International Journal of Technology Assessment in Health Care, 29:1 (2013), 53–60.

doi:10.1017/S0266462312000682

ECONOMIC EVALUATIONS AND DIAGNOSTIC TESTING: AN ILLUSTRATIVE CASE STUDY APPROACH

Sabina Sanghera, Rosanna Orlando, Tracy Roberts
Health Economics Unit, University of Birmingham

Objective: The aim of this study was to present a clear process of synthesizing test accuracy data when conducting economic evaluations of diagnostic tests for health technology assessment (HTA) assessors and health economists.
Methods: We appraised the methods advocated for using diagnostic test accuracy data in economic evaluations. We used a case study of fetal anemia in which data from a screening test are used in combination with a confirmatory test.
Results: We developed a step-by-step guide and consider two scenarios: when data on test accuracy from several studies are based on (i) the *same* test threshold for positivity and (ii) *different* test thresholds.
Conclusions: We conclude that each approach has its strengths and limitations. We show that the optimal operating point of the test should be identified to determine the true cost-effectiveness of the test. We advocate that these issues require a multidisciplinary team of health economists, decision modelers and statisticians.

Keywords: Economic evaluation, Diagnostic tests, Sensitivity and specificity, ROC curve

Tests are used to determine whether a patient has a disease and requires treatment. As tests cannot correctly identify the patient's disease status 100 percent of the time, diagnostic errors will occur. To assess if a test or combination of tests should be used in practice, to avoid an inefficient use of healthcare resources, a model-based economic evaluation is conducted to compare alternative testing strategies in terms of their relative costs and clinical effectiveness (1).

The value of any test depends on its diagnostic test accuracy, which is the ability of the test to differentiate between patients that have a disease and those that do not (2). It is commonly presented in terms of sensitivity and specificity.

A consensus toward a uniform approach for incorporating test accuracy in economic evaluations does not yet exist. Most economic evaluations of diagnostic tests consider sensitivity and specificity to be independent which is not always appropriate (3). Attempts by others to provide advice on appropriate methods have typically lacked clarity (4–6). Research into correctly identifying appropriate diagnostic tests and the threshold at which they are most cost-effective must be conducted to prevent poor healthcare resource management. Other factors such as patient compliance and skills in test performance do affect health outcomes, but test accuracy can be easily addressed.

We aim to propose a process of synthesizing evidence and conducting economic evaluations of diagnostic tests. This study is divided into two sections. First, we summarize the three main suggested approaches for handling accuracy data for economic evaluations, highlighting their advantages and disadvantages. We incorporate the main advantages to provide a simplified guide for dealing with the issues. Second, we demonstrate the use of our guide in a case study of fetal anemia.

SUMMARY OF KEY ISSUES AND TRADE-OFF BETWEEN SENSITIVITY AND SPECIFICITY

Sensitivity identifies the True Positive Rate (TPR) of a test. This is the proportion of times the test positively identifies the true positives (TPs) that do have the disease (7). Specificity is the proportion of times the test negatively identifies the True Negatives (TNs) that do not have the disease (7). This is known as the True Negative Rate (TNR).

Generally diagnostic test results are in the form of continuous, ordinal data, where results are not classified as disease positive or negative, but are categorized in terms of disease severity (8). Sensitivity and specificity depend on arbitrary cutoff (threshold) levels, chosen to categorize a result as positive and further determine whether a positive result is mild, moderate or severe (8). When single studies identify sensitivity and specificity for several cutoff values a receiver operating characteristic (ROC) curve of test accuracy for different disease severity levels could be presented. Similarly, when several studies use different cutoff levels the data can be combined using a summary ROC (SROC) curve.

First, we thank the clinicians, Professor Mark Kilby and Dr. Katie Morris, for their time and expertise in this complex disease area. Second, we also thank Dr. Pelham Barton for his input on the decision model and Professor Jon Deeks for offering advice in respect of SROC curves. The current work was carried out unfunded.

