TECHNICAL SUPPORT DOCUMENT 25: EVIDENCE SYNTHESIS OF DIAGNOSTIC TEST ACCURACY FOR DECISION MAKING

REPORT BY THE DECISION SUPPORT UNIT

NOVEMBER 2024

Efthymia Derezea[1], AE Ades[1], Gabriel Rogers[2], Alex J Sutton[3], Nicola J Cooper[3], Jean Hamilton[4], Hayley E Jones[1]

[1] Population Health Sciences, Bristol Medical School, University of Bristol

[2] Manchester Centre for Health Economics, University of Manchester

[3] Complex Reviews Synthesis Unit, Department of Population Health Sciences, University of Leicester

[4] ScHARR, Division of Population Health, University of Sheffield

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street, Sheffield. S1 4DA.
Tel (+44) (0)114 222 0734

E-mail dsuadmin@sheffield.ac.uk
Website www.nicedsu.org.uk
Twitter @NICE_DSU

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) External Assessment Centre is based at the University of Sheffield with members at York, Bristol, Leicester and the London School of Hygiene and Tropical Medicine. The DSU is commissioned by The National Institute for Health and Care Excellence (NICE) to provide a research and training resource to support the Institute's Centre for Health Technology Evaluation Programmes. Please see our website for further information www.nicedsu.org.uk.

The production of this document was funded by the National Institute for Health and Care Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

NICE describes the methods it follows when carrying out health technology evaluations in its process and methods manual. This provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The manual does not provide detailed advice on how to implement and apply the methods it describes. The DSU series of Technical Support Documents (TSDs) is intended to complement the manual by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in selected topic areas. They make recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE technology evaluations, whether companies, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides. The TSDs will be amended and updated whenever appropriate. Where minor updates or corrections are required, the TSD will retain its numbering with a note to indicate the date and content change of the last update. More substantial updates will be contained in new TSDs that entirely replace existing TSDs.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Prof Allan Wailoo, Director of DSU and TSD series editor.

ACKNOWLEDGEMENTS

The authors thank Tom Trikalinos, Mark Sculpher, Marta Soares and Thomas Walker for reviewing this document and providing helpful feedback. HEJ acknowledges funding from the Medical Research Council (MRC) and National Institute for Health and Care Research (NIHR) for some of the research that led to development of this TSD (MR/M014533/1, MR/T044594/1). NJC and AJS acknowledge funding from NIHR Evidence Synthesis Group @ CRSU and support from the NIHR Applied Research Collaboration East Midlands (ARC EM) and Leicester NIHR Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

This report should be referenced as follows:

Derezea E, Ades AE, Rogers G, Sutton AJ, Cooper NJ, Hamilton J, Jones HE. Evidence synthesis of diagnostic test accuracy for decision making: Technical Support Document 25. 2024, [available from http://www.nicedsu.org.uk]

EXECUTIVE SUMMARY

This Technical Support Document (TSD) describes methods for meta-analysis of the accuracy – as quantified by sensitivity and specificity – of a diagnostic test, for use in NICE decision models.

We first describe methods for meta-analysis of a single estimate of sensitivity and specificity from each study. The data from a systematic review will typically take this form when a test produces a truly binary result or when the "threshold" used to classify results as positive or negative is implicit rather than explicit – usually due to a subjective element in reading of test results. Data may also have this structure if the test is intended to be operationalised at a particular numerical threshold and all studies have evaluated its accuracy at that threshold alone. We describe the binomial bivariate meta-analysis model for sensitivity and specificity, the equivalent hierarchical summary receiver operating characteristic (HSROC) model, and how results from this model can be presented.

We also consider the situation in which studies in the systematic review report accuracy at different explicit threshold values, with some studies reporting sensitivity and specificity at more than one threshold. We describe a flexible model to synthesise all such data in a unified analysis, producing pooled estimates of sensitivity and specificity across a range of threshold values. There may be precision gains from fitting this model, even if the decision model only evaluates cost-effectiveness at a single developer-recommended threshold.

This TSD additionally discusses use of meta-analysis of test accuracy results in decision models. We discuss accounting for correlation between estimated sensitivity and specificity, and the critical role of prevalence. We also discuss use of results from the multiple thresholds model in a decision model. For the situation in which the decision problem includes choice of threshold at which to operationalise the test in practice (which will not always be the case), we demonstrate how results from the multiple thresholds model can be used to determine the optimal threshold according to some criterion, such as maximum expected net benefit. We additionally demonstrate that this "optimal" threshold will depend heavily on assumed prevalence in the decision model.

As in previous TSDs in the evidence synthesis series, we focus primarily on a Bayesian statistical approach to synthesis, and provide examples in WinBUGS. We also discuss other software options – in particular, how the binomial bivariate model can be fitted within a frequentist framework. Synthesis can be challenging when the number of studies is small; we illustrate the advantages of Bayesian methods in this situation.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS AND DEFINITIONS

AUC	Area under curve
CrI	Credible interval
DIC	Deviance information criterion
DOR	Diagnostic odds ratio
DTA	Diagnostic test accuracy
FN	False negative
FP	False positive
FPF	False positive fraction
HSROC	Hierarchical summary receiver operating characteristic
INB	Incremental net benefit
MCMC	Markov chain Monte Carlo
NB	Net benefit
pD	Effective number of model parameters
QALY	Quality-adjusted life year
ROC	Receiver operating characteristic
SROC	Summary ROC
TN	True negative
TP	True positive
TPF	True positive fraction

1. INTRODUCTION

Synthesis of the accuracy of one or more diagnostic tests is often an important component of reviews performed for NICE, most notably for the NICE Centre for Health Technology Evaluation (CHTE) and NICE Guidelines. For evaluation of the impacts on patient outcomes of testing strategies – in combination with treatment strategies – NICE has a preference for use of evidence from “test and treat” randomised controlled trials where available (1). However, such evidence is rarely available in practice and – even where it does exist – is not generalisable to all possible uses of a test, or to alternative treatment options or pathways. Therefore, a “linked evidence” approach is commonly used instead. Here, a decision model links together evidence on test accuracy with assumptions about how test results will be used (e.g. informing treatment decisions or use of a subsequent test), and with evidence on treatment effectiveness and other parameters (2). The focus of this TSD is on evidence synthesis of data on test accuracy for incorporation into these models.

This TSD describes methods for meta-analysis of the accuracy of a single test, or of each test under evaluation separately. We use standard terminology in referring to the test(s) under evaluation as the index test(s) and assume that we are interested in the ability of the index test to discriminate between those with and those without some clearly specified target condition, which is often but not always a disease. For brevity, we will sometimes use “disease” in this document in place of “target condition”, but we emphasise that the target condition can in practice be many things other than a disease (e.g. treatment resistance (3)). The methods described could also be applied to data quantifying the ability of a test to predict a future clinical outcome (e.g. pre-term labour (4)). However, measures other than sensitivity and specificity (our focus) are often used to quantify prognostic or predictive ability.

We assume that a systematic review has identified a number of studies that have compared results on the index test with results on a “reference standard” test. Methods described in this TSD are based on an assumption that the reference standard in each study does not itself make errors, i.e. is considered to be a “gold standard” test for the target condition. (See *Section 7* for a brief discussion of the situation in which this is known not to be true.) Note that if different studies used different reference standards, each of these must be considered to be error-free. Each study, typically referred to as a Diagnostic Test Accuracy (DTA) study, produces a 2×2 table of cross-classified test results.

We focus in this TSD on methods for meta-analysis of sensitivity and specificity, defined as the proportion of individuals for which a test is correct among those with and those without the target condition, respectively. Under the assumption of an error-free reference standard, these are estimated by simple observed proportions using the 2×2 table. We describe the bivariate random effects model (5, 6) and the mathematically equivalent hierarchical summary receiver operating characteristic curve (HSROC) model (7), and how results can be summarised in receiver operating characteristic (ROC) space using summary regions and/or a summary ROC curve.

In practice, index tests often produce continuous numerical results, which could be dichotomised at different points (referred to as thresholds or cut-offs), producing different 2×2 tables. For the situation in which some studies in a systematic review report 2×2 tables for multiple thresholds, clearly stating the numerical threshold values that these tables correspond to, this TSD also describes a model to synthesise all data together (8). This produces pooled estimates of sensitivity and specificity at any numerical cut-off, including the test developer's specified threshold if there is one and thresholds not explicitly reported on in any study. If the project scope does not specify a clear threshold at which the test must be operationalised, these can be used to estimate the threshold at which the test is most cost-effective.

As in previous TSDs in the evidence synthesis series, we focus primarily on a Bayesian statistical approach to synthesis, and provide examples in WinBUGS. WinBUGS implements Bayesian analysis using Markov chain Monte Carlo (MCMC) simulation. We describe, however, how the binomial bivariate model is a standard generalised linear mixed model which can also be fitted within a frequentist framework, in software such as R or Stata. There are several advantages of the Bayesian MCMC approach, however. First, this allows us to obtain posterior samples and make statistical inference for any quantity of interest that we can write down mathematically. Second, it facilitates extensions in a straightforward manner to more complex situations, such as handling multiple thresholds (*Section 5*), and sets the base for extending to flexible models for synthesis of test comparisons or dealing with imperfect reference standards. Moreover, the Bayesian approach allows us to introduce information through informative priors, enabling us to make use of previous knowledge and/or overcome potential computational issues. Finally, use of MCMC simulation allows the export of stochastically correlated results into a probabilistic decision model (*Section 6*), avoiding normal approximations.

WinBUGS code to fit the meta-analysis models is provided in the Appendix. We assume knowledge of how to check convergence, ensure that Monte Carlo error is not too large, and identify issues such as conflict between priors and data (9). The results described in this TSD were produced via R, using R2WinBUGS. The R and WinBUGS code to reproduce all results and figures, and datasets used, is available in the following GitHub repository: https://github.com/FeniaDerezea/TSD25

The rest of this document is organised as follows. *Section 2* provides a brief introduction to commonly used measures of diagnostic test accuracy. The principal concepts around meta-analysis of sensitivity and specificity are introduced in *Section 3*. *Section 4* focuses on methods for meta-analysis of a single estimate of sensitivity and specificity from each study. We describe and demonstrate the bivariate random effects model, with discussion of appropriate outputs from this model aided by two worked examples. We also discuss possible approaches to fitting this model when the number of available studies is small (*Section 4.5*), and provide a brief overview of the many alternative software options for fitting the bivariate model (*Section 4.6*). *Section 5*, focusing on meta-analysis of data across multiple thresholds, follows a similar format. In *Section 6*, we demonstrate with a simplified decision model how outputs from the meta-analysis models described in *Sections 4* and *5* can be used to evaluate and compare the cost-effectiveness of testing strategies. We demonstrate the important role of the prevalence of the target condition in the population being tested, and how the “optimal” threshold at which to operationalise the test in a population with a particular prevalence can be evaluated. Finally, a brief discussion of topics not covered by this TSD is provided in *Section 7*.

2. MEASURES OF TEST ACCURACY

In a typical DTA study, both the index test and reference standard are carried out on a number of individuals and results compared, producing a 2×2 table of cross-classified test results as shown in *Table 1*. We use T to denote the index test result (positive: T=1, negative: T=0), and D to denote true disease state (positive: D=1, negative: D=0). As noted in *Section 1*, we assume that the reference standard correctly represents true disease state, D. The true positive (TP) and true negative (TN) cell counts are the number of individuals correctly classified by the index test as having or not having the target condition, respectively. FP is the number of individuals without the target condition who were incorrectly positive on the index test, while FN is the number of individuals with the target condition who were incorrectly negative.

Table 1: 2×2 table of cross-classified results from a diagnostic test accuracy study

		True disease state (D)	
		Positive (D = 1)	**Negative**
Index test result (T)	**Positive (T = 1)**	True Positive (TP)	False Positive (FP)
	Negative (T = 0)	False Negative (FN)	True Negative (TN)

We do not provide a comprehensive overview of measures of test accuracy in this TSD, but briefly describe some key measures in the following subsections.

2.1. SENSITIVITY AND SPECIFICITY

The sensitivity, also referred to as the true positive fraction (TPF), and specificity of the index test are defined as the probability of correct classification in those with and those without the target condition, respectively. Written mathematically:

$$\text{Sensitivity } = Pr(T = 1|D = 1)$$

$$\text{Specificity} = Pr(T = 0|D = 0)$$

From a single DTA study, sensitivity and specificity are estimated from *Table 1* by $TP/(TP + FN)$ and $TN/(TN + FP)$ respectively.

In this TSD, we will also often refer to the complement of specificity (i.e. 1 – specificity), as the False Positive Fraction (FPF):

$$\text{FPF} = Pr(T = 1|D = 0)$$

We note that the TPF and FPF are often referred to as true and false positive *rates,* respectively (or TPR and FPR). The terms can be used interchangeably, but we follow Pepe (10) in adopting "fraction" since it more accurately reflects that these quantities are probabilities, rather than rates.

2.2. PREDICTIVE VALUES

The Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are generally considered to be more interpretable and clinically relevant than the sensitivity and specificity. These are defined as the probability of the disease given a positive test result, and the probability of not having the disease given a negative test result, respectively, i.e.

$$\text{PPV} = Pr(D = 1|T = 1)$$

$$\text{NPV} = Pr(D = 0|T = 0)$$

The PPV and NPV are mathematically related to sensitivity and specificity through Bayes' rule, which gives:

$$\text{PPV} = \frac{Pr(T = 1|D = 1)Pr(D = 1)}{Pr(T = 1)}$$

$$= \frac{\pi \times \text{Sensitivity}}{[\pi \times \text{Sensitivity}] + [(1 - \pi) \times (1 - \text{Specificity})]} \qquad \textbf{(1)}$$

and

$$\text{NPV} = \frac{Pr(T=0|D=0)Pr(D=0)}{Pr(T=0)}$$

$$= \frac{(1-\pi)\times \text{Specificity}}{[(1-\pi)\times \text{Specificity}] + [\pi \times (1-\text{Sensitivity})]} \qquad \textbf{(2)}$$

where π is the prevalence or pre-test probability of the target condition in the population or individual being tested.

Although PPV and NPV can be estimated directly from *Table 1*, as $TP/(TP+FP)$ and $TN/(TN+FN)$ respectively, estimates calculated in this way are only relevant to the situation where the pre-test probability π is equal to the observed prevalence in the DTA study. PPV and NPV for other values of π can be estimated by first estimating sensitivity and specificity and then applying equations (1) and (2).

2.3. RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES

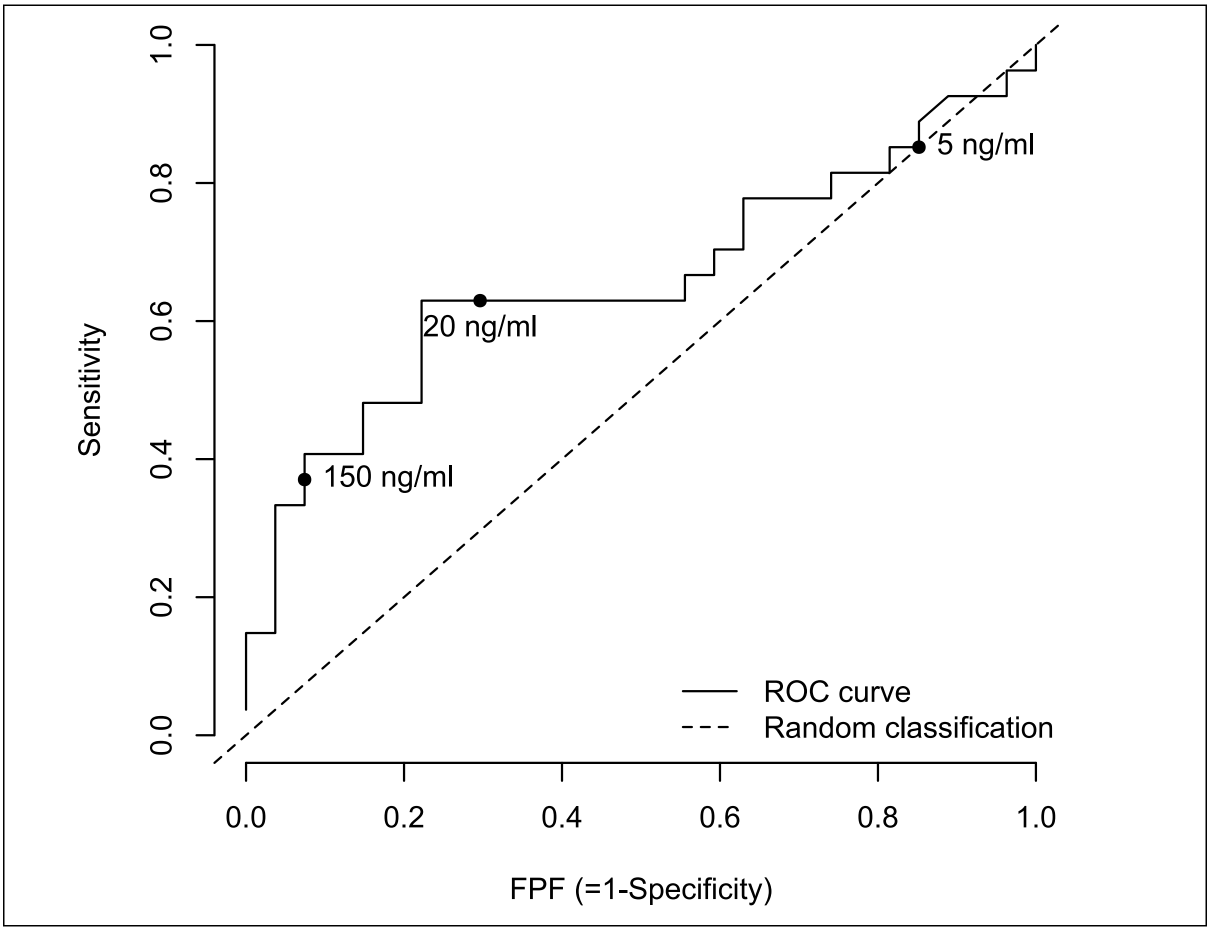
As noted in *Section 1*, many index tests are continuous, and can be operationalised at different cut-offs. For example, a test might quantify the level of a biomarker in a sample, or produce a score based on a questionnaire, such as the Patient Health Questionnaire (PHQ-9), higher scores on which are suggestive of more severe depression. Assuming that a higher value of the continuous test result is associated with increased probability of disease[1], by definition specificity can be increased by increasing the cut-off, but at the cost of a reduction in sensitivity. Similarly, sensitivity can be increased by lowering the cut-off, but specificity will fall as a result. In addition to continuous tests with explicit numerical thresholds, some index tests can be considered to have “implicit” thresholds: these are tests that do not produce a continuous numerical result but are dependent on interpretation, such as x-rays or other imaging technology, or lateral flow devices where readers might vary in their

[1] The opposite will hold for some tests, e.g. Cycle Threshold (Ct) values for many Polymerase Chain Reaction (PCR) tests, but the logic can be reversed.

interpretation of faint lines. In these instances, different interpreters may tend to be either more or less stringent in what they consider to be a positive result.

A receiver operating characteristic (ROC) curve is often used to show the relationship between sensitivity and specificity as the threshold is varied. This plots the sensitivity against the FPF for all thresholds. *Figure 1* shows an empirical ROC curve, using individual-level data from a study reporting on the accuracy of α-fetoprotein in detecting hepatocellular cancer in people with cirrhosis (11). For illustration purposes we show some of the numerical threshold values corresponding to points on the curve. The dashed unit line, representing a hypothetical test where Sensitivity = FPF is often shown as a reference. More accurate tests will have ROC curves lying further away from that dashed line: closer to the top left corner of the plot (i.e. high sensitivity and low FPF).

Figure 1: Example receiver operating characteristic (ROC) curve, plotting sensitivity against FPF (1 – specificity) across all thresholds.



Where numerical threshold values are shown on the ROC curve, as in *Figure 1*, the sensitivity and specificity corresponding to these values can be extracted from the plots using digitising software if not reported directly in the text or tables (8).

2.4. OTHER MEASURES

Other commonly reported measures include the Area Under the (ROC) Curve (AUC), the diagnostic odds ratio (DOR), and the positive and negative likelihood ratio. Both the AUC and DOR are single summary measures of accuracy, reflecting the overall ability of the test to distinguish between diseased and disease-free individuals. In *Figure 1*, the AUC is 0.67; a value of 1 would represent perfect discrimination. For details regarding its calculation see for example (12, 13). The DOR and likelihood ratios can be calculated as functions of sensitivity and specificity.

A digital interactive primer on evaluating diagnostic test accuracy, covering much of what has been outlined in *Section 2*, is available online (https://apps.crsu.org.uk/DTA-Primer/) together with an interactive graphic showing how sensitivity, specificity, threshold and ROC curves inter-relate (https://apps.crsu.org.uk/RocCurves/).

3. META-ANALYSIS OF SENSITIVITY AND SPECIFICITY: GENERAL PRINCIPLES

In DTA meta-analysis, it is widely recommended to meta-analyse sensitivity and specificity, rather than other measures of accuracy. Meta-analysis of PPV and NPV is generally discouraged due to the critical dependence of these measures on prevalence (14). Instead, pooled estimates of sensitivity and specificity can be used to estimate PPV and NPV at any given pre-test probability, and/or other summary measures that may be of interest (such as the DOR or likelihood ratios). Joint uncertainty in sensitivity and specificity is easily propagated through these calculations when the synthesis is performed using MCMC simulation: an example is provided in *Section 4*.

Three key features to consider in meta-analysis of sensitivity and specificity are:

1) There is often considerable heterogeneity in these measures across DTA studies in a meta-analysis. This would arise by definition if threshold varied

across studies – even if all studies lay on the same ROC curve. In practice, heterogeneity due to other factors is also typical.

2) Sensitivity and specificity can typically be expected to be negatively correlated across studies.
3) Small or zero cell counts in 2×2 tables (very few FPs or very few FNs in some studies) are not uncommon.

