoplasia, NSAID treatment) may raise calprotectin. This test should not be used in cases of suspected colorectal cancer.
Inova QUANTA Flash [9IF]	<16	N	
Inova QUANTA Flash [9IF]	<16.1	N	
Inova QUANTA Flash [9IF]	<16.2	N	
Inova QUANTA Flash [9IF]	19	N	Negative (<50ug/g)
Inova QUANTA Flash [9IF]	34		Result not suggestive of intestinal inflammation. If indicated suggest manage patient as per IBS pathway. Results must be interpreted within the clinical context. Do not use different calprotectin assay results interchangeably.'
Inova [2IF]	37	N	Negative (<50 ug/g)
Method development [MDV]	44	N	Faecal Calprotectin Less than or equal to 50ug/g. Not indicative of intestinal inflammation. Levels >50ug/g may be associated with organic intestinal disease. Results should be interpreted in line with clinical assessment.
Not stated, please specify	13	N	
Orgentec Alegria [2OC1]	<20		
Orgentec Alegria [2OC1]	11.6	N	

Interpretation of Faecal Calprotectin results

Participants were asked to provide a free text interpretation for each specimen based upon a combination of the result they obtained and the clinical scenario below.

A 40 year-old male visited his Family Doctor. The details on the request form were:- Alternating diarrhoea and constipation for 3 months. ?IBS.

Specimen 167C

Orgentec Alegria [2OC1]	17.6	N	
Orgentec Alegria [2OC1]	26	N	
Quantum Blue [3BU]	XPL		
Quantum Blue [3BU]	48	N	
Quantum Blue [3BU]	60	E	Values of 50-200 ug/g may indicate mild organic disease.
Quantum Blue [3BU]	67	E	60 – 100 Intermediate faecal Calprotectin result - repeat may be useful
Quantum Blue [3BU]	68	E	50-250ug/g – equivocal. Advise repeat after 6 weeks, ensuring no NSAID use for at least 4-6 weeks. If repeat >50ug/g, advise formal referral by letter to gastroenterology
Thermo EliA Calpro 2 [2KO2]	12	N	
Thermo EliA Calpro 2 [2KO2]	13		
Thermo EliA Calpro 2 [2KO2]	14	N	Negative calprotectin level. Please refer to NICE guidance CG61 for further management.
Thermo EliA Calpro 2 [2KO2]	16	N	Not indicative of IBD. Levels of >50 ug/g are associated with organic intestinal disease, but should be interpreted in line with clinical assessment.
Thermo EliA Calpro 2 [2KO2]	19	N	Result <75ug/g: Negative result. Symptoms likely to be due to IBS. Please refer to guideline: http://www.bhrhospitals.nhs.uk/pathology?smbfolder=193 Please note change of cut off as form 24.5.2017
Thermo EliA Calpro 2 [2KO2]	20	N	n/a
Thermo EliA Calpro 2 [2KO2]	22	N	Faecal Calprotectin levels within normal range. This finding is useful in excluding bowel inflammation due to infection or autoimmune disease though this finding must be interpreted in the clinical context
Thermo EliA Calpro 2 [2KO2]	23	N	calprotectin within normal limits
Thermo EliA Calpro 2 [2KO2]	24	N	Calprotectin within reference range. No evidence of active bowel inflammation. Refer to NICE Clinical Guideline 61.
Thermo EliA Calpro 2 [2KO2]	24	N	Negative
Thermo EliA Calpro 2 [2KO2]	27	N	A calprotectin result of less than 50mg/kg effectively excludes active inflammatory bowel disease. Normal values can occur in lymphocytic colitis , coeliac disease and stable diverticulosis. If symptoms persist suggest referral to gastroenterology.
Thermo EliA Calpro 2 [2KO2]	29	N	Faecal calprotectin levels within normal range. This is useful in excluding bowel inflammation due to infection or autoimmune disease, although must be interpreted in the clinical context.
Thermo EliA Calpro 2 [2KO2]	30	N	
Thermo EliA Calpro 2 [2KO2]	32	N	Borderline faecal calprotectin result, bowel inflammation not excluded. Suggest repeat if clinical suspicion warrants or investigate further as appropriate.
Thermo EliA Calpro 2 [2KO2]	34	N	Faecal calprotectin within reference interval.
Thermo EliA Calpro 2 [2KO2]	37	N	
Thermo EliA Calpro 2 [2KO2]	38	N	Faecal calprotectin within reference range (? irritable CALP bowel syndrome)
Thermo EliA Calpro 2 [2KO2]	38	N	Calprotectin within reference range not suggestive of intestinal inflammation.
Thermo EliA Calpro 2 [2KO2]	45	N	Calprotectin less than or equal to 50ug/g. Not suggestive of intestinal inflammation. Results should be interpreted in line with clinical assessment.
Thermo EliA [2KO]	15	N	
Thermo EliA [2KO]	17	N	<50 ug/g : Negative. Calprotectin level not suggestive of organic pathology.
Thermo EliA [2KO]	17	N	Result not suggestive of IBD. Probable IBS follow management pathway for IBS. If symptoms don't improve suggest refer to a gastroenterologist.
Thermo EliA [2KO]	20	N	
Thermo EliA [2KO]	22	N	
Thermo EliA [2KO]	29	N	Negative result. Symptoms likely to be due to IBS. Manage in primary care. Only refer if severe watery diarrhoea or clinical concerns.