Zweig and Campbell (9) explained that sensitivity and specificity are negatively correlated and, therefore, cannot be considered separately. At individual test thresholds for positivity both diseased and nondiseased patients would be categorized as positive (Supplementary Figure 1, which can be viewed online at www.journals.cambridge.org/thc2012073). This occurs because tests generally do not have 100 percent TPR (sensitivity) and 100 percent TNR (specificity). As the test threshold for positivity is lowered, sensitivity will increase and specificity will decrease. Hence, the number of individuals who are correctly categorized test positive will increase, but the number of nondiseased who are correctly identified as negative will decrease, resulting in the number of false positives increasing (9). This indicates that the test accuracy results obtained are linked to the threshold used in the study (9). The optimal clinical test threshold would have an appropriate balance between TPR and TNR (9).

When continuous data are available, it has been argued that health economics techniques and principles should be used to identify the test threshold for categorizing a positive result (4;6;9). This would enable the investigator to identify the test threshold belonging to the sensitivity and specificity pair at which the test is most cost-effective. This test threshold is often referred to as the Optimal Operating Point (OOP). Once the OOP is identified, it could be entered into the economic evaluation as the sensitivity and specificity pair for the test.

In reality, the correlation between sensitivity and specificity is ignored; the OOP is not used because the test threshold is often, incorrectly, based on one threshold value that has been chosen to be clinically preferable even though several cutoffs are available (8). Furthermore, to overcome the limitations of selecting one threshold, a sensitivity analysis is conducted to assess uncertainty of results when the sensitivity and specificity pair is changed. Typically, sensitivity is changed while keeping specificity constant based on the assumption that they are independent when they are in fact correlated (10;11).

Current Approaches

We identified three proposed methods of conducting economic evaluations of diagnostic tests. All the approaches agreed that the OOP for each test should be identified and entered into the economic evaluation, but the optimal method by which the OOP should be identified is unclear (4;6). The three methods identified include the ROC curve, Value of Information (VOI), and ROTS (which is not an acronym). We will briefly discuss these approaches and provide some insight as to why there is no consensus on a uniform approach. The strengths and limitations of each approach are highlighted in Table 1.

ROC Curve. ROC curves illustrate the range of sensitivity and specificity pairs that a test can achieve for every test threshold value, and demonstrates the trade-off between the two as the test threshold for positivity changes (9). The ROC curve approach involves applying a formula, which incorporates only the cost implications of test results, and uses this formula to create a tangent on the ROC curve. The point at which the ROC curve and the slope meet is the OOP for the test. For further details on the methodology, see References 4;5;9;12.

VOI. VOI is based on clinical decision analysis and determines the cost-effectiveness of a test for a range of disease prevalence values (6). A formula is applied which ensures the optimal net benefit of a test is identified for a certain prevalence (7). The results are used to create a tangent on the ROC curve and the point where the tangent and the ROC curve meet is the OOP. This process is repeated for a range of disease prevalence values. For further details on the methodology see Phelps and Mushlin (6).

Here, the VOI does not stand for the value of reducing uncertainty in cost-effectiveness analysis as understood by health economists (4;6). The VOI by Phelps and Mushlin (6) relates to minimizing uncertainty of population heterogeneity (4).

ROTS. The ROTS approach is an adaptation of the VOI approach and involves the translation of the ROC curve into the cost-effectiveness space (4). ROTS assumes a fixed prevalence and willingness-to-pay (WTP) threshold. ROTS uses nonlinear expansion to incorporate partial implementation of treatments at the same time (4). It is assumed that the mean costs and quality-adjusted life-years (QALYs) for every sensitivity and specificity pair on the ROC curve is calculated, demonstrating the changes in costs and effects as the diagnostic threshold is changed. The mean costs and QALYs of changing the test threshold *and* transferring patients from one treatment to another is then plotted in the cost-effectiveness space. Net monetary benefit (NMB) isoquants for £30,000 WTP are also plotted (4). The OOP is identified as the greatest net benefit, similar to the VOI approach of Phelps and Mushlin. It is asserted that alternative tests can then be compared based on the cost-effectiveness results at their OOP (4). For further details on the methodology, see Laking et al. (4).