All synthesis methods described in this TSD allow for these three features where possible, by:

1) Incorporating random effects to allow for heterogeneity in sensitivity and specificity across studies.
2) Capturing the anticipated between-study correlation in sensitivity and specificity.
3) Modelling exact binomial likelihoods, rather than relying on normal approximations.

4. META-ANALYSIS OF ONE ESTIMATE OF SENSITIVITY AND SPECIFICITY PER STUDY

In this section we describe methods for meta-analysis of a single estimate of sensitivity and specificity per study. This may be the case for a truly binary test, a test with no explicit threshold[2], or if meta-analysis has been restricted to accuracy at a particular explicit threshold value.

It is now well recognised that the binomial bivariate random effects model for sensitivity and specificity (14) is equivalent to another proposed approach, the hierarchical summary receiver operating characteristic (HSROC) model (7), in the simple situation with no study-level covariates (15, 16). Originally the bivariate model was used to produce correlated estimates of "average"[3] sensitivity and specificity, typically represented by a transformed ellipse in ROC space, while the HSROC model was used

[2] See *Section 2.3*

[3] Mean on the logit scale, alternatively viewed as median sensitivity and specificity on the probability scale. We will also sometimes refer to these as "summary estimates" in this TSD.

primarily to produce a summary ROC curve, representing the typical trade-off between sensitivity and specificity across studies. Due to the equivalence, the HSROC curve can be derived from the bivariate model and vice versa. In this TSD, we provide code to fit the bivariate model and produce both types of output.

Note that the bivariate/HSROC model does not incorporate threshold values into the analysis. This means that, although a summary ROC curve can be estimated, it is not possible to know which point on the curve each threshold relates to. Either parameterisation is easily extended to incorporate study-level covariates, which could in principle include threshold. However, due to the specific functional form for the relationship between sensitivity, specificity and threshold – and because, where thresholds are numeric, it is common for accuracy at more than one threshold to be reported in some studies – we refer to *Section 5* for guidance on how to incorporate this information where available.

The HSROC curve is however an informative summary of the data when thresholds are implicit rather than explicit[4], as it describes the sensitivity that can be obtained at different values of specificity (and *vice versa*). Focus on a single summary point in ROC space will be most appropriate where threshold effects are considered to be minimal – for example, if all of the data relate to accuracy at the same explicit threshold value, or if the test produces a clear dichotomous result with little chance of reader variability.

Note that the HSROC curve supersedes earlier proposals on how to generate an SROC curve, such as the Moses and Littenberg approach (17, 18), which relies on normal approximations to binomial likelihoods and is discouraged now that exact approaches are available (see *Section 3*). The bivariate or HSROC model is also more general in that it allows for the scenario in which variation in threshold is not the only source of heterogeneity in sensitivity and specificity.

[4] See *Section 2.3*

4.1. BIVARIATE RANDOM EFFECTS MODEL

4.1.1. Notation

We use index $i = 1, \dots, I$ to denote study number, and index j to denote disease state, where

$$j = \begin{cases} 1 & \textit{diseased}\ (D = 1) \\ 2 & \textit{disease-free}\ (D = 0) \end{cases}$$

We denote by r_{ij} the number of participants who are positive on the index test in group j of study i. In other words, r_{i1} is the number of true positives (TP) and r_{i2} the number of false positives (FP) in study i. Further, let $p_{ij} = Pr(r_{ij} = 1)$, such that p_{i1} is the sensitivity and p_{i2} is the FPF in study i. Finally, we use N_{ij} to denote the total number of individuals in population j of study i.

4.1.2. Model specification

We assume binomial likelihoods for the number of positive test results within each disease state, in each study:

$$r_{ij} \sim \text{Binomial}(p_{ij}, N_{ij}) \qquad \textbf{(3)}$$

As sensitivity (p_{i1}) and FPF (p_{i2}) are probabilities (therefore bounded between 0 and 1), for the between-studies model a transformation (link function) is used to map these values into the real line $(-\infty, +\infty)$, most commonly the logit (log-odds) transformation:

$$\text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right).$$

Across studies, a bivariate normal distribution is assumed for either logit transformed sensitivity and specificity or – equivalently – logit transformed sensitivity and FPF. In this TSD we describe the latter version, i.e.:

$$\begin{pmatrix} \text{logit}(p_{i1}) \\ \text{logit}(p_{i2}) \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \Sigma\right)$$

where $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

Here, σ_1^2, σ_2^2 represent between-study variances of logit-transformed sensitivity and FPF respectively and ρ is the between-study correlation between these two measures, typically anticipated to be positive.

Note that a special case occurs when the correlation parameter ρ is one. In this case there is no variation in sensitivity at a given value of FPF (or *vice versa*). This means that all studies lie on the same ROC curve.

Summary estimates of sensitivity and FPF are produced by back-transforming the parameters m_1 and m_2, using the inverse-logit function, such that:

$$\text{Summary sensitivity} = \frac{e^{m_1}}{1 + e^{m_1}}$$

$$\text{Summary specificity} = 1 - \frac{e^{m_2}}{1 + e^{m_2}} = \frac{1}{1 + e^{m_2}}$$

In a Bayesian framework, this back-transformation is done at each iteration of the MCMC simulation, to produce posterior distributions for these pooled estimates. These summary estimates are correlated and uncertainty around them can be represented using a region (transformed ellipse) in ROC space, as we will demonstrate in *Section 4.4*.

Between-study heterogeneity in sensitivity and specificity is directly quantified through estimates of σ_1 and σ_2, similar to τ which is commonly used to denote the between-study heterogeneity in standard univariate random effects meta-analysis. These parameters represent standard deviations of the logit-transformed measures across studies. Similar to prediction intervals in univariate random effects meta-analysis (19), it can also be informative to derive and plot a 95% prediction region, to help visualise the extent of heterogeneity. This region represents the area in which we predict the sensitivity and specificity would lie in a hypothetical new study drawn from the bivariate random effects distribution. We can generate predictive distributions for sensitivity and FPF in a new study, allowing for their correlation, by simulating additional probabilities $logit(pred\_p_1)$ and $logit(pred\_p_2)$ from the bivariate normal distribution, and inverting the logit transformations, within the model code.

4.1.3. Prior distributions for hyperparameters

In fitting the bivariate model in a Bayesian framework, we need to ensure that the variance-covariance matrix Σ is positive semi-definite. One way to do so is to assume a vague inverse-Wishart prior for Σ. The results in this case however can be sensitive to values of hyper-parameters and might not correspond to sensible priors for ρ, σ_1, and σ_2 (20). An alternative approach is to specify the bivariate normal distribution in the following conditional normal format within the model code (20, 21):

$$\mathrm{logit}(p_{i1}) \sim N(m_1, \sigma_1^2)$$

$$\mathrm{logit}(p_{i2}) \sim N\left(m_2 + \rho\frac{\sigma_2}{\sigma_1}(\mathrm{logit}(p_{i1}) - m_1), (1-\rho^2)\sigma_2^2\right)$$

Vague prior distributions can then be placed on the parameters σ_1, σ_2 and ρ directly. Although we would typically expect ρ to be non-negative, in order to not force it to be estimated as positive we would suggest the fully vague prior $\rho \sim \mathrm{Uniform}(-1{,}1)$ as default when there are sufficient studies available to estimate this. Similarly, wide $\mathrm{Uniform}(0{,}5)$ prior distributions might be used for σ_1 and σ_2 in this situation.

Although Normal prior distributions with a large variance may seem the natural “vague” choice for m_1 and m_2, these place most of the weight on values of sensitivity and specificity close to zero or one. Uniform prior distributions on $\mathrm{logit}^{(-1)}(m_1)$ and $\mathrm{logit}^{(-1)}(m_2)$ are more attractive but can potentially slow down model fitting. As default fully vague priors for this model, we suggest use of $\mathrm{Logistic}(0{,}1)$ priors for m_1 and m_2, which are equivalent to $\mathrm{Uniform}(0{,}1)$ priors on the probability scale (i.e. for summary sensitivity and specificity).

See *Section 4.5* for discussion of choice of priors when the number of studies is small.

4.2. PRODUCING HSROC PARAMETERS AND CURVES FROM THE BIVARIATE MODEL

As noted above, the HSROC model is mathematically equivalent to the bivariate model[5], but it is parameterised differently and is conceptualised in terms of each study having its own ROC curve. In this section, we describe the HSROC model and how its parameters – and the HSROC curve – can be estimated from the bivariate model.

The HSROC model assumes the following parametric form for the logit-transformed probabilities in each study (as defined by equation (3)):

$$\mathrm{logit}\left(p_{ij}\right) = e^{-\beta X_j}(\theta_i + X_j\alpha_i)$$

where $X_1 = 1/2$ and $X_2 = -1/2$ (7). Here, θ_i, α_i and β are referred to as the cutpoint, accuracy and scale parameters respectively. The cutpoint parameter represents the point on the study's ROC curve corresponding to the reported data: as θ_i increases, both the sensitivity and the FPF increase. The scale parameter β defines the degree of asymmetry in each study's ROC curve around the diagonal line (representing Sensitivity = FPF). If β = 0 then the curve is symmetrical around this line.

To allow for heterogeneity across studies, parameters θ_i and α_i are assumed to be random effects as follows:

$$\theta_i \sim N(\Theta, \tau_\theta^2)$$

$$\alpha_i \sim N(\Lambda, \tau_\alpha^2)$$

where Θ, Λ are the means of the cutpoint and accuracy parameters respectively and τ_θ^2, τ_α^2 their respective variances. Note that because each study only provides a single point on its ROC curve, it is not feasible to allow also for varying scale parameter across studies: β is therefore assumed to be shared across all studies. This corresponds to all study-level ROC curves being parallel when plotted on the logit scale.

[5] In the absence of study-level covariates.

The HSROC curve is then defined by the equation:

$$\text{Sensitivity} = \text{logit}^{-1}\left(e^{-\beta}\text{logit}(\text{FPF}) + e^{-\beta/2}\Lambda\right)$$

evaluated across the observed range of FPF values across studies.

The parameters of the HSROC model can be written as the following functions of parameters of the bivariate model (15):

$$\beta = \log\left(\sigma_2/\sigma_1\right)$$
$$\Theta = \frac{1}{2}\left(\sqrt{\sigma_2/\sigma_1}\, m_1 + \sqrt{\sigma_1/\sigma_2}\, m_2\right)$$
$$\Lambda = \sqrt{\sigma_2/\sigma_1}\, m_1 - \sqrt{\sigma_1/\sigma_2}\, m_2$$
$$\sigma_\alpha^2 = 2\sigma_1\sigma_2(1-\rho)$$
$$\sigma_\theta^2 = \frac{1}{2}\sigma_1\sigma_2(1+\rho)$$

Note that the HSROC curve is symmetrical around the diagonal ($\beta = 0$) if $\sigma_1 = \sigma_2$. Further, if sensitivity and FPF are 100% correlated across studies ($\rho = 1$) then $\sigma_\alpha = 0$, i.e. there is no variability in “accuracy” across studies. This corresponds to there being only one single ROC curve, shared across all studies, so that the true sensitivities and FPFs all lie on the same curve.

We note that the Rutter and Gatsonis HSROC curve is only one of several possible SROC curves that is compatible with the bivariate model (16) but it is the most commonly used in practice and has the desirable feature of treating sensitivity and specificity equally.

4.3. RELATIONSHIP BETWEEN SENSITIVITY, SPECIFICITY AND PREVALENCE

Sensitivity and specificity are sometimes modelled jointly with prevalence, with some authors proposing trivariate models (22, 23) . We do not advocate this approach in general. Prevalence can be extremely variable across studies, particularly if some studies in the meta-analysis are of a diagnostic case-control or “two gate” design (24), in which case a pooled estimate of prevalence is meaningless. We would therefore avoid making unnecessary assumptions about the distribution of prevalence across studies.

We can see two main ways in which prevalence can be related to sensitivity and/or specificity. First, in cases where there is no explicit threshold, if there is an increase in the *a priori* probability of disease, a rational human observer would shift their threshold for detection, simultaneously increasing the probability of both true positives and false positives. This phenomenon has been widely studied in signal detection theory (25, 26) and would lead to a correlation between implicit threshold/cutpoint and prevalence.

A quite different situation is where higher disease prevalence is associated with the detectability of the disease itself, perhaps because a higher proportion of the tested population have the disease in a more advanced form. This could arise if, for example, studies with higher prevalence of cancer tended to have proportionally more cases of advanced cancer, which may be more likely to be detected. This would lead to a correlation between sensitivity and prevalence, but no correlation between specificity and prevalence.

If either form of systematic relationship with prevalence is suspected, we suggest this could be dealt with by including prevalence as a covariate, acting on the relevant model parameters. As with baseline risk in meta-analysis of interventions (27), we would recommend formulating this such that the covariate is “true” prevalence (after accounting for sampling variation) rather than the observed proportion. Further discussion is beyond the scope of this TSD.

4.4. CODE AND WORKED EXAMPLES

WinBUGS model code for the bivariate random effects model is provided in *Appendix A1*. This includes calculation of the HSROC parameters, facilitating plotting of the HSROC curve when appropriate. Additionally, the model code can be used to produce estimates of the PPV and NPV at any specified value of pre-test probability or prevalence, based on estimates of sensitivity and specificity from the bivariate model. Evaluating equations (1) and (2) at each iteration of the MCMC facilitates production of 95% CrIs around estimates that allow for uncertainty in both sensitivity and specificity and the correlation between the two. Note that consideration should be given to the extent to which these summary estimates are meaningful (i.e. whether significant threshold effects are anticipated) before using them in this way.

Two types of figure are commonly used to show results from meta-analysis of test accuracy: coupled forest plots of sensitivity and specificity, and plots in ROC space, often referred to as summary ROC or SROC plots. SROC plots show study-level estimates of sensitivity and specificity in ROC space, and should include either or both of the following:

- Estimates of summary or "average" estimates, with a 95% credible region representing joint uncertainty in these and (often) a 95% prediction region
- An HSROC curve, which can also be plotted with 95% credible intervals and prediction intervals. Note that we would recommend against extrapolating this curve beyond the range of the observed data.

Where summary estimates with 95% regions are shown, the credible region is usually based on a bivariate normal approximation to the posterior distribution of m_1 and m_2. An ellipse on the logit scale is estimated from this assumption, which is transformed to the probability scale for plotting. Similarly, the shape of the prediction region is usually based on a bivariate normal approximation to the joint posterior of $\mathrm{logit}(pred\_p_1)$ and $\mathrm{logit}(pred\_p_2)$. Note that the correlation parameters for these bivariate distributions are *not* equal to ρ (which represents the correlation between the logits of sensitivity and FPF from the same study) but must instead by obtained as the correlation between posterior samples for m_1 and m_2, and the correlation between posterior samples for $\mathrm{logit}(pred\_p_1)$ and $\mathrm{logit}(pred\_p_2)$ respectively.[6]

If the bivariate normal approximation to the joint posterior appears poor – which may be the case if either summary sensitivity or specificity is very close to one – analysts might consider other approaches to summarising the joint posterior for these plots, such as kernel based approximations.

[6] When running WinBUGS in the "point and click" way rather than through R2WinBUGS, these values are obtained after fitting the model by clicking "Inference" then "Correlations" and specifying the two parameters of interest.

We now describe two worked examples which differ in terms of which outputs may be considered most informative. R code to fit the bivariate model via R2WinBUGS and produce the figures shown is available at the GitHub repository.

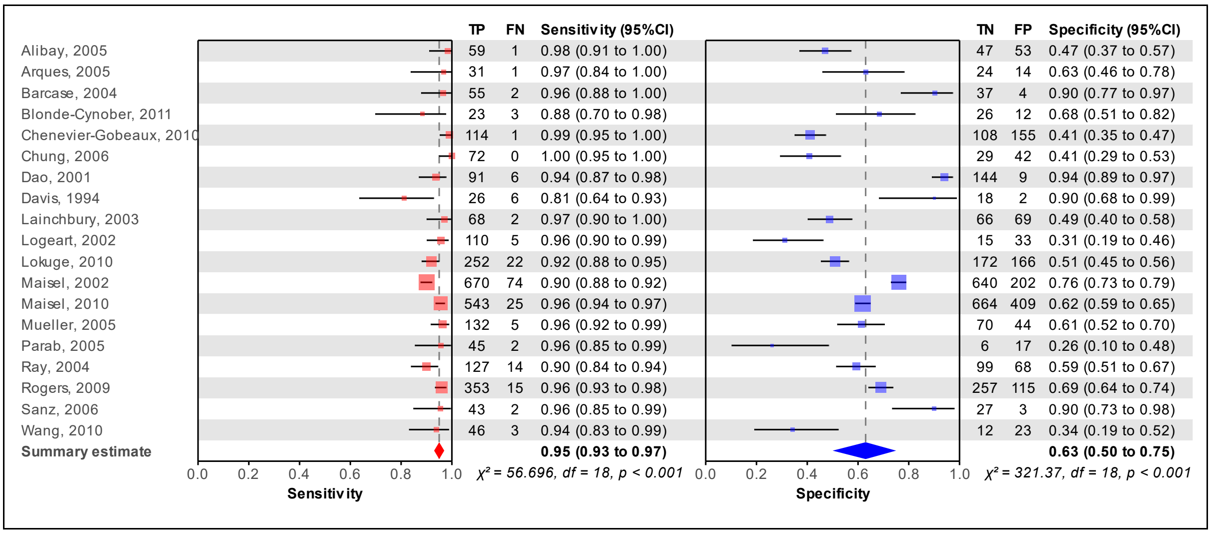
For analysts less familiar with producing plots in R, we note the availability of R package `DTAplots` (28), which can also be used to produce coupled forest plots and SROC plots using posterior outputs from a Bayesian bivariate meta-analysis (functions `Forest` and `SROC_rjags` respectively). Example code to produce slightly simplified versions of the coupled forest and SROC plots for Example 1 is also provided in the GitHub repository.

4.4.1. Example 1: Accuracy of B-type natriuretic peptide (at a threshold of 100ng/L) for diagnosis of acute heart failure

We first consider data from a systematic review of the accuracy of B-type natriuretic peptide (BNP) in diagnosis of acute heart failure, carried out to inform NICE guidelines (29). To demonstrate the bivariate model, we consider only 2×2 tables corresponding to a cut-off of approximately 100ng/L: 19 studies reported such data. Data are shown on *Figure 2*. Note that, in practice, data were also available at other thresholds for this example (8, 29), and we would therefore recommend using the model described in *Section 5* for the main synthesis, with the bivariate model fitted as a sensitivity analysis if the threshold of 100ng/L is of particular interest for the decision model (see *Section 5.2.3*).

From the bivariate model, we obtain pooled estimates of sensitivity = 0.95 (95% credible interval, CrI, 0.93, 0.97) and specificity = 0.63 (95% CrI 0.50, 0.75). The between-study standard deviation parameters for logit(sensitivity) and logit(FPF) are estimated as 0.62 (95% CrI 0.33, 1.13) and 1.09 (95% CrI 0.74, 1.67) respectively, while the between-study correlation, ρ, is estimated as 0.49 (95% CrI −0.15, 0.85). *Figure 2* shows coupled forest plots, showing both study-level and pooled estimates.

Figure 2: Coupled forest plots for the B-type natriuretic peptide (100ng/L) example



	TP	FN	Sensitivity (95%CI)	TN	FP	Specificity (95%CI)
Alibay, 2005	59	1	0.98 (0.91 to 1.00)	47	53	0.47 (0.37 to 0.57)
Arques, 2005	31	1	0.97 (0.84 to 1.00)	24	14	0.63 (0.46 to 0.78)
Barcase, 2004	55	2	0.96 (0.88 to 1.00)	37	4	0.90 (0.77 to 0.97)
Blonde-Cynober, 2011	23	3	0.88 (0.70 to 0.98)	26	12	0.68 (0.51 to 0.82)
Chenevier-Gobeaux, 2010	114	1	0.99 (0.95 to 1.00)	108	155	0.41 (0.35 to 0.47)
Chung, 2006	72	0	1.00 (0.95 to 1.00)	29	42	0.41 (0.29 to 0.53)
Dao, 2001	91	6	0.94 (0.87 to 0.98)	144	9	0.94 (0.89 to 0.97)
Davis, 1994	26	6	0.81 (0.64 to 0.93)	18	2	0.90 (0.68 to 0.99)
Lainchbury, 2003	68	2	0.97 (0.90 to 1.00)	66	69	0.49 (0.40 to 0.58)
Logeart, 2002	110	5	0.96 (0.90 to 0.99)	15	33	0.31 (0.19 to 0.46)
Lokuge, 2010	252	22	0.92 (0.88 to 0.95)	172	166	0.51 (0.45 to 0.56)
Maisel, 2002	670	74	0.90 (0.88 to 0.92)	640	202	0.76 (0.73 to 0.79)
Maisel, 2010	543	25	0.96 (0.94 to 0.97)	664	409	0.62 (0.59 to 0.65)
Mueller, 2005	132	5	0.96 (0.92 to 0.99)	70	44	0.61 (0.52 to 0.70)
Parab, 2005	45	2	0.96 (0.85 to 0.99)	6	17	0.26 (0.10 to 0.48)
Ray, 2004	127	14	0.90 (0.84 to 0.94)	99	68	0.59 (0.51 to 0.67)
Rogers, 2009	353	15	0.96 (0.93 to 0.98)	257	115	0.69 (0.64 to 0.74)
Sanz, 2006	43	2	0.96 (0.85 to 0.99)	27	3	0.90 (0.73 to 0.98)
Wang, 2010	46	3	0.94 (0.83 to 0.99)	12	23	0.34 (0.19 to 0.52)
Summary estimate			**0.95 (0.93 to 0.97)**			**0.63 (0.50 to 0.75)**

$\chi^2 = 56.696$, df = 18, p < 0.001

$\chi^2 = 321.37$, df = 18, p < 0.001

Figure 3 (left panel) shows the SROC plot. Here, the grey circles represent the observed sensitivity and specificity in each study, with studies with larger sample sizes depicted by larger circles. The red triangle shows the estimated average sensitivity and specificity, while the red dashed line depicts a 95% credible region, representing joint uncertainty in these estimates. The 95% prediction region (blue dashed line) represents joint bounds on the expected true sensitivity and specificity in a future hypothetical study, drawn from the same bivariate random effects distribution. For this data set, we see that the prediction region is very wide in the x-axis direction, indicating a large amount of heterogeneity in specificity across studies. As the threshold is the same (100ng/L) for all data included, we did not include an HSROC curve on this plot.