Appendix N
RWE analysis

N.1 FC density plots

Figure N-1 shows the population density plots for FC1 values used within the RWE analysis. Plots A and B illustrate the IBS and IBD populations in the base case analysis (setting censored FC data equal to their respective limits), and plots C and D illustrate the IBS and IBD populations in the sensitivity analysis (setting censored FC data equal to the EQA median estimates as outlined in Table 7-2). The solid black line in each plot illustrates the density of FC1 values based on the bootstrapped YFCCP dataset, with notable peaks occurring at 600 µg/g (in plots A and B) and 915 µg/g (in plots C and D) due to all right-censored 2BU data being set equal to these values within the base case and sensitivity analyses respectively. The dashed light green line illustrates the $FC1_{sim}$ values generated for the 4BU assay within the error model simulation, and the dotted purple line illustrates the $FC1_{sim}$ values generated for the 2KO2 assay. Figure N-2 provides the same plots but for the FC2 values.

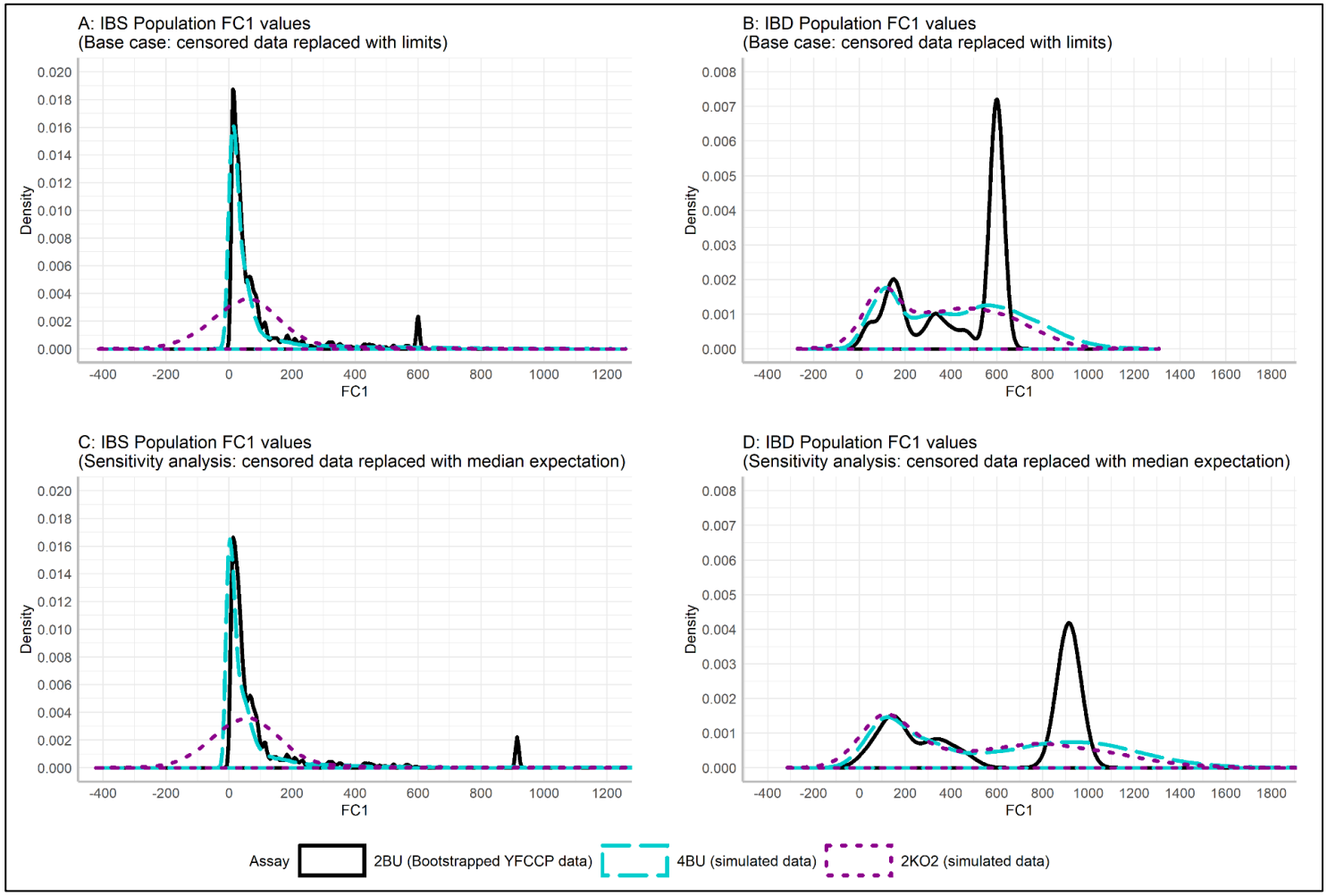


Figure N-1. RWE analysis: FC1 density plots

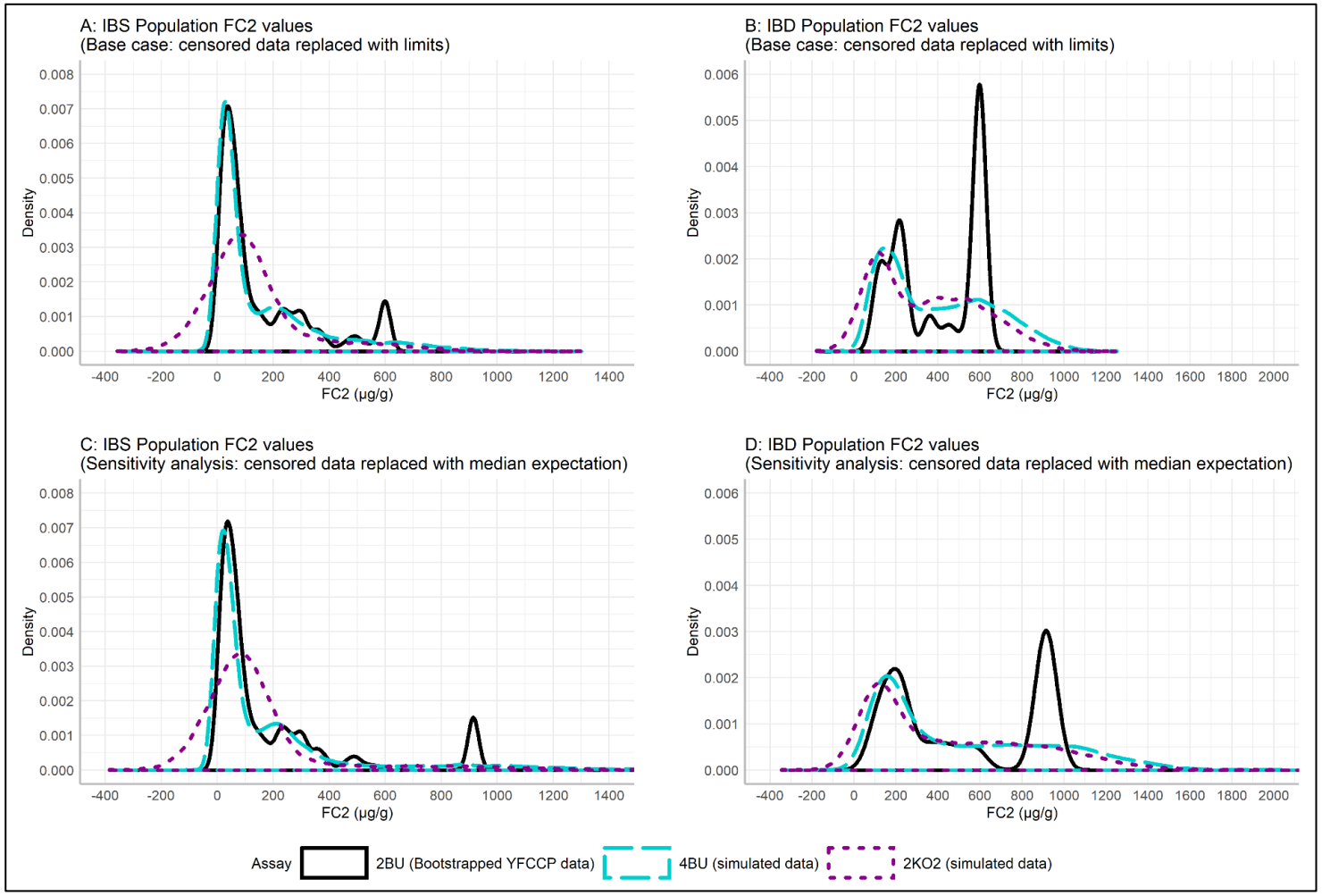


Figure N-2. RWE analysis: FC2 density plots

N.2 Post-hoc sensitivity analysis

Figure N-3 provides the FC1 density plots for the post-hoc sensitivity analysis (i.e. excluding the two extreme FC values discussed in section 7.5.1); whilst

Figure **N-4** provides the FC2 density plots for the same analysis. Comparison of these figures with the corresponding plots provided in section N.1 above, illustrates the significant impact that removal of the two extreme 2KO2 values has in terms of the simulated 2KO2 FC distributions. This difference is particularly noticeable within the IBS population FC1 and FC2 distributions.

Table N-1 provides the results of the RWE post-hoc sensitivity analysis for the 2KO2 method.

Figure N-5 provides the results of the bias correction exercise for the post-hoc sensitivity analysis, based on applying a fixed absolute correction value; and Figure N-6 provides the corresponding results of the bias correction exercise applying a proportional correction factor. The results are discussed in the main thesis text (see section 7.5.2). Note that for the proportional correction factor exercise, the range of factors simulated was shifted (from 1.0 to 2.5 in 0.05 increments in the base case) to 2.0 to 3.5 in 0.05 increments: this is due to the fact that based on initial simulations, it was found that higher proportional correction factors were required in order to capture the points of highest diagnostic yield and closest match to the 2BU assay results.