Data from Multiple Studies

The approaches discussed previously are based on data from individual studies and have not considered the incorporation of data from several studies by meta-analysis. The Bayesian approach by Sutton et al. (13) incorporates meta-analysis into the economic evaluation, but the approach is complex, involving comprehensive decision modeling. The basic principle underpinning all the available approaches is similar: In ROTS and Sutton et al., an ROC curve is generated and the OOP is identified by determining the sensitivity and specificity pair with the greatest NMB for a test. Sutton et al. (13). recognize the VOI approach in their article and discussed their own approach with the authors of ROTS but chose not to use either approach to identify the OOP, which may suggest that neither the ROTS nor VOI are appropriate when a meta-analysis is conducted.

https://doi.org/10.1017/S0266462312000682 Published online by Cambridge University Press

Table 1. Advantages and Disadvantages of Current Approaches

Approach	Advantages	Disadvantages	Comment
ROC Curve	• Optimal method of presenting sensitivity and specificity data. Illustrates range of sensitivity and specificity pairs and demonstrates trade-off	• Presented in cost benefit framework: not preferred by decision-makers • ROC curve not used to full potential to identify OOP	Appears to refer to data from a single study only and does not discuss meta-analyses Suggested that it can be adapted to net benefit framework but the method is not clearly outlined which may have led to its limited use [2].
VOI by Phelps and Mushlin	• Presented in cost-effectiveness analysis and net monetary benefit framework: preferred method by decision-makers • Comprehensive: considers prevalence, assesses cost-effectiveness and presents a method for decision-making • Utilises ROC Curve to identify OOP	• Presented in net-benefit/prevalence space: not ideal for decision-makers • Assesses cost-effectiveness according to changes in prevalence: decision-makers require results according to current prevalence of disease; better suited to sensitivity analysis • Sensitivity analysis is not discussed or self-explanatory	Appears to refer to data from a single study only, does not discuss meta-analyses We believe VOI is a good approach but may not be widely adopted due to its complexity
ROTS by Laking et al	• Presented in cost-effectiveness analysis and net monetary benefit framework: preferred method by decision-makers • Utilises ROC curve to identify OOP – the OOP should be used when available • Most intuitive approach	• Computationally complex • Assumes high level of existing knowledge in the area • Sensitivity analysis is only briefly explained • Claimed to be more in line with cost-effectiveness analysis. Promotes quality adjusted life years, which is not always appropriate for diagnostic testing	Appears to refer to data from a single study only, does not discuss meta-analyses The approach appears to address many of the limitations of the previous two approaches, but we believe it is overly complex and must be presented in a more accessible manner
Bayesian approach by Sutton et al.	• Presented in cost-effectiveness analysis and net monetary benefit framework: preferred method by decision-makers • Utilises ROC Curve to identify OOP – the OOP should be used when available	• Overly complex: uses comprehensive decision modelling and thus assumes advanced statistical knowledge	Integrates economic evaluations and meta-analyses, does not discuss single studies

Note. The table shows the advantages and disadvantages of the ROC Curve, VOI, ROTS and the approach by Sutton et al. A pragmatic approach was taken to identify these advantages and disadvantages based on possible practical reasons for why these approaches may not have been widely adopted.
ROC curve, receiver operating characteristic curve; VOI, value of Information; OOP, optimal operating point; ROTS is not an acronym.

Case Study

Fetal anemia is a condition where the fetus does not receive enough oxygen. It ranges from mild to severe and can lead to morbidity and mortality if untreated (14). A question of current interest is whether the MCA Doppler test, which is a diagnostic ultrasound test, should be incorporated into the clinical pathway. The accuracy of the MCA Doppler will determine which fetuses require confirmatory diagnosis using an invasive cordocentesis test which carries its own risk of mortality. Cordocentesis is performed on all cases that the MCA Doppler identifies as positive to confirm the need for treatment with intravascular-intrauterine transfusion (IUT) (15). By incorporating the MCA Doppler into the screening pathway for fetal anemia, the number of avoidable life-threatening cordocentesis procedures may decline, as only those patients that are identified to be positive by the Doppler would receive the cordocentesis and its associated risks.

In this section, we aim to demonstrate the process of synthesizing evidence and conducting economic evaluations of diagnostic tests. We use fetal anemia as a case study because there is no economic evidence for current practice (K. Morris, personal communication, 2010). For each scenario, we provide a step-by-step guide that explains the best method to use when data from both individual and several studies are provided. Our approach will be a simplified version of ROTS, similar to Sutton et al., as it is the most intuitive, but will avoid the complexity. Our guide to conducting economic evaluations of diagnostic tests is outlined in Table 2.

