The Evaluation and Expansion of Methodologies Relating to the Reporting and Analyses of Intermediate Test Results: Improving the Clinical Utility of Diagnostic Research

Bethany Shinkins

Wolfson College

Supervisors: Professor Rafael Perera and Dr. Richard Stevens

A thesis submitted for the degree of Doctor of Philosophy

Table of Contents

Acknowledgements

I have been incredibly fortunate over the past few years to work with such friendly and encouraging colleagues and I owe a great deal of gratitude to many members of the Department of Primary Care Health Sciences. First and foremost, I would like to thank my supervisors Professor Rafael Perera and Dr Richard Stevens. Their support, both from an academic and personal perspective, has been invaluable over the past three years. I would also particularly like to thank Professor David Mant and Dr Richard Mayon-White, whose clinical observations inspired the focus of this thesis. Dr Matthew Thompson and Dr Susan Mallett have been fantastic at providing advice on all things diagnosis-related.

I would also like to thank the National School for Primary Care Research for funding this work.

On a personal level, I would like to thank my friends and family who have provided necessary escape at times, but who have also been patient when I have had to prioritise work. My parents have been, as always, wonderfully supportive – thanks for your consistent belief in my abilities.

I dedicate this thesis to Sara, without whom completion of this thesis would not have been possible. You're the best.

Abstract

Background and objectives

It has been argued that the binary framework frequently adopted to analyse test accuracy does not represent the clinical reality of diagnostic practice, and the recognition of an intermediate category of test result could make the utility of diagnostic tests more transparent. The objective of this thesis is to explore the value of moving away from the binary framework when evaluating and interpreting quantitative diagnostic tests.

Methods

This thesis starts with an overview of the key arguments against dichotomising quantitative test results and a summary of some of the alternative methods proposed. Four distinct studies are then reported: 1) a systematic review of the methods currently used to evaluate the accuracy of quantitative cancer biomarkers, 2) a survey of GPs exploring preferences for threshold guidance 3) an evaluation of existing methods for identifying an intermediate range of test values, and 4) an assessment of the feasibility of applying these methods to the results of a meta-analysis.

Results

The binary framework remains the most common method for evaluating the accuracy of quantitative tests, despite the survey of GPs indicating that a single threshold interpretation is less helpful than identifying rule-in and rule-out thresholds. Existing methods for identifying an intermediate range of values require some adaptation to incorporate the cost trade-offs relating to different outcomes but, given complete reporting at the primary research, could be applied to the results of a meta-analysis.

Conclusion:

The 2 x 2 diagnostic framework frequently fails to capture many of the realities and complexities of clinical research questions. Standardised methods that facilitate complete reporting of test

accuracy in primary diagnostic accuracy studies are required to allow for greater flexibility when producing threshold recommendations further down the evidence pathway.

Chapter One

Thesis Introduction

1.1. Introduction

Diagnostic testing is the acquisition of clinical information for the purpose of characterising and refining the health status of a given patient (1). A diagnostic test can refer to anything from basic patient characteristics, symptoms or signs, clinical history, physical examination, laboratory tests or other technical tests such as imaging. The main objective of diagnostic accuracy research is to identify tools or features which accurately identify disease, ideally in combination with minimal cost to the patient and health care system.

1.1.1. The 2 x 2 Classification Table

In order to evaluate whether a test of interest (the index test) is accurate or not, its diagnoses need to be compared to the true disease status in a clinically relevant sample of individuals. Ideally a 'gold standard' test is available to provide this information e.g. autopsy, but in most scenarios there is no such test available and a 'reference standard' test, which may be subject to some error, has to be used instead.

A standard 2 x 2 classification matrix (also referred to as a contingency table or a confusion matrix) is commonly used to represent the four outcomes of a diagnostic accuracy study. The true disease status of the individuals, as defined by the gold or reference standard test, forms the two columns across the top of the table. The results of the index test are then typically categorised as either positive or negative to form the two rows down the side of the table (see Figure 1.1). To achieve

this for tests of an ordinal or continuous nature, their test scales must be dichotomised at a single 'optimal' threshold.

	Disease Status	
	Disease Present	**Disease Absent**
Positive Test Result	True Positives (TP)	False Positives (FP)
Negative Test Result	False Negatives (FN)	True Negatives (TN)

Figure 1.1. 2 x 2 classification matrix typically used for reporting the results of a diagnostic accuracy study

The terms 'true positive' and 'true negative' categories are used to describe 'diseased' and 'healthy' individuals who are correctly classified by the diagnostic test, respectively. A 'false negative' result refers to a diseased individual who is incorrectly classified as healthy, in contrast to a 'false positive' result where the disease is absent, but the diagnostic test misclassifies an individual as having the target disease. By summing the 'disease present' column and dividing it by the total number of patients in the table, the prevalence of disease in the sample can be estimated.

This common framework has resulted in the development of a number of different, albeit somewhat complementary, binary-based summary statistics to evaluate and compare the performance of different diagnostic tests.

1.1.2. Sensitivity, Specificity and Likelihood Ratios

Sensitivity and specificity are the probabilities of positive or negative test results given the presence or absence of disease, and are calculated by analysing across the columns of the contingency table (see Figure 1.2) (2). Yerushalmy was the first to adopt the commonly used nomenclatures to describe the accuracy of a diagnostic test, which are now recommended by the STARD statement (a guideline for improving reporting quality of diagnostic accuracy research) to be used as MESH terms to identify diagnostic accuracy studies (3, 4).

Sensitivity is the proportion of the target disease-positive population which are correctly classified i.e. the 'true positive rate'. By contrast, specificity is the proportion of the target disease-negative

population which are correctly classified i.e. the 'true negative rate'. Both measures must be available in order to get a complete picture of the accuracy of a given test (5).

	Disease Status	
	Disease Present	**Disease Absent**
Positive Test Result	True Positives (TP)	False Positives (FP)
Negative Test Result	False Negatives (FN)	True Negatives (TN)
Diagnostic Accuracy	Sensitivity TP/(TP+FN)	Specificity TN/(TN+FP)

Figure 1.2. The calculation of sensitivity and specificity from the standard 2x2 diagnostic contingency table

For a test with a dichotomous outcome, the likelihood ratio of a test can be calculated from its sensitivity and specificity [Eq. 1.1]. Its clinical interpretation is the likelihood of a given test result (positive or negative) in an individual with disease compared to the likelihood of that same result in an individual without disease.

$$\boldsymbol{LR+} = \frac{\boldsymbol{sensitivity}}{\boldsymbol{1-specificity}} \qquad\qquad \boldsymbol{LR-} = \frac{\boldsymbol{1-sensitivity}}{\boldsymbol{specificity}}$$

[Eq. 1.1]

According to Bayes theorem, the likelihood ratio of a test can be combined with the 'prior' or 'pre-test' probability of disease to calculate the adjusted probability of disease given the information provided by the test, known as the 'posterior' or 'post-test' probability. This approach acknowledges that few tests are accurate enough to 'rule in' or 'rule out' disease completely, and therefore test results should be understood as merely adjusting the probability of disease (6). The prior probability is dependent on the prevalence of disease in the target population, in addition to other diagnostic information collected prior to the test such as patient characteristics or symptoms. In order to use the likelihood ratio to adjust the probability of disease, the probability of disease must also be converted to the odds of disease [Eq. 1.2] (7). This nonlinearity can make likelihood

ratios hard to work with at times (8). For example a test with a likelihood ratio of 20 does not increase the probability of disease twice as much as a likelihood ratio of 10.

$$\textbf{\textit{Post-test odds}} = \textbf{\textit{Pre-test odds x Likelihood Ratio}}$$

[Eq. 1.2]

1.1.3. Positive and Negative Predictive Values

The predictive value of a diagnostic test quantify the probability that the test will give an accurate diagnosis, and can be calculated by analysing across the rows of the contingency table (see Figure 1.3) (9). The positive predictive value is the probability that a patient with a positive test result has the disease. It is calculated by dividing the number of true positive results by the total number of individuals with a positive test result i.e. the sum of the true positive and false positive results. By contrast, the negative predictive value is the probability that a patient with a negative test result is disease-free. It is calculated by dividing the number true negative results by the total number of individuals with a negative test result i.e. the sum of the true negative and false negative results.

	Disease Status		
	Disease Present	**Disease Absent**	**Predictive Value**
Positive Test Result	True Positives (TP)	False Positives (FP)	Positive Predictive Value TP/(TP+FP)
Negative Test Result	False Negatives (FN)	True Negatives (TN)	Negative Predictive Value TN/(FN+TN)

Figure 1.3. The calculation of positive and negative predictive values from the standard 2x2 diagnostic contingency table

Ideally, the evaluation of a diagnostic test would summarise its performance regardless of the population that the test is being studied in, allowing the results to be generalised to other clinical populations and settings (10, 11). In contrast to predictive values, Bayesian theorem implies that likelihood ratios, and consequently sensitivity and specificity, are constant properties of the test; independent of the prevalence of the population in which they were calculated (11). This, however, does not make these statistics immune to 'spectrum bias': the phenomenon that the patient mix

or spectrum of disease severity included in the patient sample can considerably influence the performance of a diagnostic test (12).

1.1.4. Intermediate Test Ranges

"The two zones demarcated by binary models are inadequate for the many clinical decisions that are trichotomous rather than dichotomous"

Feinstein, 1990

Over two decades ago, Alvan Feinstein, a leading figure in the development and evaluation of diagnostic research methods, argued that the binary 'positive–negative' framework commonly used in test accuracy research is not representative of diagnostic decision making in clinical practice (13). He claimed clinicians recognise that diagnostic tests are rarely capable of ruling in or ruling out disease and typically interpret test results as either 'positive', 'negative' or 'uncertain'. Instead of dichotomising quantitative results (results of tests that are on a continuous or ordinal scale), he proposed that two thresholds should be identified to allow for the recognition of an intermediate, 'uncertain', range of test values.

Diagnostic uncertainty is particularly rife in general practice and 'test of time' strategies are often implemented to handle patients for whom a confident diagnosis cannot be achieved (14). Alternative methodological frameworks that allow for diagnostic uncertainty to be better recognised, such as the one proposed by Feinstein, may therefore be particularly pertinent to the evaluation of tests for use in general practice.

The overarching objective of this thesis is therefore to explore this concept of an intermediate category of test result further, focusing specifically on the following objectives:

- To review existing arguments for and against the introduction of an intermediate range of test values

- To establish which methods are currently being used to analyse and report the accuracy of quantitative test results in contemporary diagnostic research
- To ascertain GP preferences regarding threshold guidance for the interpretation of quantitative diagnostic test results
- To evaluate novel methods for identifying intermediate test ranges in diagnostic accuracy research via applications to real clinical data

1.2. Outline of the thesis

Chapter 2: Background and Literature Review

This chapter starts by providing some background on the challenges specific to diagnosis in general practice. The adequacy of the binary model to summarise the accuracy of quantitative diagnostic tests is then discussed, followed by a review of some alternative methods proposed in the literature.

Chapter 3: Diagnostic Accuracy Research of Quantitative Cancer Biomarkers: A Methodological Systematic Review

A systematic review describing the methods used to analyse and report the accuracy of cancer biomarkers in contemporary diagnostic accuracy research is presented. Methodological limitations in both primary research and meta-analyses are discussed.

Chapter 4: Interpreting Quantitative Diagnostic Tests: A Survey of GPs

In this chapter, the results of an online survey of general practitioners exploring preferences for threshold guidance when interpreting quantitative diagnostic test results are reported. GPs were asked to compare a number of different interpretation formats to ascertain what depth of information would be most helpful to their clinical decision making.

Chapter 5: The Evaluation of Existing Methods for Defining an Intermediate Range of Values on a Quantitative Diagnostic Test Scale

An evaluation of existing statistical methods for identifying an intermediate test range is reported. Test results from three commonly used inflammatory markers for detecting serious bacterial infection in children were used as a case study to explore the strengths and limitations of each method.

Chapter 6: Deriving an Intermediate Range of Values from the Results of a Diagnostic Accuracy Meta-analysis

The feasibility of applying two methods for identifying an intermediate range (previously evaluated in **Chapter Five**) to the results of a diagnostic accuracy meta-analysis is explored.

1.3. References

1. Knottnerus JA. The Evidence Base of Clinical Diagnosis: BMJ Books; 2002.

2. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994 Jun 11;308(6943):1552.

3. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. Public Health Reports. 1947;62(40):1432.

4. Bossuyt PM, Reitsma JB, Standards for Reporting of Diagnostic A. The STARD initiative. 2003;361(2985213r, l0s, 0053266):71.

5. Begg CB. Biases in the assessment of diagnostic tests. Stat Med. 1987 Jun;6(4):411-23.

6. Bianchi MT, Alexander BM. Evidence based diagnosis: Does the language reflect the theory? Brit Med J. 2006 August 26;333(7565):442-5.

7. Perera R, Heneghan C, Badenoch D. Statistics toolkit: BMJ Books; 2008.

8. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet. 2005 Apr 23-29;365(9469):1500-5.

9. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994 Jul 9;309(6947):102.

10. Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Academic Radiology. 2003 Jun;10(6):670-2.

11. Coulthard MG. Quantifying how tests reduce diagnostic uncertainty. Archives of disease in childhood. 2007 May 11;92(5):404-8.

12. Ransohoff DFFAR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New England Journal of Medicine. 1978;299(17):926-30.

13. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. J Clin Epidemiol. 1990;43(1):109-13.

14. Heneghan C, Glasziou P, Thompson M, Rose P, Balla J, Lasserson D, et al. Diagnostic strategies used in primary care. BMJ. 2009;338.

Chapter Two

Background and Literature Review

2.1. Overview

This chapter consists of a review of the literature on the adequacy of the binary model to summarise the accuracy of quantitative diagnostic tests. The specific topics discussed are:

- Medical diagnosis and diagnostic uncertainty in general practice
- Accounting for diagnostic uncertainty in test accuracy research
- Reporting and analysing inconclusive test results
- The downfalls of dichotomising quantitative diagnostic test scales
- Alternative methods for analysing and interpreting quantitative diagnostic test results

2.2. Background

2.2.1. Medical Diagnosis

The factual contrary of the 'healthy' and the 'pathological', the underpinnings of the practice of medical diagnosis, infers the possibility to define each of them unambiguously (1). Diagnosis forms an essential part of routine patient evaluation, yet despite its black-and-white theoretical foundations, the clinical reality is often one of unavoidable uncertainty and error (2). The reductionist perspective of diagnosis attributes all medical uncertainty to gaps in our scientific knowledge, and claims that with greater study of the complex nature of disease, all uncertainty can eventually be eliminated (3). Diagnostic uncertainty is therefore assumed to be a product of the measurement or the test itself, and it is the quantification of this uncertainty which defines the primary objective of diagnostic accuracy research.

Traditionally, diagnostic testing is described as the acquisition of clinical information for the purpose of characterising and refining the health status of a given patient (4). In recent decades however, clinical practice has been transformed by the proliferation of new technologies made available to aid the diagnosis of patients (4). The utility of diagnostic tests is no longer limited to the traditional diagnostic task; tests are frequently implemented for other clinical purposes such as screening, treatment and dose selection, monitoring, and prognosis. As evidence-based medicine has grown in popularity and reliance on these technologies increases, there has been mounting pressure for high quality research to help clinicians understand the strengths and weaknesses of the myriad of test options now offered (5).

2.2.2. Diagnostic Uncertainty in General Practice

Diagnosis has been described as the 'Achilles heel' of general practice (6). Failures relating to diagnosis account for nearly one third of complaints in general practice (7), in addition to being the lead driver of medical errors (8).

A systematic review of research relating to diagnostic errors and delays in general practice revealed some intrinsic characteristics of the clinical setting which heighten the diagnostic challenge (9). For the majority of non-acute patients, the General Practitioner tends to be the first point of interaction with the healthcare system. This leads to the evaluation of patients who are typically in the early stages of disease, when many key diagnostic features are yet to develop. As a consequence a high proportion of patients present with generic signs and symptoms, making it difficult to distinguish between multiple possible diagnoses. Furthermore, general practice oversees a population where the probability of serious disease is typically much lower compared to more specialised, secondary clinical settings (9). Testing patients when they are in the preliminary stages of disease or when the risk of disease is minimal significantly reduces the predictive value of test results (10), further limiting the potential for a confident diagnosis.

A common example of a typical, yet challenging, patient presentation is the diagnosis of serious bacterial infection in children. The prior probability of serious infection in children is around 1% in the UK general practice (11) and patients usually present with very generic symptoms that would be typical of a self-limiting, minor infection. General Practitioners are faced with the task of ensuring that cases of serious infection are not missed, without overburdening the health system and patient with unnecessary referrals. Risk quantification in this scenario has shown evidence of a gap between the maximum levels of diagnostic certainty achievable at presentation and the minimum certainty required for further investigation or referral (12, 13).

The described host of obstacles, together with the limited range of diagnostic tests readily available to GPs, give rise to heightened diagnostic uncertainty in general practice. There are, however, a number of 'test of time' strategies which allow GPs to delay making an immediate diagnosis in many situations and deal with diagnostic uncertainty more effectively (14). Diagnostic safety-netting is one such example, where the patient is advised under what circumstances they should seek further medical advice should their condition deteriorate or fail to improve (15). General

practice may therefore be the ideal platform for exploring the topic of diagnostic uncertainty further (16).

2.3. Accounting for Diagnostic Uncertainty in Test Accuracy Research

A series of letters among experts in the field were published in the Journal of Clinical Epidemiology highlighting the inadequacy of current research methods to account for the uncertainty inherent in diagnostic decision making (17). The clinical reality of diagnostic practice is that tests are often imperfect and struggle to discriminate between patients with and without a given disease, leaving a subset for which certainty about disease status cannot be established. It has been argued that the key shortcoming of the frequently adopted dichotomous framework is that it fails to allow for an 'uncertain' diagnostic category, an outcome which is frequently unavoidable in clinical practice.

2.3.1. The 3 x 2 Classification Matrix

A 'three-zone pattern of clinical reasoning' is proposed, whereby the state of disease can be categorised as either 'present', 'uncertain' or 'absent' via a trichotomous partition of the test scale (17). If a test value falls within the 'uncertain' category, then further investigation is required, hence the role of the test becomes contributory to the diagnostic process rather than definitive. Simel and colleagues agree with this reasoning, stating that clinical decision making is often influenced by results which are neither positive nor negative (18). To accommodate this extra category of outcome when reporting the results of a diagnostic accuracy study, it is suggested that the commonly adopted 2 x 2 contingency table is extended to a 3 x 2 matrix (see Figure 2.1).

TEST RESULT	DISEASE STATUS	
	Present	**Absent**
Positive	True Positive	False Positive
Inconclusive	Further Investigation Required	
Negative	False Negative	True Negative

Figure 2.1. 3 x 2 classification matrix, allowing for the reporting of inconclusive results

2.3.2. Inconclusive Test Results

Simel et al. outline three types of inconclusive test result (18):

- Uninterpretable results are those that 'do not meet the minimum criteria constituting an adequate test'
- Intermediate test results are those that 'confer a likelihood ratio for disease that is more than that conferred by a negative result, but less than that of a positive test'
- Indeterminate test results are those that add no additional diagnostic information to the original probability of disease. In technical terms, these are test results with a likelihood ratio approximately equal to 1, meaning that knowledge of the test result does not alter the probability of disease

When these terms were incorporated into a search strategy, it became apparent that the distinction between these different types of non-positive, non-negative result is often lost, and the terms are used interchangeably in the literature. QUADAS, a tool for assessing the quality of diagnostic accuracy studies included in systematic reviews, asks reviewers to check that 'uninterpretable, indeterminate or intermediate test results' are reported (19). In a recent evaluation of the QUADAS checklist, authors of Cochrane reviews reported that this recommendation was difficult to apply due to confusion about the applicability to particular diagnostic tests (20).

On the back of these findings, I wanted to explore whether inconclusive results are consistently reported in diagnostic accuracy research. The STARD statement, a list of items encouraging high quality reporting of diagnostic accuracy studies, recommends that authors 'report how indeterminate results, missing responses and outliers of the index tests are handled' (21). The full STARD guidance expands on this statement, advocating the reporting of 'uninterpretable, indeterminate, and intermediate results' (item 22).

I identified systematic reviews which evaluate the quality of reporting via adherence to the STARD statement and calculated the proportion of primary studies that complied with this item. I searched Medline [1946 to 2012] and Embase [1974 to 2012] via OvidSP using the term 'STARD' [All fields]. The inclusion criteria used to select relevant studies was as follows:

1) systematic review of diagnostic accuracy studies,
2) reported their inclusion criteria for study selection
3) authors assessed adherence to item 22 of the STARD statement and reported the number of studies compliant

After removing duplicates, 598 items were returned and the titles and abstracts of each were read to identify articles that met the inclusion criteria. Reviews were included in the overview regardless of the publication date. The following pieces of information were extracted from each review:

- The inclusion criteria used to select individual diagnostic accuracy studies
- The number of studies included in the review
- The publication period of the individual studies
- The percentage of studies that complied with item 22 of the STARD statement

As some papers failed to provide the references for the studies included in their review, it was not feasible to rule out the possibility that a given diagnostic accuracy study featured in more than one

systematic review. One review was completely excluded because all of the studies included featured in a later review.

A full breakdown of the results for each review can be found in the appendix of this chapter. Based on 1156 primary studies included in 22 systematic reviews (published between 2005 and 2011), only one third (n=400, 35%) of studies reported the presence or absence of inconclusive results adequately (22). This finding pointed towards the need to clarify the importance of inconclusive results and provide clearer guidance to researchers on how to handle these results in the reporting and analysis of diagnostic accuracy studies. Based on the findings of this review, I developed some recommendations for the transparent reporting of inconclusive test results and reviewed some of the methods that are commonly used to analyse them. To facilitate this, I proposed a more pragmatic approach to defining inconclusive results, distinguishing between those that are valid (that is, where an adequate test result has been obtained, but the result is not clearly positive or negative) and those that are invalid (that is, where the key diagnostic feature is uninterpretable or the actual result is missing). This work, in combination with the overview of STARD systematic reviews, was published in the BMJ's 'Methods and Reporting' section and the final manuscript can be found in the publications section at the end of this thesis.

The way in which valid inconclusive results should be reported and analysed is dependent on the measurement scale of the test i.e. whether it is continuous, ordinal or categorical in nature. For quantitative tests in particular, the definition of a 'valid inconclusive test result' must be determined by the analyst. For example, valid inconclusive results for a test on a continuous scale (which I will now refer to as 'intermediate' test results in the rest of this thesis) typically lie in the range of values where the typical distributions of test results for 'disease present' and 'disease absent' individuals overlap. The degree to which they encompass this region of uncertainty, however, depends on the relative implications of false positive and false negative test outcomes. In this instance, the placement and number of thresholds is a key methodological consideration.

2.3.3. Distributional Assumptions

The distributional assumptions underlying the diagnostic task are an essential consideration when exploring appropriate methodologies for evaluating the accuracy of a test. Diagnostic theory is founded on the orthodox healthy–pathological dichotomy (two sub-populations); however this assumption does not hold for all diagnostic scenarios. As a result of greater understanding of the etiology and evolution of disease, there are clinical conditions where multiple stages of disease progression have been defined to allow for more targeted therapy. Cancer is a classic example of this: distinct sub-populations of cancer patients have been defined based on the size and spread of disease. In these scenarios, it is assumed that the distribution of test results for each of these sub-populations will present with a distinct mean and standard deviation (23).

The recognition of a transitional state of disease, often referred to as 'pre-disease', is becoming increasingly common as screening methods and preventative therapies improve. Viera outlined three conditions where the recognition of pre-disease is beneficial: 1) those diagnosed with pre-disease must be significantly more likely to develop disease than those in whom the disease has been ruled out, 2) there must be a targeted intervention that reduces the likelihood of developing disease, and 3) the benefits of intervention in this sub-population must outweigh the harms (24). A 3 x 3 classification matrix has been proposed for reporting the accuracy of test results in this scenario (25), in addition to a number of different methods for analysing the diagnostic accuracy of a test (26-29).

The described addition of further disease states is a different methodological challenge to that outlined by Feinstein and colleagues. The easiest way to distinguish between the two scenarios is to lay out both problems in terms of the diagnostic classification matrix. The recognition of additional disease states extends the number of columns in the classification matrix e.g. a 2 x 3 matrix. In contrast, the recognition of additional 'inconclusive' categories of test result requires extra rows e.g. a 3 x 2 matrix. In this latter case, the assumption of two underlying patient

distributions still holds, and the addition of a third 'uncertain' category accounts for 'results which lie in the middle between two relative certainties' (17). Skaltsa and colleagues confused these two scenarios when proposing a novel method for identifying decision thresholds in the face of multiple disease states (30). I used this as an opportunity to clarify the distinction between the two scenarios, and published a letter to the Editor of Statistics in Medicine highlighting the theoretical differences (23). The final manuscript can be found in the publications section of this thesis.

In a later letter, Feinstein does entertain the idea of a 3 x 3 framework when acknowledging that 'the test results and the definitive diagnosis will often be uncertain, inconclusive, or indeterminate' (31). In this context however, the addition of an 'uncertain' disease category accounts for the patients where the reference standard is incapable of producing a definitive diagnosis, not a disease state itself. A Bayesian approach for evaluating the performance of a test in this scenario has been suggested (32).

A further by-product of improvements in preventative medicine is the recognition of risk factors for future complication as diseases themselves. Some examples of this include hypertension, anaemia, and diabetes. Diagnostic tests are used as surrogates for defining these diseases e.g. a patient with blood pressure consistently over 140/90 mmHg is regarded as hypertensive (33). For these diseases, there is only one underlying distribution of test results and patients fall somewhere on this measurable continuum of disease. From a theoretical perspective, this presents as a slightly different diagnostic challenge since the task is to place patients accurately on that continuum, rather than determine whether they belong to one of two or more distinct disease populations.

The rest of this literature review will focus specifically on methods to analyse and interpret quantitative test results for the two-class diagnostic scenario.

2.4. Dichotomising Quantitative Diagnostic Tests

The dichotomisation of quantitative test scales has become commonplace to mirror the underlying classification problem. An 'optimal' single decision threshold is typically identified to split test results into two groups: those that provide evidence of the presence of disease and those that provide evidence against the presence of disease. Tests are rarely perfectly accurate however, and typically the distributions of test values for each disease state overlap, such as those in Figure 2.2. Consequently, the selection of an optimum single threshold involves some degree of trade-off to reach an acceptable balance of false positive and false negative test results.