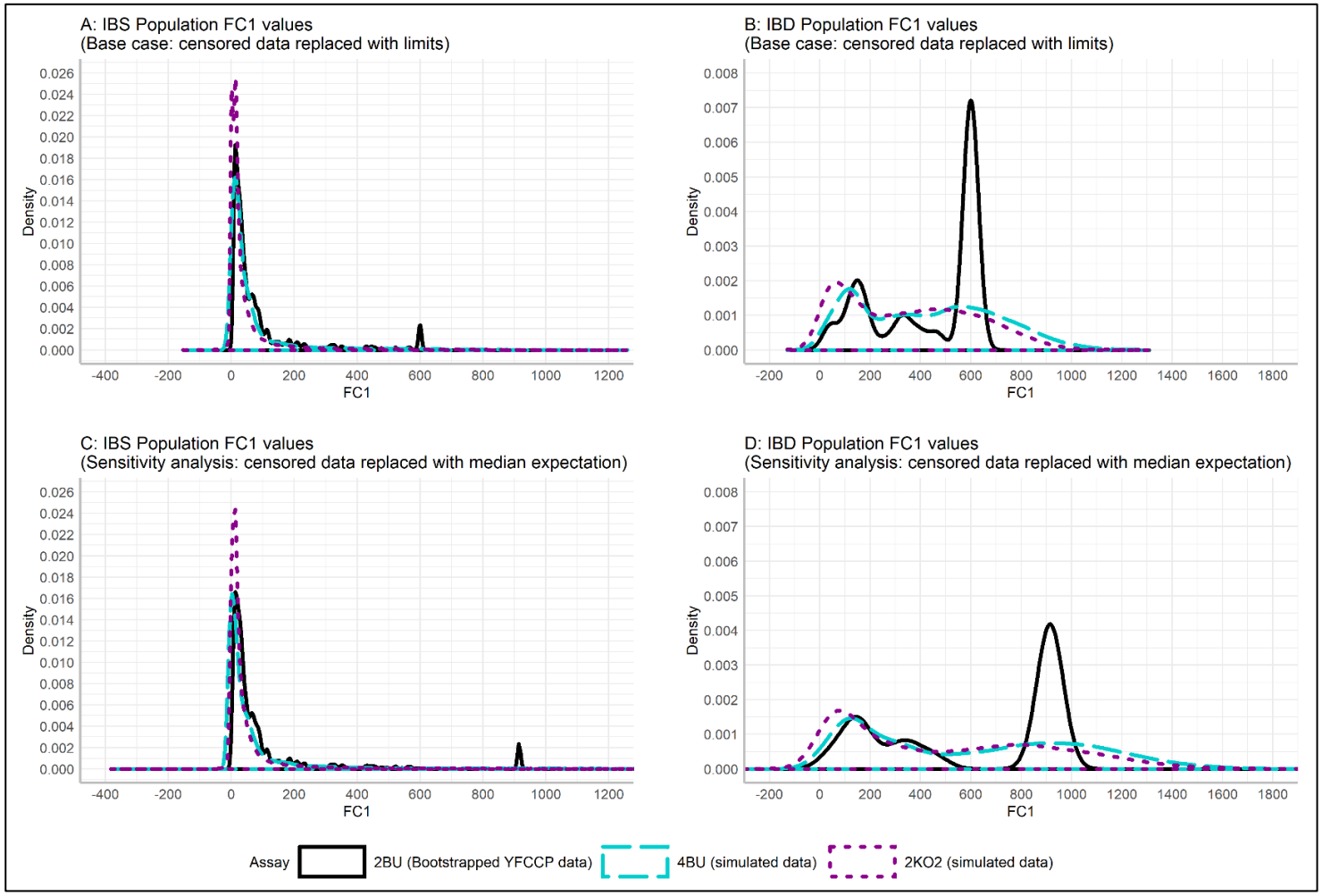


Figure N-3. RWE analysis: FC1 density plots for the post-hoc sensitivity analysis