In the right panel of *Figure 3*, we show the SROC plot on the logit scale: we see the estimated positive correlation, and how the joint uncertainty in summary measures and the joint predictive distribution are represented by ellipses on this scale.

Using the summary estimates, we also calculated the PPV and NPV with corresponding 95% CrIs (allowing for joint uncertainty in sensitivity and specificity), for all possible prevalence or pre-test probability values. “Leaf plots” (30) can be used to visualise how a positive or negative test result impacts probability of disease: an example is shown in *Figure 4* (red = PPV, blue = 1 – NPV).

Figure 3: SROC plot for the B-type natriuretic peptide (100 ng/L) example: on the probability scale (left) and logit scale (right)

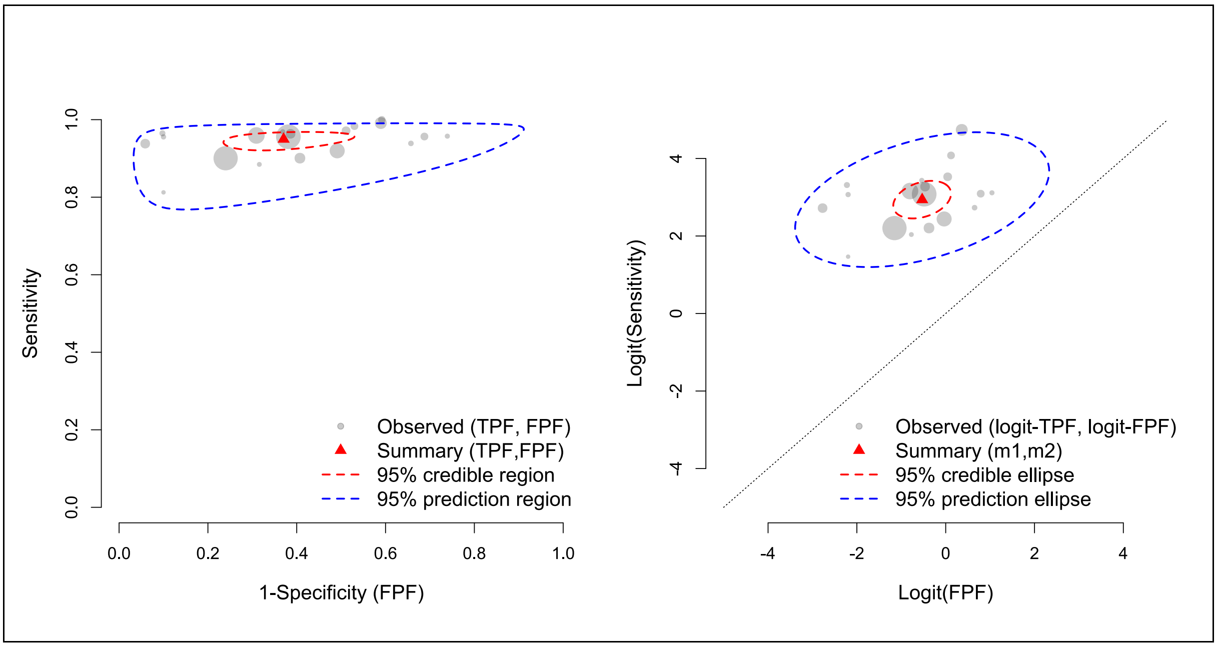
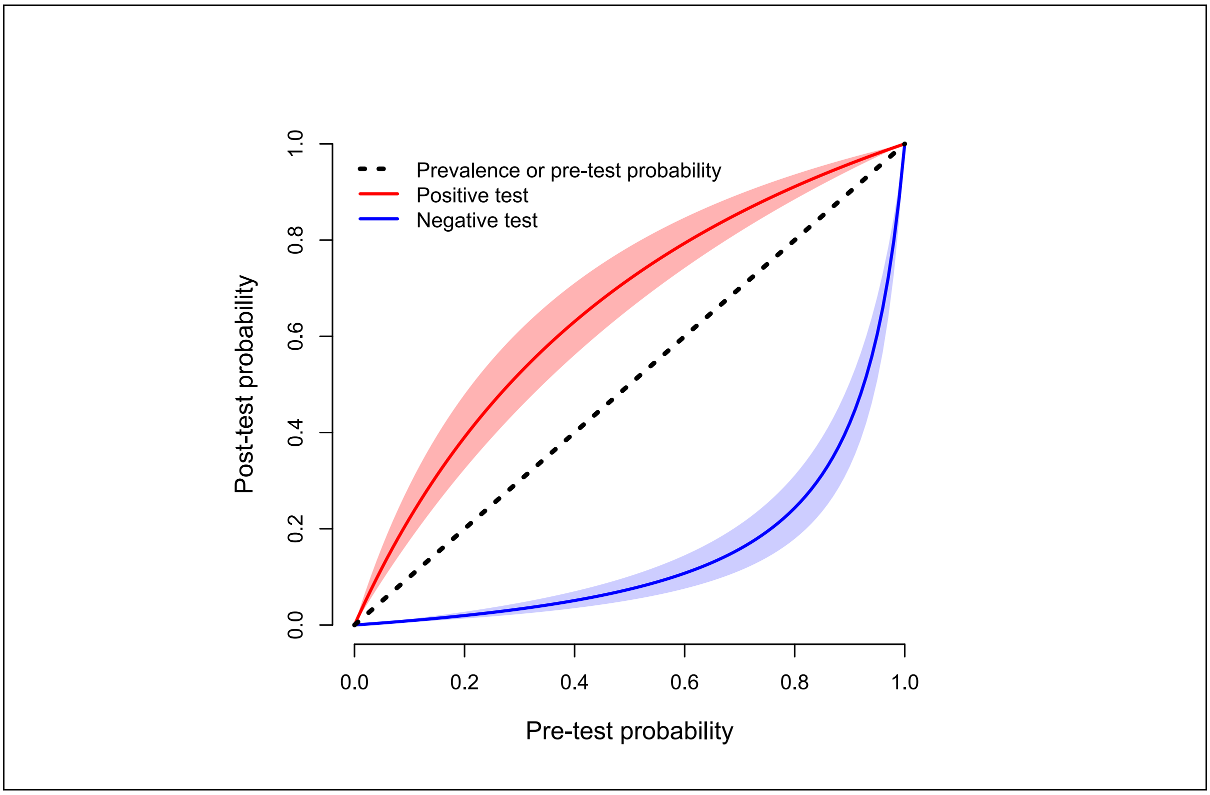


Figure 4: Leaf plot: Post-test probability given a positive or negative test result, for each pre-test probability or prevalence. Estimates based on summary sensitivity and specificity from bivariate random effects model, fitted to the B-type natriuretic peptide (100ng/L) example. Shaded areas are 95% CrIs



4.4.2. Example 2: Accuracy of radiography for detection of enamel caries

Example 2 consists of data from 45 studies reporting on the accuracy of analogue or digital radiography for detection of enamel caries, among asymptomatic patients (31). Thresholds are implicit, since the index test does not produce a numerical value, but – as is typical with imaging tests – there is some degree of subjectivity in interpretation of images as positive or negative.

We fitted the bivariate random effects model using the code provided. As some implicit threshold effects are anticipated for this test, we also calculated the HSROC parameters. Both sets of parameter estimates are shown in *Table 2*.

Table 2: Bivariate and HSROC model parameter estimates for the enamel caries example

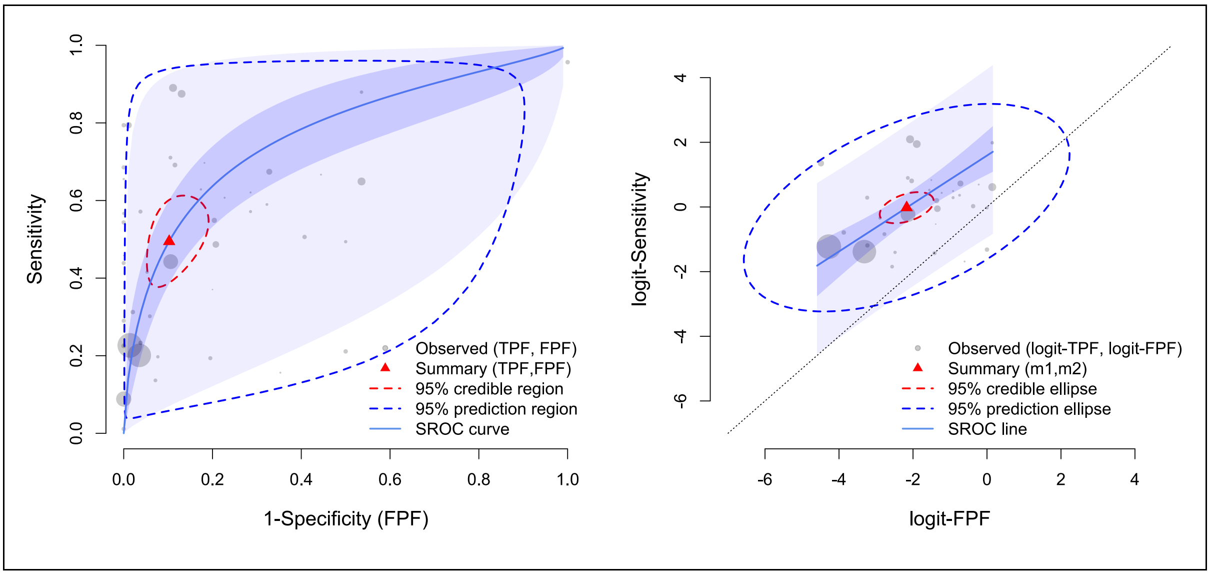
	Parameter	Interpretation	Estimate (95% CrI)
Bivariate model parameters	m_1	Mean logit-sensitivity	−0.02 (−0.41, 0.37)
	m_2	Mean logit-FPF	−2.17 (−2.78, −1.61)
	σ_1	Standard deviation of logit-sensitivity	1.27 (1.02, 1.63)
	σ_2	Standard deviation of logit-FPF	1.72 (1.28, 2.39)
	ρ	Correlation between logit-sensitivity and logit-FPF	0.52 (0.20, 0.75)
HSROC parameters	Λ	Mean accuracy parameter	1.85 (1.29, 2.42)
	Θ	Mean cutpoint parameter	−0.95 (−1.37, −0.53)
	β	Scale parameter (assumed constant across studies)	0.30 (−0.05, 0.66)
	σ_α	Standard deviation of accuracy parameters	1.44 (1.07, 1.96)
	σ_θ	Standard deviation of cutpoint parameters	1.28 (1.01, 1.68)

Figure 5 (left panel) shows the SROC plot, this time also including (unlike *Figure 3*) the HSROC curve. The darker and lighter shaded areas represent the 95% CrI and predictive intervals around the HSROC curve respectively. The HSROC curve is asymmetrical, as indicated by the point estimate of the scale parameter β not being equal to zero. A positive β indicates a larger spread of test results in the disease-free compared with diseased populations, although we note that in this example the CrI is wide and also contains zero. When plotted on the logit scale (*Figure 5*, right panel), we

see how the HSROC line is linear on this scale, with slope $e^{-\hat{\beta}} = 0.74$, a little less than the 1 (represented by the dotted black diagonal line).

For this example, we observe a very large degree of heterogeneity in both sensitivity and specificity: this is clear from the extremely wide 95% prediction region and large estimated standard deviations for sensitivity and specificity on the logit scale across studies (*Figure 5* and *Table 2*). Note that – theoretically – these large standard deviations could be consistent with a common true ROC curve across all studies. This would suggest that variation in threshold alone explained all of the heterogeneity in sensitivity and specificity. This is clearly not the case in this example, however. The CrI for the estimated correlation ρ is far from 1 (0.52, 95% CrI 0.20 to 0.75), such that we observe a large estimated standard deviation of 1.44 (95% CrI 1.07 to 1.96) in accuracy parameters across studies. This estimated spread of ROC curves across studies is also represented by the wide prediction intervals around the HSROC curve.

Figure 5: SROC plot for the radiography for enamel caries detection dataset



4.5. FITTING THE BIVARIATE MODEL WHEN THE NUMBER OF STUDIES IS SMALL

This section considers the situation in which only a small number of studies is available, but synthesis is still required in order to provide the best representation of the available evidence to inform the cost-effectiveness analysis.

As noted in *Section 3*, there is often considerable between-study heterogeneity in DTA meta-analyses and – as we saw from Example 2 – heterogeneity can be present on both parameter scales (logit-transformed sensitivity and specificity, or the HSROC parameters). The bivariate random effects model requires estimates of two between-study standard deviation parameters and the correlation parameter, ρ, in addition to the two means (m_1 and m_2).

When only a small number of studies is available for synthesis, fitting the bivariate model with fully vague prior distributions as described in *Section 4.1.3* can lead to poor or non-identifiability of some parameters. Even if all parameters are at least weakly identified, posterior distributions may be extremely wide, with little change from their vague prior distributions. As is also the case with meta-analysis of intervention effectiveness, the posterior distributions for between-study standard deviation parameters might include values which are, on reflection, infeasible (32).

In these circumstances, some simplification of the model to remove one or more parameters might be considered. If implicit threshold effects are expected, thinking in terms of the HSROC parameterisation can assist in model simplification. If data are insufficient to estimate an asymmetric HSROC curve, estimation of a symmetric curve may be a reasonable approximation (33). In the bivariate parameterisation, this corresponds to setting $\sigma_1 = \sigma_2$. The WinBUGS code provided can be easily modified to set this constraint. A further potential simplification might be to assume that *all* heterogeneity is explained by varying threshold, such that all studies lie on the same ROC curve. This corresponds to setting $\rho = 1$. Note that a single symmetric ROC curve ($\rho = 1, \sigma_1 = \sigma_2$) can be estimated from even a single 2×2 table.

Use of informative or weakly informative prior distributions, principally for standard deviation and correlation parameters, is another option within a Bayesian framework. Guidance is available for using external evidence to inform prior distributions for variance components in meta-analysis of intervention effectiveness (34-36), and the same general principles apply in DTA: prior distributions could be based on “similar” meta-analyses (e.g. alternative tests under consideration for the same decision problem, that operate similarly and produce results in the same units), or elicited from experts. It can be helpful to consider how large σ_1 or σ_2 could really be in practice. As described by Spiegelhalter *et al.* (37), if $\delta_i \sim N(m, \sigma^2)$ across studies, then the 2.5th and 97.5th percentile of $\exp(\delta)$s across studies lie $\exp(3.92\sigma)$ apart, which might be considered the "range" of odds across studies. For example, if the median sensitivity across studies is 0.60, then $\sigma_1 = 1$ would mean that the true sensitivity lies between 0.17 and 0.91 in 95% of studies. See *Table 3* for additional examples. We see that, unless median sensitivity or specificity is very close to 1, even an upper limit of 2 for standard deviation parameters would allow for the possibility of sensitivity or specificity varying across practically the entire [0,1] range.

Where only a small number of studies are available for synthesis, analysts should consider what the maximum realistic value of these standard deviation parameters could be, based on consideration of how diverse sensitivity and/or specificity could realistically be across studies (*Table 3*) and formulate appropriate prior distributions, e.g. Half-Normal, with consideration to this (37). For example, Half-Normal(0, 0.51^2), Half-Normal(0,1.02^2) and Half-Normal(0,1.53^2) place just 5% of probability on the standard deviation being greater than 1, 2 or 3, respectively.

Table 3: Interpretation of between-study standard deviation parameters in the bivariate model

σ	"Range" of odds across studies	Implied range on the probability scale		
		p = 0.6	p = 0.9	p = 0.995
0.2	2.19	0.50 and 0.69	0.86 and 0.93	0.99 to 1.00
0.5	7.10	0.36 and 0.80	0.77 and 0.96	0.99 to 1.00
1.0	50.40	0.17 and 0.91	0.56 and 0.98	0.97 to 1.00
2.0	2539.84	0.03 and 0.99	0.15 and 1.00	0.80 to 1.00
3.0	127999.79	0.00 and 1.00	0.02 and 1.00	0.36 to 1.00

The prior distributions described for ρ, m_1 and m_2 in *Section 4.1.3* are also extremely vague and could be tightened up for specific applications based on prior knowledge. For example, we would usually expect ρ to be non-negative, and would anticipate values of ρ close to one for tests with obvious threshold effects. As such, a prior such as $\rho \sim \text{Uniform}(0.5, 1)$ could be reasonable when reader variability is clearly anticipated. Where threshold effects are not expected, other factors that vary across studies can also lead to a positive correlation, such as differences in setting (e.g.

laboratory or community), study quality issues (e.g. imperfect reference standard in some studies) or population differences. For example, recall that in Example 1, where all studies used the same explicit threshold, the posterior estimate of ρ was ~0.5. A prior distribution for ρ can be centred around an initial "guess" $\hat{\rho}$ by setting the prior on Fisher's Z transformation of ρ:

$$z \sim N\left(\frac{1}{2}\log\left(\frac{\hat{\rho}+1}{\hat{\rho}-1}\right), v^2\right)$$

where v is selected to represent reasonable uncertainty on the ρ scale. This transformation can be undone within the model code as:

$$\rho = \frac{e^{2z}-1}{e^{2z}+1}$$

For example, z~ $N(0.55, 0.2^2)$ produces a prior distribution for ρ that is centred around 0.5, with 95% prior interval from 0.16 to 0.74.

Standard model fit tools can be used to explore different options for a given data set. Comparing posterior mean residual deviance in restricted and unrestricted parameter models can help show where any restrictions are in conflict with the data (21). See *Section 5.3* for an example.

4.6. OTHER SOFTWARE OPTIONS

Alternative options to the WinBUGS code provided with this report to fit the bivariate/HSROC model exist in most popular statistical analysis packages. Several options are described below for R (including online app interfaces), Stata and specialist Bayesian programs. A more in-depth coverage of this topic, including code for SAS, is available elsewhere (33).

Using the R software, packages (38, 39) to fit the bivariate model using frequentist methods are available. However, these use a normal approximation to the Binomial likelihoods and thus we do not recommend their use. The alternative is to use the `glmer` function within the general linear mixed effects package `lme4` (40). Details of how to use this for DTA meta-analysis are available elsewhere (33). An online app –

MetaDTA (41) - exists for facilitating the bivariate/HSROC analyses of DTA meta-analysis data, using the `lme4` package behind a non-technical “point and click” user interface (42). When using `lme4`, the user should be aware that parameter estimation is not conducted via a simulation method, therefore in order to be able to incorporate the anticipated between-study correlation in sensitivity and specificity in a decision model, the user has two options: a) assume bivariate normality of the parameter estimates for (logit) sensitivity and specificity; or b) estimate uncertainty in the parameter estimates via bootstrapping (e.g. using the `bootMer` command in the `lme4` package). The advantage of the latter is that, as noted in *Section 4.4*, the normal approximation can be poor when sensitivity or specificity is close to 1, and bootstrapping provides a way of providing correlated stochastic samples to inform a decision model (see later) (N.B the bootstrap analysis cannot be produced in MetaDTA currently).

Similar frequentist model fitting options also exist for Stata software. Bivariate models can be fitted in Stata using a generalised linear mixed methods approach implemented by the user-written programs `metandi` (43), `midas` (44) and `metadta` (45, 46). In a similar manner to that described above for `lme4` in R, in order to avoid normal approximations to parameter estimates, the `bootstrap` command is required to provide correlated samples for stochastic decision models.

Alternative approaches to fitting the bivariate/HSROC model via Bayesian methods also exist. The JAGS software (47) provides an alternative Bayesian MCMC computation “engine” to WinBUGS and uses very similar coding syntax; specific code is available elsewhere (33). Stan (48) is another Bayesian model estimation program (although it uses different algorithms from WinBUGS and JAGS). The online app MetaBayesDTA (49) provides a “point & click” interface to this program for fitting the bivariate meta-analysis DTA model. Additionally, two packages in R `bamdit` (50) and `meta4diag` (51) fit the bivariate model using Bayesian methods. The former uses JAGS as the computational engine, while the latter uses a further R package, `INLA` (52), which is non-simulation based so doesn't automatically provide the necessary stochastic parameter samples required for informing a decision model.

4.6.1. Software options when the number of studies is small

When analysing small numbers of studies, model convergence problems exist using frequentist approaches, and practical experience suggests these can often be overcome using Bayesian methods, even with relatively vague prior distributions. It will often not be possible to fit the versions of the model with simplified structure or informative prior distributions (described in *Section 4.5*) when using the Bayesian or frequentist packages for R and Stata, nor for the online apps cited. Finally, bootstrap sampling is not recommended generally for datasets with very small sample sizes. Therefore, when the number of studies is small, working directly with WinBUGS, JAGS (or STAN etc) code is recommended.

5. META-ANALYSIS OF THE ACCURACY OF CONTINUOUS TESTS ACROSS EXPLICIT THRESHOLDS

Where index tests provide a numerical value, it is common to find that some studies in a systematic review report on the accuracy at more than one threshold. A number of models have been proposed to synthesise data of this type and produce pooled estimates of the sensitivity and specificity at each explicit threshold (8, 53, 54). In this section we describe the model proposed by Jones *et al.* (8) as this is fitted in WinBUGS or JAGS and produces probabilistic results, facilitating use in a decision model. A brief discussion of other models for data of this type is included in *Section 5.4*.