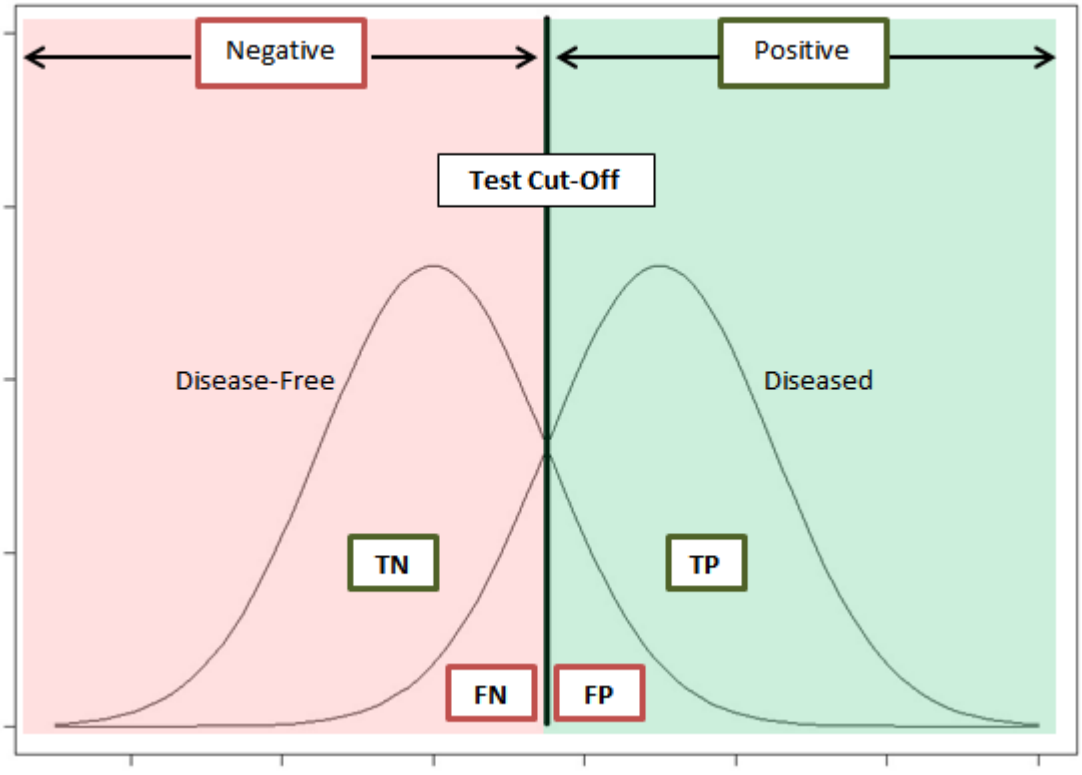


Figure 2.2. Trading off misclassifications when selecting a single diagnostic threshold. TP = 'true positive', FP= 'false positive', TN = 'true negative' and FN = 'false negative'

A wealth of research has been dedicated to establishing and improving the statistical methods used to select an 'optimal' single threshold to facilitate the dichotomisation of an ordinal or continuous test scale, and this topic could be the sole focus of an extensive review in itself.

One commonly used tool is Receiver Operating Characteristic (ROC) curve analysis, a technique originally founded on signal detection theory and used by engineers in World War II to evaluate the accuracy of radar detection (34). Every test value is taken as a potential diagnostic threshold for which sensitivity (the true-positive rate) and specificity (the true-negative rate) are calculated. The sensitivity for each threshold is then plotted along the y-axis against the false-positive rate on the x-axis (1-specificity) to form the ROC curve (Figure 2.3). Although this graph was originally proposed as a threshold-independent analytic tool, it is now frequently used to identify the point on the test scale that provides the optimal trade-off between sensitivity and specificity. The test value on the curve which is closest to the top left-hand corner of the ROC space, or that provides the maximal combined sensitivity and specificity (Youden's Index) is typically extracted as the optimal decision threshold (35).

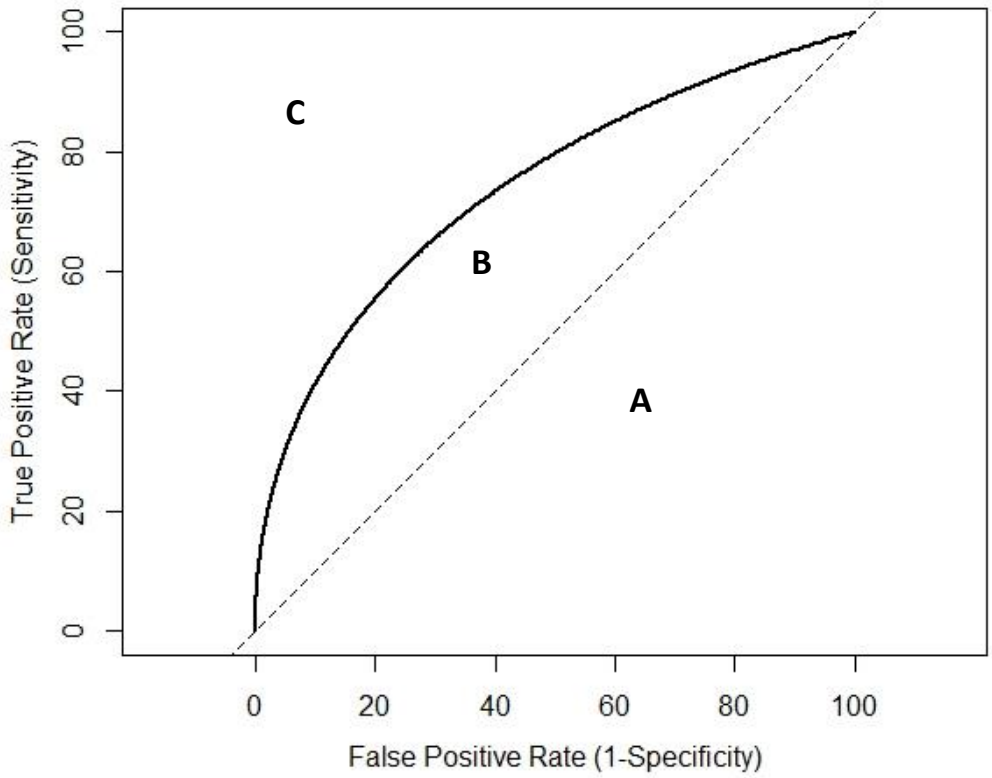


Figure 2.3. Example of a Receiver Operating Characteristic curve. The dashed line (A) depicts an uninformative test i.e. it is no more accurate than chance. The ROC curve (B) illustrates how the true-positive rate varies as the false-positive rate increases. The closer the ROC curve is to the top left-hand corner (C), the more accurate the test.

Although attractive in its simplicity, a key shortfall of basic ROC analysis for the selection of decision thresholds is that it assumes that the costs associated with each type of misclassification is equal (34). The clinical reality, however, is that the cost of a false negative result is likely to be different to that of a false positive result. Comprehensive methods have however been developed for incorporating a cost function into ROC analysis to account for the clinical settings where the consequences of different types of misclassification are unequal e.g. (36, 37).

There is extensive literature describing the problems and costs of dichotomising continuous measures e.g. (38-40). The key issue is that patients with results just either side of the threshold will be characterised illogically as very different to one another (38). These results are therefore more likely to be misclassified (41), thus ignoring the increased uncertainty relating to these results and impacting the overall perceived accuracy of the test.

2.5. Identifying an Intermediate Range of Test Values

The identification of an intermediate range of values has been argued to be more representative of how clinicians interpret quantitative test results in practice (17, 42, 43). Santuz and colleagues state that clinicians rely on an 'unconscious Bayesian reasoning' which leads to the differential weighting of low, mid-range, and high test results (42). Simel et al. also subscribe to this theory, insisting that clinicians use a trichotomous framework, rather than thinking in dichotomous or polychotomous dimensions (43).

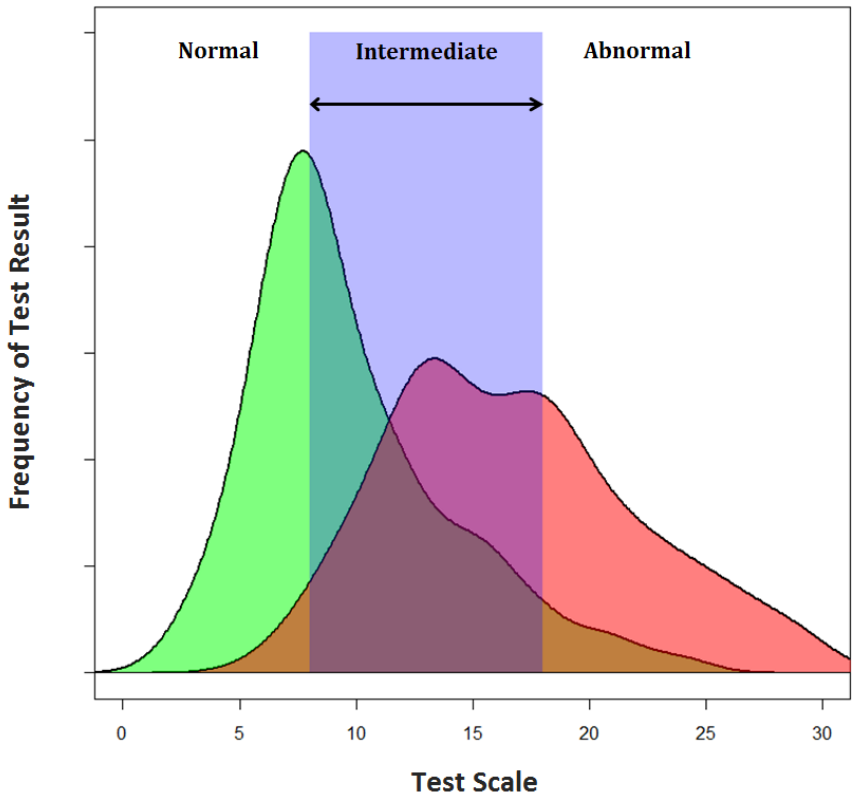


Figure 2.4. The identification of an intermediate range of test values for a quantitative diagnostic test

The addition of a third 'intermediate' category of result when interpreting a quantitative test scale necessitates the selection of two thresholds; a task either requiring the adaptation of current methodologies implemented to select a single threshold, or the development of novel approaches. Assuming a higher test value implies greater risk of disease, results beyond the upper threshold would be interpreted as 'positive' and those below the lower threshold would be 'negative'. Those results in between would fall into the 'intermediate' test category, in most cases highlighting the need for further investigation (Figure 2.4).

2.5.1. Methods for Identifying an Intermediate Range of Test Values

Greiner proposed the 'Two-Graph Receiver Operating Characteristic' method to delimit an intermediate range. Its similarities to standard ROC methodology lie in its reliance on calculating sensitivity and specificity for every point on the test scale (44, 45). In contrast, the next step is to plot both the sensitivity and the specificity (on the y-axis) against all of the values on the test scale

(on the x-axis). The two thresholds are then identified by finding the test values at which 90% (or 95%) sensitivity and specificity is achieved (see Figure 2.5).

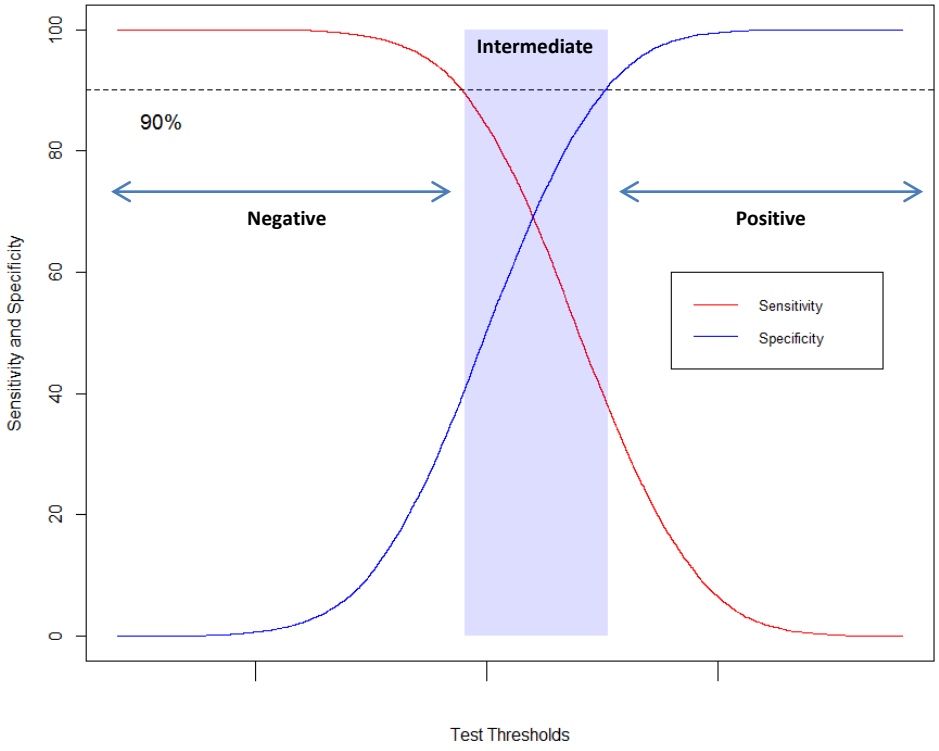


Figure 2.5. Example of a TG-ROC curve. The test values that achieve 90% sensitivity and 90% specificity form the lower and upper limits of the intermediate range, respectively.

Coste and Pouchot propose an alternative method for identifying an intermediate range of test values, described as a 'grey zone' (46). The first step of the methodology involves the estimation of the pre-test probability of disease and the post-test probabilities required to rule in and rule out the target condition. The minimal positive likelihood ratio required to rule in a positive diagnosis, and the maximal negative likelihood ratio required to rule out a positive diagnosis can then be calculated using conditional probability theory. The second step involves the calculation of the positive and negative likelihood ratio at each test value, which is then plotted as curves against the test scale (Figure 2.6). The point at which a test value achieves the required minimal positive likelihood ratio is taken as the optimal lower limit of the intermediate range, and the point at which the maximal negative likelihood ratio is achieved becomes the optimal higher limit.

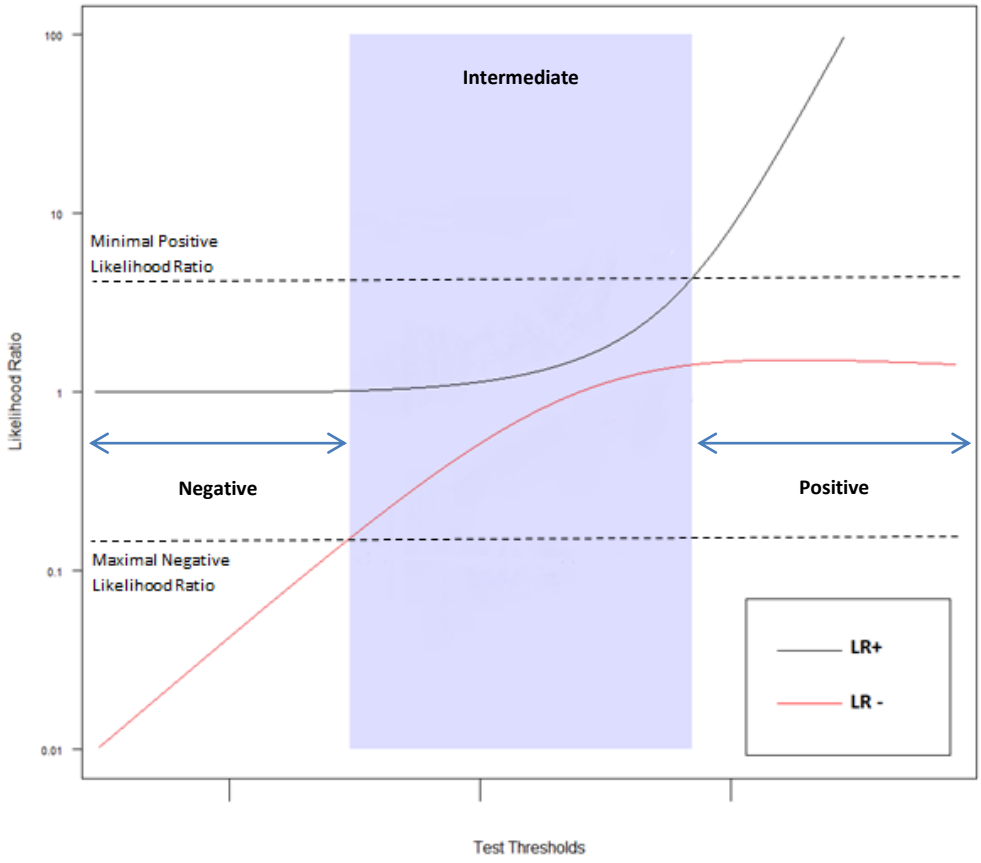


Figure 2.6. Example of selecting limits of an intermediate range based on the likelihood ratio curves, as recommended for the Grey Zone method

The simplistic approach of this method is however criticised as failing to take into account or recognise the many complex issues which are faced when this concept is actually applied to individual patients in clinical practice (47). The estimation of the pre-test probability of disease for a given patient is one such complication. For example, there is a paucity of evidence regarding the prevalence of diseases presenting in general practice, in addition to the diagnostic value of patient history, symptoms and clinical examination (48). It is proposed that the initial step of the method, the estimation of the pre-test probability and the selection of acceptable post-test probabilities, could be avoided by working directly with likelihood ratios, a prevalence-independent measure (47).

Ripley describes the general classification problem from a pattern recognition perspective and describes a model which is allowed to 'reject' records that cannot be confidently assigned to a

particular class (49), thus reducing the proportion of misclassification errors. The importance of assigning costs to rejected records is emphasised, as the implication is that these records would be subject to a 'more expensive, second tier of classification'. It is argued that the implementation of a reject option is advantageous for those scenarios where the cost of delaying or refusing to provide a decision is less than the cost of making an incorrect decision (50).

Tortorella outlines the theory for the dichotomous classification task by introducing a reject rule based on information from the ROC curve (51). The method is dependent on the estimation of costs for the four classification outcomes, in addition to the costs of rejecting a record. Two thresholds are selected to delimit the 'reject region' where the class distributions overlap along the posterior probability scale. The purpose of the reject region is to minimise the number of errors, but will inevitably encompass some correct classifications to prevent it from spanning the majority of the probability scale. Although the terminology is different, the reject region is akin to defining an intermediate range of values on a diagnostic test scale. The key difference with this method compared to those previously described is that it relies on evidence beyond the diagnostic accuracy study. Specifically, it requires the *quantification* of the costs and benefits associated with each diagnostic outcome. Irwig and colleagues state that it should be the post-test probability scale that should be trichotomised rather than the test scale, as it is the medical decision that is trichotomous in nature (52). Decision threshold theory describing the test and treatment thresholds that clinicians commonly implement is used to substantiate this argument, advocating a three-zone categorisation of the probability of disease. Thus, Ripley's pattern recognition approach could serve as a possible solution to this proposal. Furthermore, manipulation of the probability scale allows additional diagnostic factors to be included in the prediction model. Greiner's method only addresses the issue of identifying an intermediate range for a single test (44, 45). Coste and Pouchot, however, account for the multivariate scenario by stating that other diagnostic information available prior to carrying out the test should be incorporated into the pre-test

probability estimate (46). This approach, however, would not account for correlations that would likely exist among the multiple predictors.

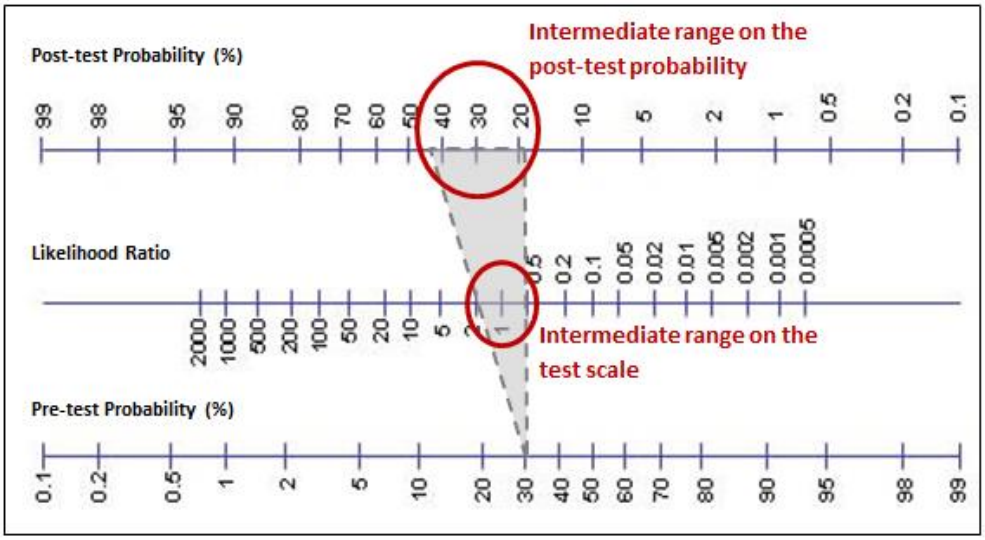


Figure 2.7. Visualising the distinction between trichotomising the test scale and trichotomising the post-test probability scale on a nomogram

2.5.2. Analysing Diagnostic Accuracy for Multiple Categories of Test Result

One of the issues with extending the 2 x 2 matrix to include a third category of test result is that many key diagnostic accuracy indices rely on a binary outcome (43). Feinstein believes that clinicians' reluctance to understand and implement common diagnostic indices in practice is because they are clinically unrealistic as they fail to recognise diagnostic uncertainty (17). As a consequence, he suggests that binary-based measures such as sensitivity and specificity are abandoned completely, and that researchers should focus on developing new analytic techniques which accommodate a trichotomous outcome.

For the three-category scenario, the proportion of patients in the intermediate range has been proposed as an effective means of describing the usefulness of a quantitative diagnostic test in practice (46). An alternative suggestion put forward is the 'valid proportion range'; the proportion of the test scale inside the intermediate range (44). This latter suggestion may be interesting to

researchers, but is unlikely to be of any clinical relevance, whereas the proportion of patients who receive an 'uncertain' outcome and require additional attention is of high clinical significance.

The likelihood ratio, typically used to summarise the relative odds of having the disease and obtaining either a positive or negative test result, can also be applied when three or more categories of test result are identified. In this context, they are typically referred to as 'multi-level' or 'stratum-specific' likelihood ratios and have been advocated as an effective means of navigating large zones of uncertainty (53-55). Multi-level likelihood ratios facilitate a richer level of interpretation, and also highlight whether a test performs asymmetrically i.e. whether the test is more powerful as a rule-in or rule-out test (56).

When evaluating the accuracy of aldosterone/PRA ratio as a screening test for aldosterone-producing adenoma, the authors set out to identify a single decision threshold but found that there was not a single test value that provided a high sensitivity alongside adequate specificity (56). Rule-in and rule-out thresholds were identified instead and category-specific likelihood ratios were calculated. Clear clinical recommendations were provided for each category of patient, with further testing using CT scans being proposed for those with intermediate test results. Santuz and colleagues also identified three categories of test result when evaluating the accuracy of procalcitonin for the diagnosis of early-onset neonatal sepsis and again used likelihood ratios to summarise test accuracy (42). In this example, the identification of an intermediate range of values helped to highlight the large proportion of patients for which procalcitonin was incapable of detecting the presence or absence of infection.

In keeping with Feinstein's proposal of an 'uncertain' category of test result (17), the clinical examples of an intermediate range identified in this review tended to include results that offer minimal adjustment of the probability of disease. A more formal definition of test values with likelihood ratios between 0.5 and 2 has been proposed (57-59), although an example where this definition has actually been used to identify the limits of an intermediate range was not found.

Despite strongly advocating the trichotomisation of quantitative test scales, Simel and colleagues concede that the addition of a greater number of categories may be useful in some instances. They contend that a multiple-category approach (i.e. more than three categories) does not invalidate their proposal of a trichotomous framework; the additional categories would simply represent 'layers' of the intermediate test range (43). Battaglia and Pewsner refer to this concept as 'shades of grey' and argue that this deconstruction of the intermediate range would facilitate a richer interpretation of the test result (47).

A multiple-category approach was used to evaluate the accuracy of various quantitative diagnostic tests for predicting exudative pleural effusions (60). Categories were defined by splitting the test scales at fixed intervals, with as many as 12 categories being identified for one of the tests. The authors strongly advocate the greater diagnostic precision achievable by identifying such a large number of categories of test results, but also describe some of the issues they faced. The stratification of the test scale into lots of categories led to small patient sizes in each category, and thus point estimates of category-specific likelihood ratios had very wide confidence intervals. In some cases, this led to non-monotonic likelihood ratios across ordinal categories of test results.

Generally, there appears to be little consistency or justification for the number of categories chosen or how thresholds have been selected in studies where multi-level likelihood ratios have been calculated. In some instances, test scales have been broken down using fixed, regular intervals e.g. (60, 61), and in others, the selection of thresholds has been driven by the ability of the test to rule in and rule out disease e.g. (56). Sonis recommends that the number of categories should be sufficiently large to offer a wide range of likelihood ratios, without compromising the stability of the likelihood ratio estimate for each category (54). An alternative proposal is to start by defining lots of categories and then collapse them until the 95% confidence intervals no longer overlap (62). For both of these methods, the number of categories would be largely driven by the sample size of the dataset being analysed.

2.6. Information Loss

Any categorisation of a quantitative test scale will result in the loss of some diagnostic information, as the probability of disease consequently has to be averaged over a range of post-test probabilities. For the interpretation of results from a single quantitative diagnostic test, however, it is generally acknowledged that some degree of categorisation is usually helpful. Information theory, a mathematical approach to quantifying information and uncertainty, has been used to explore the impact of categorising quantitative diagnostic tests to estimate the relative loss of information incurred.

By identifying three categories of test result instead of the commonly used dichotomy, it was found that 20-30% more of the information content of the test could be preserved (63). Further information is gained by allowing for more categories, although the benefit becomes insignificant once 5-7 categories of test result have been identified. Heckerling reviewed diagnostic accuracy studies published over a 5-year period in four key journals and evaluated the effect of dichotomisation on the information content of the tests (64). Sixty-seven studies were reported in sufficient detail to allow classification of test results into both dichotomous and multiple categories. It was found that the identification of multiple categories (ranging from 3-10) allowed a median of 14% more information to be retained, with a maximum increase of 109%. This wide variation in information gain is due to its dependence on the pre-test probability of disease and the discriminatory performance of the test. Information loss is most notable for tests of limited predictive value and/or scenarios with a low prior probability of disease (63). These characteristics are typical of general practice, suggesting that a move away from dichotomisation may be particularly beneficial in this clinical setting.

Simel et al. argue that the decision to adopt a dichotomous, ordinal, or continuous approach when evaluating a quantitative test scale should ultimately be a clinical decision. There is a direct trade-off between convenience and precision, and if there are specific ranges of test result that map

directly to certain patient management decisions, then an ordinal interpretation may be more pragmatic (65).

I was invited to write an editorial piece for the British Journal of General Practice regarding the downfalls of dichotomising quantitative test scales, with specific reference to handling diagnostic uncertainty in general practice. The final manuscript can be found in the publications section.

2.7. Interpreting 'Result-Specific' Accuracy Summaries

The need to categorise a quantitative test scale at all has been questioned in the literature when 'result-specific' accuracy summaries can be produced, avoiding the issue of information loss entirely.

Simel et al. provide the theory for modelling a simple continuous likelihood ratio function using logistic regression (65). This method facilitates the estimation of a likelihood ratio, alongside 95% confidence intervals, for every value on a quantitative test scale. A graphical format, the 'POD' plot (Figure 2.8), has also been proposed to facilitate the adjustment of the probability of disease for a given likelihood ratio (66). The model for deriving the result-specific likelihood ratio would ideally be embedded in a computer program which would automatically produce the POD plot, avoiding the need for the clinician to do any calculations by hand.