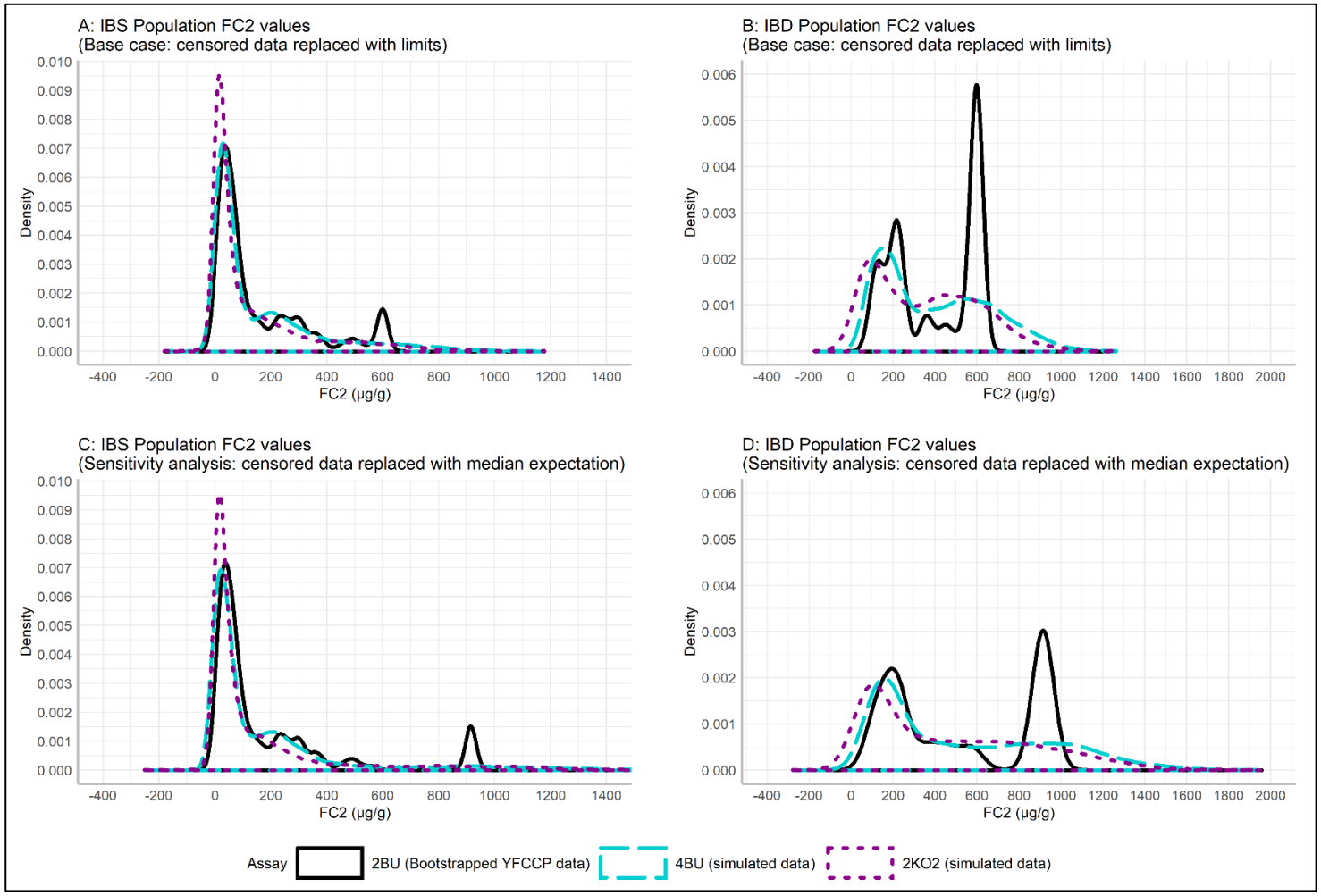


Figure N-4. RWE analysis: FC2 density plots for the post-hoc sensitivity analysis

Table N-1. YFCCP RWE analysis: outcome results for 2KO2 method, including the post-hoc sensitivity analysis

FC assay method	Diagnostic accuracy		Cost-effectiveness			
	Sensitivity	Specificity	Cost	QALY	NMB	INMB (£) YFCCP [2BU] vs. comparator
EQA analysis results						
YFCCP [2KO2] base case	0.655	0.828	£246	0.7869	£15,493	£88
YFCCP [2KO2] post-hoc sensitivity analysis	0.622	0.957	£201	0.7876	£15,551	£30
FC cost-utility model diagnostic accuracy inputs and cost-effectiveness outputs (for reference)						
YFCCP [2BU] intervention	0.938	0.920	£212	0.7896	£15,581	-
No FC (Tibble data)	0.350	0.730	£259	0.7836	£15,412	£169
No FC (NICE data)	1.000	0.790	£232	0.7879	£15,526	£55
FC testing (YFCCP, 50 µg/g cut-off)	0.960	0.600	£314	0.7836	£15,359	£222
FC testing (Tibble data)	0.900	0.800	£245	0.7860	£15,474	£107
FC testing (NICE data)	0.930	0.940	£197	0.7880	£15,562	£19