We emphasise that the NICE decision problem may not include choice of threshold. Where there is a clear test developer specified threshold, a focus on cost-effectiveness at that threshold will typically be agreed in the scope. However, even if the decision model requires only estimates of the sensitivity and specificity at one pre-specified threshold, we would recommend fitting a model incorporating data at all available thresholds, as we would typically expect to obtain more precise estimates of accuracy at this critical threshold from a unified analysis. The bivariate model (*Section 4.1*) can be fitted to the data relating to the manufacturer-recommended threshold as a sensitivity analysis, allowing identification and exploration of any inconsistencies (*Section 5.2.3*).

5.1. MULTIPLE THRESHOLDS MODEL

5.1.1. Notation and data format

As previously, we use index $i = 1, \dots, I$ to denote study number, and index j to denote disease state. Now, say that each study i reports estimates of test accuracy at explicit threshold values C_{it}, where $t = 1, \dots, T_i$. Note that both the threshold values themselves and the number of thresholds reported on can vary across studies, and T_i may equal 1 in some studies. As previously, let N_{ij} denote the total number of individuals in population j of study i.

We denote with y_{ijl} the result of the continuous index test for the lth individual in disease group j of study i. We assume that higher values of this continuous measure are associated with an increased probability of disease. Let $p_{ijt} = P(y_{ijl} \geq C_{it})$, such that p_{i1t} is the sensitivity and p_{i2t} the FPF at threshold C_{it} in study i. Let x_{i1t} and x_{i2t} denote the number of TPs and FPs respectively at threshold C_{it} in study i. Data from study i at multiple thresholds can then be written in the format shown in *Table 4*. Note that each study can contribute data at a different set of thresholds. If individual participant data are available from a study, this can be reformatted to *Table 4* format, considering each unique value of continuous test result as an additional threshold. This allows the full individual level data to contribute to the synthesis.

Table 4: Format of data required from each study, indexed i, to fit the multiple thresholds model

		Number of individuals with index test result $> \boldsymbol{C_{it}}$			
Threshold value		$\boldsymbol{C_{i1}}$	$\boldsymbol{C_{i2}}$	**…**	$\boldsymbol{C_{iT_i}}$
True Disease status	**D=1**	x_{i11}	x_{i12}	…	x_{i1t_i}
	D=2	x_{i21}	x_{i22}	…	x_{i2t_i}

Note that by definition counts x_{i1t} and x_{i2t} must decrease with increasing C_{it} (although may stay constant between thresholds that are close together, particularly in small studies).

5.1.2. Model specification

The likelihood can be specified as the following series of conditional independent binomial distributions (8):

$$x_{ij1} \sim \text{Binomial}\left(p_{ij1}, N_{ij}\right)$$

$$x_{ijt} | x_{ij,t-1} \sim \text{Binomial}\left(\frac{p_{ijt}}{p_{ij,t-1}}, x_{ij,t-1}\right), \qquad t = 2, \dots, T_i$$

We further assume that there exists a monotonic transformation $g()$ of the index test's results to a symmetrical distribution (i.e. removing skew) such that:

$$g\left(y_{ijl}\right) \sim \text{Logistic}(\mu_{ij}, \sigma_{ij})$$

Under this assumption, the following parametric relationship between sensitivity, FPF and threshold holds:

$$\text{logit}\left(p_{ijt}\right) = \frac{\mu_{ij} - g\left(C_{ij}\right)}{\sigma_{ij}}$$

The transformation $g()$ may be pre-specified based on knowledge of the distribution of the continuous test's result in the diseased and disease-free populations. For tests producing right skewed measures, e.g. biomarkers, $g() = log()$ may often be a reasonable approximation (8).

Jones *et al.* (8) additionally describe a more flexible version of the model, in which $g()$ is defined by a Box–Cox transformation:

$$g(C_{it}) = \begin{cases} \dfrac{\left(C_{it}^{\lambda} - 1\right)}{\lambda}, & if\ \lambda \neq 0 \\ \log(C_{it}), & if\ \lambda = 0 \end{cases}$$

Here, λ is an additional parameter to be estimated from the data. Note that this model simplifies to the $g() = log()$ version when $\lambda = 0$. As such, we recommend fitting the "full" version where feasible with the data available.

The means (μ_{i1} and μ_{i2}) and scale parameters (σ_{i1} and σ_{i2}) of the underlying logistic distributions are assumed to vary across studies as follows:

$$\begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \log(\sigma_{i1}) \\ \log(\sigma_{i2}) \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} m_{\mu 1} \\ m_{\mu 2} \\ m_{\sigma 1} \\ m_{\sigma 2} \end{bmatrix}, S\right)$$

where S is a 4×4 covariance matrix. Jones *et al.* (8) discuss different structures for S, with the most general case being an unstructured symmetrical matrix. Here we focus on a restricted structure for S, where correlations are assumed to be caused only by dependencies between (μ_{i1}, μ_{i2}), $(\log(\sigma_{i1}), \mu_{i1})$ and $(\log(\sigma_{i2}), \mu_{i2})$. Under this assumption we can re-write the random effects as the following series of univariate conditional normal distributions:

$$\mu_{i1} \sim \mathrm{N}\left(m_{\mu 1}, \tau_{\mu 1}^2\right)$$

$$\mu_{i2} | \mu_{i1} \sim \mathrm{N}\left(m_{\mu 2} + \rho_\mu \frac{\tau_{\mu 2}}{\tau_{\mu 1}}(\mu_{i1} - m_{\mu 1}), (1 - \rho_\mu^2)\tau_{\mu 2}^2\right)$$

$$\log(\sigma_{ij}) | \mu_{ij} \sim \mathrm{N}\left(m_{\sigma j} + \rho_{\mu\sigma} \frac{\tau_{\sigma j}}{\tau_{\mu j}}(\mu_{ij} - m_{\mu j}), (1 - \rho_{\mu\sigma}^2)\tau_{\sigma j}^2\right), j = 1,2$$

We can obtain summary estimates of sensitivity and FPF for any threshold (C) of interest within the observed range as:

$$\text{Summary Sensitivity} = \text{logit}^{-1}\left(\frac{m_{\mu 1} - g(C)}{m_{\sigma 1}}\right)$$

$$\text{Summary FPF} = \text{logit}^{-1}\left(\frac{m_{\mu 2} - g(C)}{m_{\sigma 2}}\right)$$

To enable production of prediction intervals, we additionally draw predictive values from the four random effects distributions, similar to *Section 4.1.2*:

$$pred\_\mu_1 \sim \mathrm{N}\left(m_{\mu 1}, \tau_{\mu 1}^2\right)$$

$$pred\_\mu_2 \sim \mathrm{N}\left(m_{\mu 2} + \rho_\mu \frac{\tau_{\mu 2}}{\tau_{\mu 1}}(pred\_\mu_1 - m_{\mu 1}), (1 - \rho_\mu^2)\tau_{\mu 2}^2\right)$$

$$pred\_log(\sigma_j) \sim \mathrm{N}\left(m_{\sigma j} + \rho_{\mu\sigma} \frac{\tau_{\sigma j}}{\tau_{\mu j}}(pred\_\mu_j - m_{\mu j}), (1 - \rho_{\mu\sigma}^2)\tau_{\sigma j}^2\right), j = 1,2$$

5.1.3. *Prior distributions for hyperparameters*

Very vague options for prior distributions are[7]:

$$m_{\mu j} \sim \mathrm{N}(0{,}1000),\ m_{\sigma j} \sim \mathrm{N}(0{,}1000)$$
$$\tau_{\mu j} \sim \mathrm{Uniform}(0{,}5),\ \tau_{\sigma j} \sim \mathrm{Uniform}(0{,}5)$$
$$\rho_{\mu} \sim \mathrm{Uniform}(-1{,}1),\ \rho_{\mu\sigma} \sim \mathrm{Uniform}(-1{,}1)$$

$$\lambda \sim \mathrm{Uniform}(-3{,}3)$$

See *Section 5.3* for discussion of choice of priors when the number of studies is small.

5.2. CODE AND WORKED EXAMPLE

We demonstrate the use of the multiple thresholds model with the following worked example. WinBUGS code to fit the multiple thresholds model with unknown Box–Cox transformation parameter and restricted covariance structure is shown in *Appendix A2*. Code to bring the data into the appropriate format for the model in R, call WinBUGS to run all versions of the model, and produce the figures is available from the GitHub repository.

5.2.1. *Example 3: Accuracy of OC-Sensor in detecting colorectal cancer*

We use a data set from DG56 (55), which assessed the potential use of faecal immunochemical tests (FIT) to guide colorectal cancer pathway referrals, among symptomatic patients presenting in primary care. While FIT testing (at a threshold of 10µg/g) was already recommended as a triage tool for further investigations in patients with low risk symptoms, this appraisal assessed the potential use of FIT in patients with medium and high risk symptoms. Several FIT testing strategies were explored, including varying the threshold used to denote a positive result. Data on the accuracy

[7] Note that, unlike in *Section 4.1.3*, we do not suggest logistic distributions for the 'm' parameters. This is because these are not logit-transformed probabilities in this model: they represent means and $\log(SD)$ of continuous measure on the $g()$ scale.

of the OC-Sensor test were available from 11 studies, reporting 2×2 tables relating to between 1 and 10 different thresholds, with 14 different threshold values reported on in total (between 4 and 200 µg/g).

Preliminary data investigation suggests that that logit transformed TPFs and FPFs are approximately linear with log transformed threshold (*Figure 6*). As the key assumption of the multiple thresholds model is that the relationship between these logit transformed probabilities and $g(C)$ is linear within each study, this may suggest that $g() = log()$ is a reasonable approximation for this data set.

Figure 6: Checking the linearity of the OC-Sensor data given the log transformation

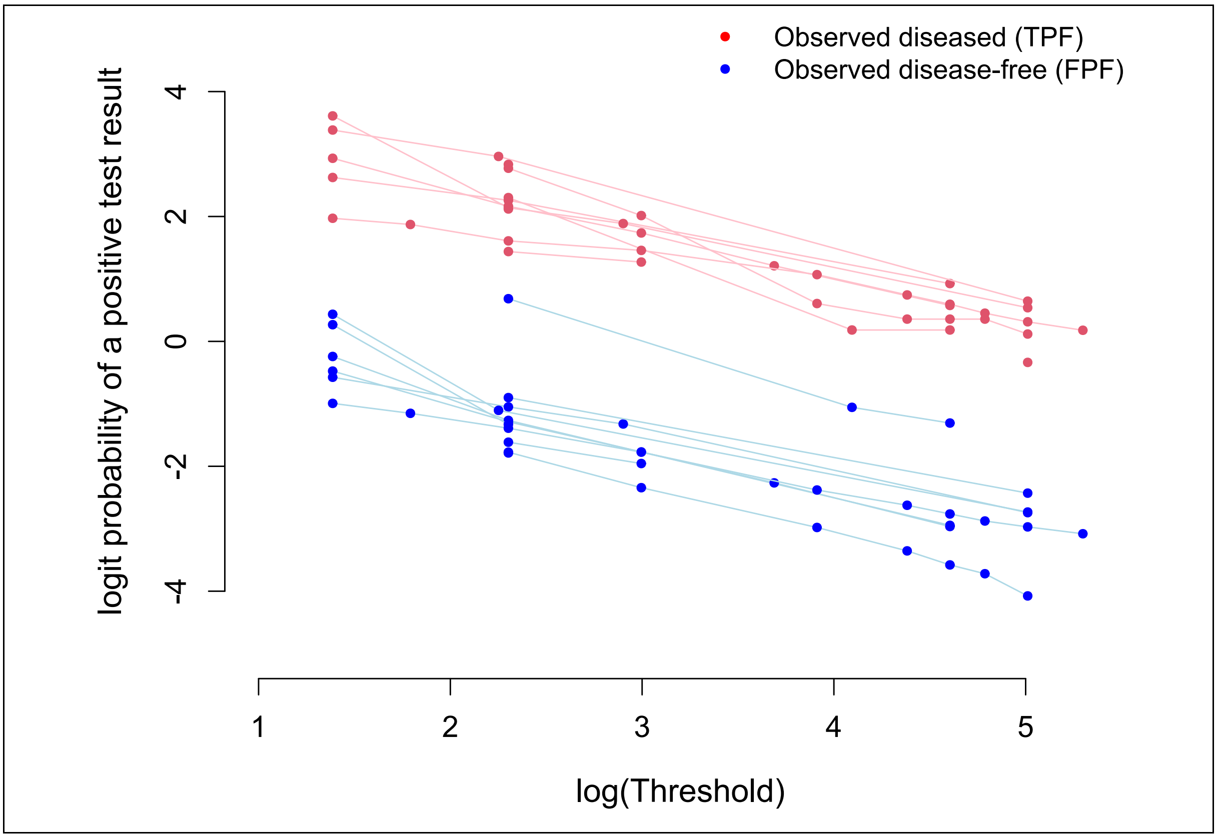


Table 5 shows the fit of the full multiple thresholds model as shown in *Section 5.1.2*, and of an alternative in which we pre-define $\lambda = 0$ (i.e. set $g() = log()$). For the full Box–Cox version of the model, we also explored two alternative structures for the between-studies correlation matrix, S: independence, (i.e. S is a diagonal matrix) and full, unstructured, as described by Jones *et al* (8). We see that the Box–Cox version

has substantially better model fit than the “log” version for this data set. The Box–Cox transformation parameter is estimated as $\lambda = -0.17$ (95% CrI −0.19, −0.15), suggesting that the continuous distributions of OC-Sensor test results are slightly more right skewed than Log-Logistic ($\lambda = 0$). There is no meaningful improvement in fit through allowing for between study correlations in random effects in this example, however.

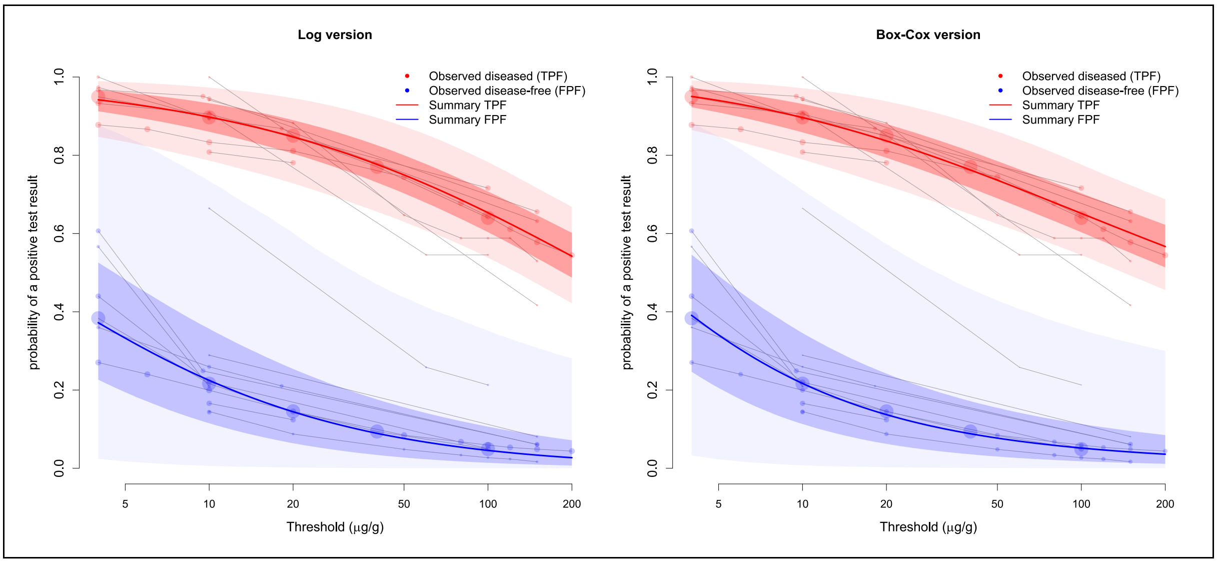
Table 5: Comparison of model fit for the OC-Sensor data set. DIC = residual deviance + pD, where pD is the effective number of parameters.

Transformation of threshold	Covariance matrix structure	Residual deviance	pD[8]	DIC
Log	Restricted	999.6	28.6	1028.2
Box–Cox	Restricted	658.7	29.1	687.7
	Independence	658.0	29.9	687.9
	Unrestricted	657.2	29.8	687.0

Notably, despite the large difference in model fit for the Box–Cox versus “log” model, summary estimates of sensitivity and FPF are seen to be very robust to this choice across the observed range of thresholds: *Figure 7.* For example, at a threshold of 10 μg/g the summary sensitivity is the same for both versions (0.90, 95% CrI 0.86 to 0.93), while summary specificity is also the same but with marginally different CrI (0.88, Box–Cox 95% CrI 0.66, 0.89 compared with “log” CrI 0.65, 0.89). Note that estimates might not agree so closely if the estimate of λ lies further from 0, and also summary estimates between the two models would differ more if we extrapolated beyond the range of the observed data. As with the HSROC curve, we recommend against extrapolating in this way.

[8] pD was calculated outside of WinBUGS based on the deviance at the mean of the fitted values, which is generally more stable than the version reported within WinBUGS: see TSD 2 (31).

Figure 7: Sensitivity and FPF across all thresholds for the OC-Sensor data set: study-level estimates, summary estimates with 95% CrIs (darker shaded regions) and 95% prediction intervals (paler shaded regions). Larger circles depict greater sample size. Estimates shown are from models with independent random effects.



5.2.2. Model critique

Analyses of datasets such as the OC-Sensor example should be accompanied by appropriate model critique, including consideration of global model fit and plots of observed versus fitted curves. The very poor global fit of even the Box–Cox version of the model for this data set (the posterior mean residual deviance 658.0 from 84 data points) would deserve further examination.

An example model critique is illustrated in *Appendix B1*. This establishes that the majority of the poor fit arises from the specificity data in studies with exceptionally large disease-free sample sizes. The difference between observed and fitted specificity in these studies is, however, found to be relatively small. For this reason, the fitted sensitivity and FPF curves can be expected to represent a robust summary of the evidence, despite the poor global model fit.

5.2.3. Estimates at a pre-specified threshold

As discussed above, if the decision model requires estimates of sensitivity and specificity only at a particular (test developer-recommended) threshold, we would recommend fitting the bivariate model (*Section 4.1*) to data relating to that threshold only, as a sensitivity analysis, to check that there are no important discrepancies. This may be particularly important for examples such as this one, where the residual deviance of the multiple thresholds model is large.

More formally, we can use node splitting, originally suggested for detecting conflict between different sources of information in Bayesian hierarchical models by O'Hagan (56), and commonly used to assess inconsistency between direct and indirect evidence in network meta-analysis (57). Here, we re-fit the multiple thresholds model after excluding data relating to the critical threshold, then compare estimates for the critical threshold with those from the bivariate model. Let ξ_{Ind} denote the parameter of interest (e.g. sensitivity) estimated from the "indirect" evidence (i.e. multiple thresholds model excluding data for the critical threshold) and ξ_{Dir} denote the parameter estimated from the direct evidence (i.e. bivariate model). Then a Bayesian p-value can be calculated as $2 * \min\left(\Pr\left(\xi_{Dir} > \xi_{Ind}\right), \Pr\left(\xi_{Dir} < \xi_{Ind}\right)\right)$. A small p-value suggests evidence of inconsistency.

If inconsistency is identified, it will be important to explore reasons for this. It is possible – especially if data are sparse – that the multiple thresholds model may be failing to adequately model the underlying distributions of continuous test results, although the risk of this should be minimised if sufficient data are available to use the full version of the model. Checks on the validity of the model might include plots similar to *Figure 6*, but with the Box–Cox transformed threshold as the x-axis, to identify possible non-linearity. It may be more likely, however, that any such inconsistencies arise due to heterogeneity in the underlying evidence, and we should be careful to understand differences in study design and characteristics that underlie such results. Following careful evaluation, it may be appropriate to meta-analyse only data corresponding to the threshold of interest; however, we do not recommend that this approach is taken without first exploring modelling of all available data.

For the OC-Sensor example, *Table 6* provides a comparison of summary results for two thresholds, including the 10µg/g threshold recommended for use with low risk patients. To check for potential inconsistencies, we compared summary sensitivity and specificity estimates between the multiple thresholds model fitted to data excluding the threshold of interest with estimates from the bivariate model fitted to data for only that threshold. We see that estimates are very similar, and the node split p-values are large, suggesting no cause for concern.

Table 6: Comparison of summary sensitivity and specificity at two thresholds: multiple thresholds model fitted to data excluding the threshold of hypothetical interest and bivariate model fitted to the data at that threshold alone. OC-Sensor example.

Threshold		Multiple thresholds model (Box–Cox version): excluding the relevant threshold data	Bivariate model fitted to relevant threshold data alone	Node split p-value
10µg/g	Sensitivity	0.89 (0.84, 0.93)	0.89 (0.85, 0.93)	0.95
	Specificity	0.75 (0.64, 0.84)	0.77 (0.65, 0.55)	0.77
100µg/g	Sensitivity	0.66 (0.60, 0.71)	0.65 (0.53, 0.74)	0.78
	Specificity	0.95 (0.88, 0.98)	0.94 (0.79, 0.98)	0.83

5.3. FITTING THE MULTIPLE THRESHOLDS MODEL WITH FEW DATA POINTS

As discussed in *Section 4.5* for the bivariate model, synthesis can be challenging when the number of studies is small, but may still be required to inform the cost-effectiveness analysis. For the multiple thresholds model, similar challenges can arise when only a small number of studies report data at more than one threshold, or where the total number of thresholds reported across in the data set is small. In any of these situations, the full model with the vague priors outlined in *Section 4* is unlikely to be appropriate.

As in the bivariate model case, a number of simplifications to the full model might be considered. Possible simplifications, which might be explored based on knowledge of the test and data set and use of standard model fit tools, might include (but are not limited to):

- The parameter λ might be set to a particular value, based on prior understanding of the distribution of continuous test results in the diseased and disease-free populations, or its range could be narrowed. This might be based on individual participant data from one study, histograms presented in a study publication, analysis of similar datasets, and/or exploration via plots like *Figure 6*.