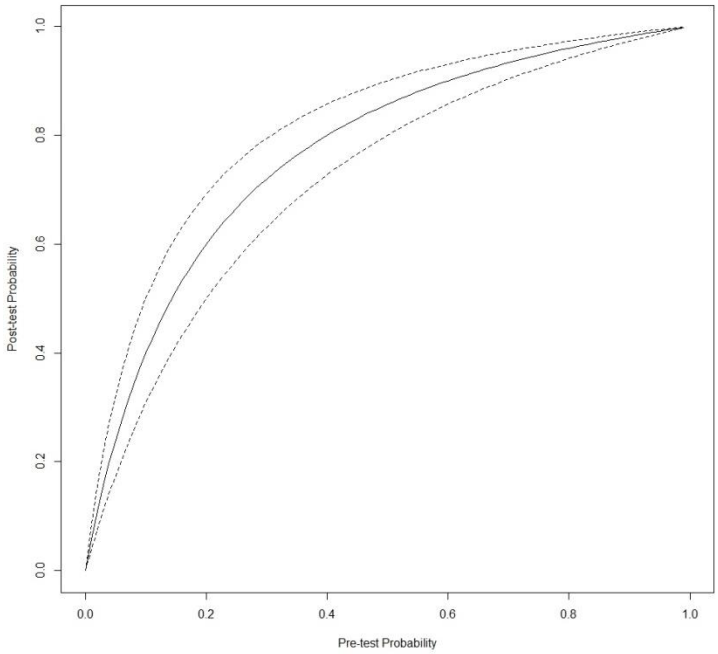


Figure 2.8. Example of a POD plot where the pre-test probability of disease can be adjusted for a specific test result (with 95% confidence intervals shown as dashed lines)

Heffner and colleagues evaluated the accuracy of multiple quantitative diagnostic tests for detecting pleural fluid exudates using binary, multi-level, and continuous likelihood ratios (67). The greater precision and richer interpretation provided by continuous likelihood ratios was illustrated, and this approach is strongly advocated by the authors as being the optimal approach for evaluating the accuracy of quantitative diagnostic tests. The results of this study are externally validated by a different research group using an alternative dataset, although the added value that continuous likelihood ratios provide in terms of improving categorisation is questioned (68). They contend that this Bayesian approach to interpreting test results is simply an explicit way of representing what clinicians already do implicitly, thus minimising the benefit of introducing a more complicated model.

2.8. Summary

Current methods in diagnostic research centre around the 2 x 2 classification table, but the evidence in this review demonstrates how this model is a waste of valuable diagnostic information

for tests on a quantitative test scale. The identification of an intermediate test range is argued to be more representative of how clinicians interpret test results in practice, and facilitates the explicit recognition of diagnostic uncertainty. Alternatively, multiple-category or result-specific analyses would facilitate the extraction of a richer depth of information, prioritising precision over simplicity. The helpfulness of these alternative methods to clinicians, however, is as an area that still needs to be explored.

Diagnostic uncertainty is an unavoidable feature of general practice, but it is also unique in that there are patient management strategies available for handling it. Revisiting the methods used to analyse and interpret test results to allow for the explicit recognition of diagnostic uncertainty will help to maximise the clinical utility of diagnostic accuracy research in general practice.

2.9. References

1. Lemoine M. The meaning of the opposition between the healthy and the pathological and its consequences. Med Health Care Phil. 2009 Aug;12(3):355-62.

2. Brenner H. How independent are multiple 'independent' diagnostic classifications? Statistics in Medicine. 1996;15(13):1377-86.

3. Quill TE, Suchman AL. Uncertainty and control: learning to live with medicine's limitations. Humane medicine. 1993 Apr;9(2):109-20.

4. Knottnerus JA. The Evidence Base of Clinical Diagnosis: BMJ Books; 2002.

5. Reid MC, Lachs MS, Feinstein AR. Use of Methodological Standards in Diagnostic Test Research: Getting Better but Still Not Good. JAMA. 1995;274(8):645-51.

6. Howie JG. Diagnosis--the Achilles heel? The Journal of the Royal College of General Practitioners. 1972 May;22(118):310-5.

7. Summerton N. Diagnosis and general practice. The British journal of general practice : the journal of the Royal College of General Practitioners. 2000 Dec;50(461):995-1000.

8. Sandars J, Esmail A. The frequency and nature of medical error in primary care: understanding the diversity across studies. Fam Pract. 2003 Jun;20(3):231-6.

9. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care--a systematic review. Fam Pract. 2008 Dec;25(6):400-13.

10. Knottnerus JA. Interpretation of diagnostic data: an unexplored field in general practice. The Journal of the Royal College of General Practitioners. 1985 Jun;35(275):270-4.

11. Thompson MJ, Van den Bruel A. Diagnosing serious bacterial infection in young febrile children. BMJ (Clinical research ed). 2010 April 20;340:c2062.

12. Van den Bruel A, Haj-Hassan T, Thompson M, Buntinx F, Mant D, European Research Network on Recognising Serious Infection i. Diagnostic value of clinical features at presentation to identify serious infection in children in developed countries: a systematic review. Lancet. 2010 Mar 6;375(9717):834-45.

13. Buntinx F, Mant D, Van den Bruel A, Donner-Banzhof N, Dinant GJ. Dealing with low-incidence serious diseases in general practice. The British journal of general practice : the journal of the Royal College of General Practitioners. 2011 Jan;61(582):43-6.

14. Heneghan C, Glasziou P, Thompson M, Rose P, Balla J, Lasserson D, et al. Diagnostic strategies used in primary care. BMJ. 2009;338.

15. Almond S, Mant D, Thompson M. Diagnostic safety-netting. The British journal of general practice : the journal of the Royal College of General Practitioners. 2009 Nov;59(568):872-4; discussion 4.

16. Green C, Holden J. Diagnostic uncertainty in general practice. A unique opportunity for research? The European journal of general practice. 2003 Mar;9(1):13-5.

17. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. J Clin Epidemiol. 1990;43(1):109-13.

18. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making. 1987 Apr-Jun;7(2):107-14.

19. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003 Nov 10;3:25.

20. Whiting PF, Bossuyt PMM, Sterne JAC, Deeks JJ, Reitsma H, Leeflang M, et al. Updating QUADAS: Evidence to inform the development of QUADAS-2. 2011.

21. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ. 2003 Jan 4;326(7379):41-4.

22. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013 2013-05-16 11:27:50;346:f2778.

23. Shinkins B, Stevens R, Perera R. 'Optimum threshold estimation based on cost function in a multistate diagnostic setting' by K. Skaltsa, L. Jover, D. Fuster and J. L. Carrasco. Statistics in Medicine. 2012 May 20;31(11-12):1110-2.

24. Viera AJ. Predisease: when does it make sense? Epidemiol Rev. 2011 Jul;33(1):122-34.

25. Sappenfield RW, Beeler MF, Catrou PG, Boudreau DA. Nine-cell diagnostic decision matrix. A model of the diagnostic process; a framework for evaluating diagnostic protocols. Am J Clin Pathol. 1981 Jun;75(6):769-72.

26. AlaviMajd H, Borumandnia N, Khadem Mabudi AA, Kariman N, Ardabili NS, Hajifathali A. Selection of two cut-off points via generalized Youden index and receiving operating characteristic surface to predict preeclampsia using the hemoglobin levels in the first trimester of pregnancy. Koomesh. 2012;14(3):321-6.

27. Dong T, Tian L, Hutson A, Xiong C. Parametric and non-parametric confidence intervals of the probability of identifying early disease stage given sensitivity to full disease and specificity with three ordinal diagnostic groups. Stat Med. 2011 Dec 30;30(30):3532-45.

28. Mossman D. Three-way ROCs. Med Decis Making. 1999 Jan-Mar;19(1):78-89.

29. Nakas CT, Alonzo TA, Yiannoutsos CT. Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. Statistics in Medicine. 2010 Dec 10;29(28):2946-55.

30. Skaltsa K, Jover L, Fuster D, Carrasco JL. Optimum threshold estimation based on cost function in a multistate diagnostic setting. Stat Med. 2012 May 20;31(11-12):1098-109.

31. Feinstein AR. Diagnostic-Tests Are Not Always Black or White - or, All That Glitters Is Not (a) Gold (Standard) - Response. Journal of Clinical Epidemiology. 1991;44(9):970-.

32. Matchar DB, Simel DL, Geweke JF, Feussner JR. A Bayesian Method for Evaluating Medical Test Operating Characteristics When Some Patients Conditions Fail to Be Diagnosed by the Reference-Standard. Medical Decision Making. 1990 Apr-Jun;10(2):102-12.

33. Hypertension: management of hypertension in adults in primary care. NICE guideline. 2011 August 00. Report No.

34. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Preventive veterinary medicine. 2000 May 30;45(1-2):23-41.

35. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006 Apr 1;163(7):670-5.

36. Metz CE. Basic principles of ROC analysis. Seminars in nuclear medicine. 1978 Oct;8(4):283-98.

37. Skaltsa K, Jover L, Carrasco JL. Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. Biometrical journal Biometrische Zeitschrift. 2010 Oct;52(5):676-97.

38. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006 May 6;332(7549):1080.

39. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006 Jan 15;25(1):127-41.

40. Streiner DL. Breaking up is hard to do: The heartbreak of dichotomizing continuous data. Can J Psychiat. 2002 Apr;47(3):262-6.

41. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997 May 15;16(9):981-91.

42. Santuz P, Soffiati M, Dorizzi RM, Benedetti M, Zaglia F, Biban P. Procalcitonin for the diagnosis of early-onset neonatal sepsis: a multilevel probabilistic approach. Clin Biochem. 2008 Oct;41(14-15):1150-5.

43. Simel DL, Matchar DB, Feussner JR. Diagnostic tests are not always black or white: or, all that glitters is not [a] gold [standard]. J Clin Epidemiol. 1991;44(9):967-70; discussion 70-1.

44. Greiner M, Sohr D, Gobel P. A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. J Immunol Methods. 1995 Sep 11;185(1):123-32.

45. Greiner M. Two-graph receiver operating characteristic (TG-ROC): a Microsoft-EXCEL template for the selection of cut-off values in diagnostic tests. J Immunol Methods. 1995 Sep 11;185(1):145-6.

46. Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. Int J Epidemiol. 2003 Apr;32(2):304-13.

47. Battaglia M, Pewsner D. Commentary: black and white or shades of grey? Int J Epidemiol. 2003 Apr;32(2):314-5.

48. McCowan C, Fahey T. Diagnosis and diagnostic testing in primary care. The British journal of general practice : the journal of the Royal College of General Practitioners. 2006 May;56(526):323-4.

49. Ripley BD. Pattern recognition and neural networks: Cambridge Univ Pr; 2008.

50. Ha TM. Optimum tradeoff between class-selective rejection error and average number of classes. Eng Appl Artif Intel. 1997 Dec;10(6):525-9.

51. Tortorella F. A ROC-based reject rule for dichotomizers. Pattern Recogn Lett. 2005 Jan 15;26(2):167-80.

52. Irwig L, Glasziou P. Trichotomous decisions do not imply trichotomous tests. J Clin Epidemiol. 1991;44(11):1279-80.

53. Youngstrom EA, Youngstrom JK. Evidence-based assessment of pediatric bipolar disorder, Part II: Incorporating information from behavior checklists. J Am Acad Child Adolesc Psychiatry. 2005 Aug;44(8):823-8.

54. Sonis J. How to use and interpret interval likelihood ratios. Family Medicine Journal. 1999;31(6):432-7.

55. Straus SE RW, Glasziou P, Haynes BR. Evidence-based medicine: how to practice and teach EBM: Elsevier/Churchill Livingstone; 2005.

56. Hirohara D, Nomura K, Okamoto T, Ujihara M, Takano K. Performance of the basal aldosterone to renin ratio and of the renin stimulation test by furosemide and upright posture in screening for aldosterone-producing adenoma in low renin hypertensives. J Clin Endocrinol Metab. 2001 Sep;86(9):4292-8.

57. Guyatt G, Bass E, Brill-Edwards P, Holbrook A, Jaeschke R, Elizabeth Juniper M, et al. Users 'Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test: I B. What Are the Results and Will They Help Me In Caring for My Patients? Journal ofAmerican Medical Association, 271 (9). 1994:703-7.

58. Teitelbaum JS, Eliasziw M, Garner M. Tests of motor function in patients suspected of having mild unilateral cerebral lesions. Can J Neurol Sci. 2002 Nov;29(4):337-44.

59. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med. 2003 Jul;29(7):1043-51.

60. Heffner JE, Sahn SA, Brown LK. Multilevel likelihood ratios for identifying exudative pleural effusions(*). Chest. 2002 Jun;121(6):1916-20.

61. Khairy M, Clough A, El-Toukhy T, Coomarasamy A, Khalaf Y. Antral follicle count at down-regulation and prediction of poor ovarian response. Reprod Biomed Online. 2008 Oct;17(4):508-14.

62. Peirce JC, Cornell RG. Integrating stratum-specific likelihood ratios with the analysis of ROC curves. Med Decis Making. 1993 Apr-Jun;13(2):141-51.

63. Rifkin RD. Maximum Shannon information content of diagnostic medical testing. Including application to multiple non-independent tests. Med Decis Making. 1985 Summer;5(2):179-90.

64. Heckerling PS. Information content of diagnostic tests in the medical literature. Methods Inf Med. 1990 Jan;29(1):61-6.

65. Simel DL, Samsa GP, Matchar DB. Likelihood ratios for continuous test results--making the clinicians' job easier or harder? J Clin Epidemiol. 1993 Jan;46(1):85-93.

66. Tandberg D, Deely JJ, O'Malley AJ. Generalized likelihood ratios for quantitative diagnostic test scores. The American journal of emergency medicine. 1997 Nov;15(7):694-9.

67. Heffner JE, Highland K, Brown LK. A meta-analysis derivation of continuous likelihood ratios for diagnosing pleural fluid exudates. Am J Respir Crit Care Med. 2003 Jun 15;167(12):1591-9.

68. Porcel JM, Pena JM, de Vera CV, Esquerda A, Vives M, Light RW. Bayesian analysis using continuous likelihood ratios for identifying pleural exudates. Resp Med. 2006 Nov;100(11):1960-5.

Chapter Two Appendix

Author (Year)	Inclusion Criteria	Publication Period	No. of Studies	No. Compliant to item 22.	%
Areia et al., 2010	All diagnostic endoscopy-related studies	1998 to 2008	110	23	21%
Coppus et al., 2006	Diagnostic or prognostic studies published in two key reproductive medicine journals: journal Fertility and Sterility and Human Reproduction	1999	24	1	4%
		2004	27	4	15%
Fontela et al., 2009	Diagnostic accuracy studies of commercial tests for Tuberculosis	2004-2006	45	8	18%
	Diagnostic accuracy studies of commercial tests for HIV		18	0	0%
	Diagnostic accuracy studies of commercial tests for Malaria		27	7	26%
Freeman et al., 2009	Diagnostic accuracy of non-invasive prenatal diagnostic (NIPD) tests	1996 - 2006	27	17	63%
Hing et al., 2009	Diagnostic accuracy studies of McMurray's test with a gold standard of knee arthroscopy or MRI	1980-2007	11	0	0%
Jahromi et al., 2005	Studies that compared the accuracy of color duplex ultrasound with angiography for the diagnosis of internal carotid artery stenosis	1989-2004	47	15	32%
Johnson et al., 2007	Studies evaluating Optical Coherence Tomography for the diagnosis of Glaucoma	2001-2006	30	12	40%
Krzych et al., 2009	Diagnostic accuracy studies of BNP and NT-proBNP	2004-2007	28	5	18%
Legare et al., 2007	Studies of self-administered instruments that evaluate physicians' perceptions of the decision-making process	Before 2007	11	2	18%
Lumbreras et al., 2006	Diagnostic accuracy studies of genetic, molecular and proteomic tests published in JAMA, Lancet, NEJM, Cancer Research, and Clinical Cancer Research	2002-2005	44	5	11%
Mahoney et al., 2007	Diagnostic accuracy studies of handheld glucose monitors	2002-2006	52	17	33%
Miller et al., 2009	Studies on the magnetic resonance imaging assessment of juvenile idiopathic arthritis	1992-2007	18	0	0%

Paranjothy et al., 2007	Diagnostic accuracy studies of Scanning Laser Polarimetry for Glaucoma	1997-2005	20	4	20%
Perry et al., 2010	Accuracy studies of screening instruments to identify the risk of suicide and self-harm behaviour in offenders	1988-2005	5	1	20%
Roposch et al., 2006	Studies on the use of ultrasonography for the diagnosis of developmental dysplasia	1985-2001	11	0	0%
Selman et al., 2005	Studies assessing the accuracy of Technetium 99-labeled sulphur colloid for the identification of groin node metastases in women with squamous cell cancer of the vulva.	1997-2003	11	1	9%
Selman et al., 2011	Studies included in systematic reviews of minimal and non-invasive tests to determine the lymph node status in gynaecological cancers	Pre 2004	85	19	22%
		Post 2004	17	6	35%
	Studies included in systematic reviews of Down's serum screening markers and uterine artery Doppler to predict small for gestational age in obstetrics	Pre 2004	85	54	63%
		Post 2004	17	11	66%
Shunmugam et al., 2006	Studies assessing the accuracy of the Heidelberg Retina Tomograph for the diagnosis of Glaucoma	1993-2004	29	8	28%
Siddiqui et al., 2005	Diagnostic accuracy studies published in five leading ophthalmic journals: AJO, Archives, BJO, (IOVS) and Ophthalmology	2002	16	5	31%
Siva Rama et al., 2006	Diagnostic accuracy studies published in three leading general orthopaedic journals: JBJS (American), JBJS (British) and CORR	2002-2004	26*	13	50%
Smidt et al., 2006	Diagnostic accuracy studies published in 12 journals – all with an impact factor of 4 or higher.	2000	124	73	59%
		2004	141	80	57%
Zafar et al., 2008	Reports of diagnostic accuracy for Diabetic Retinopathy screening	2000-2004	76	9	12%

* The review included 37 studies in total, however item 22 was deemed 'not applicable' for 11 of the studies

Chapter Three

Diagnostic accuracy research of quantitative cancer biomarkers: A methodological systematic review

3.1. Overview

- **Research Objective:** How is the diagnostic accuracy of quantitative cancer biomarkers analysed and reported in primary research and meta-analyses?
- **Background:** Alternatives to the binary model to evaluate the accuracy of quantitative diagnostic tests have been proposed; however the extent to which they are implemented in contemporary research is unclear.
- **Methods:** Medline and Embase were systematically searched for diagnostic accuracy meta-analyses of quantitative cancer biomarkers published since 2009. The methods used to analyse diagnostic accuracy in the primary studies featuring in each meta-analysis were extracted, in addition to the methods implemented in the meta-analyses themselves.
- **Results:** Quantitative diagnostic test scales are still commonly dichotomised, although paired accuracy statistics at multiple binary thresholds are reported in many primary studies. Meta-analyses pooled the results from just one 2 x 2 table per study. Exploration of the optimal threshold at the meta-analysis level was generally infeasible due to limited reporting at the primary study level.
- **Conclusion:** For primary studies, methods which convey how the test performs at different thresholds in a format that can be extracted for inclusion in a meta-analysis are needed. This will facilitate the exploration of optimal thresholds at the more powerful meta-analysis level.

3.2. Introduction

Traditional methods for evaluating the performance of a diagnostic test typically centre on a binary framework. Although this approach is logical for a test which naturally produces positive and negative results, it has been demonstrated extensively that the dichotomisation of quantitative diagnostic test scales to fit this model results in a significant loss of useful information e.g. (1, 2).

As described in depth in **Chapter Two**, a number of alternative methods have been proposed to summarise the accuracy of quantitative diagnostic test results e.g. intermediate test ranges or multiple-category and result-specific approaches. The introduction of an intermediate test outcome is argued to better represent how clinicians interpret quantitative test results (3, 4), whereas multiple-category and result-specific methods facilitate a richer interpretation of the test result and minimise information loss (5-7). There was little evidence available to help assess the extent to which these alternative methods are being implemented in contemporary diagnostic accuracy research for quantitative tests. The main objective of this chapter is therefore to review the methods implemented in current research to evaluate the accuracy of quantitative diagnostic tests.

In addition to primary research, meta-analyses are also of great interest as this 'higher' level of evidence is given much more weight in the guideline development process and is therefore likely to have the greatest clinical impact. This systematic review will therefore assess both the methods used in diagnostic meta-analyses and those used in the primary studies featuring in these reviews.

Diagnostic accuracy studies are notoriously difficult to identify in the literature (8), and the one relevant MESH term available consists of methodological terms (sensitivity and specificity). Using this MESH term as a means of identifying research relevant to this review is likely to bias our results significantly and therefore I was advised by an information specialist to restrict the review to a specific clinical area. Mallett and colleagues evaluated the methods and reporting of systematic

reviews of diagnostic tests for cancer (9). This review, however, only included four meta-analyses of tests with continuous outcomes, all of which reported only dichotomous results. To expand on this work, this systematic review will therefore assess the methods used in the evaluation of quantitative cancer biomarkers specifically.

3.3. Methods

3.3.1. Literature Search

Systematic literature searches of Medline (1946 to 2013, May Week 4) and Embase (1974 to 2013, Week 22) were carried out using the Ovid SP interface. The search string consisted of four parts: the Cochrane Cancer Network string to identify cancer studies (10), the terms 'biomarker*' or 'marker*' or 'blood' or 'serum' or 'sera' or 'plasma' or 'antibodies' or 'urin*' or 'assay*' to identify studies relating to quantitative markers, the terms 'diag*' or 'prediag*' or 'screen*' or 'detect*' or 'predict*' to identify diagnostic accuracy studies and the terms 'meta*analysis' or 'systematic review' to restrict the search to reviews rather than primary research.

As many of the novel meta-analysis methods for pooling results from quantitative diagnostic tests have only been published in the last few years, results were restricted to reviews published from 2009 onwards.

3.3.2. Inclusion Criteria

Meta-analyses were included if they evaluated a quantitative diagnostic test for the presence or absence of cancer, including cancer staging, metastasis and recurrence. In keeping with Mallett et al.'s original review (9), markers of prognosis, risk factors, treatment response and survival were excluded. Conference abstracts were excluded as they typically contain insufficient information for a full methodological assessment. The inclusion and exclusion criteria for the selection of meta-analyses can be found in Figure 3.1.

Based on the assumption that the primary studies will have already undergone relatively strict selection criteria to be included at the meta-analysis level, very few restrictions were placed on the inclusion of primary studies in this review. Non-English studies had to unfortunately be excluded due to interpretation and accessibility issues, and conference abstracts were also excluded due to their limited content. If there were any primary studies that featured in more than one meta-analysis, these duplicates were removed so that the methods used in these studies were only included in the results once.

Inclusion criteria for meta-analyses:

- Systematic review
- Assess at least one diagnostic test for the presence or absence of cancer/cancer staging/recurrence of cancer (tests for cancer risk factors such as pre-cancerous lesions, prognosis and treatment response were not included)
- Human participants
- Reported the references for the primary studies included in the review

Exclusion criteria for meta-analyses:

- Did not carry out a meta-analysis of the diagnostic accuracy of the test(s)
- Either the meta-analysis or all of the included primary studies were not published in English
- Conference abstracts
- Not a quantitative biomarker

Figure 3.1. Inclusion and exclusion criteria for the selection of meta-analyses

3.3.3. Data Extraction

I (from now on referred to as BS) screened the titles and abstracts of the search results according to the inclusion criteria (Figure 3.1) to identify relevant meta-analyses. The full texts were obtained for those that met the inclusion criteria, and the exclusion criteria applied to select the final meta-

analyses to be included in the review. The full texts for the primary studies included in each of the selected meta-analyses were then obtained and assessed for eligibility.

Two reviewers (BS and JJ) carried out the data extraction. Agreement was reached by consensus or by reference to a third reviewer (BN) in the case of any differences. Customised Excel spread sheets were developed for the data extraction process for both the meta-analyses and primary studies. A table of the information extracted from the included meta-analyses and primary studies can be found in the Appendix of this chapter (3B and 3C). For primary studies that evaluated the accuracy of more than one diagnostic test, information was only extracted for the test(s) assessed in the meta-analysis.

On completion of the extraction process, the meta-analysis methods were further evaluated to assess whether all of the available data in the primary studies was extracted and included. In scenarios where only partial data had been extracted, we assessed whether the authors of the meta-analyses provided any justification for their data selection.

3.4. Results

3.4.1. Search Results

Figure 3.2 provides the results of the literature search and the selection of the meta-analysis studies to be included in this review based on the inclusion criteria. A list of included meta-analyses and key details can be found in the Appendix (3A).

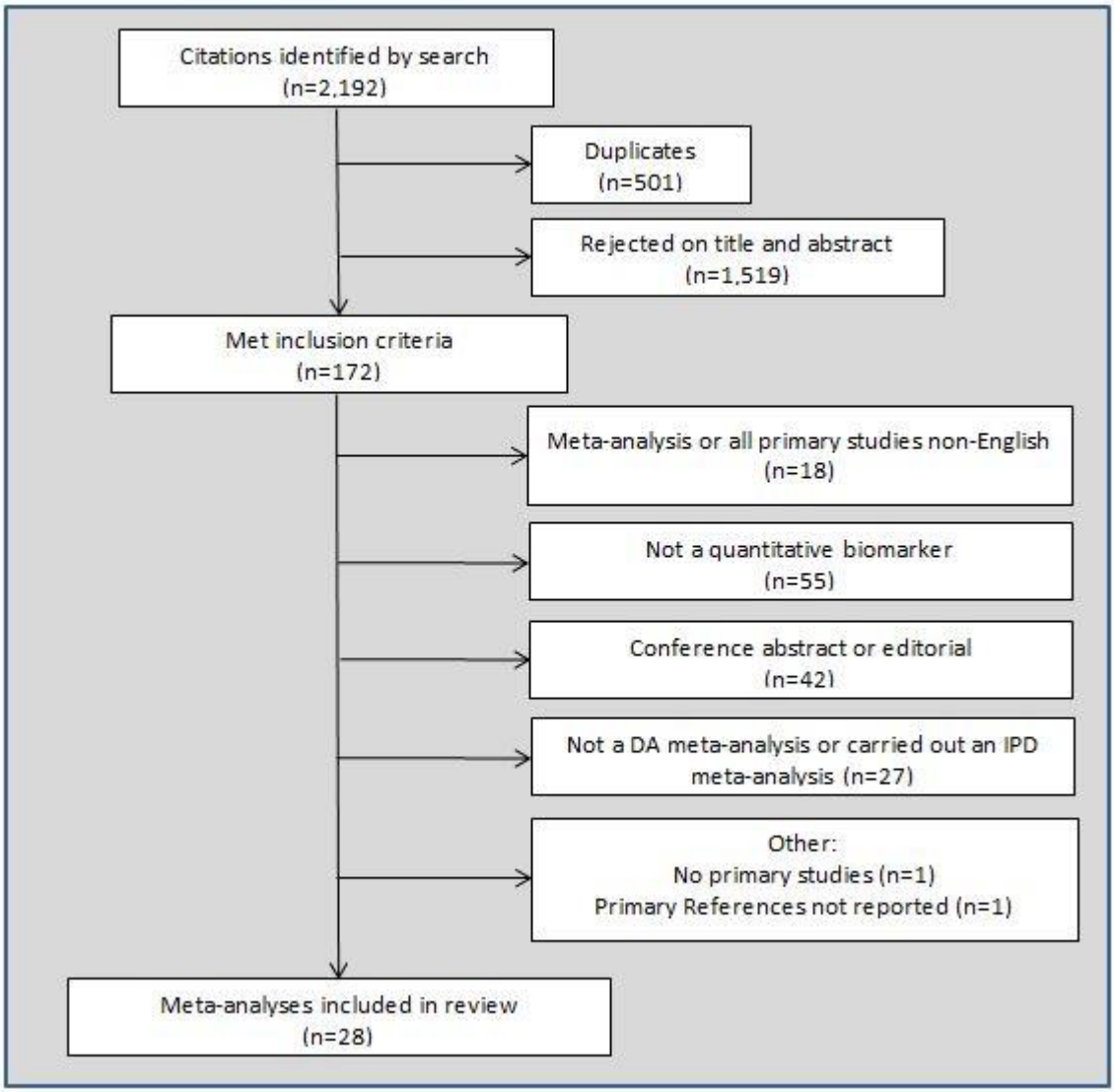


Figure 3.2. Flowchart of the selection process of meta-analyses

Many authors stated that they had to exclude eligible primary studies from the meta-analysis because there was insufficient data reported to extract the 2 x 2 table (n=19/28). The number of primary studies that had to be excluded for this reason was reported for 16 of these reviews. Based on these reported figures, a median of five primary studies had to be excluded for this reason per meta-analysis (range = 1 to 37 studies). Other common reasons that authors reported for excluding primary studies from meta-analyses were: duplication of data across multiple publications (n=7/28), small sample size (n=6/28), sample selection biases e.g. a control group of healthy

individuals or no control group (n=5/28). In five meta-analyses, the authors failed to report whether any primary studies had been excluded.