Diagnostic Testing. The MCA Doppler is a noninvasive ultrasound device. A test threshold value of 1.29 multiples of the median (MOM) identifies mild fetal anemia, 1.5 MOM identifies moderate anemia and 1.55 MOM identifies severe anemia (17). A

https://doi.org/10.1017/S0266462312000682 Published online by Cambridge University Press

Table 2. Guide to Conducting Economic Evaluations of Diagnostic Tests

	Single study	Multiple studies	Comment
Study uses one test threshold to identify abnormal result	• Treat sensitivity and specificity independently • Cannot identify OOP • Enter available sensitivity and specificity pair for each test into decision model as convention	• Treat sensitivity and specificity independently • Cannot identify OOP • Collate data from each study using traditional meta-analysis techniques • Enter overall test accuracy result from meta-analysis into decision model • Compare against alternative test	
Study uses several different test thresholds to identify abnormal result	• Correlation between sensitivity and specificity is relevant • Identify the OOP using ROC curve: Sensitivity and specificity pair with the greatest NMB • Enter test accuracy at OOP into decision model • Compare against OOP of alternative test	• Correlation between sensitivity and specificity is relevant • Collate data from each study using either bivariate meta-analysis or Hierarchical Summary Receiver Operating Characteristic ('HSROC') • Identify the OOP using ROC curve: Sensitivity and specificity pair with the greatest NMB • Enter test accuracy at OOP into decision model • Compare against OOP of alternative test	Bivariate meta-analysis or HSROC are mathematically equivalent and take into account the correlation between sensitivity and specificity, unlike traditional meta-analysis methods. [16]

Note. The table details our guide to conducting economic evaluations of diagnostic tests when data from several or single studies are available and when data on single or several diagnostic thresholds are available.
OOP, optimal operating point; ROC curve, receiver operating characteristic curve; HSROC, hierarchical summary receiver operating characteristic curve; NMB, net monetary benefit.

test threshold of >1.5 MOM is commonly used and is able to detect moderate-severe levels of fetal anemia (18). The tests ability to detect fetal anemia increases as the severity of anemia increases; therefore, the likelihood of incorrect test results decreases.

METHODS

The model-based economic evaluations are cost-effectiveness analyses and were conducted from a UK National Health Service (NHS) perspective in secondary care. Outcomes are expressed as "cost per survival" due to difficulties obtaining utility values in a fetal disease.

Model Structure

A decision tree was developed following consultation with clinical experts. We compared the outcomes of an at-risk population when they receive the strategy of including the MCA Doppler screening test against a strategy of invasive cordocentesis testing only. The decision model (Supplementary Figure 2, which can be viewed online at www.journals.cambridge.org/thc2012074) included the costs and impact of treatment.

Assumptions

We considered one cycle of testing and treatment. We assumed that every fetus has the same risk of developing fetal anemia and those that were identified to have fetal anemia underwent an IUT. We obtained an estimate of the prevalence of at-risk women from previous literature (15). The model parameters and risks associated with IUT and cordocentesis are outlined in Supplementary Table 1, which can be viewed online at www.journals.cambridge.org/thc2012075 (15;18–23).

Resource Use and Costs

We obtained the cost of conducting the MCA Doppler from the Birmingham Women's Hospital by using a uterine Doppler as a proxy, due to difficulties obtaining direct cost data for an MCA Doppler. The cost includes the time spent by the nurse while conducting the test. As cordocentesis and IUT are more complex, we assumed that a consultant performs these procedures, and the cost reflects this. We used the cost of an adult blood transfusion as a proxy for IUT due to difficulties obtaining data. All costs are expressed in 2009. Discounting was not applied, as the time horizon was shorter than a year.