- The structure of covariance matrix S, representing the relationships between the four sets of random effects, could be simplified to an independence structure (as in *Table 5*).
- We note that setting $\sigma_{i2} = c \times \sigma_{i1}$, for some constant c would correspond to an assumption that the degree of asymmetry of the ROC curve is constant across studies (producing parallel lines in logit space), which may often be a reasonable approximation.
- A common slopes model (i.e. $\tau_{\sigma j} = 0,\ j = 1,2$) might also be considered, particularly if only one or two studies report data at more than one threshold value (allowing within-study slope parameters to be estimated).

As discussed in *Section 4.5*, informative or weakly informative prior distributions might also be formulated based on external data or expert opinion.

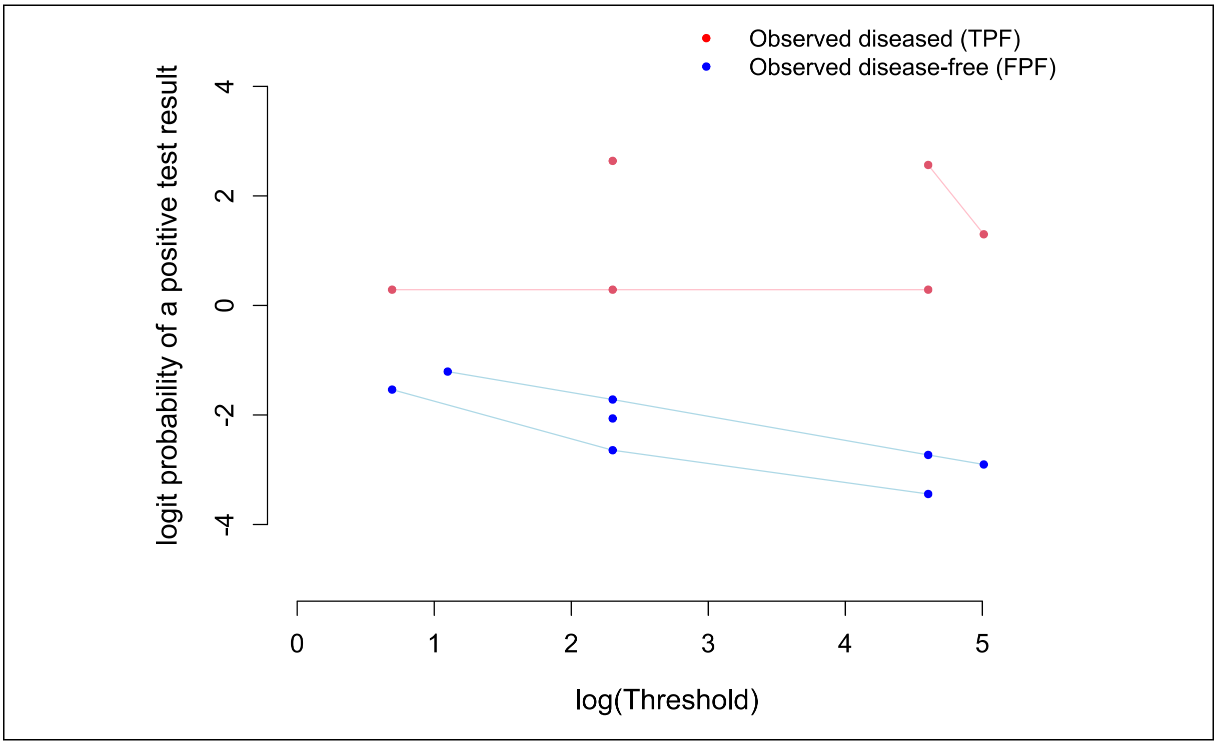
5.3.1. Example 4: Accuracy of FOB Gold in detecting colorectal cancer

To demonstrate the use of the multiple thresholds model with few data points, we consider an additional data set from DG56 (55). The appraisal considered eight different FIT tests including OC-Sensor (described in *Section 5.2*) and FOB Gold.

However, only three studies reported data on the accuracy of FOB Gold. These three studies reported data at between one and four thresholds, with the full dataset providing a total of 8 pairs of sensitivity and specificity, relating to thresholds between 2 and 150µg Hb/g. In total across the three studies, there were 51 and 4,084 individuals with and without colorectal cancer, respectively – therefore there are very limited data available from which to estimate sensitivity.

Due to the sparsity of the data, we pre-specified the transformation $g() = \log()$ prior to synthesis: i.e. $\lambda = 0$. Based on the limited data available, there were no obvious deviations from linearity on this scale (*Figure 8*). Further, OC-Sensor and FOB Gold are similar tests: both are quantitative FIT tests using immunoturbidimetric methods to measure haemoglobin concentrations. As estimates from the OC-Sensor analysis were robust to setting $\lambda = 0$, this approximation seemed reasonable. To further slightly reduce the number of parameters to be estimated, an independence structure was assumed for the covariance matrix S of the four sets of random effects.

Figure 8: Checking the linearity of the FOB Gold data given the log transformation[9]



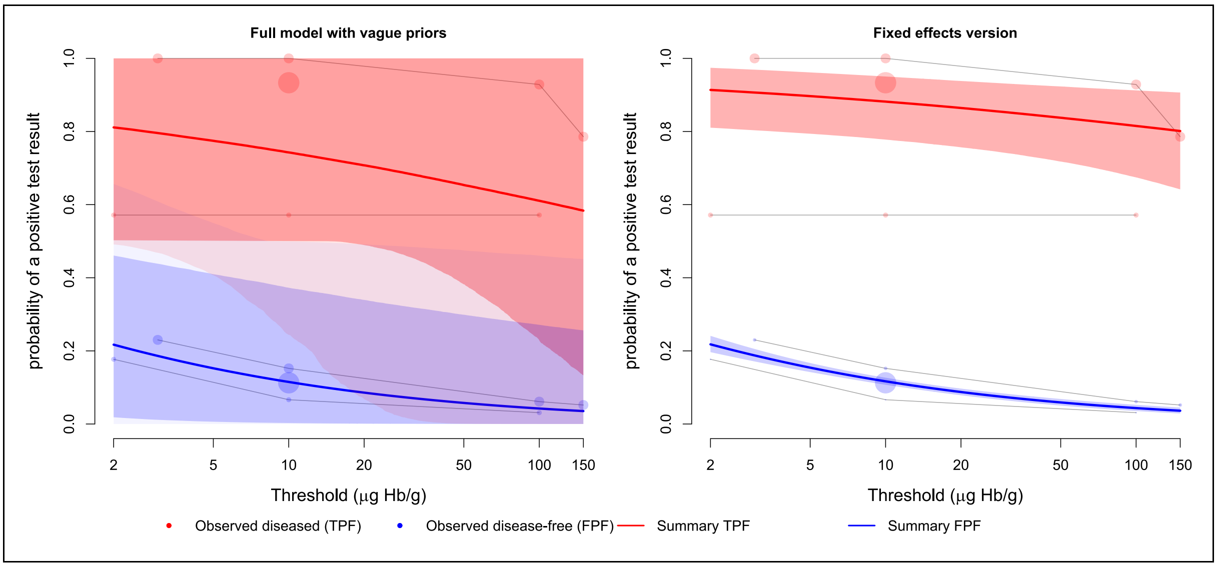
The multiple thresholds model with vague prior distributions as specified in *Section 5.1.3* for all other hyperparameters produced summary estimates of sensitivity and FPF, but with very wide CrIs, as shown in *Figure 9* (left panel). Inspecting the posterior density plots for the four between-study standard deviation parameters ($\tau_{\mu j}, \tau_{\sigma j}, j = 1{,}2$: *Figure 10*, red lines), we observe that these have not been sufficiently updated from the prior distributions, particularly the two parameters relating to the diseased population. There is not enough information to estimate the between-study heterogeneity from the data alone.

The right panel of *Figure 9* shows results from fitting a fixed effect model, i.e. setting $\tau_{\mu j} = \tau_{\sigma j} = 0,\ j = 1{,}2$. In contrast to the left panel, the 95% CrIs around estimated sensitivity and FPF here are unrealistically narrow, particularly for FPF. Use of these

[9] Two datapoints are missing from this plot because observed TPF=1, such that logit(TPF) is undefined.

results would suggest greater certainty in the results for FOB Gold than for OC Sensor (shown in *Section 5.2*), for which more data were available.

Figure 9: Sensitivity and FPF across all thresholds for the FOB Gold data set: study-level estimates, summary estimates with 95% CrIs (darker shaded regions) and 95% prediction intervals (paler shaded regions). Larger circles depict greater sample size. Left panel: with U(0,5) priors for standard deviations; right panel: fixed effect model



As an alternative between these two extremes, we used weakly informative Half-Normal(0,1) prior distributions for the four between-study standard deviation parameters. From comparison with posterior distributions for these parameters from the OC Sensor analysis, these appeared sufficiently wide. Results from an analysis using these prior distributions are shown in *Figure 11.* A comparison of summary estimates from the three analyses at two example thresholds is also shown in *Table 7.*

Figure 10: Posterior densities for standard deviations of random effects, for the multiple thresholds model fitted to the FOB Gold dataset: comparison of results with vague U(0,5) priors and weakly informative HN(0, 1) priors. HN(0,1) prior densities are shown as grey shaded regions, for comparison.

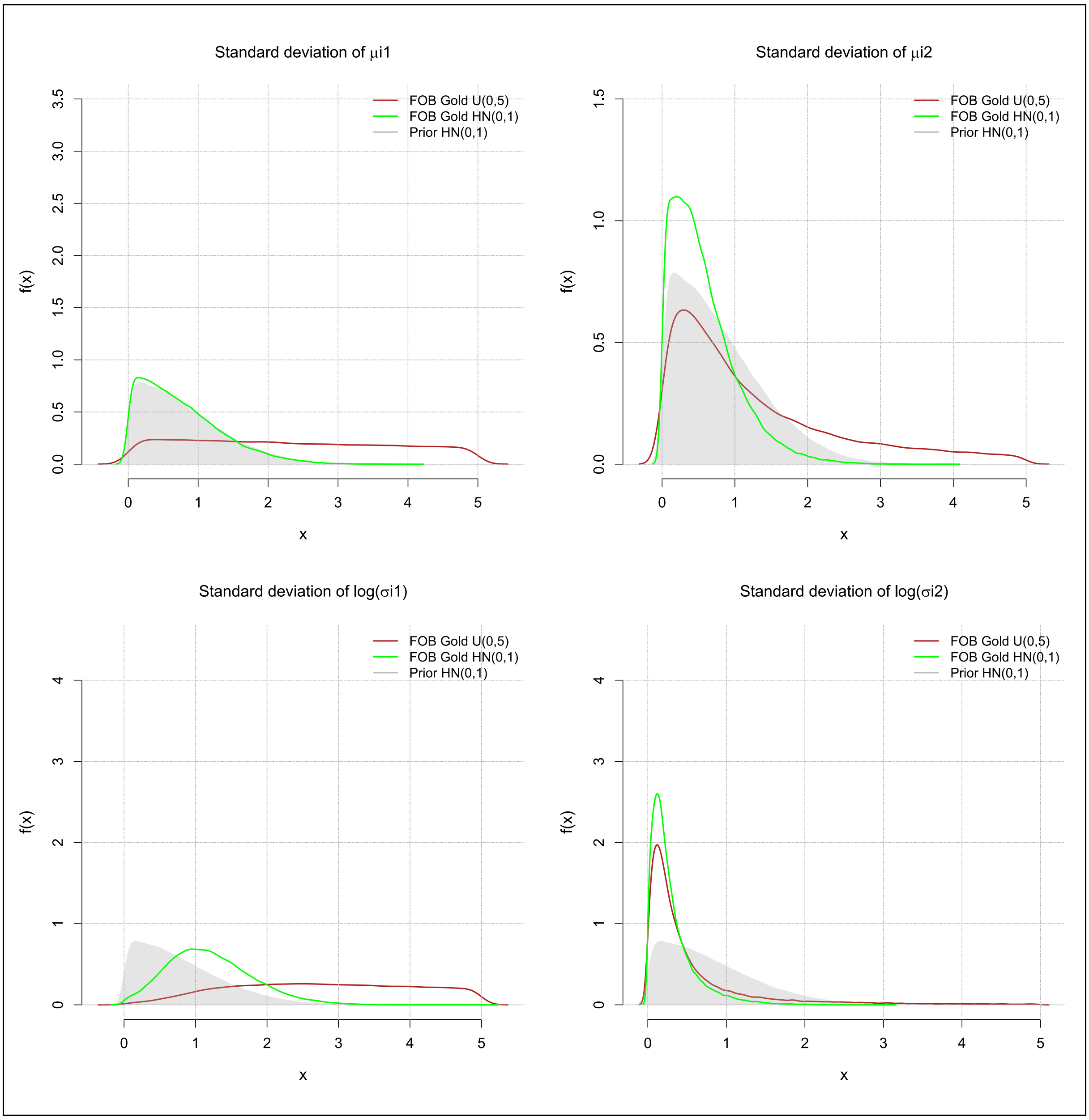
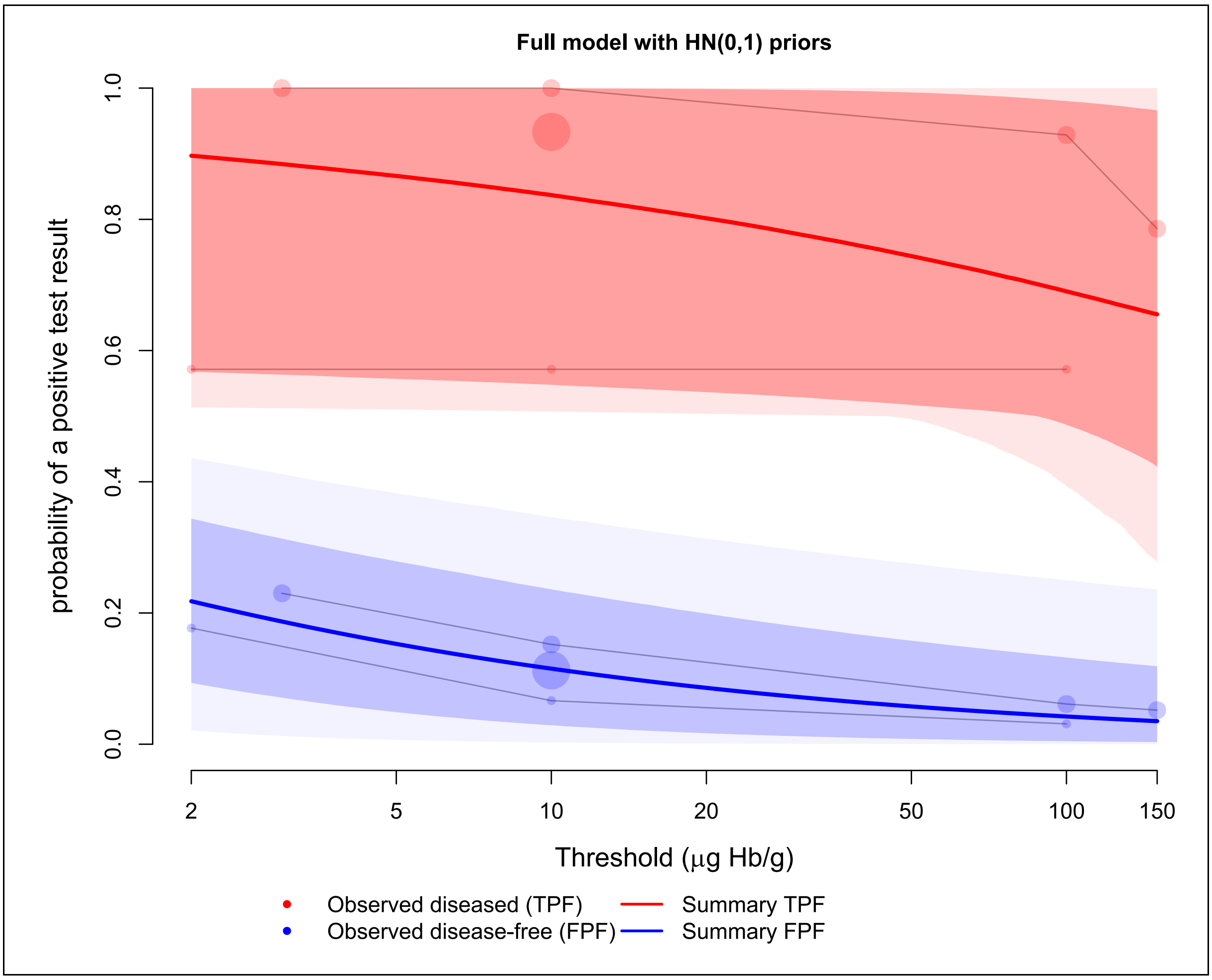


Figure 11: Sensitivity and FPF across all thresholds for the FOB Gold data set, with HN(0,1) prior distributions for between-study standard deviations: study-level estimates, summary estimates with 95% CrIs (darker shaded regions) and 95% prediction intervals (paler shaded regions). Larger circles depict greater sample size.



There are some differences in point estimates across the three analyses (e.g. at a threshold of 10µg/g, estimates of sensitivity range from 0.74 to 0.88). The widths of the CrIs from the model with weakly informative priors for standard deviations lie between those from the other two models. The residual deviance for the fixed effect model is much larger than the number of data points (=16), suggesting poor model fit. In contrast, the residual deviance for the analysis with vague priors is less than 16, which may indicate overfitting. As we would anticipate, the residual deviance for the analysis with weakly informative priors lies between the two. We would select this analysis on the basis that it provides more precise estimates by ruling out unrealistic posterior values, while still being consistent with the evidence.

Table 7: Summary estimates of sensitivity and specificity (with 95% CrIs) at thresholds of 10µg/g and 100µg/g, comparing different options for the between study heterogeneity parameters. FOB Gold dataset.

Threshold		Random effects model with vague priors	Fixed-effect model	Random effects model with HN(0,1) priors for between-study standard deviations
10µg/g	Sensitivity	0.74 (0.50, 1.00)	0.88 (0.78, 0.95)	0.84 (0.55, 1.00)
	Specificity	0.89 (0.63, 1.00)	0.88 (0.87, 0.89)	0.88 (0.76, 0.97)
100µg/g	Sensitivity	0.61 (0.23, 1.00)	0.82 (0.67, 0.91)	0.69 (0.49, 0.98)
	Specificity	0.96 (0.73, 1.00)	0.96 (0.95, 0.96)	0.96 (0.86, 1.00)
Model fit	Residual deviance	13.3	35.0	15.2
	pD	3.5	2.8	6.7
	DIC	16.8	37.8	21.9
Priors	$\tau_{\mu 1}$	Uniform(0,5)	Set to zero	HN(0,1)
	$\tau_{\mu 2}$	Uniform(0,5)	Set to zero	HN(0,1)
	$\tau_{\sigma 1}$	Uniform(0,5)	Set to zero	HN(0,1)
	$\tau_{\sigma 2}$	Uniform(0,5)	Set to zero	HN(0,1)

Sensitivity analyses in which we used HN(0,0.5^2) and HN(0,2^2) priors produced similar summary estimates but with posterior mean residual deviance slightly further from the number of data points (see *Appendix B2*), suggesting the HN(0,1) choice is reasonable.

5.4. OTHER MODELS AND SOFTWARE

We have focused in this section on the multiple thresholds model proposed by Jones *et al.* (8), but note that several alternative models have also been proposed for data of this structure. Zapf *et al.* describe four approaches and compare estimates in a case study (58). The model proposed by Steinhauser *et al.* (53) is similar to that described in *Section 5.1* and can be implemented using the R package `diagmeta` (59). Key differences relative to the Jones model include use of normal approximations to the likelihoods and a requirement to pre-specify the transformation $g()$. The parametric form also differs, such that the random effects have different interpretations. In some circumstances involving little within-study information about the relationship with threshold (i.e. very few studies reporting at multiple thresholds) and the observed relationship across studies being counterintuitive, the Steinhauser model can estimate this relationship in the "wrong" direction: e.g. estimating sensitivity to increase with threshold for a test in which we know higher values represent increased probability of disease. Since the model of Jones *et al.* is the only one using simulation to estimate parameter distributions, this is the only one that can directly supply (i.e. without requiring bootstrapping) correlated samples from the estimated test accuracy parameter distributions, avoiding the need for normality assumptions to be made.

6. USING THE SYNTHESIS OUTPUTS IN A DECISION MODEL

In this section, we describe how results from a DTA meta-analysis can be used in a decision model, and provide a simplified example to illustrate some key considerations that are likely to arise in practice.

6.1. INTEGRATING DIAGNOSTIC SYNTHESES IN A DECISION MODEL

Probabilistic decision modelling (60) is recommended for all NICE programmes (1). This has the effect of propagating the uncertainty in our estimates of sensitivity and specificity appropriately throughout the decision model, integrating this with our uncertainty relating to long-term consequences of each of the four outcomes (TP, FN, TN, FP). As discussed earlier, posterior estimates of sensitivity and specificity from the evidence synthesis models described in *Section 4* and 5 are likely to be correlated. A deterministic approach, in which posterior estimates of sensitivity and FPF are plugged into the model, will fail to propagate this correlation, and will, in any case, produce incorrect results if the decision model in non-linear in sensitivity or FPF.

One advantage of a fully Bayesian approach to synthesis, for example using WinBUGS, is that we can conveniently embed the posterior simulation outputs of these quantities in the decision model, thereby preserving underlying parameter uncertainty, correlation, and uncertainty in correlation (61, 62).

While in principle the decision model can be incorporated into the Bayesian MCMC model code, estimating cost effectiveness jointly alongside the synthesis, this will be unnecessarily computationally expensive for all but the simplest decision models. A practical alternative is to take posterior samples from the Bayesian MCMC ("coda") and export these to software of choice for the decision model. For example, modellers may import the posterior samples to an Excel workbook and use a randomly selected row in each iteration of their probabilistic analysis. In the sections that follow, we use R2WinBUGS functionality by exporting posterior samples into R for the decision model.