A total of 386 primary studies were included in the 28 meta-analyses reviewed. Some primary studies featured in more than one meta-analysis; these duplicates were removed so that they were only counted in the results once (60 studies in total). The full details of the selection of primary studies can be found in Figure 3.3. The number of primary studies that had to be excluded from each meta-analysis and reasons are available in the Appendix (3A).

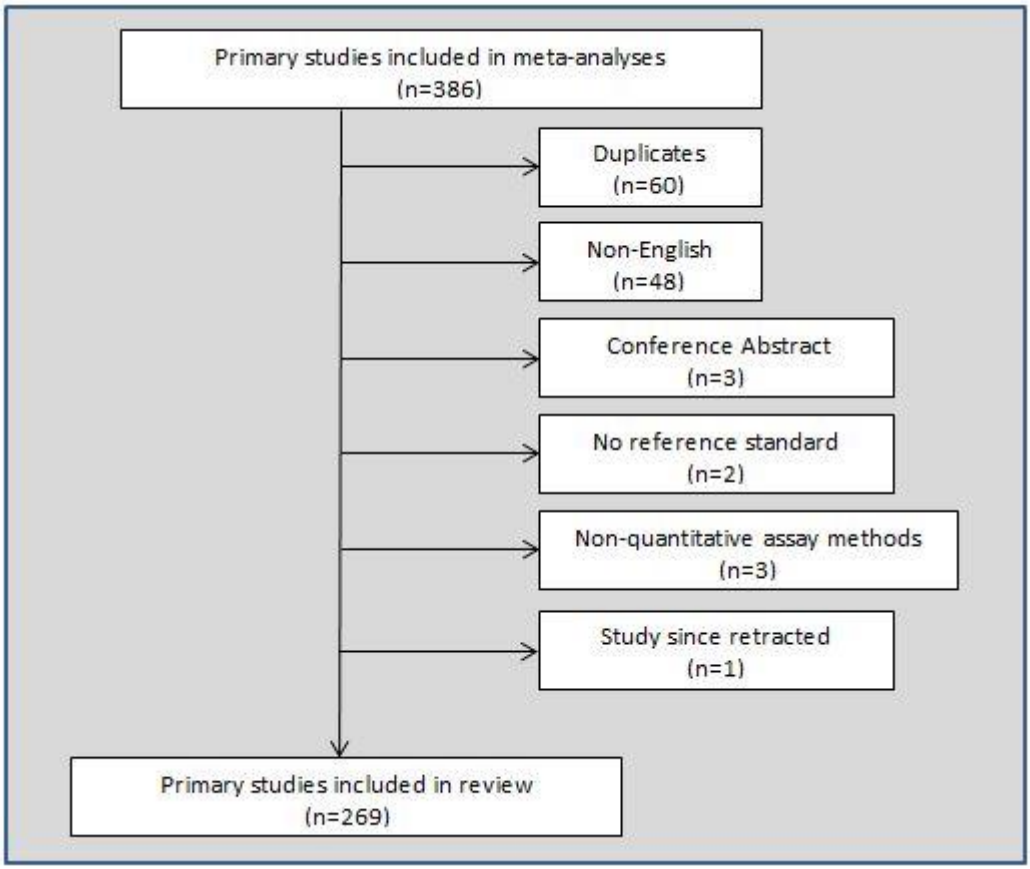


Figure 3.3. Flowchart for the selection process of primary studies

Table 3.1 shows the breakdown of the studies included in the review by type of cancer and the diagnostic test under evaluation. Some of the included primary studies and meta-analyses evaluated the accuracy of more than one test. For example, there were 2 meta-analyses that evaluated both Cancer Antigen 125 and Human Epididymis Protein 4. Ovarian cancer was the most

popular focus of diagnostic research, with 7 meta-analyses focusing on biomarkers for this type of cancer. The primary studies included in the meta-analyses overlapped significantly as many evaluated the same diagnostic tests.

Type of Cancer	Number of Meta-analyses	Number of Primary Studies
Prostate	**3**	**35**
Prostate-Specific Antigen	2	21
Prostate-Specific Antigen 3	1	14
Oesophageal	**1**	**13**
Serum p53	1	13
Ovarian	**7***	**53**
Cancer Antigen 125	4	47
Human Epididymis Protein 4	5	25
Bladder	**2**	**14**
Survivin mRNA	2	14
Mesothelioma	**2**	**36**
Soluble Mesothelin-related Peptides	1	11
Multiple Biomarkers (see (11))	1	35
Lung	**5**	**47**
Circulating Cell-free DNA	1	10
Pro-gastrin-releasing peptide	2	17
Circulating MicroRNA	1	13
Vascular Endothelial Growth Factor	1	7
Colorectal	**3**	**26**
Faecal Tumour M2-Pyruvate Kinase	2	8
Carcino-embryonic Antigen	1	18
Pancreatic	**2**	**14**
MicroRNA	1	7
Serum Immunoglobulin G4	1	7
Liver	**3**	**31**
Des-Gamma Carboxyprothrombin	1	19
Golgi Protein 73	1	6
Multiple Biomarkers (see (12))	1	7
Total	**28**	**269**

Table 3.1. Studies included in the review, grouped by target cancer.*Two meta-analyses evaluated both Cancer Antigen 125 and Human Epididymis Protein 4

3.4.2. Methods used in Primary Studies

Table 3.2 shows a summary of the methods used to analyse and report accuracy in the included primary studies. The binary model was implemented to analyse accuracy in the majority (96%) of primary studies, with very few adopting a multiple-category or result-specific model (5% and less than 1%, respectively). Of those implementing the binary model, accuracy was reported at more than one threshold in 34% of the primary studies. ROC curves which describe the accuracy of a test across all possible values were presented in half of the primary studies reviewed (51%).

Sensitivity and specificity were the most commonly reported accuracy statistics (84% and 81%, respectively), followed by the area under the ROC curve which was reported in just over half of the primary studies (54%). Likelihood ratios were reported very rarely (4%).

Assessment Item	No. of unique primary studies (%) out of 269 assessed
Plots reported	
Raw results broken down by disease status	126 (47%)
ROC curve	137 (51%)
Accuracy statistics reported	
Sensitivity	225 (84%)
Specificity	217 (81%)
Positive and negative likelihood ratio	10 (4%)
Positive and negative predictive value	91 (34%)
Overall accuracy	45 (17%)
Area under the ROC curve	146 (54%)
Adopted a binary model to analyse accuracy	**257 (96%)**
Reported accuracy at 1 threshold	178 (66%)
Reported accuracy at 2 thresholds	40 (15%)
Reported accuracy at 3 thresholds	16 (6%)
Reported accuracy at 4 thresholds	10 (4%)
Reported accuracy at 5 or more thresholds	13 (5%)
Range of number of binary thresholds reported	1 - 20
Adopted a multi-category model to analyse accuracy	**13 (5%)**
Reported accuracy for 3 categories	7 (3%)
Reported accuracy for 4 categories	3 (1%)
Reported accuracy for 5 or more categories	3 (1%)

Table 3.2. Methods used to analyse accuracy in primary studies

Of those studies that implemented the binary model to analyse accuracy (96%, n=257/267), the most common method for determining at which threshold(s) to report accuracy was through ROC analysis (n=75/257, 29%). A fifth of primary studies (n=55, 21%) reported accuracy at particular levels of sensitivity and/or specificity (11%) and a further 13 (5%) used reference range methods to select thresholds (2 or 3 standard deviations from the mean in the healthy population). Accuracy was reported at pre-determined, rather than data-driven thresholds, in around a fifth of these primary studies (22%, n=57/257). Fourteen (5%) of these reported accuracy at laboratory or manufacturer recommended thresholds and in 19 (7%) the threshold was based on those reported in previous relevant literature. A substantial proportion failed to report their threshold selection method or selected arbitrary thresholds at which to report accuracy (28%, n=73/257).

Of the primary studies that reported accuracy for more than 2 categories of test result, the selection of thresholds was arbitrary or unjustified in the majority (n=8/13). Five of the primary studies that reported test results as a trichotomy used pre-determined thresholds to delimit their test ranges.

3.4.3. Methods used in Meta-analyses

All of the meta-analyses carried out a search of at least one electronic database, with the vast majority (27/28) searching more than one database. Five of the meta-analyses (18%) did not state their search strategy. In those that did, frequently used terms relating to diagnostic methodology included: ‘sensitivity and specificity’ (n=11, 39%), ‘accuracy’ (n=6, 21%) and ‘diagnosis’ (n=6, 21%). Twenty-five meta-analyses assessed the quality of included primary studies (89%), 23 of which used QUADAS and 2 used independent criteria. Four of the meta-analyses that used QUADAS also used the STARD checklist to assess reporting quality. Publication bias was assessed in 13 of the meta-analyses (46%), of which 11 implemented the Deeks Funnel plot (13), 5 used the Egger test (14), and 1 used the Begg test (15).

The majority of the meta-analyses reviewed (n=24/28) included at least one primary study that reported accuracy at more than one binary threshold. In all of these reviews however, just one 2 x 2 table was extracted from each primary study. The authors rarely justified their data extraction selection or acknowledged that there were multiple 2 x 2 tables available. It appeared in most cases that 2 x 2 tables were selected to minimise the variability in thresholds across primary studies.

Two-thirds of meta-analyses (n=19/28) reported the positivity thresholds from each primary study at which accuracy was meta-analysed. The majority of meta-analyses faced the problem of having to pool accuracy results across more than one positivity threshold (n=26/28). The presence of a threshold effect, which could be due to implicit differences between study designs and/or explicit differences in positivity thresholds, was explored in just over half of the meta-analyses(n=15/28), of which 12 calculated the Spearman's correlation coefficient to look for evidence of an inverse relationship between sensitivity and specificity. Only 5 meta-analyses reported evidence of a threshold effect. General study heterogeneity was explored in 25 of the meta-analyses reviewed, with the Cochran-Q and I^2 statistic being the most commonly implemented methods.

Table **3**.3 summarises the analytic methods used to meta-analyse test accuracy. The most commonly pooled accuracy statistics were sensitivity and specificity, the diagnostic odds ratio, and positive and negative likelihood ratios. A basic random effects model (the DerSimonian Laird method (16)) was the most popular method for pooling statistics (n=14/28), with the bivariate model which takes into account the correlation between paired summary statistics being employed in 6 meta-analyses. Multivariable adjustments for other factors were made in 10 studies, and subgroup analyses were carried out in 19 studies. Summary ROC curves and the area under the sROC curve were also frequently implemented to produce threshold-independent summaries of pooled accuracy. Moses and Littenberg's linear regression model (17) was used to construct sROC curves in 14 meta-analyses, and the Hierarchical sROC model (18) implemented in 4 meta-analyses.

MetaDisc was the most popular statistical software for carrying out these analyses (n=21), with Stata (n=9) and SAS (n=2) being used to run more complex models.

Assessment Item	No. of meta-analyses (%) out of 28 assessed
Pooled accuracy statistics reported	
Sensitivity	24 (86%)
Specificity	23 (82%)
Positive and negative likelihood Ratio	19 (68%)
Positive and negative predictive value	2 (7%)
Diagnostic odds ratio (DOR)	20 (71%)
Area under the ROC curve	3 (11%)
Methods used for Pooling statistics	
Fixed effects	3 (11%)
Univariate	1 (4%)
Multivariate	2 (7%)
Random Effects	14 (50%)
Univariate	10 (36%)
Multivariate	4 (14%)
Bivariate Random Effects Model	6 (21%)
Univariate	2 (7%)
Multivariate	4 (14%)
Mean ± SD	1 (4%)
Unclear	4 (14%)
Graphs reported	
Forest plots	21 (75%)
Forest plot of sensitivity and specificity	18 (64%)
Forest plot of likelihood ratios	3 (11%)
Forest Plot of DOR	5 (18%)
Forest Plot of AUC	1 (4%)
Summary ROC curve	24 (86%)
Statistics Derived from sROC curve	
Q* (maximal pooled sensitivity and specificity)	15 (54%)
Area under sROC	20 (71%)
Methods used to estimate sROC curve	
Moses and Littenberg method	14 (50%)
HSROC	4 (14%)
Unclear	6 (21%)

Table 3.3. Analytic methods used in meta-analyses

3.5. Discussion

3.5.1. Methods for Analysing Accuracy

This review suggests that the binary model continues to be the most commonly adopted approach to analysing the accuracy of quantitative cancer biomarkers at both the primary study and meta-analysis level. The analytic methods employed at both levels were very similar to those identified in a previous review of cancer diagnostic test evaluations (9), which primarily consisted of tests with binary or ordinal outcomes. This demonstrates that a 'one size fits all' approach is being adopted, despite extensive literature showing that the binary model is not suitable for evaluating tests with a continuous outcome.

Multi-level and result-specific approaches (19, 20), where accuracy is summarised for multiple ranges of test results or modelled across the whole test scale, were implemented in very few primary studies. Although these methods offer a far richer interpretation of test accuracy, they also require more complex statistical analyses compared to the simple derivation of summary statistics from a 2 x 2 table. For example, multi-level likelihood ratios require the calculation of likelihood ratios for lots of categories rather than just one, and the most appropriate number of categories and the optimal placement of thresholds must also be considered. The calculation of result-specific accuracy statistics involves modelling, which requires a greater level of expertise in statistics and some familiarity with statistical software. If authors are naïve to the information loss incurred by implementing the binary model, these greater analytical demands, in addition to producing a more complex analysis for the reader to understand, may be key deterrents to employing these alternative methods in practice. The STARD statement, a checklist providing guidance on how diagnostic accuracy studies should be reported, could be one means of conveying the importance of these more appropriate analytic methods for quantitative test results.

Despite accuracy statistics being available at multiple thresholds in the few studies that did report multi-level or result-specific results, the methods employed in the meta-analysis dictated that the

analyses were restricted to a single 2 x 2 table from each study. These results show that the promotion of multi-level or result-specific methods for analysing accuracy is somewhat fruitless if this additional information is then ignored at the meta-analysis level. Irwig and colleagues outline methods for meta-analysing results that have been reported for more than two categories or as a continuum (21). These methods are only suitable however when all of the primary studies have adopted exactly the same methodological approach i.e. they have all reported accuracy for the same number of categories or they have all used the same model to predict accuracy across the whole test scale. The results of this review show that this is not realistic as, in the vast majority of cases, a range of methods will have been adopted to analyse accuracy in the relevant primary research. This type of analysis would therefore only be useful when individual patient data is available.

3.5.2. The Problem of Thresholds

For tests on a quantitative scale, binary summary statistics are highly dependent on the threshold at which accuracy is evaluated. There were a wide variety of methods implemented to select threshold(s) to dichotomise test scales; the most commonly used being ROC analysis. This approach is data-driven, and therefore the 'optimal' threshold will vary from study-to-study due to heterogeneity in study design, sampling variability and measurement error.

This variability in thresholds is particularly challenging at the meta-analysis level when attempting to pool accuracy statistics. 93% of the meta-analyses included in this review pooled accuracy statistics calculated at different positivity thresholds. Evidence of a threshold effect was explored in many of the meta-analyses, typically by calculating the Spearman's Correlation coefficient for sensitivity and specificity. The method determines whether there is a statistically significant inverse correlational relationship between sensitivity and specificity, as would be expected across the spectrum of thresholds. Of the studies that employed this method (N=12), less than half found evidence of a threshold effect. This is likely to be due to the fact that this statistical test lacks power

in this context as a result of the small number of primary studies included in the majority of meta-analyses (21, 22). On finding evidence of a threshold effect, many of the authors employed a random effects model to pool sensitivity and specificity. Although this method allows for the variability in sensitivity and specificity to be better represented, it still fails to account for their correlation (23).

As the pooling of summary statistics will generally lead to an underestimation of the accuracy of a test when a threshold effect is evident (24), a more appropriate method for evaluating accuracy in this scenario is the summary ROC curve. This method depicts how sensitivity and specificity varies between studies, in contrast to a standard ROC curve which shows how these measures vary across different thresholds on the test scale within a single study. Summary ROC (sROC) curves are intended to overcome the problem of a threshold effect and threshold-independent summaries of accuracy, such as the diagnostic odds ratio, can also be derived from them.

A key drawback of the sROC curve as a means of summarising the accuracy of a quantitative test, however, is that it cannot be used to directly determine the most appropriate positivity threshold; this information has been lost in the analytical process. This is because the variation in accuracy between studies may not be entirely driven by different positivity thresholds being employed; it could also be in part a result of implicit differences in the studies that cause variation in the proportion of patients testing 'positive', such as disease spectrum differences or partial verification (25). This limitation was evident in the current review, with the majority of authors unable to make threshold recommendations based on the results of their meta-analyses, and many listing it as a key area for future research. Thus, although it can be useful for comparing the head-to-head accuracy of competing tests, the sROC curve is limited in its utility to practising clinicians (24).

Some novel methods have been proposed to estimate sROC curves when multiple categories and/or inconsistent thresholds have been used to analyse accuracy in primary studies. Hamza and colleagues recommend a bivariate random effects model to calculate a sROC curve for scenarios

where primary studies have reported accuracy at multiple binary thresholds (26). This method is most suited to cases where accuracy has been reported at multiple, but consistent, thresholds, as the only means of overcoming the absence of certain thresholds in studies is to regard them as missing data. Dukic and Gatsonis go one step further and propose a hierarchical ordinal regression method which allows the sROC curve to be estimated when accuracy is reported for an unequal number of 'non-aligned' categories (27). Implementation of this method in some of the meta-analyses in this review would have reduced the amount of information loss and facilitated the incorporation of all of the data reported in the primary studies. This method has the additional attractive quality that pooled accuracy statistics can be extracted for any test value. Further methodological work is required to highlight the benefits of using this method compared to current binary approaches, in addition to the development of statistical packages to make it more accessible to the non-statistician.

These novel methods both involve the reconstruction of the ROC curves for each study, recognising that it is preferable to meta-analyse the accuracy of quantitative tests across all possible thresholds (26, 27). This process, however, is computationally intensive and requires the use of complex statistical methods which are highly dependent on distributional assumptions to impute this information. The accuracy of these estimates will also be highly dependent on the fit of the model, and will unavoidably be subject to some degree of modelling error and uncertainty.

3.5.3. Information Loss

This loss of information over the course of the research pathway is a direct result of the restrictive nature of the methods being used to report test accuracy in primary research. In the present review, ROC curves, which depict the accuracy across the whole test scale, were reported in around half of the primary studies. However, these plots do not allow you to extract the sensitivity and specificity at specific values of the test scale. Analyses are consequently restricted to the thresholds at which accuracy has been explicitly reported at the meta-analysis level and exploration of

alternative thresholds is unfeasible. This issue was faced in a review evaluating the accuracy of endometrial layer thickness to predict cancer (28). The authors hypothesised that this test would perform best at a threshold 3mm, however were unable to assess this due to the lack of studies reporting accuracy at this threshold.

One way to overcome these issues relating to categorisation and threshold inconsistencies is to carry out an individual-patient data meta-analysis. This would not only negate the issue of information loss, but also allow for the adjustment of other possible causes of heterogeneity between primary studies. This is not always feasible however due to researchers' reluctance to share their data or the loss of old datasets (29). Only two of the meta-analyses identified in the original search analysed individual-patient data. A possible solution would be to make the reporting of methods that convey accuracy across all thresholds standard practice in primary research for quantitative tests. The implementation of result-specific methods may be one means of achieving this.

3.5.4. Limitations

Ideally, this review would not have been restricted to a specific clinical area and would provide a snapshot of diagnostic research in general. However, the Cochrane search string for identifying diagnostic accuracy studies could not be used to identify meta-analyses as it included methodological terms which may have biased the results of the review. In particular, methodological terms which rely on the dichotomisation of test results, such as 'sensitivity' and 'specificity' could not be used. After seeking advice from an information specialist, the most feasible approach was to limit the review to a broad area of clinical research. Cancer was selected due to the wealth of new biomarkers in production and the overwhelming attention that they have received in the diagnostic literature. By focusing the review on cancer, the results of this review also expand on existing evidence regarding the methods and reporting used generally in cancer diagnostic research (9).

Despite this more focused approach to the review, the effectiveness of the final search strategy may still have been hindered by the fact that these key diagnosis-related search terms could not be included. An issue that could not be avoided was the fact that these terms were often used in the search strategies of the systematic reviews themselves. This could have introduced bias in terms of our evaluation of the categorisation methods being adopted in primary studies. Where possible, the papers excluded on the basis that they did not facilitate the extraction of results as a 2 x 2 table were reviewed to ensure that they were not adopting multiple-category or result-specific methods. There was no evidence of this bias found.

Three of the four meta-analyses that only had accuracy available at a single binary threshold across all primary studies evaluated microRNA tests. The primary studies included in these reviews were mainly phase I type diagnostic accuracy studies and therefore lacked many of the qualities we would expect from a diagnostic accuracy study (30). For example, diagnostic accuracy indices were not typically calculated and thresholds were not reported. Many studies had very small sample sizes and consisted of analyses of laboratory samples rather than standard patient recruitment. Although the inclusion of these studies is an accurate depiction of our review findings, it was surprising to see such early phase research included in a meta-analysis.

3.5.5. Conclusion

Systematic reviews of primary research offer the highest level of evidence (31), and are consequently the optimal evidence-base for the development of clinical guidelines. It is therefore imperative that the methods used to pool the results from primary research facilitate the production of clinically useful recommendations. The meta-analysis of primary diagnostic accuracy research is still a relatively novel field and methods are still undergoing extensive development (23). The Cochrane group established a work stream for diagnostic accuracy research in 2007, and to date there are only 14 published systematic reviews in the Cochrane library. However 69 protocols are in progress, highlighting the recent surge in the production of this type of evidence.

The findings of this review demonstrate a number of issues with the methods currently being used to evaluate the accuracy of quantitative diagnostic tests.

The ability to meta-analyse accuracy across the full quantitative test scale would facilitate a richer understanding of the discriminatory performance of quantitative diagnostic tests and allow greater flexibility in the exploration of clinically useful decision thresholds. Authors of primary studies need to consider whether the methods that they are using to report accuracy will allow their results to be easily extracted for meta-analysis. For tests on a quantitative scale, this involves identifying methods which minimise the loss of information about how the test performs at different thresholds of the test scale. In order for primary research to still be useful to the clinical reader, this may require the use of more than one method in the analysis. The literature review in **Chapter Two** revealed no evidence regarding GP preferences for quantitative test result interpretation, and consequently this will be the focus of the next chapter.

What this chapter adds:

- Describes the methods currently implemented to evaluate the accuracy of quantitative diagnostic tests, both at the primary study and meta-analysis level
- Demonstrates that meta-analyses are notably limited by incomplete reporting in primary diagnostic accuracy research
- If individual patient data could be made readily available, meta-analysing accuracy across the full quantitative test scale would allow greater flexibility in the exploration of clinically useful decision thresholds

3.6. Acknowledgements

I would like to thank Dr. Jing Jin, Dr. Brian Nicholson, Nia Roberts and Dr. Tim James for their contribution to this chapter. Dr. Jin independently extracted the data from both the primary studies

and the meta-analyses. Dr. Nicholson acted as a mediator, and resolved any differences in the findings between the two reviewers. Nia Roberts, an information specialist, helped to develop the search strategy. Dr. James, a Clinical Biochemist, provided advice on studies where it was unclear whether the laboratory methods would produce results on a quantitative test scale.

3.7. References

1. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006 May 6;332(7549):1080.

2. Heckerling PS. Information content of diagnostic tests in the medical literature. Methods Inf Med. 1990 Jan;29(1):61-6.

3. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making. 1987 Apr-Jun;7(2):107-14.

4. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. J Clin Epidemiol. 1990;43(1):109-13.

5. Furukawa TA, Goldberg DP, Rabe-Hesketh S, Ustun TB. Stratum-specific likelihood ratios of two versions of the general health questionnaire. Psychol Med. 2001 Apr;31(3):519-29.

6. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet. 2005 Apr 23-29;365(9469):1500-5.

7. Sonis J. How to use and interpret interval likelihood ratios. Family Medicine Journal. 1999;31(6):432-7.

8. Beynon R, Leeflang MM, McDonald S, Eisinga A, Mitchell RL, Whiting P, et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev. 2013;9:MR000022.

9. Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. BMJ. 2006 Aug 26;333(7565):413.

10. Lodge M. The Cochrane cancer network. 2005.

11. van der Bij S, Schaake E, Koffijberg H, Burgers JA, de Mol BA, Moons KG. Markers for the non-invasive diagnosis of mesothelioma: a systematic review. Br J Cancer. 2011 Apr 12;104(8):1325-33.

12. Witjes CD, van Aalten SM, Steyerberg EW, Borsboom GJ, de Man RA, Verhoef C, et al. Recently introduced biomarkers for screening of hepatocellular carcinoma: a systematic review and meta-analysis. Hepatology international. 2013 Mar;7(1):59-64.

13. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol. 2005 Sep;58(9):882-93.

14. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997 Sep 13;315(7109):629-34.

15. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994 Dec;50(4):1088-101.

16. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986 9//;7(3):177-88.

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

18. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001 Oct 15;20(19):2865-84.

19. Irwig L. Modelling result-specific likelihood ratios. J Clin Epidemiol. 1992 Nov;45(11):1335-8.

20. Bowden SC, Loring DW. The diagnostic utility of multiple-level likelihood ratios. J Int Neuropsychol Soc. 2009 Sep;15(5):769-76.

21. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. Journal of clinical epidemiology. 1995 January;48(1):119-30; discussion 31-2.

22. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol. 2006;6:31.

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

24. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ. 2001 Jul 21;323(7305):157-62.

25. 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 Oct;58(10):982-90.

26. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. BMC Med Res Methodol. 2009;9:73.

27. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003 Dec;59(4):936-46.

28. Foy R, Warner P. About time: diagnostic guidelines that help clinicians. Qual Saf Health Care. 2003 Jun;12(3):205-9.

29. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221.

30. Knottnerus JA. The Evidence Base of Clinical Diagnosis: BMJ Books; 2002.

31. Straus SE RW, Glasziou P, Haynes BR. Evidence-based medicine: how to practice and teach EBM: Elsevier/Churchill Livingstone; 2005.

Chapter Three Appendix

3A. Meta-analyses included in the Review

Systematic Review	Cancer Biomarker(s) Reviewed	Type of Cancer	Number of Studies Meeting Inclusion Criteria	Primary Study Exclusions
Ruiz-Aragón et al (2010)	PCA3	Prostate	14	0
Zhang et al (2010)	Circulating cell free DNA	Lung	10	0
Luoa et al (2010)	Soluble mesothelin-related peptides	Mesothelioma	11	0
Zhou et al (2012)	Golgi protein 73	Liver	7	1 – retracted due to plagiarism
Tang et al (2011)	ProGRP	Lung	12	10 – non-English language
van der Bij et al (2011)	Multiple markers	Mesothelioma	35	1 – non-English language
Medeiros et al (2009)	CA-125	Ovarian	17	0
Gu et al (2009)	CA-125	Ovarian Recurrence	12	3 – non-English language
Ferrero et al. (2013)	CA-125 and Serum human epididymis protein 4	Ovarian	16	0
Gao et al. (2012)	Des-carboxy prothrombin (DCP)	Liver	19	1 – non-English language
Li et al. (2012)	Faecal tumour M2-pyruvate kinase	Colorectal	6	0
Lin et al. (2013)	Human epididymis protein 4	Ovarian	11	0
Lippi et al. (2012)	PSA	Prostate	11	0
Tonus et al. (2012)	Faecal tumour M2-pyruvate kinase	Colorectal	8	9 – 4 non-English language, 3 conference abstracts, 2 – no reference standard
Witjes et al. (2013)	Multiple biomarkers	Liver	7	0
Yang et al. (2011)	ProGRP	Lung	10	1 – non-English language

Wu et al. (2012)	HE4	Ovarian	9	0
Yu et al. (2012)	HE4	Ovarian	10	2 – non-English language
Li et al. (2012)	CA-125 and Serum human epididymis protein 4	Ovarian	8	0
Harvey et al. (2009)	PSA	Prostate	10	0
Morselli-Labate et al (2009)	serum IgG4	Pancreatic	7	0
Tan et al (2009)	CEA	Colorectal	19	1 – non-English language
Shen et al. (2012)	Vascular endothelial growth factor	Lung	7	3 – non-quantitative laboratory methods
Wan et al (2012)	MicroRNA	Pancreatic	7	0
Zhang et al. (2012)	Serum p53 Antibody	Oesophageal	13	2 – non-English language
Xia et al. (2010)	Survivin mRNA	Bladder	7	22 – non-English language
Ku et al. (2012)	Survivin mRNA	Bladder	13	1 – non-English language
Shen et al. (2013)	Circulating microRNAs	Lung	13	0
Total number of primary studies reviewed			**329**	

3B. Data Extracted from Meta-analyses

Basic study details:

- Target cancer and diagnostic test(s) under evaluation
- Whether an electronic search of more than one database was carried out and the diagnostic methodological terms included in the search strategy
- Reasons (and numbers where available) for the exclusion of primary studies
- Whether a quality assessment was carried out of the included primary studies and, if so, the criteria used

Analyses:

- Level of data extracted from primary studies e.g. a single 2 x 2 table, multiple 2 x 2 tables, or accuracy statistics for multiple categories

- Whether related positivity thresholds were also extracted and reported
- Which statistics were pooled and the methods was used to calculate them
- Whether forest plots were produced
- Whether the presence of a threshold effect and general study heterogeneity was explored and, if so, the methods were used to assess these
- Whether sub-group or adjusted analyses were carried out to account for possible drivers of heterogeneity
- Whether an sROC curves was produced and, if so, the method was used to produce it and the statistics derived from it
- Whether evidence of publication bias was explored and, if so, the methods used to achieve this
- The software used to carry out the analyses

3C. Data Extracted from Primary Studies

- Whether a table of patient descriptives in each disease group was reported
- Whether a graph showing the distribution of test values in each disease group was reported
- Whether an ROC curve and AUC has been reported
- The model(s) used to analyse accuracy i.e. binary, multi-level or result-specific
    - If a binary model has been used, the number of different thresholds at which accuracy has been reported and these thresholds were selected
    - If a multi-level model has been used, the number of categories for which accuracy has been reported and how categories were selected
- The summary statistics used to analyse the accuracy of the test
- Whether a multivariate analysis was conducted to assess the predictive capabilities of the biomarker having adjusted for key patient characteristics or other diagnostic tests available

Chapter Four

Interpreting quantitative diagnostic test results: a survey of GPs

4.1. Overview

Research Objective: To establish whether GPs prefer to interpret quantitative diagnostic test results based on a single threshold, multiple thresholds or result-specific predictive values for clinical decision making.

Methods: An online survey of a UK representative sample of 202 GPs was carried out. Respondents compared five alternative formats for guiding test result interpretation: 1) a single threshold 2) a rule-in and a rule-out threshold 3) a narrow range of 'uninformative' values 4) multiple categories 5) a graph of result-specific predictive values.

Results: Despite being current practice, the single threshold was preferred by only 12% of GPs. Rule-in and rule-out thresholds were considered the most helpful for clinical decision making (34% overall preference), with the result-specific predictive values and the multi-category formats also being popular (preferred by 23% and 21%, respectively).

Conclusion: Results suggest that GPs would find it helpful to have more information than offered by a single threshold interpretation of quantitative diagnostic test results.

4.2. Introduction

As demonstrated in the previous chapter, the binary model is typically adopted to evaluate the accuracy of quantitative diagnostic tests. A number of eminent researchers have brought this practice to question however, e.g. (1, 2) claiming that the dichotomisation of quantitative test scales is not representative of how clinicians interpret results in practice and that, in many clinical scenarios, methods that convey a greater depth of information for interpreting test results are required (3-5). The methods for interpreting quantitative diagnostic test results reviewed in **Chapter 2** provide viable alternatives to the single threshold approach.

One of these methods is to identify a third 'intermediate' category of test result (1). Although this approach still results in some information being lost (6), it improves on current practice by highlighting the test values that are incapable of informing a certain diagnosis. The exact definition of an intermediate test range is unclear in the literature. In some cases, it has been used to define a very limited range of test results that provide minimal impact on the probability of disease (7-9). More often, it is used to encompass a much wider range of results that fall between thresholds capable of ruling in and ruling out a diagnosis (10, 11).

Another approach, which conveys an even greater depth of information, is to identify multiple (>3) categories of test ranges where results are interpreted using category-specific accuracy summaries. For example, calculating likelihood ratios for each category, known as multi-level likelihood ratios, provides the opportunity to evaluate how the odds of disease evolve over the spectrum of the scale (3, 12). The use of multiple (between 5 and 7) categories has been shown to preserve the majority of the original data (13).

Another option is to avoid categorising the test scale completely and instead implement result-specific interpretations through the use of graphs or even electronic 'risk' calculators. This approach overcomes all concerns regarding data loss, but has the potential to confuse clinicians if they are uncomfortable interpreting graphs or handling probabilities.

Although these alternative methods undoubtedly convey a greater depth of information than a single threshold interpretation, the evidence base for the level of guidance that clinicians would prefer to interpret test results is lacking. The objective of this chapter is to gather evidence via a survey about GP preferences for different methods for interpreting quantitative diagnostic test results.

4.3. Methods

4.3.1. Data Collection and Sampling

Quotes were obtained from a number of market research agencies specialising in surveys of health practitioners for varying questionnaire lengths and sample sizes. To meet the restrictions of my budget, a survey length of 15 minutes and a sample size of 200 GPs was agreed upon. If 25% or 75% of the sample prefer the rule-in and rule-out thresholds to the single threshold option (current practice), a 95% confidence interval of ±6% would be obtained. A 50% preference would provide 95% confidence intervals with a 7% margin of error.

Doctors.net were commissioned to carry out the survey as they provided the most affordable quote, a fast turnaround of results, and access to an active community of practising GPs. As an online community of health practitioners rather than a traditional market research agency, the sampling procedure used by Doctors.net was slightly different to the commonly used invitation-response method. A general newsletter is emailed to all of the GPs on their membership list (close to 200,000 individuals) on a regular basis. The newsletter includes a research section with links to the surveys that they are currently running. Respondents received incentives for completing the survey (2000 ‘eSR’ points) which members can exchange for money on collection of enough points.

To participate in the survey, respondents had to be currently working in the UK as a general practitioner. They were consequently screened out of the survey if they did not work in the UK, if they were retired, or if they were a hospital doctor rather than a GP.

Quotas were set to ensure a regionally representative sample across the strategic health authorities (see Appendix B for breakdown of respondents by region).

4.3.2. Ethics

Doctors.net carries out all of their research in full accordance with the BHBIA (British Healthcare Business Intelligence Association) and the MRS (Market Research Society) code of ethics. At the start of the survey, respondents are warned not to disclose any information regarding specific patients and have to consent to being reported to the MHRA (Medicines and Healthcare products Regulatory Agency) if they are found to be in violation of this rule. Doctors.net were responsible for removing all respondent identifier details prior to releasing the dataset.

As it was not a requirement for respondents to be practising for the NHS, ethical approval from the NHS itself was not necessary.

4.3.3. Questionnaire Development

Choice of clinical scenario

To avoid GPs referring to existing clinical knowledge of a test, a fictional diagnostic test for the target disease was invented. A disease for which there is no specific laboratory test used in general practice was also therefore a priority, as we did not want respondents drawing comparisons to a test that is already available.

Selecting the disease for the clinical scenario was difficult as there were a number of possible options. The target disease needed to be sufficiently serious and rapid in its progression to require an immediate clinical decision from the GP: either to send the patient home or to refer them to secondary care. The early symptoms of the disease needed to be relatively non-specific and perhaps indicative of many conditions to warrant the use of a laboratory test in the diagnostic work-up.

Following advice from GPs in our department, appendicitis met these criteria; if left untreated, appendicitis can result in sepsis (peritonitis) which in some cases can cause patients to go into circulatory shock and die (14). Patients with suspected appendicitis can present with a range of symptoms, the most common complaint however being severe abdominal pain. There are also currently no laboratory tests used in general practice for appendicitis, with the key diagnostic criteria being clinical examination.

A brief overview of the literature on the diagnosis of appendicitis was carried out to inform the details of a simple clinical vignette of a patient presenting with suspected appendicitis. The diagnostic accuracy of existing clinical examination and symptoms was also identified in the literature (14-17) and, to ensure that the test would be worth carrying out compared to existing tests available, the accuracy of the fictional quantitative test was set to be slightly more accurate than currently used clinical signs. Based on existing literature, the probability of appendicitis in young males presenting with abdominal pains, fever and nausea is approximately 10% (16, 18, 19).

Based on these test characteristics and a pre-test probability of 10%, quantitative test results were simulated. The vignette was intentionally brief to prevent respondents from getting too caught up in the details of the clinical scenario, but had sufficient information for the respondent to decide whether the method of test result interpretation would be useful or not.

Options for test result interpretation

Five different formats for interpreting the results from a quantitative test were included for comparison based on the findings of the literature review in **Chapter Two**. Given that the priority in general practice is to rule out disease (20), this was prioritised when determining where the thresholds should be placed in options 1-4.

The decision to use post-test probabilities in the interpretation options was based on suggestions that clinicians would find them easier to interpret and apply in clinical practice than other test

characteristics (21). For example, sensitivity and specificity are not intuitive to use in a clinical setting and fail to account for possible bias in disease spectrum (22). Likelihood ratios were a viable alternative; however given their limited use in the primary diagnostic research reviewed inChapter Three, likelihood ratios may be unfamiliar to clinicians. The use of natural frequencies to convey the post-test probability of appendicitis instead of percentages was based on extensive evidence suggesting that clinicians find natural frequencies easier to interpret than probabilities (23-25). I initially considered using qualitative interpretations of the categories such as 'rule-in', 'rule-out', 'inconclusive', but this risked potentially biasing interpretation preferences as the terms used would be subject to debate.

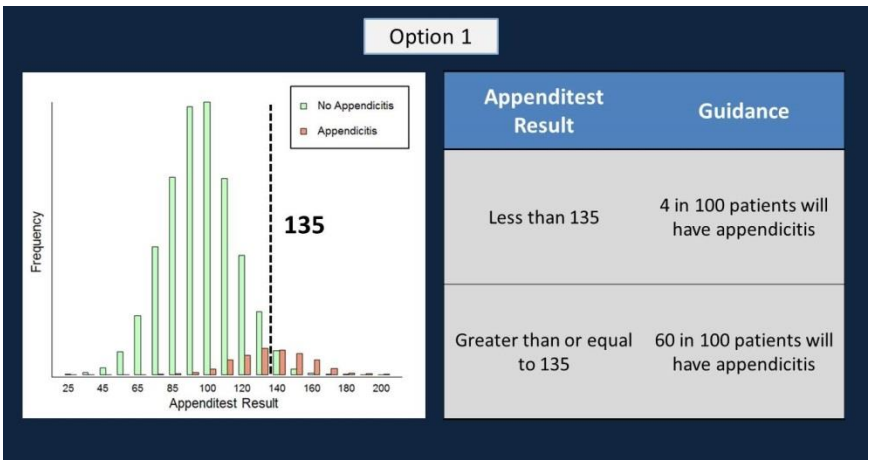


Figure 4.1. Option 1: A single 'optimal' threshold for ruling out appendicitis

The systematic review in Chapter Three demonstrated that the selection of a single 'optimal' threshold is current practice for evaluating the accuracy of quantitative diagnostic tests. Opinions regarding this format were therefore of particular interest, and respondents were asked to grade this format (using a 7-point Likert scale) in terms of its helpfulness when deciding how to manage a patient with suspected appendicitis. Respondents were also asked to describe what they liked and disliked about this format (Q1).

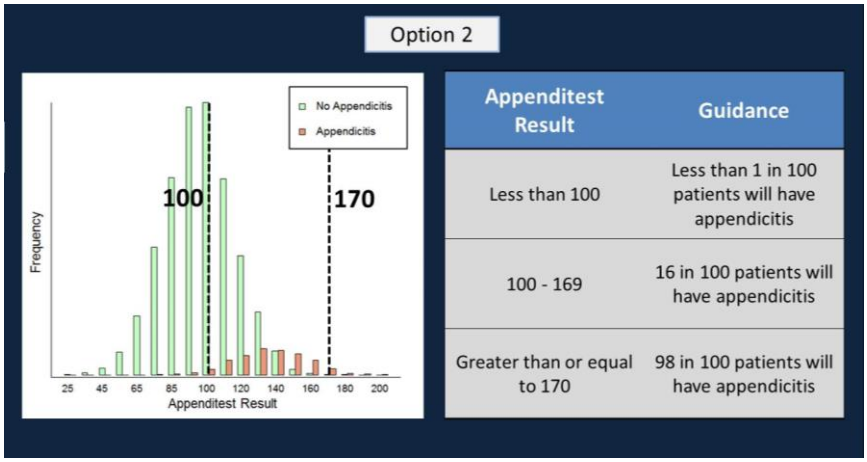


Figure 4.2. Option 2: Three categories of test result: A rule-in and a rule-out threshold delimiting an intermediate range of values

Option 2 represents calls in the literature for the introduction of an intermediate range of values (26, 27). This format requires the identification of two thresholds; an upper threshold sufficiently high to rule in disease and a lower threshold sufficiently low to rule out disease (28). As the primary focus of this thesis is to explore the value of using intermediate test ranges, our primary outcome for this survey was the proportion preferring this rule-in/rule-out format to the currently used dichotomous format.

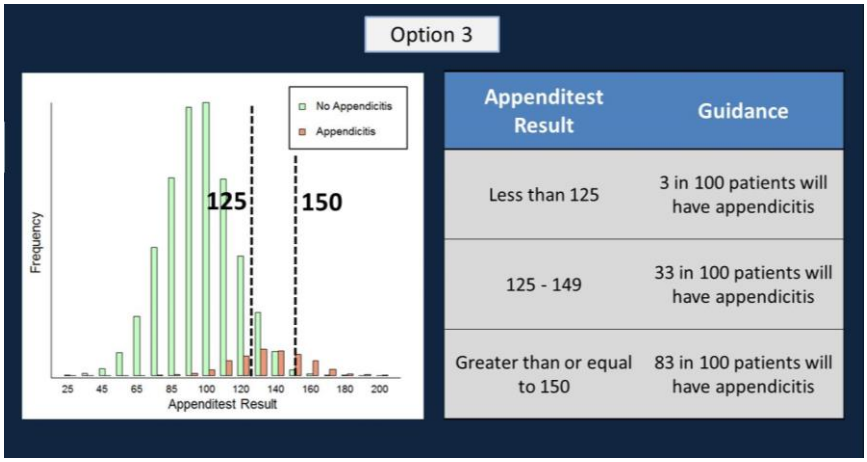
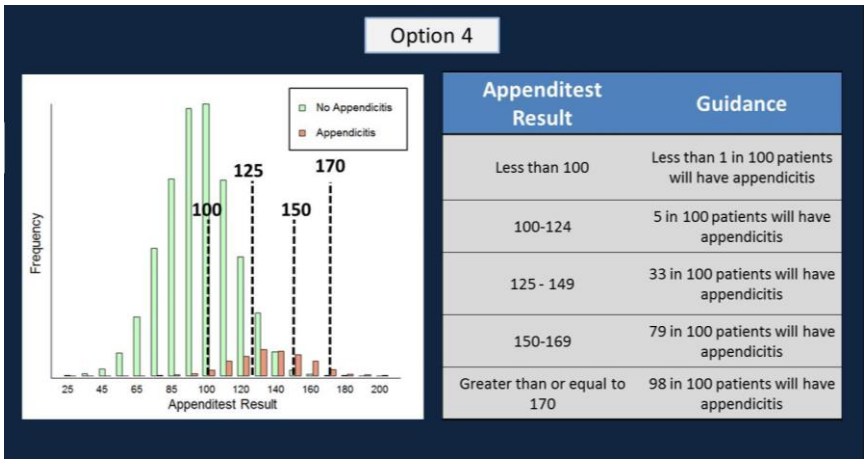


Figure 4.3. Option 3: Three categories of test result: a narrow intermediate range of values which hold little predictive value

Option 3 represents an alternative interpretation of an intermediate range identified in the literature. Intermediate test results are defined as a narrower range of values that are relatively uninformative regarding the disease status of the patient (7, 9, 29).



Appenditest Result	Guidance
Less than 100	Less than 1 in 100 patients will have appendicitis
100-124	5 in 100 patients will have appendicitis
125 - 149	33 in 100 patients will have appendicitis
150-169	79 in 100 patients will have appendicitis
Greater than or equal to 170	98 in 100 patients will have appendicitis

Figure 4.4. Multiple categories: the rule-in and rule-out categories from option 2 and the narrow intermediate range from option 3

Option 4 represents the increase in popularity, particularly in the evidence-based medicine field, for the use of multiple categories of test result to provide a greater depth of information regarding the accuracy of the test.

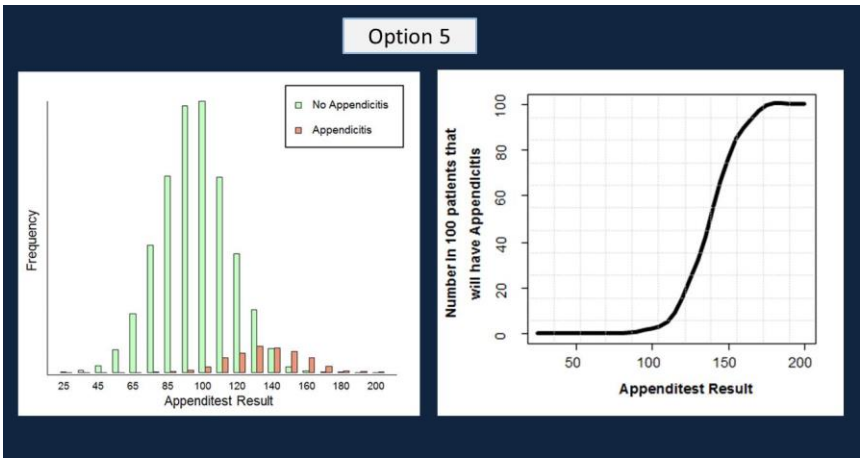


Figure 4.5. Option 5: No categories: a continuous plot of natural frequencies

Option 5 does not involve any categorisation of the test scale, and presents the post-test probability of appendicitis across all thresholds in a graphical format. Risk assessment charts and calculators are frequently used in clinical practice and therefore could feasibly be used in diagnostic test

interpretation (21). This would negate the need to determine how many thresholds should be used and where they should be placed.

Piloting the questionnaire

To ensure that the questionnaire was fully tested before carrying out the survey, 20 GPs working at the Department of Primary Care Health Sciences, University of Oxford were invited via email to participate in a pilot study, of which 17 consented. These GPs are part-time academic researchers and have varying levels of expertise in diagnostic methodology. The pilot survey was designed and distributed using Surveymonkey.com, a web-based online survey tool.

Initially, 5 GPs completed the survey with the researcher (BS) present and were asked to talk through their thoughts at each question of the survey. The questionnaire was revised based on their feedback (see questionnaire development section for details), and then the same 5 GPs completed the questionnaire again to provide feedback on any remaining issues.

The remaining 12 GPs completed the survey independently. An additional section was added to the end of the survey to allow respondents to feedback any issues that they had with the questionnaire itself. The survey was sent out in batches so that any changes made as a result of the pilot findings could then be tested by the remaining respondents. Before agreeing on the final questionnaire, all results of the pilot were sense-checked and changes were made where needed.

Outline of the final questionnaire

The key sections of the questionnaire will now be described, along with the revisions made as a result of the pilot study. The final questionnaire with each section highlighted can be found in Appendix 4A.

Introductory text: Doctors.net have a standardised introduction that is used for all of their surveys which details their privacy policy, the incentives available for completing the survey, and explanation of the respondent's responsibilities concerning patient confidentiality. Following this,

respondents are required to provide their consent for participating in the survey. The purpose of the study is then described. This text was revised slightly during the initial pilot phase to make sure respondents were aware that they were going to be asked to compare five different formats which differ in terms of the threshold information presented.

Screening questions: Initial screening questions (S1-S5) were required to ensure respondents met the inclusion criteria, followed by some questions obtaining basic demographic details. Information regarding respondent age, gender, region, and role in general practice was obtained.

Clinical Scenario: The clinical scenario and fictional quantitative test is introduced, alongside a histogram of typical test results broken down by the 'healthy' population and 'diseased' population. Originally, the histogram was only presented at the beginning of the survey. However, in the pilot study, respondents reported that they would find it useful to have the histogram to refer to throughout the survey to help them differentiate between and understand the different formats.

Comparison of the alternative interpretation methods: In the pre-pilot version of the questionnaire, respondents were asked to simply think about the quantitative tests that they use regularly in general practice when comparing the options. However, the initial pilot phase revealed that this approach failed to engage respondents and made it very difficult for them to answer the questions without a clinical scenario as a reference. The questionnaire was then revised so that respondents were presented with a specific clinical scenario to provide context for the task (see 'Clinical Scenario' in questionnaire). This section of the questionnaire required respondents to compare the alternative interpretations and state which format they found most helpful. Ideally, respondents would have compared each option head-to-head, however due to restrictions in survey length and respondent fatigue, this was not appropriate. Instead, the questionnaire was structured so that all respondents compared current practice (option 1) to the rule-in/rule-out thresholds, and then their preferred interpretation was carried forward for comparison with the

remaining options (see Figure 4.6). At each stage, respondents were again asked to describe what they liked and disliked about the interpretations.

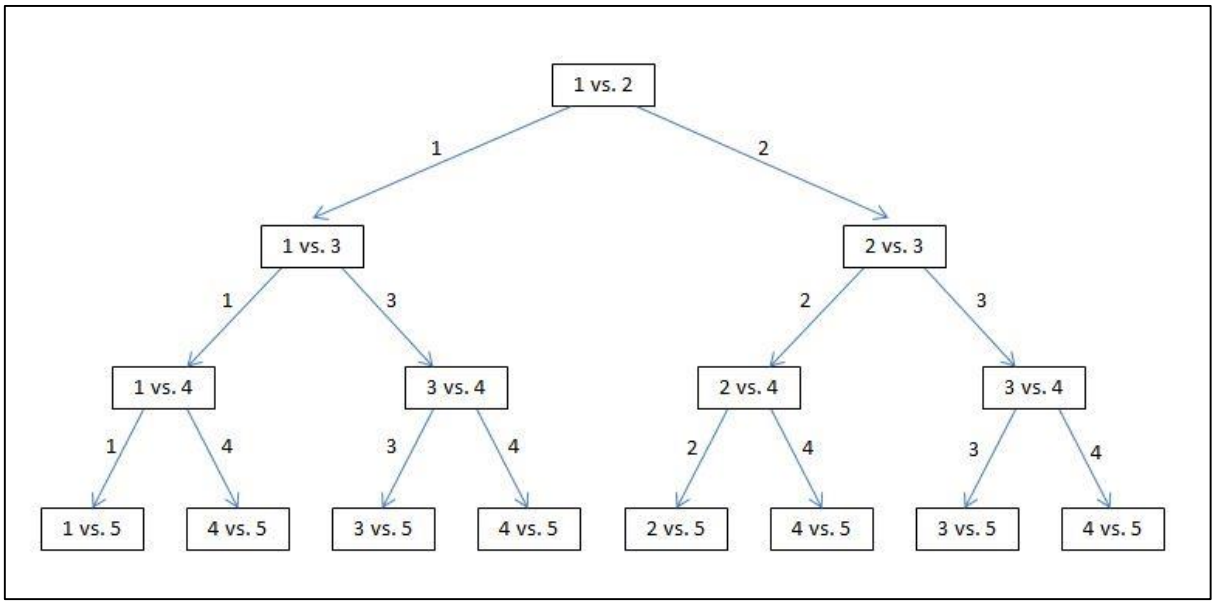


Figure 4.6. Flowchart of interpretation options presented to respondents based on their preference selections

Possible influential factors: The remainder of the questions were included to try to ascertain whether there were any aspects of the interpretation formatting and/or clinical scenario that may have influenced preferences. These questions related to the influence of having the histogram presented alongside each guidance format (Q12 and Q12a), how specific respondent preferences were to the clinical scenario presented (Q13), whether the respondent could think of a preferable format that had not been included in the options presented (Q14), whether they would have also liked information regarding the probability of the patient not having the disease (Q15), and whether natural frequencies or percentages are preferable (Q16).