ANALYSIS

Scenario 1

In scenario 1, data from several studies that use the same test threshold to determine a positive result are used. The test threshold of 1.5 MOM is considered across all studies in the review, as shown in Table 2, there is no trade-off between sensitivity and specificity, and they may be treated independently. However, heterogeneity across studies (i.e., disease frequency in the populations) should be noted as this may introduce bias. A

https://doi.org/10.1017/S0266462312000682 Published online by Cambridge University Press

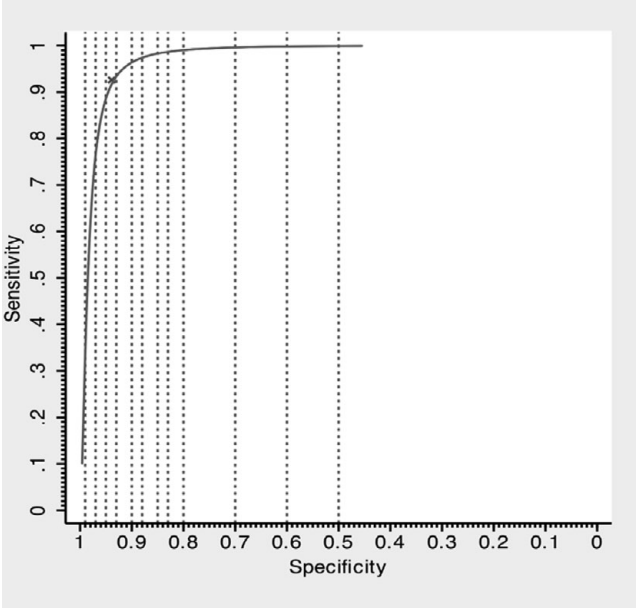

Figure 1. The summary receiver operating characteristic (SROC) curve.

traditional random-effects meta-analysis can be conducted and the pooled meta-analysis results obtained for sensitivity and specificity at the 1.5 MOM threshold can be used as the input in the decision model and compared against cordocentesis alone. We obtained data on the MCA Doppler test accuracy from the most up-to-date systematic review for fetal anemia (18).

In this scenario, as only one test threshold is used, the true cost-effectiveness of the test is not determined because the OOP could not be identified. If evaluations of diagnostic tests are conducted without identifying the OOP it is likely that a test may be incorrectly proven to not be cost-effective because its OOP was not entered into the evaluation.

Scenario 2

In scenario 2, data from several studies that use different test thresholds are used. As discussed in Table 2, the range of sensitivity and specificity pairs that the test can meet are provided and the correlation is relevant. We performed the bivariate meta-analysis using STATA. The SROC curve, generated from the bivariate meta-analysis, was used to identify a range of sensitivity and specificity pairs that the Doppler can meet as a function of changing test thresholds (see Figure 1).

As shown in Table 2, to identify the OOP of the Doppler, specificity values were chosen and the corresponding sensitivity values were read off the SROC curve. Specificity values were chosen mainly around the region where the SROC curve bends, as the changes in sensitivity are greater within this region. The identified pairs were entered into the decision tree and compared against each other, instead of against cordocentesis (using the same model arm for MCA Doppler test as previously), to identify the OOP for the MCA Doppler as the pair with the greatest NMB. The OOP could then be entered into the decision model and compared against cordocentesis alone as previously to determine whether the introduction of the MCA Doppler at its OOP would be more cost-effective than cordocentesis alone. In contrast to scenario 2, the true cost-effectiveness of the test can be determined as the OOP is identified. This scenario is illustrative as some data were not derived from a systematic review.

Similarly, when *one* study uses several thresholds, the above steps are repeated but a meta-analysis is not required and consequently a ROC curve is generated (discussed in Table 2). A practical example could not be provided due to difficulties obtaining patient-level data. In this case, although the true cost-effectiveness can be determined, there is a lack of reliability as the OOP is identified from just one study.

RESULTS

Scenario 1

The results of scenario 1, where multiple studies estimated one threshold, show that the strategy including the Doppler costs £60 with an effectiveness of 0.988 and an overall NMB of £19,670. The strategy with cordocentesis alone costs £280, with an effectiveness of 0.976 and an overall NMB of £19,250 (see Supplementary Table 2, which can be viewed online at www.journals.cambridge.org/thc2012076). The strategy including the 1.5 MOM Doppler dominates the cordocentesis alone strategy, as it is less costly and more effective. The NMB results demonstrate that the optimal strategy incorporates the MCA Doppler as it generates the greatest NMB.