We note that an alternative, compact, approach to using posterior samples from the bivariate model would be to approximate the joint posterior distribution of logit-transformed sensitivity and FPF with a bivariate normal distribution: this is the same approximation as typically used to create credible or prediction regions on SROC plots. However, as noted in *Section 4.4*, this approximation can be poor, especially when estimated sensitivity or specificity is very close to one; therefore, we do not recommend this approach unless the analyst is confident that joint normality holds.

In *Sections 4.6* and *5.4*, where we discussed software choices for the synthesis model, we briefly outlined how to embed evidence synthesis results in probabilistic decision models, including options for using results from frequentist syntheses. For a more comprehensive discussion of this topic see TSD 6 (63).

6.2. CHOOSING THE EVIDENCE SYNTHESIS OUTPUT FOR USE IN DECISION MODELLING

In *Section 6.1* we described use of a joint posterior distribution for sensitivity and FPF in a decision model. It is important to recognise, however, that – as with any random effects model – the synthesis models described in *Sections 4* and *5* produce multiple outputs, each of which could be used in the decision model. The "summary" estimates (i.e. sensitivity and FPF at the means of the random effects distribution) and predictive distributions are two of the options available, but there are others.

To select the appropriate output, it is necessary to consider the relationship between the decision-maker's target population and use scenario and the populations and use scenarios sampled in the synthesised studies. This problem is discussed in depth elsewhere: for random effects meta-analysis of relative treatment effects (27, 64-67) and also for test accuracy synthesis (61, 68). Here we list the main options.

Random effect means. This is the meta-analysis output most commonly used in decision models. However, its appropriateness is contingent on an assumption that the sensitivity and specificity of the test in the decision scenario is located at the means of the random effects distributions. It is difficult to envisage a scenario where this is truly realistic in practice.

Predictive distribution. The joint predictive distribution for sensitivity and specificity in a "new" study is a natural alternative, allowing for the heterogeneity observed in the meta-analysis. This relies only on an assumption that the accuracy of the test in the target scenario is exchangeable with those in the meta-analysis. However, as we have seen in the worked examples, prediction intervals can be extremely wide in test accuracy syntheses. If we are able to understand at least some of the reasons for heterogeneity, these intervals likely over-state the true variation in sensitivity and specificity that might be expected in practice.

Predictions from a synthesis with covariates. Where synthesis has suggested a relationship between one or more study-level covariates and sensitivity and specificity, model-based predictions most closely relating to the decision question can be used. For example, we discussed the potential for a relationship between sensitivity, specificity and prevalence in *Section 4.3*: if this is observed and modelled, then predictive distributions corresponding to the assumed prevalence in the decision population can be used.

Study-specific estimate. If the target population/scenario is well represented by one of the studies in the synthesis, then the joint posterior distribution for sensitivity and FPF for that study would be an appropriate input for the decision model. These model-based or shrinkage estimates should not be confused with the observed values of sensitivity and FPF in the study in question. The assumption here, implicit in the fitting of a random effects model, is that all the studies included in the synthesis are similar, such that the posterior estimates are appropriately drawn in to some degree towards the random effects means, borrowing strength from the other study estimates[10].

Random effects distribution. If the between-studies variation observed in the included studies is believed to represent the between-centre variation that would be observed if the diagnostic test was “rolled out” for routine use, then it would be appropriate to integrate the decision function over the entire random effect distribution. This approach results in a *per capita* net benefit (66). However, this is based on an assumption that all heterogeneity in the meta-analysis will remain inevitable in routine practice. In the context of NICE decision-making, it may be hoped that effective national guidance may attenuate some sources of variation.

In the following subsections we will demonstrate use of random effects means and predictive distributions. However, in any real example the analyst should carefully consider the potential sources of heterogeneity and choose the output considered most appropriate for the specific decision problem.

6.3. A SIMPLIFIED DECISION MODEL

To demonstrate use of synthesis results in a decision model, we introduce a simplified hypothetical decision model to evaluate the cost effectiveness of introducing a screening test for a chronic condition. This simple model consists of a decision tree only, as shown in *Figure 12*. For ease of exposition, let us assume we already know – with certainty – the payoffs associated with each possible outcome from the 2×2 table, relative to the baseline outcome of no disease and no screening test. We emphasise that this is a simplified hypothetical example: modelling of the downstream consequences of each outcome will usually be required (see Discussion). We express these as hypothetical net benefits (NB), representing the joint impact of costs and effects, when the two are put on a common scale assuming some monetary value for

[10] If, on the other hand, it was felt that the relevant study was completely different from the other studies, a synthesis should not have been conducted; under this circumstance the analyst can discard the irrelevant data and directly use the observed data from the study instead.

each unit of effect (for example, a cost-effectiveness norm such as £20,000 per QALY gained). These are as follows:

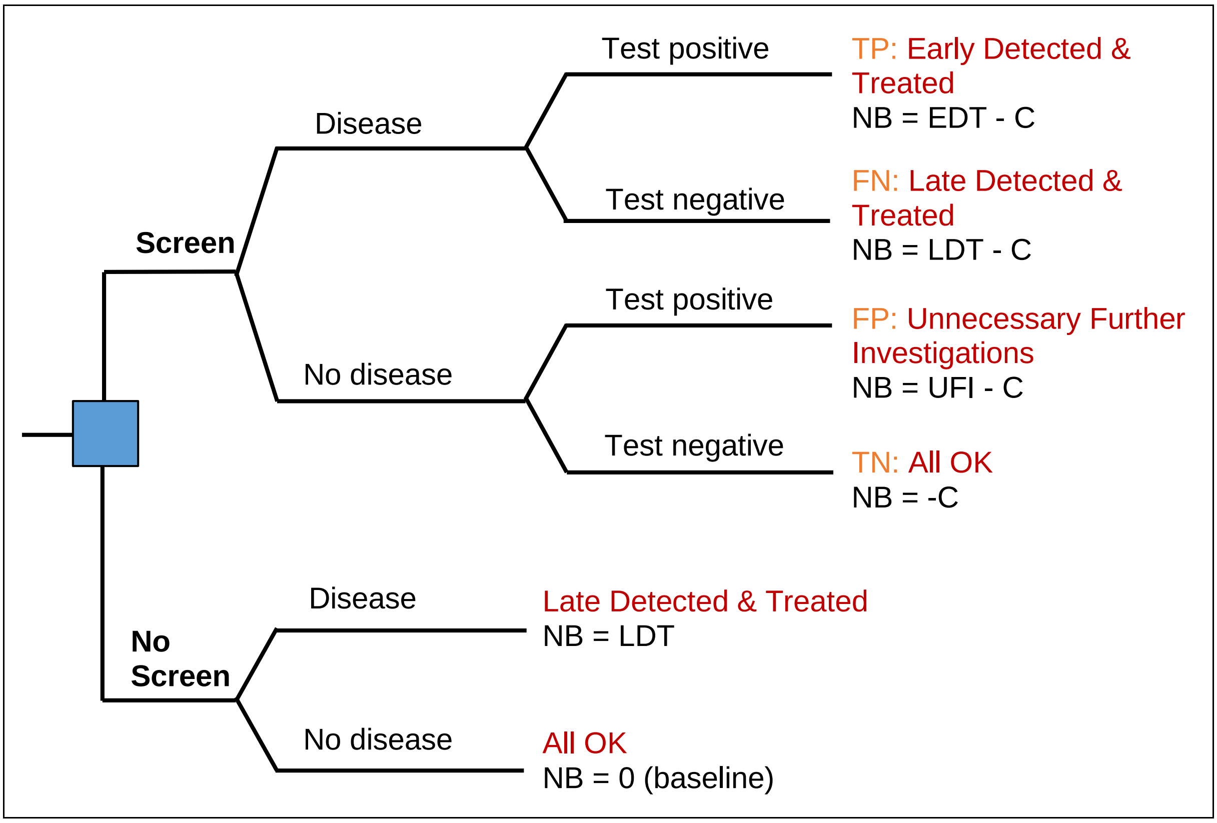
- True positive (TP) test result: early detected and treated. $NB = EDT - C$
- False negative (FN): late detected and treated. $NB = LDT - C$
- False positive (FP): unnecessary further investigations. $NB = UFI - C$
- True negative (TN). $NB = -C$

where C is the cost of the screening test.

The corresponding outcomes without screening are assumed to be:

- Diseased: late detected and treated. $NB = LDT$
- No disease (baseline). $NB = 0$

Figure 12: Simplified hypothetical decision tree evaluating introduction of a screening test



The overall net benefit of screening is then equal to:

$$NB_{screen} = \pi[Sensitivity \times EDT + (1 - Sensitivity)LDT]$$
$$+(1 - \pi)[(1 - Specificity)UFI \ + (Specificity \times 0)] - \ C$$

where π is the disease prevalence, while the net benefit of not screening is calculated as:

$$\begin{aligned} NB_{no\_screen} &= \pi \times LDT + (1 - \pi) \times 0 \\ &= \pi \times LDT \end{aligned}$$

such that the incremental net benefit (INB) of screening versus not screening is:

$$\begin{aligned} INB \ &= \ NB_{screen} \ - \ NB_{no\_screen} \\ &= \pi(EDT - LDT)Sensitivity + (1 - \pi)UFI(1 - Specificity) - \ C \end{aligned} \qquad \mathbf{(4)}$$

For the examples to follow we assume hypothetical known values of EDT = −100, LDT = −200, UFI = −50 and C = 10.

Note that, in this example, the “no screening” approach is equivalent to a screening strategy with sensitivity of 0.00 and specificity of 1.00 (that is, no TPs and no FPs) and no up-front costs. This equivalence can simplify coding, rather than having separate pathways for each simulated approach. In similar decision problems, it can also be helpful to include an arm representing “refer everyone” with sensitivity of 1.00 and specificity of 0.00 (i.e. no FNs and no TNs), in order to explore whether there are any circumstances under which the optimal approach would be to assume everyone at risk has the target condition (69) .

6.4. THE ROLE OF PREVALENCE

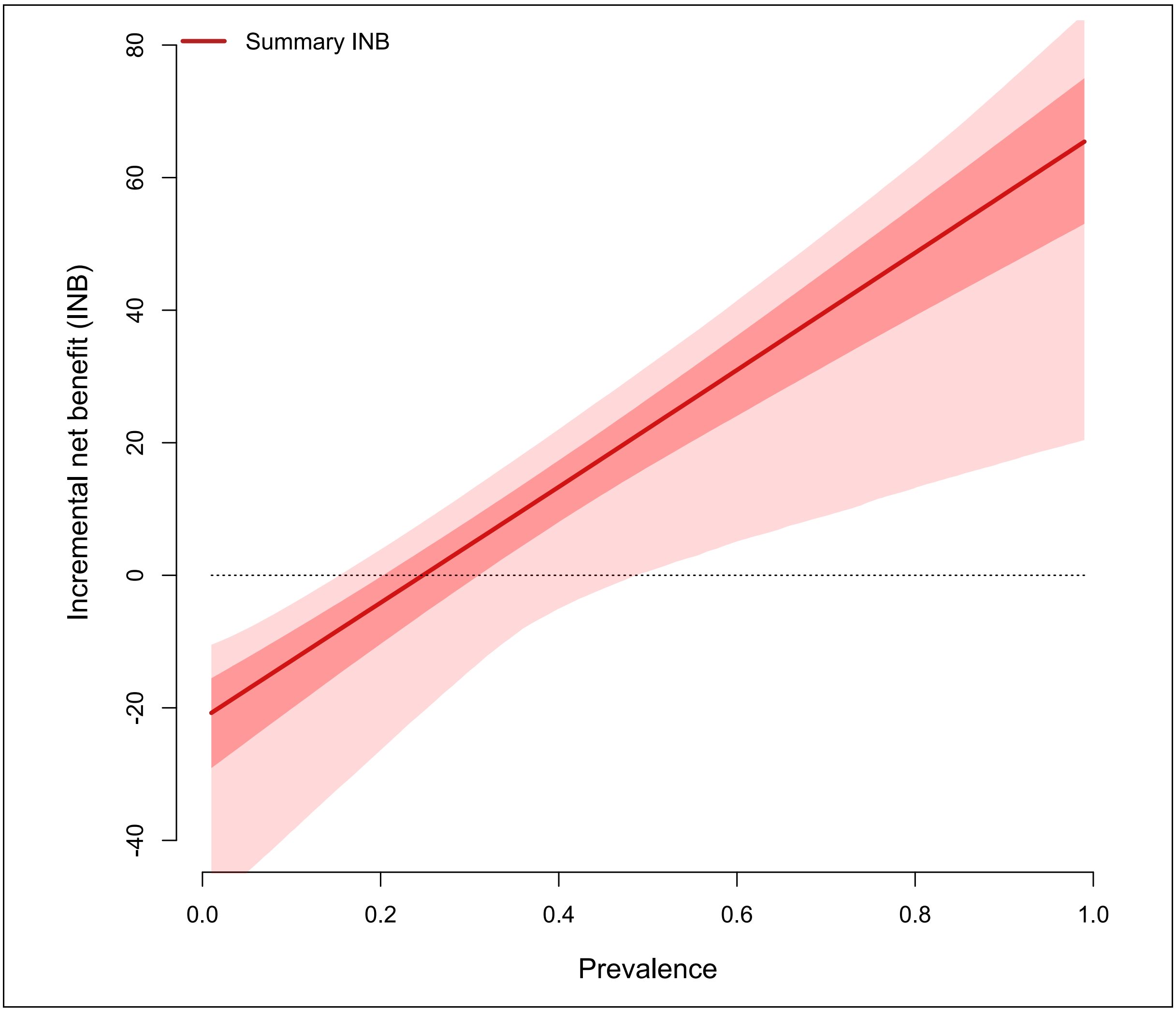
As is clear from equation (4), one of several parameters that the INB depends on is the prevalence of the disease in the population in which the test is being considered for use. To demonstrate the critical role of prevalence, we first consider the situation of a truly dichotomous screening test, or a test for which all accuracy data correspond to the same threshold.

We fitted the bivariate model to (artificial) data from 10 studies. Summary sensitivity and specificity estimates were 0.76 (95% CrI 0.64 to 0.86) and 0.77 (95% CrI 0.60 to

0.87) respectively, with posterior correlation between these estimates on the logit scale of -0.60. After accounting for the estimated between-study heterogeneity in sensitivity and specificity, 95% prediction intervals for a new study were 0.31 to 0.96 (sensitivity) and 0.18 to 0.98 (specificity) respectively.

We exported correlated posterior samples of sensitivity and specificity, and of the predictions for sensitivity and specificity in a new study, to R using R2WinBUGS. We evaluated the INB (equation (4)) at each possible value of prevalence between 0% and 100%, using each of these two sets of synthesised results. *Figure 13* shows the INB for each value of prevalence. Shaded areas represent 95% intervals when the "summary" estimates of sensitivity and specificity were used (darker shading) and when the predictive distributions were used (lighter shading).

Figure 13: Relationship between incremental net benefit (INB) and prevalence: results from a simplified hypothetical decision model



We see that the cost-effectiveness of the hypothetical screening depends heavily on prevalence, with increased net benefit of screening in a population with higher prevalence.

We can also obtain an estimate of the prevalence above which screening becomes cost-effective (π_{ce}), by solving equation (4) with respect to prevalence and setting INB=0, i.e.

$$\pi_{ce} = \frac{C - \text{UFI}(1 - \text{Specificity})}{(\text{EDT} - \text{LDT})\text{Sensitivity} - \text{UFI}(1 - \text{Specificity})}$$

By evaluating π_{ce} at each posterior sample, we obtain an estimate with uncertainty: screening is estimated to be cost effective if the prevalence is above 0.25 (95% CrI 0.20 to 0.31).

6.5. DETERMINING THE OPTIMAL THRESHOLD

6.5.1. In the absence of a numerical threshold

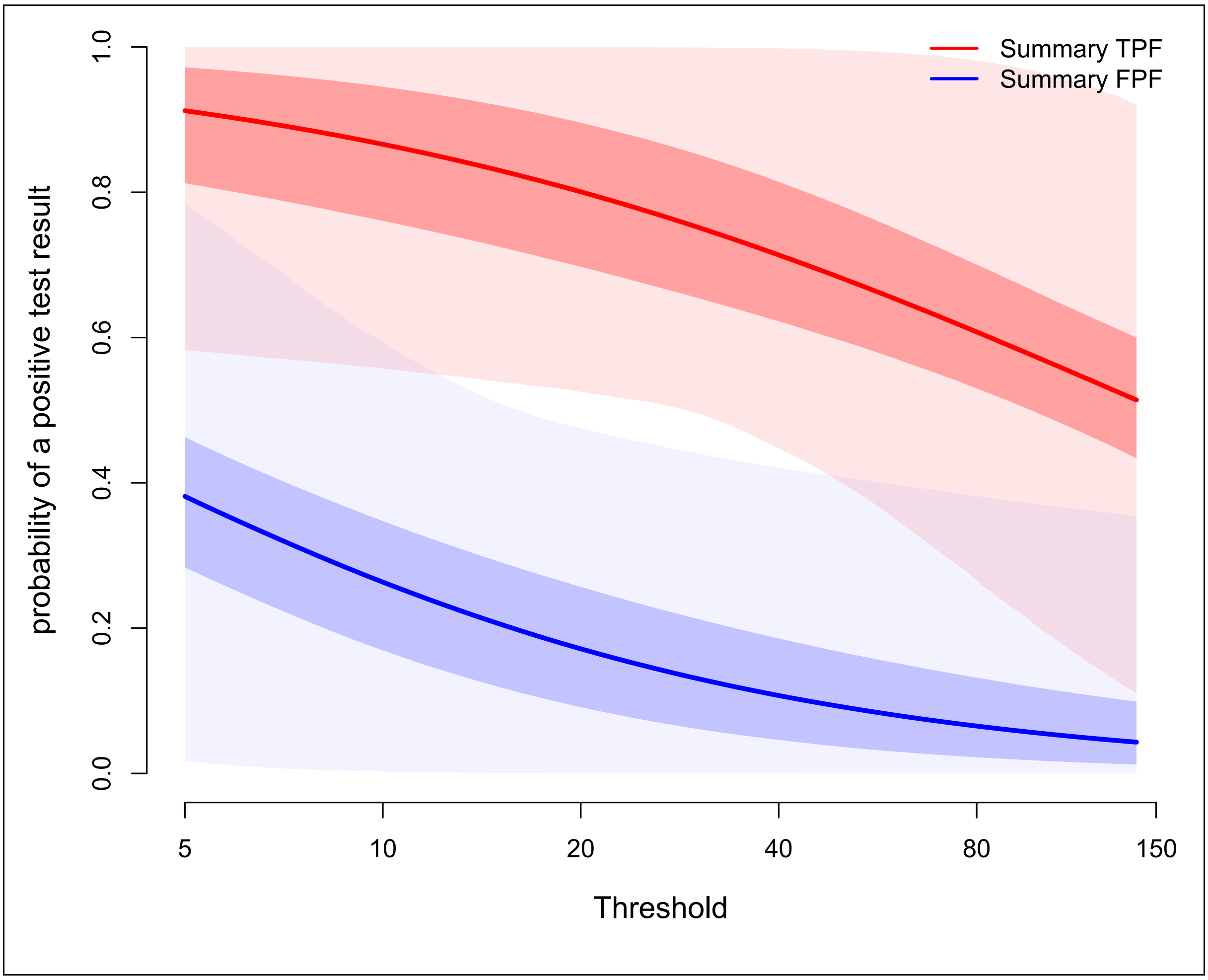
As discussed in *Sections 4.2* and *4.4.2*, when implicit thresholds are known to be present, an HSROC curve is typically considered to be a more appropriate summary of the meta-analysis results. The decision model can be evaluated at all points on the HSROC curve, producing (for example) an INB for each point. Sutton *et al.* (61) demonstrate how the point on the HSROC curve that maximises the INB, at a given prevalence, can be identified using this approach. The difficulty, however, is that – since there is no numerical threshold value to map that point on the HSROC curve to – it is not possible to choose to operationalise the test at that threshold in practice (61).

6.5.2. Determining the optimal numerical threshold

Where thresholds are explicit numerical values, and choice of threshold is to be explored in the decision model, full results from the multiple thresholds model (*Section 5*) can be used. The INB, or any other criteria of interest, may vary substantially across thresholds. In this section, we demonstrate how to find the optimal threshold at a given prevalence based on the INB, and the role of prevalence in determining this value.

For demonstration, we fitted the multiple thresholds model to artificial data from 20 studies, with data for up to 10 thresholds per study. In this demonstration we set *λ = 0*, but the approach can be easily extended to also accommodate uncertainty in *λ. Figure 14* shows summary estimates of sensitivity and FPF across thresholds, with 95% CrIs and 95% prediction intervals.