Closing Questions: The respondents were then asked if there was any further feedback they had regarding the survey and whether they could be contacted for further clarification of responses if necessary.

4.3.4. Analyses

Proportions with 95% confidence intervals were calculated for all questions. For questions on a Likert scale, averages were reported. Responses to Q8 to 11 were summarised to ascertain respondents' most preferred interpretation.

A full qualitative analysis was not carried out as free text questions were optional but, where possible, key themes were extracted from the free text responses at each comparison to provide reasoning for respondents' likes and dislikes about each format.

4.4. Results

A full breakdown of the survey results can be found in the Appendix 4B.

4.4.1. Respondent Characteristics

202 respondents completed the survey, with the quotas on region ensuring a sample of GPs that were regionally representative of the United Kingdom (see Appendix 4B for breakdown). Table 1 gives key demographic details for the sample. Respondents of each gender were well represented and a reasonable spread of ages was obtained. Over half (64%) of respondents were principal GPs. A response rate for the survey cannot be reported as targeted invitations were not sent out.

	% (N) out of 202 respondents
Gender	
Male	**59% (119)**
Female	**41% (83)**
Age	
Under 30 years	**1% (2)**
20-39 years	**42% (84)**
40-49 years	**33% (66)**
50-59 years	**20% (41)**
60 or over years	**5% (9)**
GP Type	
GP principal	**64% (130)**
Locum GP	**12% (24)**
Salaried GP	**24% (48)**

Table 4.1. Characteristics of Respondents

4.4.2. Interpretation Preferences

Respondents overall preferences in terms of which interpretation format they thought would be most helpful to their clinical decision making are shown in Table 4.2. The least popular interpretation was the uninformative range of test values (9%, n=19/202), closely followed by the single threshold (12%, n=24/202). Around a third of respondents (34%, n=69/202) found the rule-in and rule-out thresholds the most helpful, with just over a fifth of respondents preferring either the multiple categories (21%, n=43/202) or the graph of natural frequencies (23%, n=47/202). Just under half of those who preferred the graph of result-specific natural frequencies overall always selected the format with the most thresholds prior to that (n=20/202), suggesting that there was a subset of respondents that always opted for a greater depth of information.

Interpretation Options	% (N)	95% CI
Option 1: Single threshold (current Practice)	12% (24)	7.5% - 16.5%
Option 2: Rule-in and rule-out thresholds	34% (69)	27.5% - 40.5%
Option 3: Uninformative range of test values	9% (19)	5.1% - 13.0%
Option 4: Multiple thresholds	21% (43)	15.4% - 26.6%
Option 5: Graph of result-specific predictive probabilities	23% (47)	17.2% - 28.8%

Table 4.2. Overall preferences for interpretation options

Option 1: Single threshold (current practice)

When asked to grade the helpfulness of the single threshold interpretation at the start of the survey on a 7-point scale ranging from 'extremely unhelpful' to 'extremely helpful', half of the respondents stated that they found this interpretation 'somewhat helpful' (49.5%). The rest of the respondents gave a range of responses, with a fifth of respondents stating that they found the interpretation unhelpful to some extent (19.8%, n=40/202).

"There is a large false positive rate for this test so patients and myself would be worried by it rather than reassured" Q1a, practitioner ID 36. Selected "Not sure" at Q1

"Very clear and easy to interpret result" Q1a, practitioner ID 28. Selected "Very helpful" at Q1

"Not specific and certainly not sensitive" Q1a, practitioner ID 150. Selected "Not very helpful" at Q1

Of the participants that preferred the single threshold interpretation to the rule-in and rule-out thresholds (18.3%, 37/202), the key justification was that the single threshold interpretation was simpler to understand and the trichotomous format was too complicated. Another comment that frequently came up amongst these respondents is that the use of rule-in and rule-out thresholds introduced more uncertainty.

"Option 2 just increases uncertainty" Q2a, practitioner ID 192

Option 2: Rule-in and rule-out thresholds

A majority of 74.3% (150/202) respondents stated that the rule-in and rule-out thresholds were more helpful than having a single threshold. Much of the free text at Q2a among these respondents related to their preference for a greater depth of information to aid test result interpretation. More specifically, respondents found it helpful to have the areas of the test scale permitting a more confident diagnosis demarcated (the extremes), allowing appendicitis to be effectively ruled in or ruled out.

"I like the additional category in option 2, which can then quickly be interpreted as 'unlikely, possible and likely, I think this is more realistic" Q2a, practitioner ID 191

"I like the more definite bands at each end, with grey area in the middle - easy to understand and apply" Q2a, practitioner ID 89

Respondents also found it useful to know the section of the test scale that does not provide such clear-cut diagnostic guidance so that they can fully understand what a result in this 'grey area' means and when they need to rely more on alternative diagnostic clues.

"Option 2 gives a better idea what a mid-range result might mean" Q2a, practitioner ID 223

"Like: clarity of interpretation of results. Any test will have a 'grey area' displaying results in this manner I find shows exactly where the grey area is quite nicely" Q2a, practitioner ID 24

Some respondents stated that the use of rule-in and rule-out thresholds resulted in improved sensitivity and specificity, predictive capabilities, or a decrease in the number of misclassifications compared to the binary option.

"Less false positive and false negative with option 2, so is more reliable" Q2a, practitioner ID 177

"Division into 3 sub-groups gives better specificity and sensitivity" Q2a, practitioner ID 70

The free text at Q6a and Q9a was examined to explore why this interpretation was preferable to those that provided more detail for 34% of respondents. The key explanation for preferring the three-category format to the five-category format is that the latter is too complicated, with information superfluous to the diagnostic decision making process.

"Option 2 is clear. The added information in option 4 does not help my decision making" Q6a, practitioner ID 138

"There is rather too much info in option 4. Option 2 much simpler and just as informative and helpful" Q6a, practitioner ID 11

Option 3: Uninformative range of test values

Option 3 was preferred overall by the fewest proportion of respondents (9%, n=19/202). When comparing the two intermediate range interpretations (options 2 and 3), reasons for preferring for rule-in and rule-out thresholds typically related to its ability to provide stronger evidence regarding the presence of disease, whereas the categories delineated in the uninformative range option did not lead to any clear conclusions.

"Option 2 thresholds show both likely 'rule-in' and 'rule-out' thresholds. The thresholds in option 3 are not clear cut enough in any direction" Q4a, practitioner ID 11

"Option 2 is more useful for giving a definitive answer, the boundaries in 3 still leave a lot of uncertainty"

Q4a, practitioner ID 206

Option 4: Multiple thresholds

Of those respondents who preferred the multiple category interpretation overall (21.3%, 43/202), the key reason provided for finding the multiple categories more helpful than just the rule-in and rule-out thresholds was that a greater depth of information is available, but that it still remained straightforward to understand. In particular, the further division of the wide intermediate range between the rule-in and rule-out thresholds provided more clarity. Their aversion to the result-specific graphical interpretation was due to a lack of confidence in being able to correctly interpret the graph.

"I still think they're both simple to understand, and the option 4 would be more helpful when trying to interpret those intermediate results" Q6a, practitioner ID 100

"In option 4, there is increased complexity, but there is also an increased stratification which is good"

Q6a, practitioner ID 125

"Graphical illustration has more information but perhaps there is a risk of misinterpretation at a quick glance" Q11a, practitioner ID 126

Option 5: Graph of continuous natural frequencies

A common reason for preferring the graph of continuous natural frequencies option was that it allowed GPs to read off a disease probability specific to each test result. Some GPs thought that the visual representation would also be useful when explaining test results to patients.

"In option 5, no dilemma about working out the exact probability, so all values can be accurately considered" Q11a, practitioner ID 100

"Useful to show patients a graph, and seems more logical than ranges of figures" Q11a, practitioner ID 68

Interestingly, there was a subset of respondents who, despite finding the graphical interpretation the most helpful, also called for additional threshold guidance or who tried to use the shape of the

graph to create categories of test results. Presuming it is feasible to provide both the graph of continuous natural frequencies and some threshold guidance, our results indicate that 41% of respondents (n=82/202) would find the rule-in and rule-out thresholds the most helpful, followed by 31% (n=63/202) who would find the multi-category interpretation most helpful.

"Area where line is steep is the "greyer area" and need to consider other ways to diagnose, but the flatter parts of the curve are good for predicting likely/ unlikely diagnosis" Q10a, practitioner ID 71

"I like option 5 but would draw lines for the cut off boundaries in 2. This gives an idea for the presumably frequent readings between 100 and 169" Q9a, practitioner ID 190

Similar themes emerged when the rule-in and rule-out thresholds were preferred to the graph of result-specific probabilities. Many respondents stated that the graph was too confusing and difficult to interpret, with a few being concerned about misreading it.

"I still think option 2 is the easiest to use, but option 5 would be next as it does give more information, but might be easier to misread" Q9a, practitioner ID 223

"Option 2 much easier to use and use quickly. Easier if wanted to share the info with a patient as well" Q9a, practitioner ID 11

Other aspects of the interpretation formatting and/or clinical scenario

Respondents generally found the histogram more helpful than unhelpful (59.4% vs. 21.8%), with around a 35% of respondents stating that their preferences in terms of interpretation format may have been different if the histogram had not been included to illustrate the results.

Around a third of respondents felt that their preferences for the interpretation formats may have been different if they had been presented with a clinical scenario with a rarer disease (n=62/202, 30.7%), a scenario with a more common disease (n=73/202, 37.1%), a scenario with a less critical condition (n=65/202, 32.2%) or a scenario with a more critical condition (n=75/202, 37.1%). Interestingly however, only 8.4% (17/202) thought the prevalence of disease in general would be

of impact, and only 10.4% (n=21/202) thought the seriousness of the disease in general would be influential.

Opinion was divided regarding whether the additional reporting of the number of patients that would not have appendicitis would have been useful (42.6% in favour). The majority of respondents had no preference for the disease probability to be presented as a percentage or a natural frequency (48.5%, n=98/202), with around a quarter preferring one option over the other (28.7% preferred natural frequencies and 22.8% preferred percentages).

4.5. Discussion

The results of this survey question the methods currently being used to interpret quantitative diagnostic tests. The majority of general practitioners (88%, 95% CI: 83% - 92%) reported that they would find a greater depth of information helpful in their clinical decision making than is provided by a single threshold. The most popular method of interpretation was the identification of rule-in and rule-out thresholds (34%, 95% CI: 28% - 41%), but there were also notable proportions of respondents finding the multiple categories and the graph of result-specific predictive values the most helpful (21% and 23%, respectively).

These findings have implications not only on clinical guidelines, but also on the methods used in the research underpinning them. The systematic review in the previous chapter demonstrated that current practice in diagnostic research is to evaluate the accuracy of quantitative tests at a single or multiple binary thresholds. The findings in this survey raise the question therefore of whether the methods we use in research allow us to produce evidence that meets the preferences of clinicians for test result interpretation in practice.

Respondents reported that they preferred the rule-in and rule-out format due to its ability to allow them to make a definitive diagnostic decision. The uninformative range of test results was not considered helpful on its own (only 9% preferred this format overall). However, when presented alongside the rule-in and rule-out thresholds to form multiple categories, highlighting this range of test result was found to be useful. Respondents who preferred this multi-category interpretation found it beneficial to have the wide intermediate range further broken down. A substantial proportion of those who preferred the categorical interpretations reported that they were not comfortable reading probabilities from a graph. In contrast, many who preferred the graphical presentation overall stated that the availability of result-specific probabilities outweighed this extra complexity. One possible solution that would meet the requirements of the majority of

respondents would be to provide both the result-specific probabilities, in addition to some categorical guidance that links directly to their patient management options.

A surprising theme that emerged in the free text was a general dissatisfaction with the accuracy of the fictional test results. Existing evidence was obtained to ensure that the simulated accuracy was better than any of the current clinical tests available for appendicitis. These opinions support findings that clinicians tend to overestimate the accuracy of tests that are frequently used, and struggle to fully comprehend accuracy statistics such as sensitivity and specificity (30, 31). The enhanced understanding in the context of this survey could be down to a number of factors: the visual representation of the performance of the test as a histogram, the use of post-test probabilities instead of other metrics of diagnostic accuracy, or the greater depth of information available in some of the interpretation formats.

Although it was not originally intended that the survey would be focused on a single clinical scenario, the feedback from the pilot phase indicated that this was a necessary adjustment. Respondents insisted that it was impossible to ascertain if guidance is generally helpful as there are so many anomalies in general practice. This is an interesting finding in itself as it indicates that it is unlikely that there is a single solution appropriate to all clinical scenarios. The results of Q13 support this finding, with the majority of respondents (93%) selecting at least one of the hypothetical alternatives as being potentially influential to their preferences for test interpretation. This perhaps highlights inadequacies in primary diagnostic accuracy research as we saw in Chapter Three that the same methods are commonly used regardless of the clinical scenario. This more focused approach to the survey does, however, limit the generalisability of the results, and further research needs to be carried out based on alternative scenarios to ensure that the findings of this study are not particular to the example of appendicitis.

The use of post-test probabilities throughout the survey means that these figures were specific to the pre-test probability for that particular patient population. One potential issue with this

approach is that there is a lack of evidence regarding prevalence estimates for specific sub-groups of patients (32). However, with the increasing availability of national databases of clinical records, such as the CPRD (Clinical Practice Research Datalink), this issue could be overcome. Ordinarily, confidence intervals should be included to show the degree of uncertainty associated with each post-test probability estimate. These were omitted in the present survey as the test results used in the survey were simulated, and therefore the width of the confidence intervals would have only been representative of the size of the simulated dataset. In practice, confidence intervals could be drawn from meta-analysis results, but research would also need to be carried out to see if this would compromise GPs ability to understand the guidance.

The inclusion of the histogram throughout the survey was also a much debated decision. The respondents in the pilot stated that they found it helpful in distinguishing between the different formats; however this visual representation would not ordinarily feature in a clinical guideline. Only a third of respondents thought that their preferences would have been different had the histogram not been included, and nearly 60% found the histogram useful for interpreting test results.

Given the survey was targeted at general practitioners; the thresholds were positioned to prioritise ruling out appendicitis. There are currently no standardised methods for identifying an intermediate range or multiple categories of test values on a quantitative test scale, and therefore the selection of thresholds were based on clinical reasoning rather than validated statistical methods. There is the possibility that the location of the thresholds had an influence on respondent preferences, and this is something that would need to be explored in a follow-up study.

We hope to use the results of this survey to justify a much larger piece of work exploring alternative clinical scenarios. If there are cases where a single threshold is the most helpful format, priorities in terms of ruling in or ruling out disease need to be identified and ideally quantified so that the location of this threshold represents the diagnostic challenge. Decision analysis is a common method for this latter type of analysis. In cases where rule-in and rule-out thresholds are found to

be useful, the appropriate location of these thresholds needs to be ascertained i.e. how extreme do clinicians want these thresholds to be? Furthermore, it would be interesting to see what clinicians do with the intermediate results – whether they are simply ignored or used to trigger the implementation of certain patient management strategies?

There are a number of different ways that the interpretation formats could have been presented and further research is required to determine whether the approach adopted was optimal. Reassuringly, respondents did not propose any viable alternatives at Q14 (bar one suggestion for an electronic calculator). There are, however, some additional pieces of information that may have been useful. For example, the inclusion of the proportion of patients estimated to fall in each category may provide further insight. Additionally, some qualitative guidance regarding patient management for those falling in each category of test result may also be useful.

At the end of the survey, respondents expressed their interest in and enthusiasm for the topic. This is clearly an issue that clinicians feel strongly about and more work needs to be done to make guidelines regarding diagnostic tests more useful. The results of this survey in general highlight the importance of engaging with clinicians when producing threshold guidance for test interpretation. This research should precede diagnostic test evaluations so that evidence for the accuracy of the test based on these threshold preferences can be obtained. There are currently no standardised methods for identifying more than one threshold on a diagnostic test scale, and the next chapter will compare alternative statistical approaches to defining an intermediate range of test values.

What this chapter adds:

- This survey explores what information general practitioners would like to help them interpret quantitative diagnostic tests results
- The results suggest that general practitioners would find a greater depth of information helpful in their clinical decision making than is provided by a single threshold
- The most popular method of interpretation was the identification of rule-in and rule-out thresholds, although other formats were also popular suggesting that it is likely to be dependent on the clinical scenario

4.6. References

1. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. J Clin Epidemiol. 1990;43(1):109-13.

2. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making. 1987 Apr-Jun;7(2):107-14.

3. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet. 2005 Apr 23-29;365(9469):1500-5.

4. Straus SE RW, Glasziou P, Haynes BR. Evidence-based medicine: how to practice and teach EBM: Elsevier/Churchill Livingstone; 2005.

5. Sonis J. How to use and interpret interval likelihood ratios. Family Medicine Journal. 1999;31(6):432-7.

6. Heckerling PS. Information content of diagnostic tests in the medical literature. Methods Inf Med. 1990 Jan;29(1):61-6.

7. Guyatt G, Bass E, Brill-Edwards P, Holbrook A, Jaeschke R, Elizabeth Juniper M, et al. Users 'Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test: I B. What Are the Results and Will They Help Me In Caring for My Patients? Journal ofAmerican Medical Association, 271 (9). 1994:703-7.

8. Teitelbaum JS, Eliasziw M, Garner M. Tests of motor function in patients suspected of having mild unilateral cerebral lesions. Can J Neurol Sci. 2002 Nov;29(4):337-44.

9. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med. 2003 Jul;29(7):1043-51.

10. Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. Int J Epidemiol. 2003 Apr;32(2):304-13.

11. Greiner M, Sohr D, Gobel P. A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. J Immunol Methods. 1995 Sep 11;185(1):123-32.

12. Bowden SC, Loring DW. The diagnostic utility of multiple-level likelihood ratios. J Int Neuropsychol Soc. 2009 Sep;15(5):769-76.

13. Rifkin RD. Maximum Shannon information content of diagnostic medical testing. Including application to multiple non-independent tests. Med Decis Making. 1985 Summer;5(2):179-90.

14. Bundy DG, Byerley JS, Liles EA, Perrin EM, Katznelson J, Rice HE. Does this child have appendicitis? JAMA. 2007 Jul 25;298(4):438-51.

15. Ashdown HF, D'Souza N, Karim D, Stevens RJ, Huang A, Harnden A. Pain over speed bumps in diagnosis of acute appendicitis: diagnostic accuracy study. BMJ. 2012;345:e8012.

16. Kwok MY, Kim MK, Gorelick MH. Evidence-based approach to the diagnosis of appendicitis in children. Pediatric emergency care. 2004 Oct;20(10):690-8; quiz 9-701.

17. Barraclough K, Du Toit J, Budd J, Raine JE, Williams K, Bonser J. Avoiding Errors in General Practice: John Wiley & Sons; 2012.

18. Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The epidemiology of appendicitis and appendectomy in the United States. Am J Epidemiol. 1990 Nov;132(5):910-25.

19. van Randen A, Bipat S, Zwinderman AH, Ubbink DT, Stoker J, Boermeester MA. Acute appendicitis: meta-analysis of diagnostic performance of CT and graded compression US related to prevalence of disease. Radiology. 2008 Oct;249(1):97-106.

20. Heneghan C, Glasziou P, Thompson M, Rose P, Balla J, Lasserson D, et al. Diagnostic strategies used in primary care. BMJ. 2009;338.

21. Foy R, Warner P. About time: diagnostic guidelines that help clinicians. Qual Saf Health Care. 2003 Jun;12(3):205-9.

22. Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Academic Radiology. 2003 Jun;10(6):670-2.

23. Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Academic medicine: journal of the Association of American Medical Colleges. 1998 May;73(5):538-40.

24. Hoffrage U, Kurzenhäuser S, Gigerenzer G. [How can one improve the understanding and communication of the importance of medical test results?]. Zeitschrift für ärztliche Fortbildung und Qualitätssicherung. 2000 October;94(9):713-9.

25. Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. Communicating statistical information. Science. 2000;290(5500):2261.

26. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. Journal of clinical epidemiology. 1990;43(1):109-13.

27. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Medical Decision Making. 1987 April 00;7(2):107-14.

28. Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. International Journal of Epidemiology. 2003;32(2):304-13.

29. Knottnerus JA. The Evidence Base of Clinical Diagnosis: BMJ Books; 2002.

30. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. BMJ. 2002 Apr 6;324(7341):824-6.

31. Bobbio M, Fubini A, Detrano R, Shandling AH, Ellestad MH, Clark J, et al. Diagnostic accuracy of predicting coronary artery disease related to patients' characteristics. J Clin Epidemiol. 1994 Apr;47(4):389-95.

32. McCowan C, Fahey T. Diagnosis and diagnostic testing in primary care. The British Journal of General Practice. 2006;56(526):323.

Chapter Four Appendix

4A. Final Questionnaire

Introductory text

PAGE 1： Doctors.net Intro

Doctors.net.uk invites you to take part in a short survey commissioned by an academic researcher to investigate guideline formats for using the results from quantitative diagnostic tests subjects.

The survey should take up to 10 minutes to complete and all members completing the survey in full will receive 2,000 eSR points.

Please read the following text, which further explains the key aspects of this research:

- I understand that this research is commissioned by an academic researcher and is being carried out within the code of conduct of the Market Research Society and the British Healthcare Business Intelligence Association
- Doctors.net.uk will comply with all UK laws protecting your personal data and the British Healthcare Business Intelligence Association and Market Research Society guidelines
- The research is not intended to be promotional and any information presented is done so solely to explore reactions to such information
- Your responses will be totally anonymous and confidential
- Respondents have the right to withdraw from the interview at any time during the interview process and to withhold information as they see fit
- The aggregated findings of this research may be used for promotional purposes, such as Public Relations publications, however at no stage will it be possible to identify any participants

All results will be anonymised in accordance with Doctors.net.uk's zero-tolerance privacy policy and the Market Research Society Code of Conduct.

If you wish to see Doctors.net.uk privacy policy (which includes reference to the use of cookies), please click here.

http://about.doctors.net.uk/Assets/Privacy-Policy

If you wish to contact us about this survey, here are our contact details.

Researcher: Peter Constable Email: SurveyHelp@mess.doctors.org.uk

Contact reference for inclusion in the email subject: 05205095

Please confirm that you have read and understood this information

Yes

No ***THANK AND CLOSE***

PAGE 2: Consent

You are about to enter a market research interview. We are now being asked to pass on to our client details of adverse events and / or product quality complaints or potential counterfeits relating to their products that are mentioned during the course of market research interviews. Although this is an online market research interview and how you respond will, of course, be treated in confidence, should you mention an adverse event relating to a specific patient or group of patients, a product quality complaint or a potential counterfeit product, we will need to report

this, even if it has already been reported by you directly to the company or the UK regulatory authority (MHRA) using the MHRA's 'Yellow Card' system. In such a situation you will be contacted to ask whether or not you are willing to waive the confidentiality given to you under the Market Research Code of Conduct specifically in relation to that adverse event and / or product quality complaint. Everything else you contribute during the course of the interview will continue to remain confidential, and you will still have the option to remain anonymous if you so wish.

TC2 Are you happy to proceed with the interview on this basis?

- I would like to proceed and protect my anonymity
- I would like to proceed and give permission for my contact details to be passed on to the Drug Safety department of the company if an adverse event and/or product quality complaint is mentioned by me during the survey
- I don't want to proceed and wish to end the interview here **THANK AND CLOSE**

PAGE 3: Description of study

Thank you very much for participating in this study.

The aim of the following survey is to investigate guideline formats for using the results from quantitative diagnostic tests (i.e. tests that are either on a continuous scale, such as blood glucose measured in mg/L, or tests that are on an ordinal scale, such as the Beck's Depression Score). The diagnostic test that will be used throughout this survey is fictional.
You will first be presented with a clinical scenario and then be asked to compare 5 different guideline formats for interpreting a quantitative diagnostic test. These guidelines will differ in terms of the number of thresholds reported.

Screening questions

Ask all screeners on separate page

S1 Speciality confirmation

Are you a...

Single question, drop down list

GP
Hospital Doctor **THANK AND CLOSE**

S2 Region - Where are you currently practicing?
Single question, drop down list

North West SHA
North East SHA
Yorkshire & Humber SHA
East Midlands SHA
West Midlands SHA
East of England SHA
London SHA

South East Coast SHA
South Central SHA
South West SHA
Scotland
Wales
Northern Ireland
Retired THANK AND CLOSE
Not working in the UK THANK AND CLOSE

S3 GP type

Which of the following best describes your role?

Single question, drop down list

GP Principal
Salaried GP
GP Registrar
Locum GP
Other THANK AND CLOSE

S4 Gender

Are you...

Single question, drop down list

Male
Female

S5 Age

Are you …

Single question, drop down list

Under 30
30 – 39
40 – 49
50 – 59
60 or over

Clinical Scenario

Show on separate page

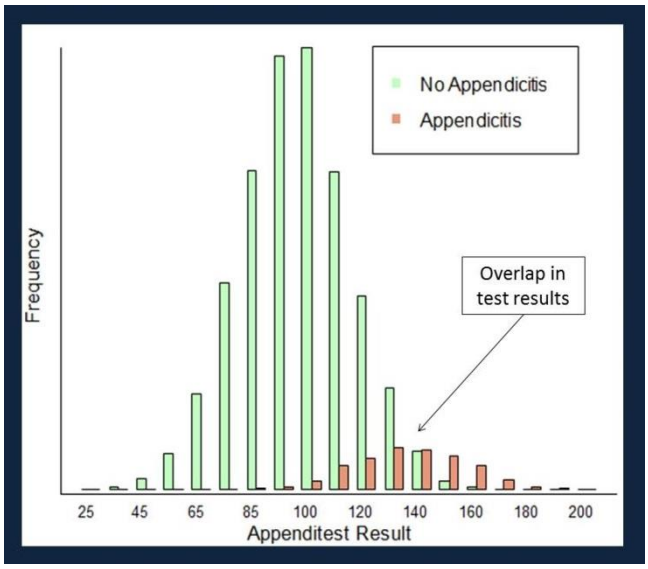
A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

A new point of care serum biomarker, 'Appenditest', has been recommended for the diagnosis of appendicitis and the device has just been introduced at your practice.
You carry out the test and refer to clinical guidelines to help you decide how to use the test result in your decision making.

New page

You are now going to be asked about 5 different guideline formats to help you use the result from Appenditest in your patient management decision.

To illustrate the difference between each format, you will be shown the following histogram of typical results from Appenditest for patients with and without appendicitis. When considering each option, please think about which format would be most useful to you when deciding how to manage a patient with suspected appendicitis in general practice.