Scenario 2

In scenario 2, where multiple studies estimated different thresholds, the results for each sensitivity and specificity pair of the MCA Doppler are presented in Table 3. NMB measures were calculated to easily identify the most cost-effective sensitivity and specificity pair. The effectiveness values only vary slightly because a comparison between tests is not made. Little difference in effectiveness is expected because there is a trade-off between sensitivity and specificity. The results indicate that the OOP could be within a range of values. Either a threshold with sensitivity of 0.925 and specificity of 0.94 or a sensitivity of 0.88 and specificity of 0.95 is the OOP for the MCA Doppler, as these values both generate the greatest NMB of approximately £19,710. As there is a minimal difference between the two NMB, but a large difference between sensitivity (0.925 or 0.88), the greatest test accuracy should be chosen. In this case, the optimal pair would be 0.925 sensitivity and 0.94 specificity. Although we cannot identify the OOP with great precision, the loss from not being too far away is not too great. For the economic evaluation, the OOP would be entered into the decision model and compared against the OOP of an alternative test.

The results are presented in the cost-effectiveness plane (Supplementary Figure 3, which can be viewed online at www.journals.cambridge.org/thc2012077). It illustrates that

https://doi.org/10.1017/S0266462312000682 Published online by Cambridge University Press

Table 3. Results for Sensitivity and Specificity Values from SROC Curve

Sensitivity	Specificity	Cost	Effectiveness	NMB
1	0.5	£170	0.983	£19,490
1	0.6	£140	0.984	£19,540
0.93	0.7	£120	0.985	£19,580
0.9	0.8	£90	0.986	£19,630
0.985	0.83	£80	0.987	£19,660
0.98	0.85	£80	0.987	£19,670
0.972	0.88	£70	0.988	£19,680
0.96	0.9	£60	0.988	£19,690
0.94	0.93	£60	0.988	£19,700
0.925	0.94	£50	0.988	**£19,710**
0.88	0.95	£50	0.988	**£19,710**
0.755	0.97	£40	0.987	£19,700
0.265	0.99	£30	0.985	£19,660

Note. The table shows the cost and effectiveness (proportion survived) results for the range of sensitivity and specificity pairs identified from the SROC curve for scenario 2. The NMB is also presented and the emboldened NMB values in the boxes both represent the sensitivity and specificity pairs with the greatest NMB. A £20,000 decision-maker WTP value for an additional unit of benefit is used to generate the NMB.

NMB, net monetary benefit; SROC curve, summary receiver operating characteristic curve; WTP, willingness-to-pay. Monetary values have been rounded to the nearest 10 for appropriateness and presentation purposes

sensitivity 0.925, specificity 0.94 is most cost-effective as it is the most effective and least costly and demonstrates the range of values that are marginally less costly and less effective, representing the minimal difference between some test accuracy measures. As the WTP threshold changed from £0 to £50,000 per additional unit of benefit the 0.925, 0.94 pair remained optimal, except at £10,000 and £15,000 WTP.

DISCUSSION

This study discussed how to use sensitivity and specificity appropriately in an economic evaluation; a step-by-step guide to conducting economic evaluations of diagnostic tests was provided (Table 2); and the approach was illustrated. We have shown that when several studies estimate one test threshold (scenario 1), the correlation between sensitivity and specificity is not significant and a traditional random-effects meta-analysis can be conducted. The pooled sensitivity and specificity results can be entered directly into the decision model and compared against an alternative test.

We further showed that when several studies estimated different test thresholds (scenario 2) the OOP of the test should be identified. The correlation between sensitivity and specificity is significant and a bivariate meta-analysis should be used to account for the correlation. An SROC curve should be used to identify a range of sensitivity and specificity pairs, which would be compared against each other in a decision model to identify the OOP of the test, which is the sensitivity and specificity pair with the greatest NMB. The OOP should then be entered into the decision model and compared against the OOP of an alternative test. We believe that this is the optimal approach because the "true" cost-effectiveness of the test can be identified as the OOP is used in economic evaluations. Finally, we briefly discussed that, when a single study estimates different test thresholds, the approach would be similar to scenario 2, but as data from one study are considered, a meta-analysis is unnecessary and an ROC curve would be constructed.