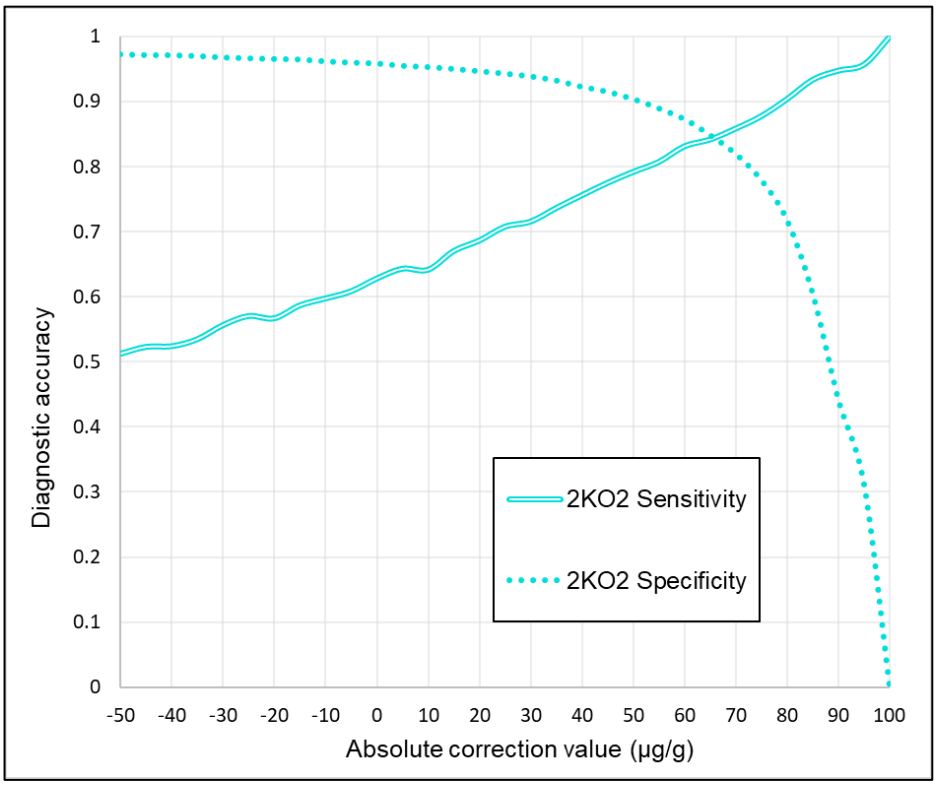


Figure N-5. RWE analysis: plot of absolute correction value vs. diagnostic accuracy for 2KO2 FC assay (post-hoc sensitivity analysis results)

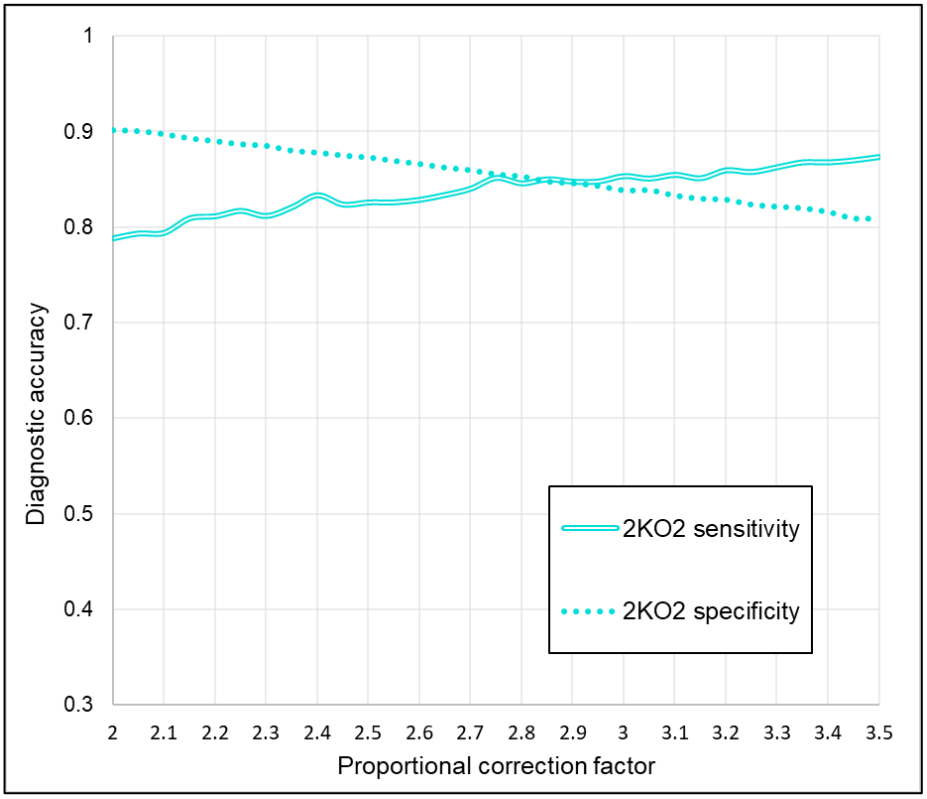


Figure N-6. RWE analysis: plot of proportional correction factor vs. diagnostic accuracy for 2KO2 FC assay (post-hoc sensitivity analysis results)