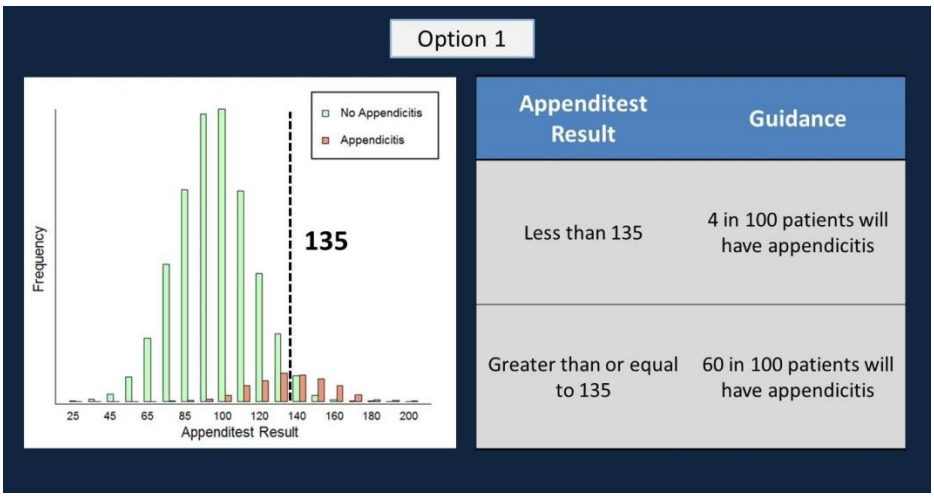


Comparison of the alternative interpretation methods

Q1 How helpful is the guideline format

A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

On the following scale, please rate how helpful this guideline format would be when deciding how to manage a patient with suspected appendicitis.



Please select one option

- Extremely unhelpful
- Very unhelpful
- Somewhat unhelpful
- Neither helpful or unhelpful
- Somewhat helpful
- Very helpful
- Extremely helpful

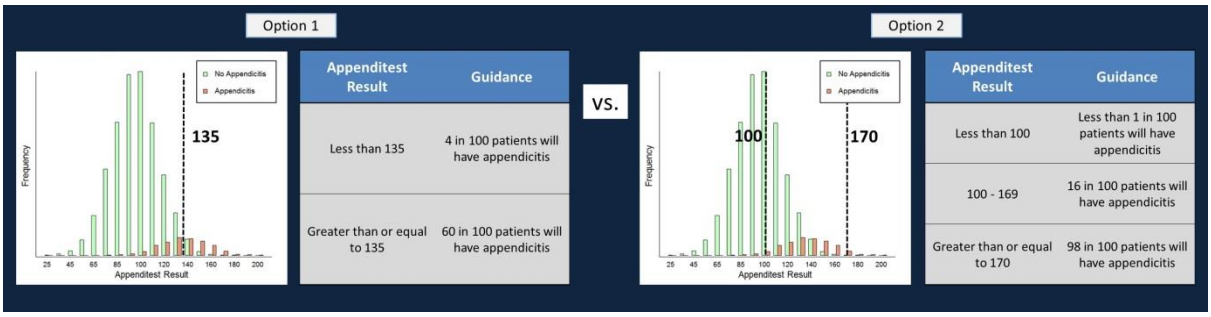
Exclusive, must answer

Q1a Likes & dislikes of guideline format (Open Text)
In the previous question you answered that the guideline format was <<pipe in from Q1>>. In a couple of sentences can you please say why.

Q2 Option 1 vs Option 2

A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 2, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?

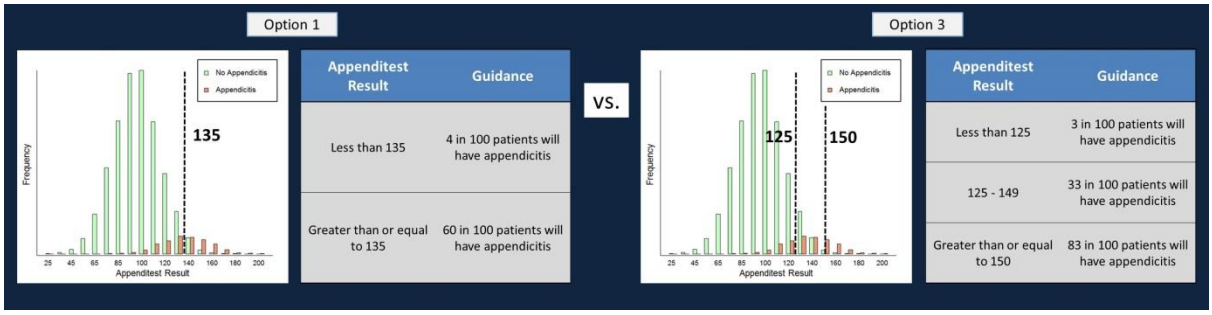


- Option 1
- Option 2

Q2a (Open Text)
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

Q3 Option 1 vs Option 3
Ask Q3 if Q2 = 1
A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

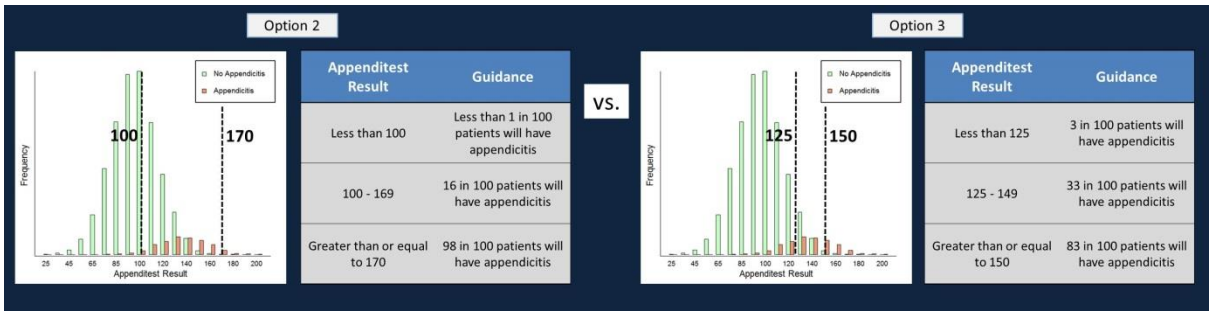
Taking into account both the additional information available in Option 3, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



- Option 1
- Option 3

Q3a (Open Text)
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

Q4 Option 2 vs Option 3
Ask Q4 if Q2 = 2
A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.
Which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (The difference between Option 2 and Option 3 is the placement of the thresholds)



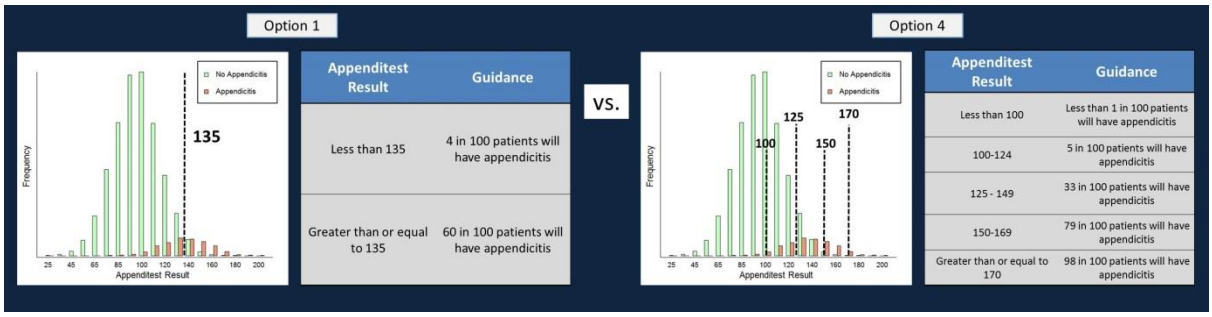
- Option 2
- Option 3

Q4a (Open Text)
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

HQ1 - Hidden Question
If Q3 = 1, populate 1
If Q3 = 3, populate 3
If Q4 = 2, populate 2
If Q4 = 3, Populate 3

Q5 Option 1 vs Option 4
Ask Q5 if HQ1 = 1
A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 4, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



- Option 1
- Option 4

Q5a (Open Text)

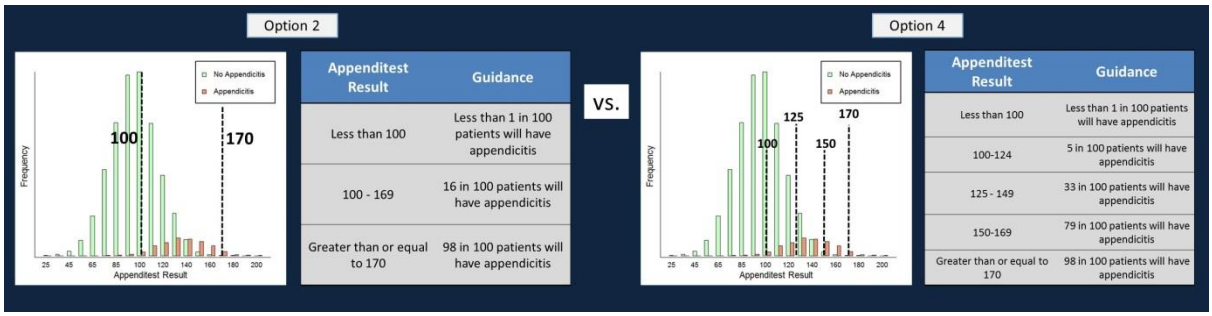
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

Q6 Option 2 vs Option 4

Ask Q6 if HQ1 = 2

A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 4, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



- Option 2
- Option 4

Q6a (Open Text)

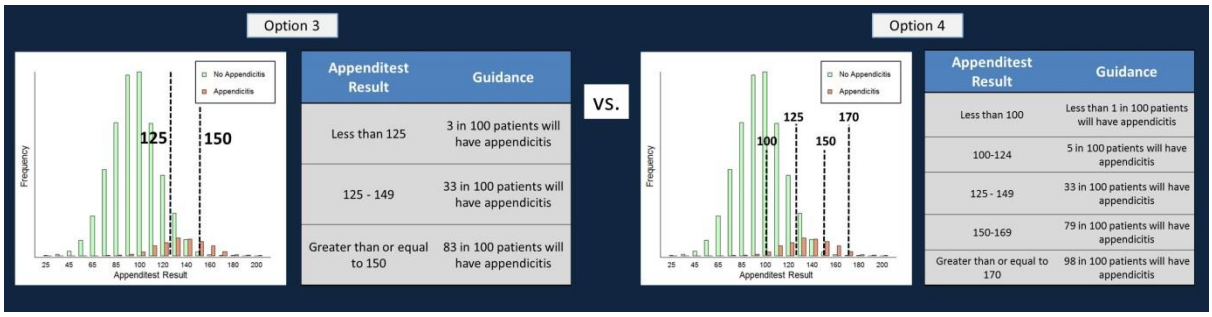
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

Q7 Option 3 vs Option 4

Ask Q7 if HQ1 = 3

A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 4, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



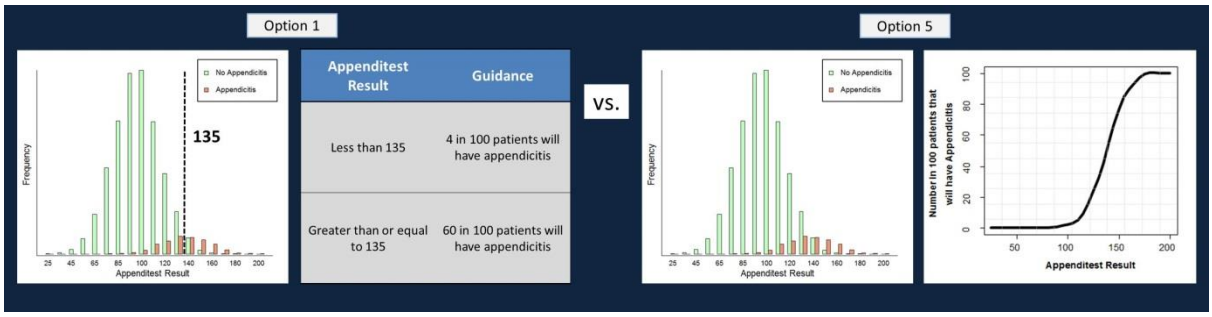
- Option 3
- Option 4

Q7a (Open Text)
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

HQ2 - Hidden Question
If Q5 = 1, populate 1
If Q5 = 4, populate 4
If Q6 = 2, populate 2
If Q6 = 4, Populate 4
If Q7 = 3, populate 3
If Q7 = 4, populate 4

Q8 Option 1 vs Option 5
Ask Q8 if HQ2= 1
A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



- Option 1
- Option 5

Q8a (Open Text)

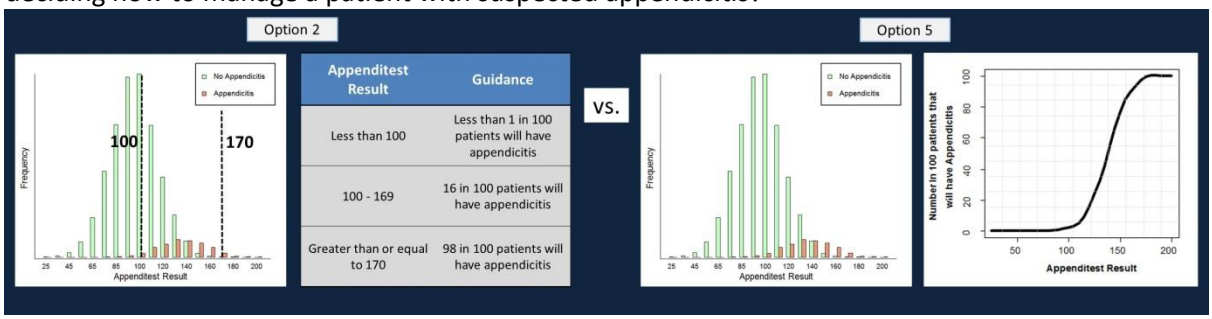
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

Q9 Option 2 vs Option 5

Ask Q9 if HQ2 = 2

A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



- Option 2
- Option 5

Q9a (Open Text)

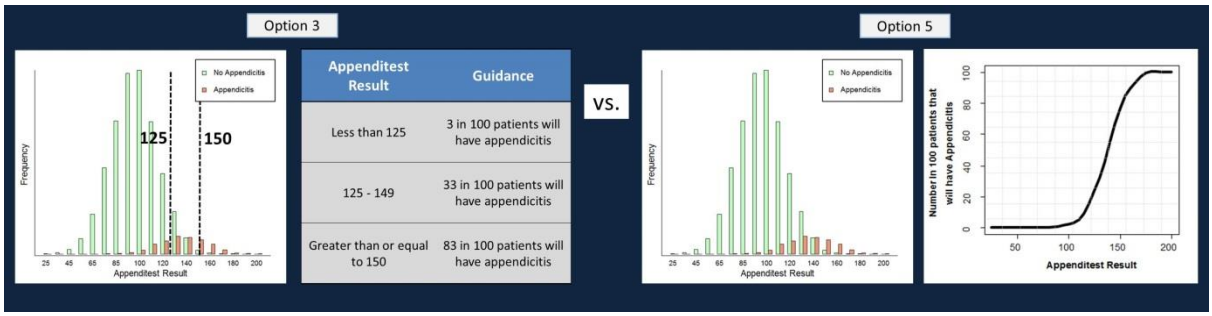
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

Q10 Option 3 Vs Option 5

Ask Q10 if HQ2 = 3

A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



- Option 3
- Option 5

Q10a (Open Text)

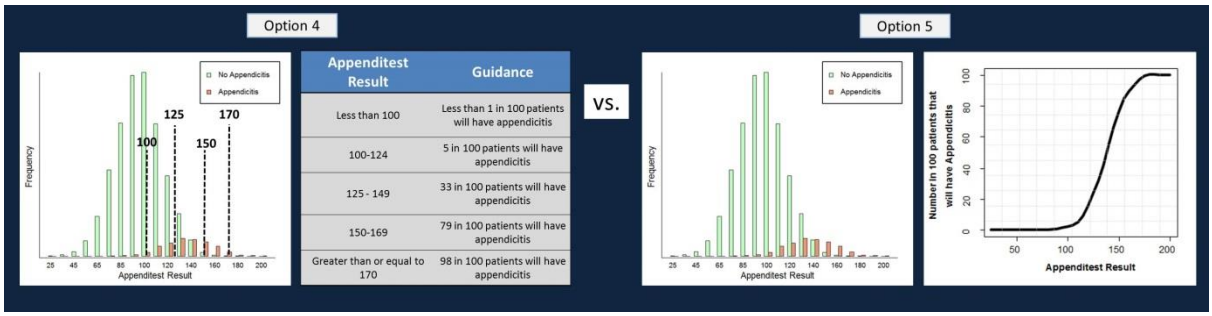
In a couple of sentences, please tell us what you like and dislike about these guideline formats:

Q11 Option 4 Vs Option 5

Ask Q11 if HQ2 = 4

A 12 year old male patient presents to you with symptoms of lower abdominal pain, nausea and fever. You carry out a physical examination to check for appendicitis, but the results are inconclusive and you are undecided from this whether to refer the patient.

Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis?



- Option 4
- Option 5

Q11a (Open Text)

In a couple of sentences, please tell us what you like and dislike about these guideline formats:

HQ3 - Hidden Question

If Q8 = 1, populate 1
If Q8 = 5, populate 5
If Q9 = 2, populate 2
If Q9 = 5, Populate 5
If Q10 = 3, populate 3
If Q10 = 5, populate 5
If Q11 = 4, populate 4
If Q11 = 5, populate 5

Possible influential factors

Q12 How useful are the additional histograms

We have illustrated each of the guideline format options using a histogram of the test results, however ordinarily in guidelines only the table of thresholds is presented. How helpful do you find the additional histogram?

Please select one option

- Extremely unhelpful
- Very unhelpful
- Somewhat unhelpful
- Neither helpful or unhelpful
- Somewhat helpful
- Very helpful

- Extremely helpful

Single response, must answer

Q12a Would preferences been different without histograms

Show Q12a on same page as Q12, show Q12a once Q12 answered

Do you think your preferences in terms of guideline format would have been different if we had not included the histogram?

- Yes
- No

Single response, must answer

Q13 Which format is preferred

Which of the following, if any, do you think would change your preference of guideline format?

Please select all that apply

- A scenario with a rarer disease (i.e. lower disease prevalence)
- A scenario with a more common disease (i.e. higher disease prevalence)
- A scenario with a less critical condition (e.g. a 70 year old with suspected arthritis)
- A scenario with a more critical condition (e.g. a baby with suspected meningitis)
- Other (please specify)

Multi

Q14 Is there an alternative format (Open Text)

Is there an alternative format that you would prefer that was not included in the options presented? If so, please describe your preferred format.

Q15 Useful to report the patients without appendicitis

Would you find it useful to also report the number of patients that will not have appendicitis, as shown?

Appenditest Result	Guidance
Less than 125	3 in 100 patients will have appendicitis 97 in 100 patients will not have appendicitis
125 - 149	33 in 100 patients will have appendicitis 67 in 100 patients will not have appendicitis
Greater than or equal to 150	83 in 100 patients will have appendicitis 17 in 100 patients will not have appendicitis

- Yes
- No

Single response, must answer

Q16 Ease of understanding

Which of the following guidelines do you find easier to understand?

- 23 in 100 patients will have appendicitis

- 23% of patients will have appendicitis

Single response, must answer

Closing Questions

F1- Final feedback (Open Text)
You have now reached the end of this survey. Is there any further feedback that you would like to give about the topic of the survey, or about the survey itself?

Please write in full

F2 Re-contact

May we re-contact you for clarification on any of your answers or for data quality purposes?

Yes
No
Single response, must answer

THANK AND CLOSE

4B. Full Results

Respondent Characteristics	% (N)
Age	
Under 30	1% (2)
30 - 39	41.6% (84)
40 - 49	32.7% (66)
50 - 59	20.3% (41)
60 or over	4.5% (9)
Gender	
Female	41.1% (83)
Male	58.9% (119)
Region	
East Midlands	5.9% (12)
East of England	9.4% (19)
London	11.4% (23)
North East	10.9% (22)
North West	4.5% (9)
Northern Ireland	3.0% (6)
Scotland	11.4% (23)
South Central	6.4% (13)
South East Coast	6.9% (14)
South West	8.9% (18)
Wales	5.0% (10)
West Midlands	7.9% (16)
Yorkshire & the Humber	8.4% (17)
GP Type	
GP principal	64.4% (130)
Locum GP	11.9% (24)
Salaried GP	23.8% (48)

Q1: On the following scale, please rate how helpful this guideline format would be when deciding how to manage a patient with suspected appendicitis (Option 1 shown – single threshold interpretation)

Extremely unhelpful (1)	Very unhelpful (2)	Somewhat unhelpful (3)	Not Sure (4)	Somewhat helpful (5)	Very helpful (6)	Extremely helpful (7)	Average Rating
0.5% 1	**4.5%** 9	**14.9%** 30	**10.4%** 21	**49.5%** 100	**18.8%** 38	**1.5%** 3	**4.66**

Q2: Taking into account both the additional information available in Option 2, but also the added complexity, which of the following guidance formats do you think would be more helpful when

deciding how to manage a patient with suspected appendicitis? (Option. 1 vs. Option. 2 shown – single threshold interpretation vs. rule-in and rule-out thresholds)

Answer Choices	% (N)
I think Option 1 would be more helpful	25.7% (52)
I think Option 2 would be more helpful	74.3% (150)

Q3: Taking into account both the additional information available in Option 3, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 1 vs. Option. 3 shown – single threshold interpretation vs. uninformative range)

Answer Choices	% (N)
I think Option 1 would be more helpful	71.2% (37)
I think Option 3 would be more helpful	28.8% (15)

Q4: Taking into account both the additional information available in Option 3, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (The difference between Option 2 and Option 3 is the placement of the thresholds) (Option. 2 vs. Option. 3 shown – rule-in and rule-out thresholds vs. uninformative range)

Answer Choices	% (N)
I think Option 2 would be more helpful	78.0% (117)
I think Option 3 would be more helpful	22.0% (33)

Q5: Taking into account both the additional information available in Option 4, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 1 vs. Option. 4 shown – rule-in and rule-out thresholds vs. uninformative range)

Answer Choices	% (N)
I think Option 1 would be more helpful	81.1% (30)
I think Option 4 would be more helpful	18.9% (7)

Q6: Taking into account both the additional information available in Option 4, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 2 vs. Option. 4 shown – rule-in and rule-out thresholds vs. multiple categories)

Answer Choices	% (N)
I think Option 2 would be more helpful	70.1% (82)
I think Option 4 would be more helpful	29.9% (35)

Q7: Taking into account both the additional information available in Option 4, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 3 vs. Option. 4 shown – uninformative range vs. multiple categories)

Answer Choices	% (N)
I think Option 3 would be more helpful	78.0% (117)
I think Option 4 would be more helpful	22.0% (33)

Q8: Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 1 vs. Option. 5 shown – single threshold interpretation vs. graph of continuous natural frequencies)

Answer Choices	% (N)
I think Option 1 would be more helpful	78.0% (117)
I think Option 5 would be more helpful	22.0% (33)

Q9: Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 2 vs. Option. 5 shown – rule-in and rule-out thresholds vs. graph of continuous natural frequencies)

Answer Choices	% (N)
I think Option 2 would be more helpful	80.0% (24)
I think Option 5 would be more helpful	20.0% (6)

Q10: Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 3 vs. Option. 5 – uninformative range vs. graph of continuous natural frequencies)

Answer Choices	% (N)
I think Option 3 would be more helpful	70.4% (19)
I think Option 5 would be more helpful	29.6% (8)

Q11: Taking into account both the additional information available in Option 5, but also the added complexity, which of the following guidance formats do you think would be more helpful when deciding how to manage a patient with suspected appendicitis? (Option. 4 vs. Option. 5 – multiple categories vs. graph of continuous natural frequencies)

Answer Choices	% (N)
I think Option 4 would be more helpful	65.2% (43)
I think Option 5 would be more helpful	34.9% (23)

Q12: We have illustrated each of the guideline format options using a histogram of the test results, however ordinarily in guidelines only the table of thresholds is presented. How helpful do you find the additional histogram?

Extremely unhelpful (1)	Very unhelpful (2)	Somewhat unhelpful (3)	Not Sure (4)	Somewhat helpful (5)	Very helpful (6)	Extremely helpful (7)	Average Rating
0.5% **1**	**6.9%** 14	**14.4%** 29	**18.8%** 38	**32.2%** 65	**23.3%** 47	**4.0%** 8	**4.61**

Q12a: Do you think your preferences in terms of guideline format would have been different if we had not included the histogram?

Options	% (N)
Yes	35.1% (71)
No	64.9% (131)

Q13: Which of the following, if any, do you think would change your preference of guideline format?

Options	% (N)
A scenario with a rarer disease (i.e. lower disease prevalence)	30.7% (62)
A scenario with a more common disease (i.e. higher disease prevalence)	37.6% (73)
A scenario with a less critical condition (e.g. a 70 year old with suspected arthritis)	32.2% (65)
A scenario with a more critical condition (e.g. a baby with suspected meningitis)	37.1% (75)
Other (please specify)	6.9% (14)

Q15: Would you find it useful to also report the number of patients that will not have appendicitis, as shown?

Options	% (N)
Yes	42.6% (86)
No	57.4% (116)

Q16: Which of the following guidelines do you find easier to understand?

Options	% (N)
23 in 100 patients will have appendicitis	28.7% (58)
23% of patients will have appendicitis	22.8% (46)
No Preference	48.5% (98)

Chapter Five

The Evaluation of Existing Methods for Defining an Intermediate Range of Values on a Quantitative Diagnostic Test Scale

5.1. Overview

Research Objective: To compare existing methods for identifying an intermediate range of values on a quantitative diagnostic test scale.

Methods: The 'TG-ROC' method (1, 2) which is based on sensitivity and specificity, and the Grey Zone method (3) which is based on predictive values and likelihood ratios, were applied to test results from three commonly used inflammatory markers for detecting serious bacterial infection in children: white blood cell count, procalcitonin and C-reactive protein. Extending these methods to allow for an additional range of 'uninformative' results is explored as a means of defining multiple categories of test result (4-6).

Results: Difficulties were encountered when applying the Grey Zone method to real clinical data, as it failed to produce lower limits of the intermediate range for all three tests. The TG-ROC method was very easy to apply and produced reasonable intermediate ranges; however the clinical relevance of placing thresholds at fixed accuracy levels across all clinical scenarios is questioned.

Conclusion: If a more considered approach is adopted when specifying clinically relevant accuracy, the TG-ROC provides a simple and practical method for identifying an intermediate range of values on a quantitative test scale. The uninformative range can provide further clarity, particularly for tests of poor accuracy.

5.2. Introduction

The dichotomous 'positive-negative' framework is often far too restrictive and fails to adequately recognise the uncertainty that plagues diagnostic decision making (7-9). It was evident in the systematic review in **Chapter Three** that the dichotomous approach remains the most commonly adopted approach to evaluating the accuracy of quantitative diagnostic biomarkers. However, the findings from the GP survey presented in **Chapter Four** indicated that GPs would find a greater depth of information useful, with a preference for rule-in and rule-out thresholds to facilitate test result interpretation.