The strength of our study is that we have provided a simple yet comprehensive step-by-step guide to conducting economic evaluations of diagnostic tests. We have explained explicitly when sensitivity and specificity can be treated independently and when it is appropriate to use each approach depending on the available data, which has not been done before. We have taken the strengths of other approaches, avoided their complexities, and provided a recommendation that is readily accessible to health economists.

We were unable to demonstrate our approach when one study estimates different thresholds due to a lack of data, but we believe that our explanation of the method was sufficient to enable a health economist to implement it. Although we could address the use of probabilistic sensitivity analysis (PSA) to assess the uncertainty of results in scenario 1, our investigations have shown that PSA cannot be easily implemented when several studies are used. In these cases, the correlation between sensitivity and specificity is relevant and must be taken into account when specifying distributions for test accuracy parameters. Although we criticized the Bayesian approach by Sutton et al. (13) for being too complex, we have not been able to identify a better method for incorporating PSA. Their approach remains the recommended method when studies estimate a range of test thresholds. When one study estimates several thresholds, the comprehensive decision model approach discussed by Sutton et al. (13) could be applied. This has not been applied before and was not discussed by Sutton et al. (13) as data from multiple studies were the focus of their study. Our initial reservation of the approach by Sutton et al. (13) was due to it requiring *advanced* statistical expertise, which may not be within the skill set of a typical health economist.

A limitation of our approach and that of Sutton et al. (13) is that, when data on multiple thresholds are combined from multiple studies, we are unable to translate the results of the most cost-effective sensitivity and specificity pair into the clinical setting. Although the pair can be identified, the numerical test threshold that the test would need to be set to in the clinical setting to detect a positive result cannot be reliably identified (13). This is because data from different studies are collated to generate the SROC curve. There will be unquantified

https://doi.org/10.1017/S0266462312000682 Published online by Cambridge University Press

variability, such as differing study population and sample sizes, in addition to the variability that we are interested in, of differing test thresholds (Deeks J, personal communication, 2010). The reliability would be improved if each study estimated a range of test thresholds, allowing the ROC curves to be combined. Thus, the analysis would not be reliant on a single study for just one result. This issue does not arise when one study estimates a range of test thresholds because the sensitivity and specificity pairs for each clinical diagnostic test threshold will have been used to construct the ROC curve and, therefore, can be identified. If the OOP lies near one of the known test threshold points on the curve, further research on test accuracy could be conducted within this new range of test thresholds.

In summary, we argue that economic evaluations of diagnostic tests should not be treated as economic evaluations of interventions because these evaluations have issues that require a specific skill set. We attempted to simplify the method for conducting economic evaluations of diagnostic tests and provide one method that can be used for alternative scenarios but reach the conclusion that the method outlined by Sutton et al. (13) should be used when studies consider multiple test thresholds. The method is best approached using a multidisciplinary team. In conclusion, the design of the framework for evaluating diagnostic tests will need further refinement. This study is meant to provide the first step for further investigation into economic evaluations of diagnostic tests.

This study focuses on test accuracy data from a screening test which is followed by a confirmatory test. Further research should outline the appropriate method when several tests are used in clinical practice, but the method is unlikely to deviate much from that discussed here. Further research should involve determining whether it could be possible to identify the numerical test threshold that matches the sensitivity and specificity pair on an SROC curve with certainty when bivariate meta-analysis is conducted. The methods of sensitivity analysis for diagnostic tests should also be researched to ensure that the methods are accessible to health economists.

POLICY IMPLICATIONS

If analysts conducting economic evaluations of diagnostic tests ignore the correlation between sensitivity and specificity and fail to identify the OOP of a test, the cost-effectiveness results will be spurious and policy makers will be misinformed. As highlighted in our case study, if the analysis is conducted incorrectly, it could lead to a course of action, such as an invasive test, that has a high mortality risk. Therefore, the data must be handled accurately to ensure that correct implications for policy makers are provided.

SUPPLEMENTARY MATERIAL

Supplementary Figure 1: www.journals.cambridge.org/thc2012073

Supplementary Figure 2: www.journals.cambridge.org/thc2012074

Supplementary Figure 3: www.journals.cambridge.org/thc2012075

Supplementary Table 1: www.journals.cambridge.org/thc2012076

Supplementary Table 2: www.journals.cambridge.org/thc2012077

CONTACT INFORMATION

Sabina Sanghera (BMedSci, MSc), (sxs574@bham.ac.uk); **Rosanna Orlando (MPharm, HEMSc)**; **Tracy Roberts (BSc (Econ), MPhil, PhD),** Professor of Health Economics, Health Economics Unit, School of Health and Population Sciences, Public Health Building, University of Birmingham, UK, B15 2TT

CONFLICTS OF INTEREST

The authors report no potential conflicts of interest.