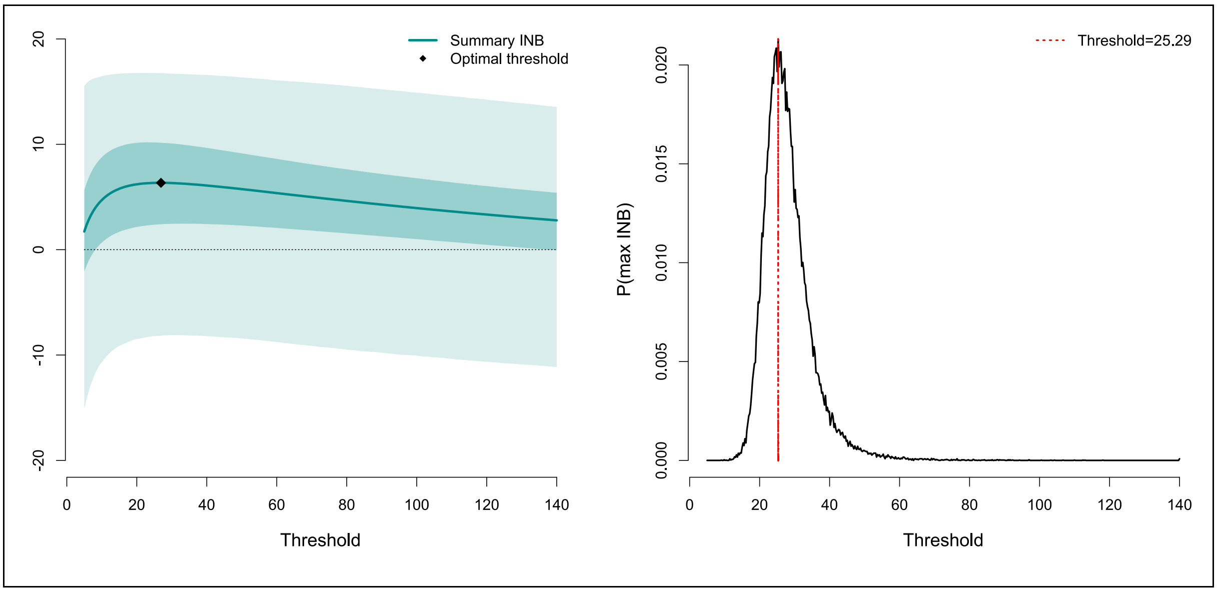
Figure 14: Summary sensitivity and FPF estimates across thresholds: results from analysis of artificial data set



In a similar manner as described in *Section 6.4*, we can process use posterior samples of the parameters from the multiple thresholds model in R, to obtain estimates of any quantity of interest with uncertainty. The parameters required are means and scale parameters of the assumed logistic distributions in the diseased and disease-free populations, and *λ* if the full Box–Cox version is used. For example, if the “summary” estimates from the model are to be used, then we require only the posterior samples of the parameters $m_{\mu 1}$, $m_{\mu 2}$, $m_{\sigma 1}$, $m_{\sigma 2}$ and λ.[11] Test sensitivity and specificity at all thresholds are calculated from these, with uncertainty, within the R code.

For a simplified demonstration, we used the same decision model as above (*Figure 12*) and assumed a known disease prevalence of 28%[12]. *Figure 15* (left panel) shows the estimated INB for each threshold, with shaded areas representing 95% intervals when the "summary" estimates of sensitivity and specificity were used (darker shading) and when the predictive distributions were used (lighter shading). For this example, the threshold with the highest point estimate of INB is 27. This is represented by a black diamond in *Figure 15* (left panel).

Figure 15: Finding the optimal threshold based on INB for a specific prevalence: artificial example



We can see, however, that other threshold values produce a similar INB, with considerable overlap in 95% CrIs across a range of thresholds. To allow for this, we can produce a 95% CrI around this estimate of the threshold that maximises the INB.

[11] If, instead, predictive distributions were to be used in the decision model (for example), then we would require posterior samples for $pred\_\mu_1$, $pred\_\mu_2$, $pred\_log(\sigma_1)$ and $pred\_log(\sigma_2)$.

[12] For simplicity, we assume a known fixed value for prevalence in this worked example. However, this can be replaced with a distribution to allow for uncertainty in prevalence among the decision population.

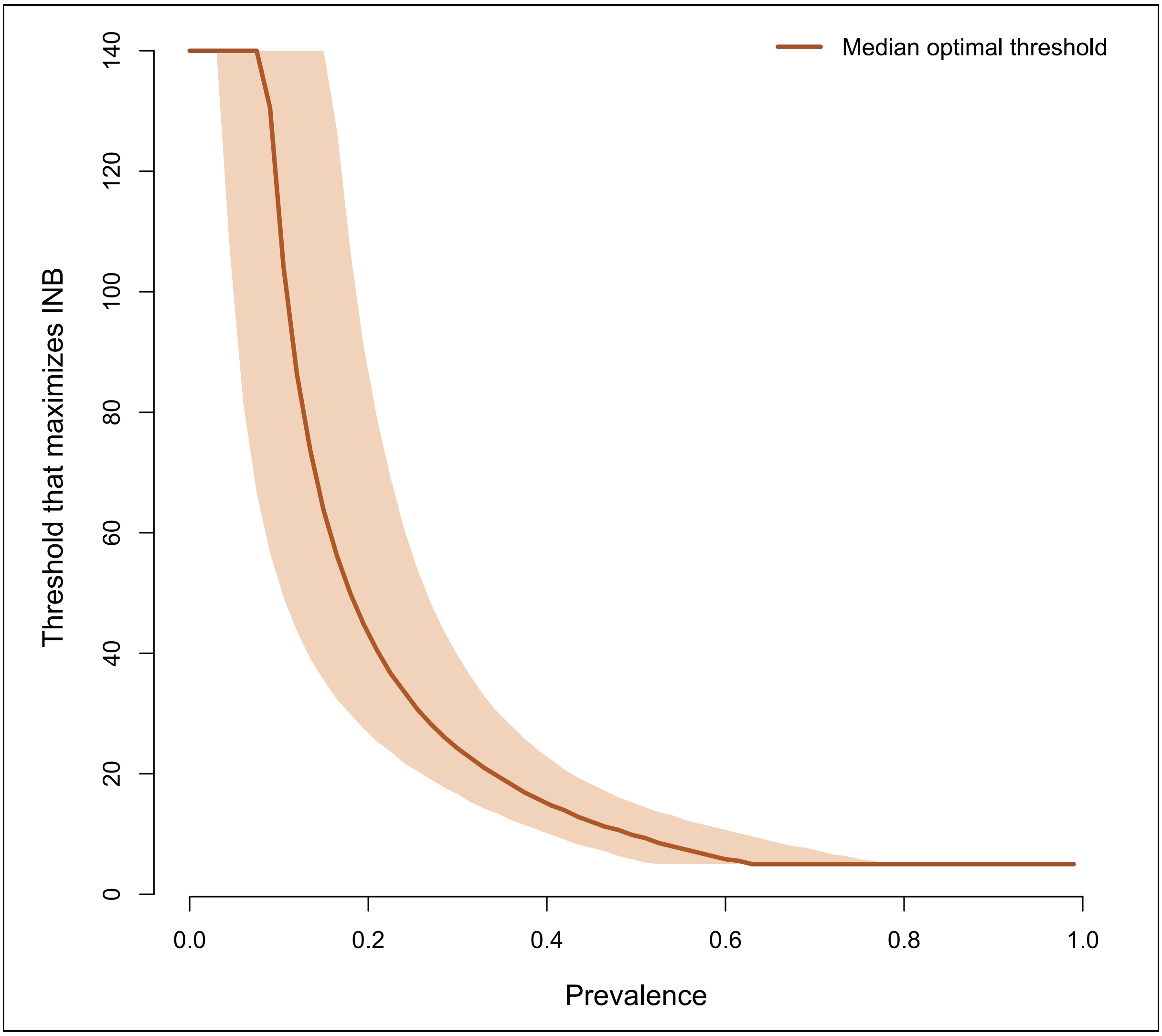
To obtain this interval, we find the threshold that maximises the INB at each iteration of the posterior samples, and then summarise across iterations using the 2.5th and 97.5th quantiles. For this example, we estimate the “optimal” threshold is 27, with 95% CrI 18 to 45. It is straightforward also to calculate the probability that each threshold produces the maximum INB. This is shown in the right panel of *Figure 15*. The threshold with the highest probability of maximising the INB is 25 in this example.

Importantly, regardless of the criterion chosen for optimising the threshold[13], the estimated optimal threshold will depend on the disease prevalence. The “optimal” threshold identified above is only valid for the assumed disease prevalence of 28%. In *Figure 16* we show how the threshold that maximises the expected net benefit varies with prevalence, with the shaded area showing the corresponding 95% CrI around optimal threshold. As prevalence increases, it is more cost-effective to use a lower threshold. This is clear from equation (4): with higher prevalence, the more weight in the INB is given to the (positive) contribution of sensitivity relative to the (negative) contribution of the FPF. For example, at a higher prevalence of 50%, the optimal threshold reduces to 10 (95% CrI 6 to 15) whereas, at a lower prevalence of 15%, the optimal threshold increases to 64 (95% CrI 36 to 140).

Note that we performed these calculations only across the range of thresholds observed in the data (in this case: 5–140). The flat areas of *Figure 16* at the extremes are a consequence of this restriction.

[13] We have focused throughout this section on maximising the INB but – even without a cost-effectiveness model – the same approach could be used to identify the threshold that maximises some other measure, such as the Youden index or proportion of people correctly diagnosed.

Figure 16: Estimate of the threshold that maximises the INB, with 95% CrI, for each prevalence



7. DISCUSSION

This is the first TSD on evidence synthesis of diagnostic test accuracy. We have focused on methods to synthesise data on the accuracy of a single diagnostic test that has been evaluated against a "gold standard" reference test in all studies being pooled across. We also demonstrated how synthesised estimates from these models can be used within a decision-making framework.

We note that there exist other sources of guidance with overlapping content to this TSD (70, 71). The methodological approaches we follow in this document are broadly in agreement with these. Here however, we have focused on performing all analyses in a Bayesian framework. One advantage of this is the ability to obtain meaningful estimates of the accuracy of tests reported on in very few studies, by using weakly informative or informative prior distributions for hyperparameters, and/or by careful reduction of the number of parameters to be estimated, guided by standard tools for model critique and comparison. Another key advantage of the Bayesian approach is the ease of full propagation of parameter uncertainty. This makes it straightforward to calculate both credible and predictive intervals around any function of parameters – whether HSROC curves, PPV and NPV at any assumed prevalence, or outputs of the decision model such as INB. We additionally demonstrated how the “optimal” diagnostic threshold based on some criterion can be selected, again with uncertainty.

This TSD sets the base for meta-analysis of diagnostic test accuracy data, and does not cover many issues that have arisen in applied evaluations of diagnostic tests. In particular, this document is not intended as a guide on how to evaluate the cost effectiveness of one or more diagnostic tests. We have illustrated the use of test accuracy synthesis in decision making using a simplified hypothetical model, in which all the outcomes downstream of the diagnostic test have been rolled up into net benefit contributions, which we further assumed were known with certainty. In practice these values are of course uncertain, and will typically need to be estimated by modelling the clinical pathways downstream of the test; the main focus may in fact be on modelling the treatment options. Simulating the consequences of FN findings additionally requires modelling of the natural history of undiagnosed disease. Soares *et al* provide a comprehensive overview of challenges encountered in assessing the value of diagnostic tests (69).

Besides this necessary simplification of the decision context, important issues in synthesis of DTA not covered in this document include:

1) Exploring between-study heterogeneity

Although we placed special focus on modelling how test accuracy varies with threshold, we did not cover incorporation of other – study-level – covariates in meta-analysis models. Taking account of study characteristics that might affect test accuracy can both reduce the amount of unexplained heterogeneity and increase precision of estimates. Study-level covariates can be incorporated in extensions of either the bivariate or HSROC model parameterisations (7, 15)[14]. Jones *et al.* also demonstrate how the multiple thresholds model can be extended to include study-level covariates, acting on the mean and/or scale parameters of underlying logistic distributions for (transformed) continuous test results (8). If there is evidence for test accuracy varying by patient characteristics, this may also be important for the decision analysis.

2) *Comparing the accuracy of tests, network meta-analysis of DTA, and estimating the accuracy of test sequences*

We additionally have not described methods for comparing the accuracy of two or more tests, network meta-analysis of test accuracy, or methods to estimate the net accuracy of a test applied sequentially, or of two or more tests in sequence or together (which may be required for many decision models, given how tests are used in practice). This methodology is an evolving field, with the presence of within-study, as well as between-study, correlations being a key challenge. A number of parameterisations to account for between-study correlations have been proposed, some arm-based (72-76) and others contrast-based (77). Most are based on extensions to the bivariate parameterisation, while others extend the HSROC parametrisation (78). A small number of proposed approaches also account for within-study correlations that arise from head-to-head comparisons of tests in a single study population (79, 80), and are also critical to the accuracy of tests applied in sequence (81). These different approaches have not been formally compared and further guidance is needed on the synthesis of such data (82).

3) *Multiple disease states*

Throughout this document we focused on the accuracy of tests for diagnosing a dichotomous target condition where there are either only two true disease states (i.e. an individual either has or does not have the target condition), or only two states are relevant from a clinical or a decision-making point of view. We acknowledge that diagnostic testing often leads to a classification into multiple disease states, for

[14] Note that the equivalence between these two models no longer holds with the inclusion of covariates.

example: disease-free, early stage, advanced stage, etc (83). Such classifications may be based on ordinal data, or on a continuous score with multiple cutpoints. Several methods exist for estimating the accuracy of a diagnostic test for multiple target condition states from a single study (12, 13), but the topic has received less attention in the meta-analysis literature. In some cases it is necessary to synthesise studies that have used different numbers of categories or different cut-points (84).

4) Accounting for lack of, or incomplete use of, a gold standard

Another key methodological challenge for test accuracy synthesis, not covered in this TSD, is how to relax the “gold standard” assumption. Often a true gold standard test either does not exist or has not been applied in most studies: for example, this may be the case for a highly invasive test that comes with associated risks. A number of “latent class” meta-analysis models have been proposed, in which disease prevalence is estimated jointly alongside sensitivity and specificity (85, 86). In some circumstances, it may be the case that the reference test used in studies has been evaluated against a gold standard in another study. In this special instance, estimates of sensitivity and specificity of the index test can be adjusted for bias prior to synthesis (87, 88). However, results depend on assumptions made about disease prevalence, and the extent of correlation between the index and reference test in the diseased and non-diseased populations. A related issue is verification bias, which occurs when only a subset of study participants are tested with the gold standard, conditional on results of one or more index tests. Some methods to adjust for verification bias have been proposed (10, 87). Finally, we note that synthesis of individual participant data on DTA has received relatively little attention to date. Modelling of individual participant data facilitates exploration of how accuracy may vary by individual level characteristics, while also directly informing within-study correlation parameters. As such, individual participant data meta-analysis – well known to have a number of advantages over modelling of aggregate data in general – may prove to be the best way of tackling the methodological challenges described above (89).

REFERENCES

1. NICE. NICE health technology evaluations: the manual. www.nice.org.uk/process/pmg36. 2023.
2. Snowsill T. Modelling the Cost-Effectiveness of Diagnostic Tests. Pharmacoeconomics. 2023;41(4):339-51.
3. NICE. Diagnostic Guidance: [DG59] CYP2C19 genotype testing to guide clopidogrel use after ischaemic stroke or transient ischaemic attack. www.nice.org.uk/guidance/dg59. 2024.
4. NICE. Diagnostic Guidance: [DG33] Biomarker tests to help diagnose preterm labour in women with intact memebranes. https://www.nice.org.uk/guidance/dg33. 2018.
5. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982-90.
6. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. 2006;59(12):1331-2; author reply 2-3.
7. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865-84.
8. Jones HE, Gatsonsis CA, Trikalinos TA, Welton NJ, Ades AE. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. Stat Med. 2019;38(24):4789-803.
9. Ntzoufras I. Bayesian modeling using WinBUGS. Hoboken, N.J.: Wiley; 2009. xxiii, 492 p. p.
10. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford ; New York: Oxford University Press; 2003. xvi, 302 p. p.
11. Sun C, Chen P, Chen Q, Sun L, Kang X, Qin X, et al. Serum paraoxonase 1 heteroplasmon, a fucosylated, and sialylated glycoprotein in distinguishing early hepatocellular carcinoma from liver cirrhosis patients. Acta Biochim Biophys Sin (Shanghai). 2012;44(9):765-73.
12. Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York: Wiley-Interscience; 2002. xv, 437 p. p.
13. Nakas CT, Bantis LE, Gatsonis C. ROC analysis for classification and prediction in practice. First edition. ed. Boca Raton: CRC Press; 2023. pages cm. p.
14. Trikalinos TA, Balion CM, Coleman CI, Griffith L, Santaguida PL, Vandermeer B, et al. Chapter 8: meta-analysis of test performance when there is a "gold standard". J Gen Intern Med. 2012;27 Suppl 1(Suppl 1):S56-66.
15. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.
16. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making. 2008;28(5):621-38.
17. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993;12(14):1293-316.

18. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993;13(4):313-21.

19. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172(1):137-59.

20. Bujkiewicz S, Achana F, Papanikos T, Riley RD, Abrams KR. NICE DSU Technical Support Document 20: Multivariate meta-analysis of summary data for combining treatment effects on correlated outcomes and evaluating surrogate endpoints. 2019; available from http://www.nicedsu.org.uk.

21. Lunn D. The BUGS book : a practical introduction to Bayesian analysis. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2013. xvii, 381 pages p.

22. Chu H, Nie L, Cole SR, Poole C. Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: alternative parameterizations and model selection. Stat Med. 2009;28(18):2384-99.

23. Hoyer A, Kuss O. Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas. Stat Med. 2015;34(11):1912-24.

24. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51(8):1335-41.

25. Tanner WP, Jr., Swets JA. A decision-making theory of visual detection. Psychol Rev. 1954;61(6):401-9.

26. Green DM, Swets JA. Signal detection theory and psychophysics. New York,: Wiley; 1966. xi, 455 p. p.

27. Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias -adjustment. 2011; last updated April 2012; available from http://www.nicedsu.org.uk.

28. Schiller I, Dendukuri N. "DTAplots". 1.0.2.5 ed. CRAN2021. p. Functions to use posterior samples from Bayesian DTA meta-analysis models.

29. Roberts E, Ludman AJ, Dworzynski K, Al-Mohammad A, Cowie MR, McMurray JJ, et al. The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. BMJ. 2015;350:h910.

30. Coulthard MG, Coulthard T. The leaf plot: a novel way of presenting the value of tests. Br J Gen Pract. 2019;69(681):205-6.

31. Walsh T, Macey R, Ricketts D, Carrasco Labra A, Worthington H, Sutton AJ, et al. Enamel Caries Detection and Diagnosis: An Analysis of Systematic Reviews. J Dent Res. 2022;101(3):261-9.

32. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. 2011; last updated September 2016; available from http://www.nicedsu.org.uk.

33. Takwoingi Y, Dendukuri N, Schiller I, Rücker G, Jones H, Partlett C, et al. Chapter 10: Undertaking meta-analysis. Draft version (4 October 2022) for inclusion in: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y, editor(s). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2. London: Cochrane.

34. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: A Generalised Linear Modelling Framework for Pairwise and Network Meta-Analysis of Randomised Controlled Trials. 2011; last updated September 2016; available from

http://www.nicedsu.org.uk.
35. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. Int J Epidemiol. 2012;41(3):818-27.
36. Ren S, Oakley JE, Stevens JW. Incorporating Genuine Prior Information about Between-Study Heterogeneity in Random Effects Pairwise and Network Meta-analyses. Med Decis Making. 2018;38(4):531-42.
37. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health care evaluation. Chichester ; Hoboken, NJ: Wiley; 2004. xiv, 391 p. p.
38. Gasparrini A, Armstrong B, Kenward MG. Multivariate meta-analysis for non-linear and other multi-parameter associations. Stat Med. 2012;31(29):3821-39.
39. Doebler P, Holling H, Sousa-Pinto B. Meta-Analysis of Diagnostic Accuracy with mada.
40. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software. 2015;67(1):1-48.
41. MetaDTA. https://apps.crsu.org.uk/MetaDTA/.
42. Freeman SC, Kerby CR, Patel A, Cooper NJ, Quinn T, Sutton AJ. Development of an interactive web-based tool to conduct and interrogate meta-analysis of diagnostic test accuracy studies: MetaDTA. BMC Med Res Methodol. 2019;19(1):81.
43. Harbord R, Whiting P. metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. Stata Journal 2009;9:211-29.
44. Dwamena B. midas: Stata module for meta-analytical integration of diagnostic test accuracy studies. Statistical Software Components S456880. Department of Economics, Boston College; 2007. ideas.repec.org/c/boc/bocode/s456880.html.
45. Nyaga VN, Arbyn M. Metadta: a Stata command for meta-analysis and meta-regression of diagnostic test accuracy data – a tutorial. Arch Public Health 2022;80:95.
46. Nyaga VN, Arbyn M. Comparison and validation of metadta for meta-analysis of diagnostic test accuracy studies. Res Synth Methods. 2023;14(3):544-62.
47. Plummer M, editor JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. 3rd International Workshop on Distributed Statistical Computing (DSC 2003); 2003 March 2003; Vienna.
48. Stan Modeling Language Users Guide and Reference Manual, Version 2.35. https://mc-stan.org. 2024.
49. Cerullo E, Sutton AJ, Jones HE, Wu O, Quinn TJ, Cooper NJ. MetaBayesDTA: codeless Bayesian meta-analysis of test accuracy, with or without a gold standard. BMC Med Res Methodol. 2023;23(1):127.
50. Verde PE. bamdit: An R Package for Bayesian Meta-Analysis of Diagnostic Test Data. Journal of Statisticsl Software. 2018;86(10):1-32.
51. Guo J, Riebler A. meta4diag: Bayesian Bivariate Meta-Analysis of Diagnostic Test Studies for Routine Practice. Journal of Statistical Software. 2018;83(1):1-31.
52. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B. 2009;71(2):319-92.
53. Steinhauser S, Schumacher M, Rucker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC Med Res Methodol. 2016;16(1):97.

54. Hoyer A, Hirt S, Kuss O. Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. Res Synth Methods. 2018;9(1):62-72.

55. NICE. Diagnostic Guidance: [DG56] Quantitative faecal immunochemical testing to guide colorectal cancer pathway referral in primary care. https://www.nice.org.uk/guidance/dg56 2023.

56. O'Hagan A. HSSS model critisism. In Green PJ, Hjort NL, Richardson, S. (eds). Highly Structured Stochastic Systems. Oxford University Press, 2003. 423-444.

57. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010;29(7-8):932-44.

58. Zapf A, Albert C, Fromke C, Haase M, Hoyer A, Jones HE, et al. Meta-analysis of diagnostic accuracy studies with multiple thresholds: Comparison of different approaches. Biom J. 2021;63(4):699-711.

59. Rücker G, Steinhauser S, Kolampally S, Schwarzer G. diagmeta: meta-analysis of diagnostic accuracy studies with several cutpoints. R package version 0.4-0; 2020. CRAN.R-project.org/package=diagmeta.