REFERENCES

1. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. *Med Decis Making.* 2009;29:E22-9.
2. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. *Biometrics.* 2003;59:936-46.
3. Novielli N, Cooper NJ, Abrams KR, Sutton AJ. How is evidence on test performance synthesized for economic decision models of diagnostic tests? A systematic appraisal of health technology assessments in the UK since 1997. *Value Health.* 2010;13:952-7.
4. Laking G, Lord J, Fischer A. The economics of diagnosis. *Health Econ.* 2006;15:1109-20.
5. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med.* 1978;8:283-98.
6. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. *Med Decis Making.* 1988;8:279-89.
7. Altman DG, Bland JM. Statistics notes: diagnostic tests 1: sensitivity and specificity. *BMJ.* 1994;308:1552.
8. Mandrekar JN, Mandrekar SJ. Statistical Methods in Diagnostic Medicine using SAS Software. [online] 2005 [cited 01.12.10]. Available from: www2.sas.com/proceedings/sugi30/211-30.pdf.
9. Zweig MH, Campbell G. Receiver Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem.* 1993;4:561-77.
10. Meads CA, Cnossen JS, Juarez-Garcia A, et al. Methods of prediction and prevention of pre-eclampsia: a systematic review and evaluation of methods of assessing urinary incontinence. *Health Technol Assess.* 2008;12:1-249.
11. Whiting P, Westwood M, Bojke L, et al. Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model. *Health Technol Assess.* 2006;10:1-154.
12. Chen L, Liao C, Chang S, Lai HC, Chen TH. Cost-effectiveness analysis for determining the optimal cut-off of immunochemical faecal occult blood test for population-based colorectal cancer screening. *J Med Screen.* 2007;14:191-9.

https://doi.org/10.1017/S0266462312000682 Published online by Cambridge University Press

13. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. *Med Decis Making.* 2008;28:650-67.
14. Kumar S, Regan F. Management of pregnancies with RhD alloimmunisation. *BMJ.* 2005;330:1255-8.
15. Bullock R, Martin WL, Coomarasamy A, Kilby MD. Prediction of fetal anemia in pregnancies with red-cell alloimmunization: comparison of middle cerebral artery peak systolic velocity and amniotic fluid OD450. *Ultrasound Obstet Gynecol.* 2005;25:331-4.
16. Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med.* 2008;149:889-97.
17. Mari G, Deter RL, Robert L, et al. Noninvasive diagnosis by Doppler ultrasonography of fetal anaemia due to maternal red cell alloimmunisation. *N Engl J Med.* 2000;342:9-14.
18. Pretlove SJ, Fox CE, Khan KS, Kilby MD. Noninvasive methods of detecting fetal anaemia: a systematic review and meta-analysis. *BJOG.* 2009;116:1558-67.
19. Bricker L, Garcia J, Henderson J, et al. Ultrasound screening in pregnancy: a systematic review of the clinical effectiveness, cost-effectiveness and women's view. *Health Technol Assess.* 2000;4:1-193.
20. Office for National Statistics. Infant mortality rates, England and Wales, 1975–2000. [online] 2005 **[cited 25.08.10]** Available from: http://www.statistics.gov.uk/STATBASE/xsdataset.asp?vlnk=4288&More=Y.
21. Schumacher B, Moise KJ. Fetal Transfusion for red blood cell alloimmunisation in pregnancy. *Obstet Gynecol.* 1996;88:137-50.
22. Varney SJ, Guest JL. The annual cost of blood transfusions in the UK. *Transfus Med.* 2003;13:205-18.
23. Weisz B, Rosenbaum O, Chayen B, et al. Outcome of severely anaemic fetuses treated by intrauterine transfusions. *Arch Dis Child.* 2009;94:F201-4.

https://doi.org/10.1017/S0266462312000682 Published online by Cambridge University Press