60. Doubilet P, Begg CB, Weinstein MC, Braun P, McNeil BJ. Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. Med Decis Making. 1985;5(2):157-77.

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

62. Ades AE, Claxton K, Sculpher M. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. Health Econ. 2006;15(4):373-81.

63. Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 6:Embedding evidence synthesis in probabilistic cost-effectiveness analysis: software choices. 2011; last updated April 2012; available from http://www.nicedsu.org.uk.

64. Ades AE, Lu G, Higgins JP. The interpretation of random-effects meta-analysis in decision models. Med Decis Making. 2005;25(6):646-54.

65. Welton NJ, White IR, Lu G, Higgins JP, Hilden J, Ades AE. Correction: interpretation of random effects meta-analysis in decision models. Med Decis Making. 2007;27(2):212-4.

66. Welton NJ, Ades AE, Caldwell DM, Peters TJ. Research prioritization based on expected value of partial perfect information: a case-study on interventions to increase uptake of breast cancer screening. . Journal of the Royal Statistical Society Series A: Statistics in Society. 2008;171(4):807-41.

67. Welton NJ, Soares MO, Palmer S, Ades AE, Harrison D, Shankar-Hari M, et al. Accounting for Heterogeneity in Relative Treatment Effects for Use in Cost-Effectiveness Models and Value-of-Information Analyses. Med Decis Making. 2015;35(5):608-21.

68. Zgodic A, Schmid CH, Olkin I, Trikalinos TA. Different evidence summaries have implications for contextualizing findings of meta-analysis of diagnostic tests. J Clin Epidemiol. 2019;109:51-61.

69. Soares MO, Walker S, Palmer SJ, Sculpher MJ. Establishing the Value of Diagnostic and Prognostic Tests in Health Technology Assessment. Med Decis Making. 2018;38(4):495-508.

70. Deeks JJ, Bossuyt P, Leeflang MM, Takwoingi Y, Cochrane Collaboration. Cochrane handbook for systematic reviews of diagnostic test accuracy. Hoboken, NJ London: Wiley-Blackwell ; The Cochrane Collaboration; 2023. pages cm p.
71. Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Also published as a special supplement to the Journal of General Internal Medicine, July 2012. .
72. Nyaga VN, Aerts M, Arbyn M. ANOVA model for network meta-analysis of diagnostic test accuracy data. Stat Methods Med Res. 2018;27(6):1766-84.
73. Owen RK, Cooper NJ, Quinn TJ, Lees R, Sutton AJ. Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. J Clin Epidemiol. 2018;99:64-74.
74. V NN, Arbyn M, Aerts M. Beta-binomial analysis of variance model for network meta-analysis of diagnostic test accuracy data. Stat Methods Med Res. 2018;27(8):2554-66.
75. Ma X, Lian Q, Chu H, Ibrahim JG, Chen Y. A Bayesian hierarchical model for network meta-analysis of multiple diagnostic tests. Biostatistics. 2018;19(1):87-102.
76. Nikoloulopoulos AK. An One-Factor Copula Mixed Model for Joint Meta-Analysis of Multiple Diagnostic Tests. Journal of the Royal Statistical Society Series A: Statistics in Society. 2022;185(3):1398-423.
77. Menten J, Lesaffre E. A general framework for comparative Bayesian meta-analysis of diagnostic studies. BMC Med Res Methodol. 2015;15:70.
78. Lian Q, Hodges JS, Chu H. A Bayesian Hierarchical Summary Receiver Operating Characteristic Model for Network Meta-analysis of Diagnostic Tests. J Am Stat Assoc. 2019;114(527):949-61.
79. Trikalinos TA, Hoaglin DC, Small KM, Terrin N, Schmid CH. Methods for the joint meta-analysis of multiple tests. Res Synth Methods. 2014;5(4):294-312.
80. Nikoloulopoulos AK. Joint meta-analysis of two diagnostic tests accounting for within and between studies dependence. Stat Methods Med Res. 2024:9622802241269645.
81. Novielli N, Sutton AJ, Cooper NJ. Meta-analysis of the accuracy of two diagnostic tests used in combination: application to the ddimer test and the wells score for the diagnosis of deep vein thrombosis. Value Health. 2013;16(4):619-28.
82. Welton NJ, Phillippo DM, Owen R, Jones HJ, Dias S, Bujkiewicz S, et al. DSU Report. CHTE2020 Sources and Synthesis of Evidence; Update to Evidence Synthesis Methods. March 2020.
83. Faria R, Soares MO, Spackman E, Ahmed HU, Brown LC, Kaplan R, et al. Optimising the Diagnosis of Prostate Cancer in the Era of Multiparametric Magnetic Resonance Imaging: A Cost-effectiveness Analysis Based on the Prostate MR Imaging Study (PROMIS). Eur Urol. 2018;73(1):23-30.
84. Corbett M, Duarte A, Llewellyn A, Altunkaya J, Harden M, Harris M, et al. Point-of-care creatinine tests to assess kidney function for outpatients requiring contrast-enhanced CT imaging: systematic reviews and economic evaluation. Health Technol Assess 2020;24(39).
85. Chu H, Chen S, Louis TA. Random Effects Models in a Meta-Analysis of the Accuracy of Two Diagnostic Tests Without a Gold Standard. J Am Stat Assoc. 2009;104(486):512-23.

86. Dendukuri N, Schiller I, Joseph L, Pai M. Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. Biometrics. 2012;68(4):1285-93.
87. Lu Y, Dendukuri N, Schiller I, Joseph L. A Bayesian approach to simultaneously adjusting for verification and reference standard bias in diagnostic test studies. Stat Med. 2010;29(24):2532-43.
88. Walter SD, Macaskill P, Lord SJ, Irwig L. Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Stat Med. 2012;31(11-12):1129-38.
89. Riley RD, Levis B, Takwoingi Y. IPD Meta-Analysis for Test Accuarcy Research. Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research2021. p. 387-420.

APPENDIX A WINBUGS CODE

A1 PROGRAM 1: BINOMIAL BIVARIATE RANDOM EFFECTS MODEL

```
model{   #program starts

for (i in 1:ns) { #loop through all studies
  r[i,1] ~ dbin(p[i,1],N[i,1]) #Binomial likelihood for TP
  r[i,2] ~ dbin(p[i,2],N[i,2]) #Binomial likelihood for FP
  logit(p[i,1]) <- delta[i,1] # model for linear predictor of TPF
  logit(p[i,2]) <- delta[i,2] # model for linear predictor of FPF
  delta[i,1] ~ dnorm(theta[1],prec[1]) # random effects on TPF
  delta[i,2] ~ dnorm(condmean[i],condprec) # random effects on TPF
  condmean[i] <-theta[2] +rho*(sd[2]/sd[1])*(delta[i,1]-theta[1])
  #calculate conditional mean
}#loop through studies ends

theta[1] ~ dlogis(0,1) # Prior for mean logit(TPF)
theta[2] ~ dlogis(0,1) # Prior for mean logit(FPF)
sd[1] ~ dunif(0,5) # Prior for between-studies SD in logit(FPF)
sd[2] ~ dunif(0,5) # Prior for between-studies SD in logit(FPF)
rho ~ dunif(-1,1) # Prior for between-studies correlation
prec[1] <- pow(sd[1],-2) # define logit(TPF) precision
prec[2] <- pow(sd[2],-2) # define logit(FPF) precision
condprec <- 1/((1-pow(rho,2))*pow(sd[2],2))# conditional precision
sumtpf <- exp(theta[1])/(1+exp(theta[1])) # summary TPF
sumfpf <- exp(theta[2])/(1+exp(theta[2])) # summary FPF
spec <- 1-sumfpf # summary specificity
beta <- log(sd[2]/sd[1])# beta parameter for HSROC model

# THETA threshold parameter for HSROC model:
Theta <- ((sqrt(sd[2]/sd[1]) )*theta[1]
+(sqrt(sd[1]/sd[2]) )*theta[2])*(1/2)

# LAMBDA accuracy parameter for HSROC model:
Lambda <- (sqrt(sd[2]/sd[1]) )*theta[1]-(sqrt(sd[1]/sd[2]) )*theta[2]

vartheta <- (1/2)*sd[1]*(sd[2]+sd[2]*rho)#variance of theta HSROC
varalpha <- 2*sd[1]*(sd[2]-sd[2]*rho) #variance of alpha HSROC
sdtheta <- sqrt(vartheta)# standard deviation of theta HSROC parameter
sdalpha <- sqrt(varalpha)# standard deviation of alpha HSROC parameter

# Predictive distributions:
predtpr ~  dnorm(theta[1],prec[1]) # logit sensitivity predictive value
predfpf ~  dnorm(condmeanpred,condprec) # logit FPF predictive value
condmeanpred <- theta[2]+rho*(sd[2]/sd[1])*(predtpr-theta[1])
#predictive cond mean
logit(pred_tpr) <- predtpr # predictive sensitivity
logit(pred_fpf) <- predfpf   # predictive fpf
predspec <- 1-pred_fpf  # predictive specificity
predlambda<-(sqrt(sd[2]/sd[1]))*predtpr-(sqrt(sd[1]/sd[2]))*predfpf
# predictive Lambda

}# program ends
```

A2 PROGRAM 2: MULTIPLE THRESHOLDS MODEL (RESTRICTED COVARIANCE STRUCTURE)

Note that because of the inclusion of likelihoods x[i,j,t] ~ Binomial(x[i,j,t-1], p[i,j,t]), the code will not run if there are consecutive x[i,j,] counts equal to zero. If there is a sequence of zero counts in the data within a study, all zero counts after the first should be removed prior to reading the data into WinBUGS. For example, if for a particular study *i* the corresponding counts for the *j*th patient group x[i,j,] are equal to (13,7,7,1,0,0,0,0), then they should be replaced as follows, to avoid computational errors: (13,7,7,1,0,NA,NA,NA). Note that, due to this requirement, “Tc” (number of thresholds) values within a study may be different between the diseased and disease-free populations. R code for replacing these consecutive zero counts is provided in the GitHub repository.

```
model{#program begins
for(i in 1:ns){ #loop over studies
  for(j in 1:2){ #loop over disease status
    n[i,j,1] <- N[i,j] #no of participants for the 1st thres
    p[i,j,1] <- pr[i,j,1]#probability of a positive test for 1st thres
    for(t in 2:Tc[i,j]){ #loop over thresholds
      n[i,j,t] <- x[i,j,t-1] #no of participants at t thres
      p[i,j,t] <-  pr[i,j,t] / pr[i,j,t-1]
      #probability of a positive test at threshold t
    } #end threshold loop
  }#end disease status loop
  for(t in 1:Tc[i,2]){  #loop over thresholds
    q[i,t] <-((pow(C[i,t],lambda)-1)/lambda)*(1-equals(lambda,0)) +
    log(C[i,t])*equals(lambda, 0) #Box–Cox transformation
  }#end threshold loop
  for(j in 1:2){ #loop over disease status
    for(t in 1:Tc[i,j]){#loop over thresholds
      x[i,j,t] ~ dbin(p[i,j,t], n[i,j,t]) # Likelihood
      #logit probability of a positive test
      d[i,j,t] <- (mu[i,j] - q[i,t] ) / s[i,j]
      logit(pr[i,j,t]) <- min(10, max(-10, d[i,j,t]))
      xhat[i,j,t] <- p[i,j,t]*n[i,j,t] # Fitted values
      dev[i,j,t]<-2*(x[i,j,t]*(log(x[i,j,t])-log(xhat[i,j,t]))
      +(n[i,j,t]-x[i,j,t])*(log(n[i,j,t]-x[i,j,t])
      - log(n[i,j,t] - xhat[i,j,t])))
      # Residual deviance contribution
    }#end threshold loop
  }#end disease status loop

  # Distributions of correlated random effects:
  mu[i,1] ~ dnorm(mean[1], prec[1])
  mu[i,2] ~ dnorm(cond.mean.mu[i], cond.prec.mu)
  cond.mean.mu[i] <- mean[2] + (rho_mu*sd[2]/sd[1])*(mu[i,1] - mean[1])
  for(j in 1:2){
    cond.mean.s[i,j] <- mean[j+2]+(rho_mu_sigma*sd[j+2]/sd[j])
                        *(mu[i,j]-mean[j])
    logs[i,j] ~ dnorm(cond.mean.s[i,j], cond.prec.s[j])I(-5,)
    s[i,j] <- exp(logs[i,j])
  }
  rd[i]<-sum(dev[i,1,1:Tc[i,1]])+sum(dev[i,2,1:Tc[i,2]])
  #Residual deviance study i
}

# Predictive distributions for random effects:
mupred[1] ~ dnorm(mean[1], prec[1])
cond.mean.mu.pred <- mean[2] + (rho_mu*sd[2]/sd[1])*(mupred[1] - mean[1])
mupred[2] ~ dnorm(cond.mean.mu.pred, cond.prec.mu)
for(j in 1:2){
  cond.mean.s.pred[j] <- mean[j+2] +(rho_mu_sigma*sd[j+2]/sd[j])
                         *(mupred[1] - mean[1])
  logspred[j] ~ dnorm(cond.mean.s.pred[j], cond.prec.s[j])
}

# Priors:
lambda ~ dunif(-3,3)
for(r in 1:4){
  mean[r] ~ dnorm(0, 0.001)
  sd[r] ~ dunif(0,5)
  prec[r] <- pow(sd[r], -2)
}
rho_mu ~ dunif(-1,1)
rho_mu_sigma ~ dunif(-1,1)

# define conditional precisions
cond.var.mu <-  (1- pow(rho_mu,2))*pow(sd[2], 2)
cond.prec.mu <- 1/cond.var.mu
for(j in 1:2){
  cond.var.s[j]<-  (1- pow(rho_mu_sigma,2))*pow(sd[j + 2], 2)
  cond.prec.s[j] <- 1/cond.var.s[j]
}
resdev <- sum(rd[])    # Total residual deviance
}#Program ends
```

APPENDIX B SUPPLEMENTARY ANALYSES FOR EXAMPLES 3 AND 4

B1 MODEL CRITIQUE FOR THE OC-SENSOR EXAMPLE

In *Section 5.2.1*, we observed that the posterior mean residual deviance (658.0) was very high relative to the number of data points (84) for the OC-Sensor example.

Examination of residual deviance contributions by study and disease group (*Table 8*) shows that model fit is good overall for sensitivity but poor for specificity in several studies. The large residual deviance contributions are from studies with large disease-free populations and reporting data at three or more thresholds.

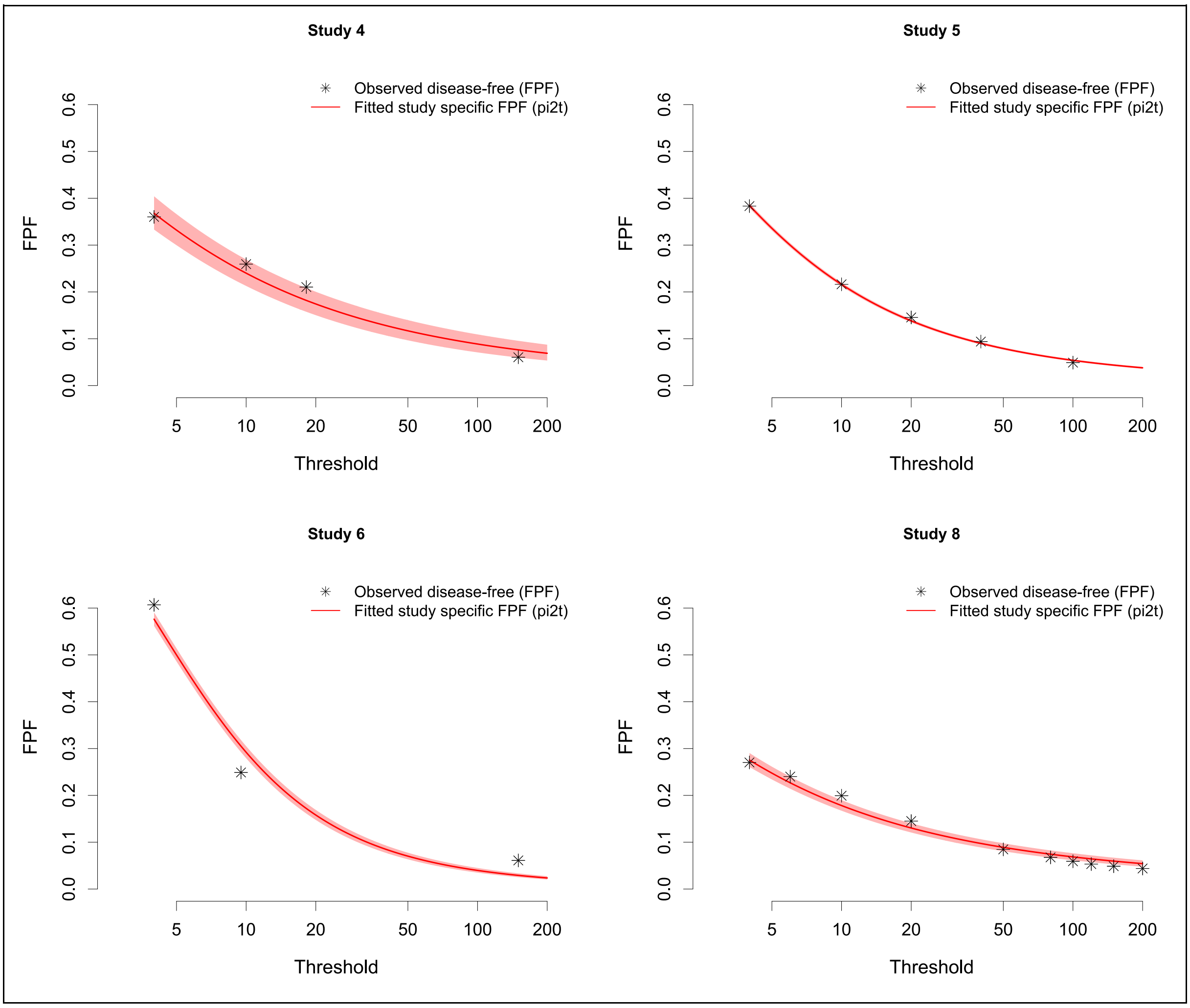
Table 8: Residual deviance contributions from all studies according to the Box–Cox (independence structure) version: OC-Sensor data

	Diseased population				Disease-free population			
Study	Number of individuals	Data points	Residual deviance	Residual deviance per data point	Number of individuals	Data points	Residual deviance	Residual deviance per data point
1	11	3	3.3	1.1	155	3	3.8	1.3
2	17	7	5.2	0.7	2,875	7	22.7	3.2
3	74	3	3.3	1.1	5,267	3	14.9	5.0
4	38	4	2.0	0.5	694	4	21.2	5.3
5	514	5	4.8	1.0	33,180	5	108.6	21.7
6	61	3	2.3	0.8	4,126	3	354.9	118.3
7	54	1	1.1	1.1	3,408	1	1.0	1.0
8	90	10	10.2	1.0	3,506	10	79.2	7.9
9	12	2	5.5	2.8	346	2	1.9	1.0
10	28	2	4.1	2.0	722	2	2.1	1.1
11	73	2	4.0	2.0	4,470	2	2.1	1.0

In *Figure 17* we compare observed versus fitted study-specific FPFs (i.e. p_{i2t}) for the four studies with the greatest residual deviance contributions. We see that the study-specific estimates are very precise, driven by large sample sizes. As a result, differences in observed versus fitted values that are small in magnitude are leading to large residual deviance contributions. In the most striking case of this (Study 5), the maximum absolute difference between observed and fitted FPF is 0.01 but the five data points contribute total residual deviance of 108.6. The differences are slightly larger (maximum of 0.06), but still not substantial, in Study 6.

Noting that the model assumes a linear relationship between λ-transformed threshold and logit FPF, it is perhaps not surprising that studies contributing large denominators and three or more thresholds may demonstrate some deviations from this. Given the overall good visual fit of the summary estimates and robustness of these to choice of transformation (*Figure 7),* the model seems to be providing a reasonable synthesis of the data. It is also worth mentioning that sample sizes in practice will rarely be this large, making this a perhaps unusual example.

Figure 17: Observed vs Fitted FPF for the four studies with the highest residual deviance contribution: OC-Sensor data. Shaded regions depict 95% CrIs around fitted values



B2 SENSITIVITY ANALYSIS TO PRIOR CHOICE FOR THE FOB-GOLD EXAMPLE

For the FOB-Gold example (*Section 5.3.1*) we performed additional analyses to assess sensitivity to choice of HN(0,1) as a weakly informative prior for the four between-study standard deviation parameters, exploring results when these were made either more (HN(0,0.5^2)) or less (HN(0,2^2)) precise. *Table 9* shows results.

Table 9: Sensitivity analysis around choice of prior distribution for between-studies standard deviation parameters: FOB-Gold example

Prior distribution for between-study SDs	Threshold				Model fit		
	10µg/g		100µg/g				
	Sensitivity	Specificity	Sensitivity	Specificity	Residual deviance	pD	DIC
Uniform(0,5)	0.74 (0.50,1.00)	0.89 (0.63,1.00)	0.61 (0.23,1.00)	0.96 (0.73,1.00)	13.3	3.5	16.8
Set to zero (fixed effect model)	0.88 (0.78,0.95)	0.88 (0.75,0.97)	0.82 (0.56,0.93)	0.96 (0.86,1.00)	35.0	2.8	37.8
HN(0, 1)	0.84 (0.55,1.00)	0.88 (0.76,0.97)	0.69 (0.49,0.98)	0.96 (0.87,1.00)	15.2	6.7	21.9
HN(0, 0.5^2)	0.86 (0.64,0.99)	0.88 (0.80,0.95)	0.75 (0.55,0.94)	0.96 (0.90,0.99)	18.3	7.0	25.3
HN(0, 2^2)	0.79 (0.51,1.00)	0.88 (0.71,0.99)	0.64 (0.38,1.00)	0.96 (0.82,1.00)	13.8	5.7	19.5