Despite the reporting of intermediate test results featuring as a recommendation in the STARD statement (10), we saw in **Chapter Two** that compliance to this recommendation is poor (8). One obstacle to making this standard practice in primary diagnostic accuracy research may be the lack of guidance on the statistical methods that should be used for identifying an intermediate range on a quantitative test scale.

The objective of this chapter is to provide a comprehensive theoretical and applied evaluation of existing methods for identifying an intermediate range on a quantitative test scale (2, 3). Both methods avoid the need for any complex statistical modelling and rely solely on the calculation of familiar test characteristics. Greiner developed the 'Two-Graph Receiver Operating Characteristic' (TG-ROC) which employs the popular diagnostic accuracy indices sensitivity and specificity to delimit an intermediate test range (1, 11). In contrast, Coste and Pouchot propose the Grey Zone method, which is instead based on predictive values and likelihood ratios (3). These methods will be applied to a dataset of diagnostic test results for three commonly used biomarkers for serious bacterial infection.

The identification of multiple thresholds (i.e. more than two) to interpret test results was also considered useful by a notable proportion of GPs in the survey presented in the previous chapter.

A statistical definition of a range of results that are 'rarely clinically significant' or 'uninformative' (4-6) was identified in the literature review, and will therefore be explored as a possible means of adapting these existing methods to further break down the intermediate range and allow for multiple categories of test result.

5.3. Methods

5.3.1. Dataset

700 children (3 months – 16 years old, median age=3 years old) with suspected serious bacterial infection were consecutively recruited from a Paediatric Assessment Unit at a UK hospital. Vital signs and clinical features were recorded by a triage nurse and parents completed a 22-item checklist of presenting symptoms on arrival. The results from diagnostic biomarkers were recorded if carried out as part of routine diagnostic work-up, the three most commonly used being white blood cell count, procalcitonin and C-reactive protein (measured in 365, 259, 149 children, respectively). The final diagnosis for each patient was agreed upon by senior paediatricians. A panel of clinicians grouped the final diagnoses into two categories: no or mild infection (n=387, 55.3%) vs. intermediate or serious infection (n=313, 44.7%).

5.3.2. Methods for Identifying an Intermediate Test Range

The Two-Graph Receiver Operating Characteristic method

The TG-ROC method uses pre-defined values of sensitivity and specificity to identify the limits of an intermediate range of test results (1). Greiner recommends that the lower 'rule-out' and upper 'rule-in' thresholds are placed at the values on the test scale that achieve 90% (or an optional 95%) sensitivity and specificity. By definition, 10% (or 5%) of those in the negative and positive categories will therefore be misclassified.

The TG-ROC method is similar to standard ROC analysis in that it relies on the calculation of sensitivity and specificity for every point on the test scale. However, unlike a standard graph of the

ROC curve, sensitivity and specificity are both plotted on the y-axis against the full range of test thresholds on the x-axis. This makes it possible to relate the observed accuracy of the test to specific test values and identify the range of test values that fail to achieve 90% sensitivity and 90% specificity; the intermediate test range (Figure 5.1). The point at which the sensitivity and specificity curves cross is the estimated optimal single threshold, given that the costs associated with a false positive result and a false negative result are equal.

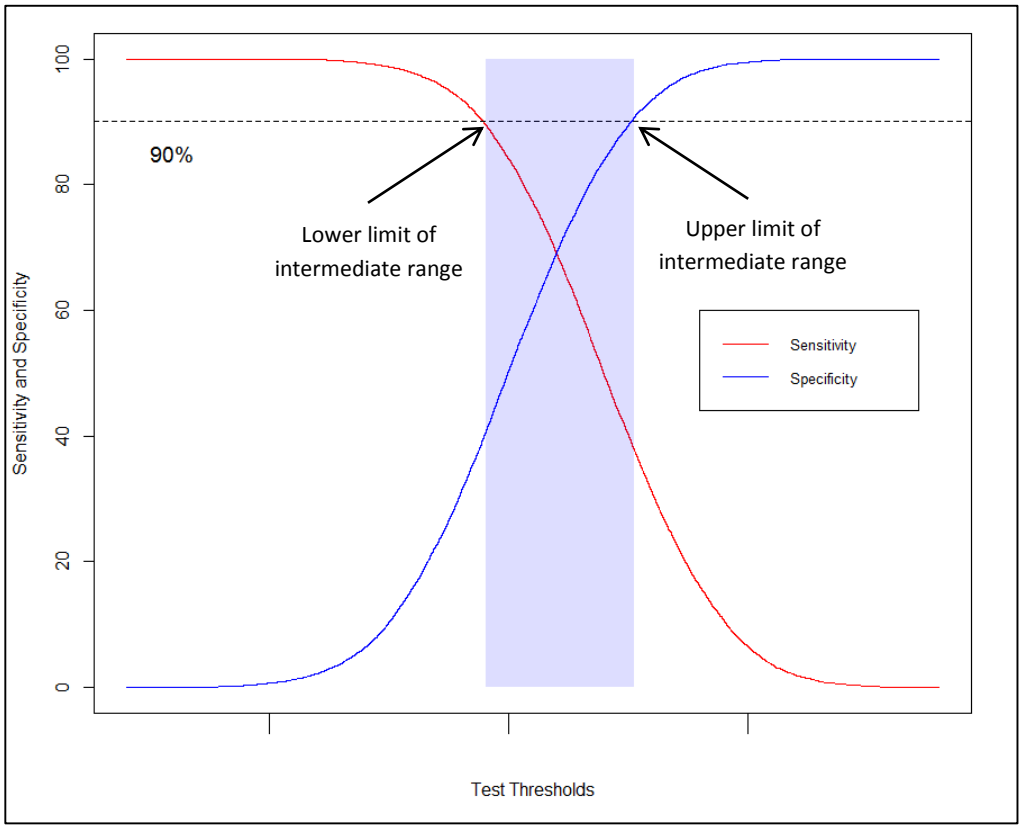


Figure 5.1. Example TG-ROC plot with intermediate range highlighted in grey (hypothetical data. N=20,000)

The Grey Zone method

To apply the Grey Zone method (3), the 'desired' accuracy levels at which to place the thresholds are derived from the following three pieces of information: 1) an estimate of the pre-test probability of disease in the clinical setting for which the test is being evaluated, 2) knowledge of the distributions of test results for the 'diseased' and 'healthy' populations, and 3) an estimate of the post-test probabilities required to rule in and rule out the disease in question.

Using conditional probability theory, the pre- and post-test probability estimates facilitate the calculation of the minimal positive likelihood ratio required to rule in disease, and the maximal negative likelihood ratio required to rule out disease.

The value at which the test achieves the required maximal negative likelihood ratio is the lower limit of the intermediate range, and the value at which the test achieves the minimal positive likelihood ratio is the upper limit. The authors recommend plotting the positive and negative likelihood ratio curves against the test scale to help identify the test values associated with the derived likelihood ratios. For the hypothetical example in Figure 5.2 (based on the same simulated data using in Figure 5.1), the pre-test probability was based on a disease prevalence of 50%, and positive and negative post-test probabilities (PPV and NPV) of 0.9 and 0.1 were arbitrarily selected.

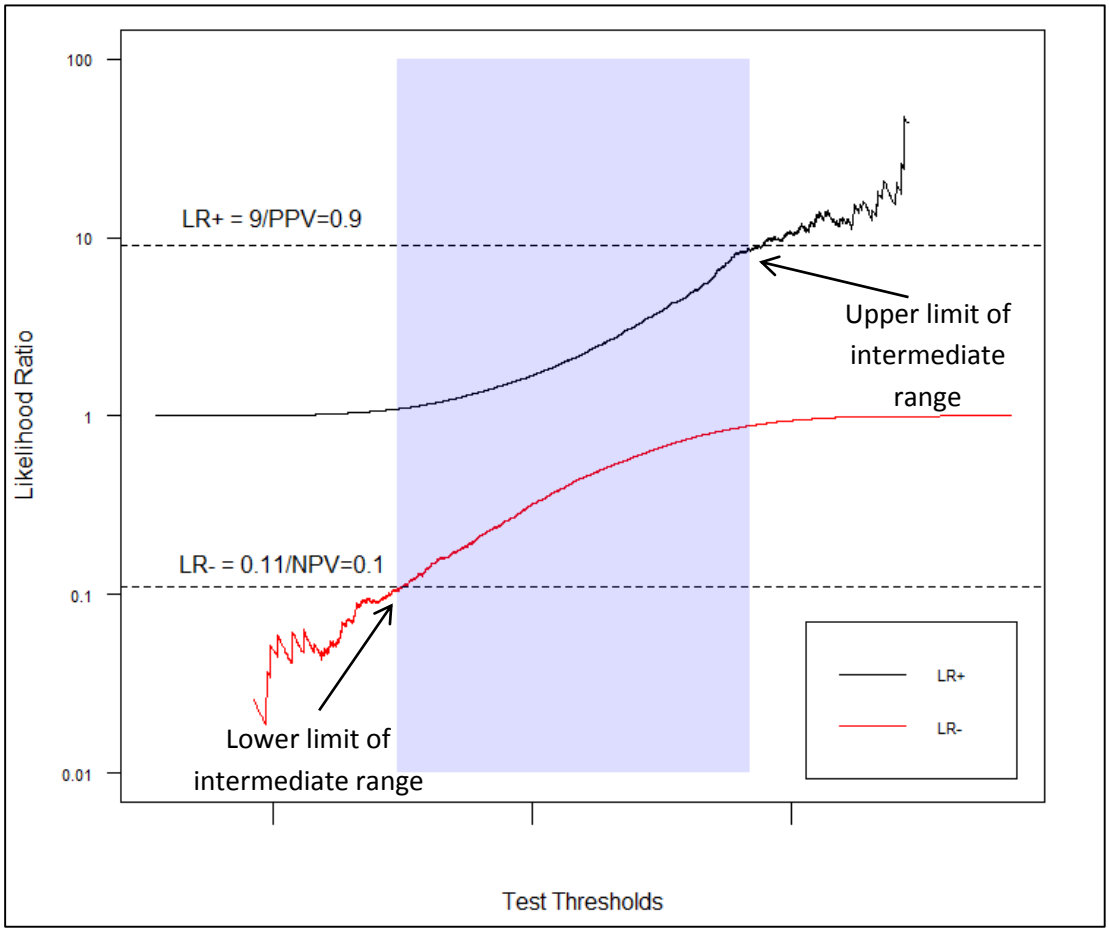


Figure 5.2. Example of a Grey Zone plot, with intermediate range highlighted (based on the same simulated data used in Figure 5.1.)

Uninformative Range

A range of uninformative values, defined as results that have a negative likelihood ratio greater than or equal to 0.5 and a positive likelihood ratio less than or equal to 2 (4-6), was also identified for all three tests. The results from the survey presented in **Chapter Four** indicate that clinicians are unlikely to find this interpretation of test results useful in isolation, but may find it helpful if presented in addition to rule-in and rule-out thresholds. This category of test result is therefore presented as an extension of the previous methods to demarcate multiple categories of test result.

5.3.3. Statistical Analyses

Diagnostic accuracy statistics were calculated for white blood cell count, procalcitonin and C-reactive protein. Sensitivity, specificity, and positive and negative likelihood ratios were calculated at the threshold that provided the maximal Youden's index. Receiver Operating Characteristic curves were plotted and the area under each curve calculated. All statistics are reported with 95% confidence intervals.

Greiner's' TG-ROC method and Coste and Pouchot's Grey Zone method (5.3.2) were then applied to identify intermediate test ranges for these biomarkers for serious bacterial infection. For the TG-ROC method, we calculated intermediate ranges based on both 90% and 95% sensitivity and specificity. For the Grey Zone method, the pre-test probability of serious bacterial infection was estimated using the prevalence of serious bacterial infection in the whole dataset (45%). As the patients were consecutively recruited, this was assumed to be a reasonable estimate of disease prevalence for this clinical setting and population. In the absence of any guidance regarding appropriate post-test probabilities, a range of probabilities were explored: 80 – 95% positive post-test probability and 5-20% negative post-test probability, each in intervals of 5%.

The test values that provide a positive likelihood ratio less than 2 and a negative likelihood ratio greater than 0.5 were then identified (4-6), and these thresholds were combined with the results from the evaluated methods to produce multiple categories of test result.

All statistical analyses were conducted in R (12).

5.4. Results

5.4.1. Diagnostic Accuracy

The prevalence of serious bacterial infection was highest in patients tested for C-reactive protein (n=193/259, 74.5%), followed by white blood cell count (n=234/365, 64.1%) and then procalcitonin (n=87/149, 58.4%). The sensitivity, specificity, likelihood ratios and predictive values for each test at the threshold that provided the maximal Youden's index can be found in Table 5.1.

	White Blood Cell Count	Procalcitonin	C-Reactive Protein
'Optimal' Simple Threshold	12.03 ($x10^9$/L)	0.45 ($x10^9$/L)	49.5 (mg/L)
Sensitivity	0.56 (0.5, 0.63)	0.45 (0.34,0.56)	0.52 (0.45,0.59)
Specificity	0.76 (0.68, 0.83)	0.94 (0.84,0.98)	0.74 (0.62,0.84)
Positive Likelihood Ratio	2.38 (1.72,3.31)	6.95 (2.62,18.44)	2.01 (1.31,3.10)
Negative Likelihood Ratio	0.57 (0.48,0.68)	0.59 (0.48,0.72)	0.65 (0.53,0.80)
Positive Predictive Value	0.81 (0.74, 0.87)	0.91 (0.78,0.97)	0.85 (0.78,0.91)
Negative Predictive Value	0.5 (0.42, 0.57)	0.55 (0.45,0.64)	0.34 (0.27,0.43)
Maximal Youden's Index	0.33 (0.18,0.46)	0.38 (0.18,0.54)	0.26 (0.07,0.43)

Table 5.1. Diagnostic accuracy at a single threshold (determined by Youden's Index)

The maximal Youden's Index suggests that procalcitonin has marginally better overall accuracy than white blood cell count and C-reactive protein, however the confidence intervals demonstrate that this is not a significant finding. All three tests offer very modest sensitivity indicating that these tests alone would not be useful for ruling out serious bacterial infection. The confidence intervals across all three tests for each parameter overlap, indicating that for a sample of this size, there is insufficient evidence that one test is superior in accuracy to the others.

Receiver Operating Characteristic curves and the area under the curves for all three tests can be found in Figure 5.3. The point estimates for the area under the curves for white blood cell count and procalcitonin are the same (0.70), with C-reactive protein being slightly lower (0.63), although the 95% confidence intervals for all three AUC values overlap notably (see Figure 5.3).

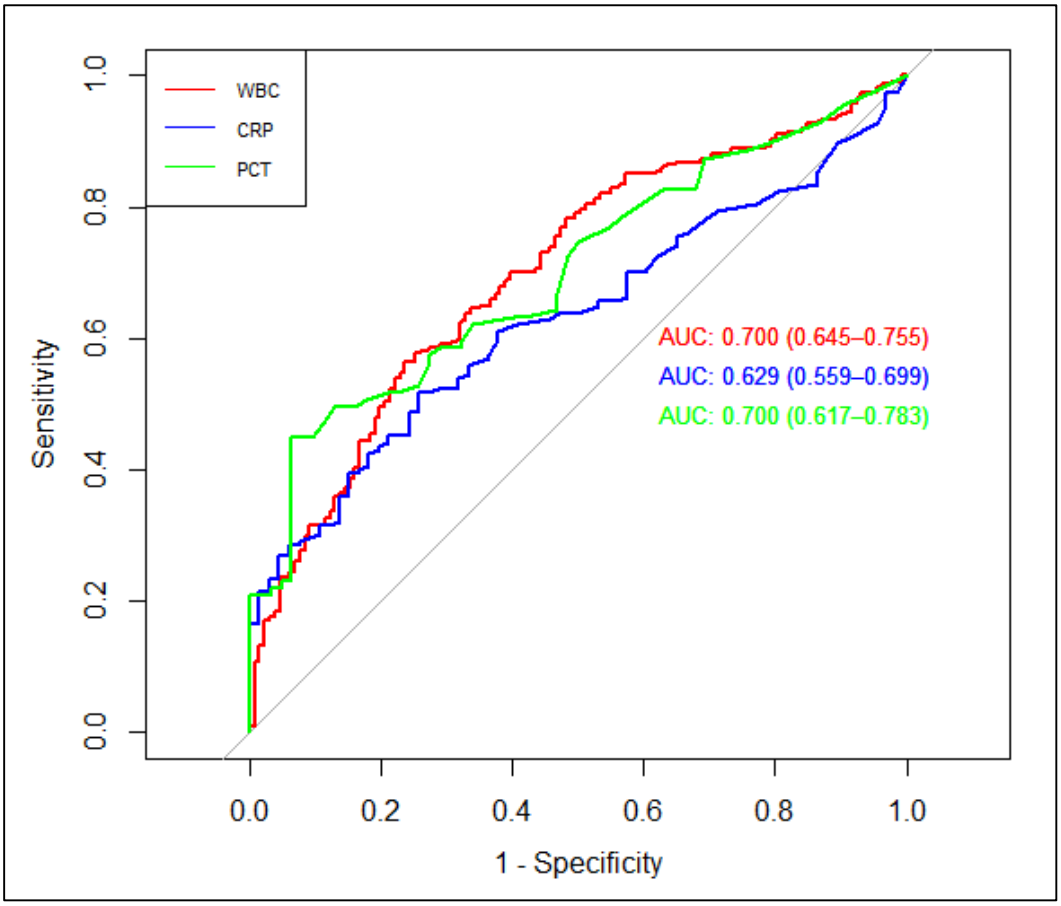


Figure 5.3. Receiver Operating Characteristic curves for white blood cell count, procalcitonin, C-reactive protein

5.4.2. The TG-ROC Method

The thresholds at which 90% sensitivity and 90% specificity were achieved for white blood cell count, C-reactive protein and procalcitonin were identified. For white blood cell count, 90% sensitivity was achieved at a threshold of 6.47 $x10^9$/L and 90% specificity was achieved at a threshold of 16.03 $x10^9$/L, forming the limits of the intermediate range. Over half of the patients (62%) fell within these limits, with just under a quarter (23.8%) falling in the 'positive' category and 14% in the 'negative' category. These results can be found in Table 5.2 for all three tests.

TG-ROC Method: Sensitivity and Specificity = 90%					
Test	**Lower Limit**	**Upper Limit**	**% Negative**	**% Intermediate**	**% Positive**
White Blood Cell Count ($x10^9$/L)	6.47	16.03	14.0%	62.2%	23.8%
Procalcitonin ($x10^9$/L)	0.12	0.42	14.8%	55.0%	30.2%
C-Reactive Protein (mg/L)	8	99	9.7%	64.1%	26.3%

Table 5.2. Intermediate range limits using the 'TG-ROC' method and the percentage of results in each category

The TG-ROC plot for white blood cell count is shown in Figure 5.4 and the equivalent plots for procalcitonin and C-reactive protein can be found in Appendix 5A.

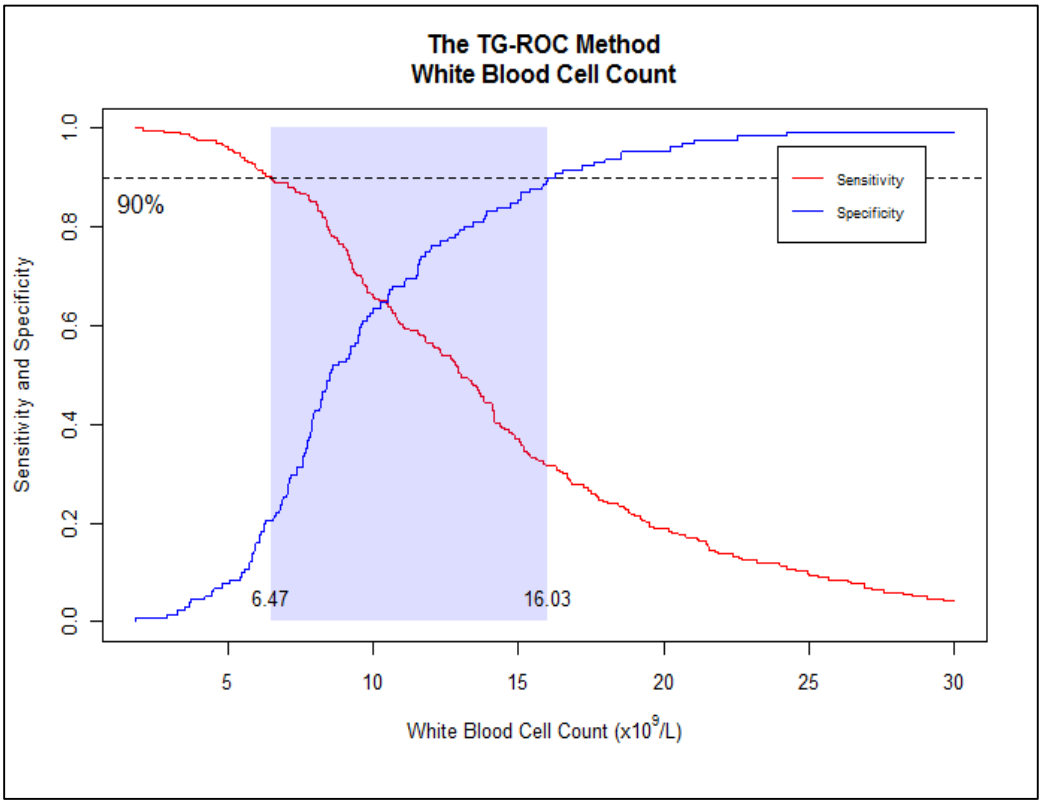


Figure 5.4. TG-ROC plot for white blood cell count with the intermediate range highlighted

5.4.3. The Grey Zone Method

In accordance with the Grey Zone method, we estimated the pre-test probability of serious bacterial infection and 'desirable' post-test probabilities. The pre-test probability was estimated

using the prevalence of serious bacterial infection in the study sample (44.7%, 95% CI: 41.0% – 48.5%). The results for a positive and negative post-test probability of 90% and 10%, respectively, are reported (Table 5.3).

Grey Zone Method: PPV = 0.9 and NPV = 0.1					
Test	**Lower Limit**	**Upper Limit**	**% Negative**	**% Intermediate**	**% Positive**
White Blood Cell Count ($x10^9$/L)	No solution	24.24	0.0%	92.6%	7.4%
Procalcitonin ($x10^9$/L)	No solution	2.90	0.0%	86.6%	13.4%
C-Reactive Protein (mg/L)	No solution	170	0.0%	83.4%	16.6%

Table 5.3. Intermediate range limits using the Grey Zone method and the percentage of results in each category

A plot of the likelihood ratio curves for white blood cell count and the intermediate range is shown in Figure 5.5. As can be seen, the negative likelihood ratio curve never achieves a negative predictive value of 0.1 and therefore the method fails to find a lower limit to the intermediate range. This was the case for all three of the tests under evaluation; the best negative predictive values achieved for white blood cell count, procalcitonin and C-reactive protein were 0.22, 0.25 and 0.34, respectively, and therefore if negative predictive values smaller than these were selected, then the method failed to identify a lower limit.

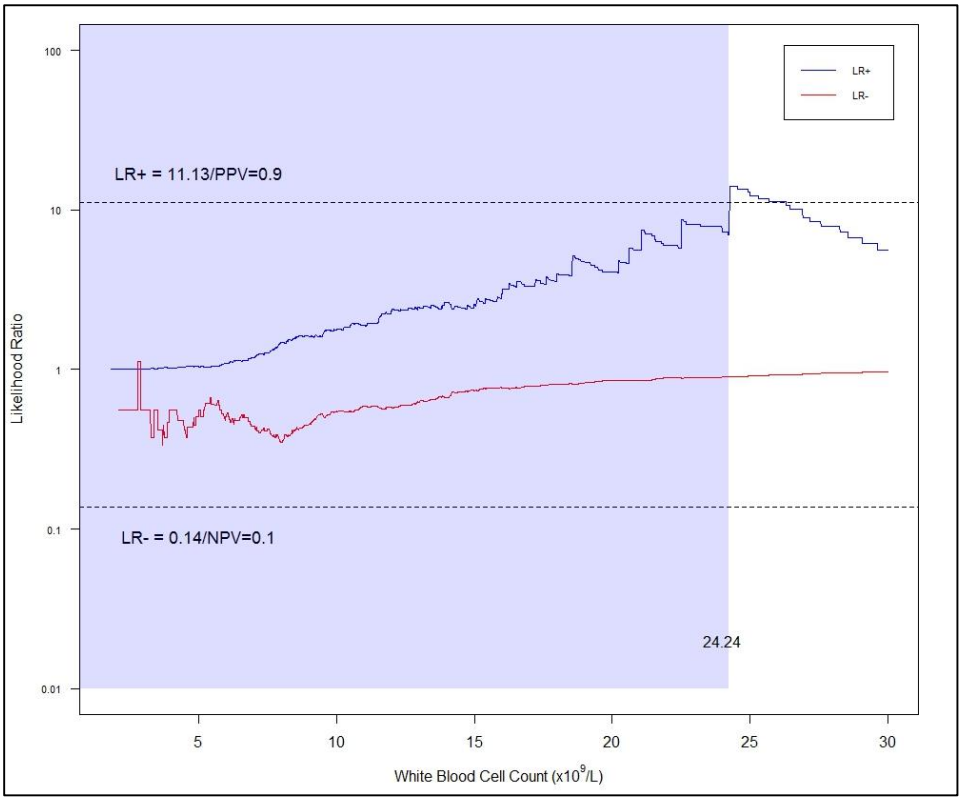


Figure 5.5. The likelihood ratio curves for white blood cell count with the intermediate range highlighted

In Figure 5.5, the positive likelihood ratio curve crosses the LR+ = 11.13 dashed line twice due to the non-monotonicity of the estimated likelihood ratio curves towards the extremes of the test scale. As a general rule, for the lower limit of the intermediate range, we took the highest threshold that achieved the desired negative likelihood ratio and the lowest threshold that achieved the desired positive likelihood ratio for the upper limit of the intermediate range.

5.4.4. Adding an 'Uninformative' Range

The results in Table 5.4 provide the lower and upper limits of the 'uninformative' range of values between which test values have a negative likelihood ratio greater than 0.5 and a positive likelihood ratio smaller than 2.

LR - ≥ 0.5 and LR + ≤ 2					
Test	**Lower Limit**	**Upper Limit**	**% Negative**	**% Uninformative**	**% Positive**
White Blood Cell Count ($x10^9$/L)	9.6	11.51	40.8%	10.7%	48.5%
Procalcitonin ($x10^9$/L)	0.16	0.28	25.5%	26.8%	47.7%
C-Reactive Protein (mg/L)	22	49	32.4%	22.0%	45.6%

Table 5.4. Uninformative range limits and the percentage of results in each category

For the white cell count results, identifying limits at a negative likelihood ratio of 0.5 and a positive likelihood ratio of 2 is equivalent to identifying an intermediate range based on values that provide sensitivities and specificities less than 70%. This narrower range of uninformative values is displayed on the TG-ROC plot in addition to the rule-in and rule-out thresholds identified at 90% sensitivity and specificity according to the TG-ROC (Figure 5.6).

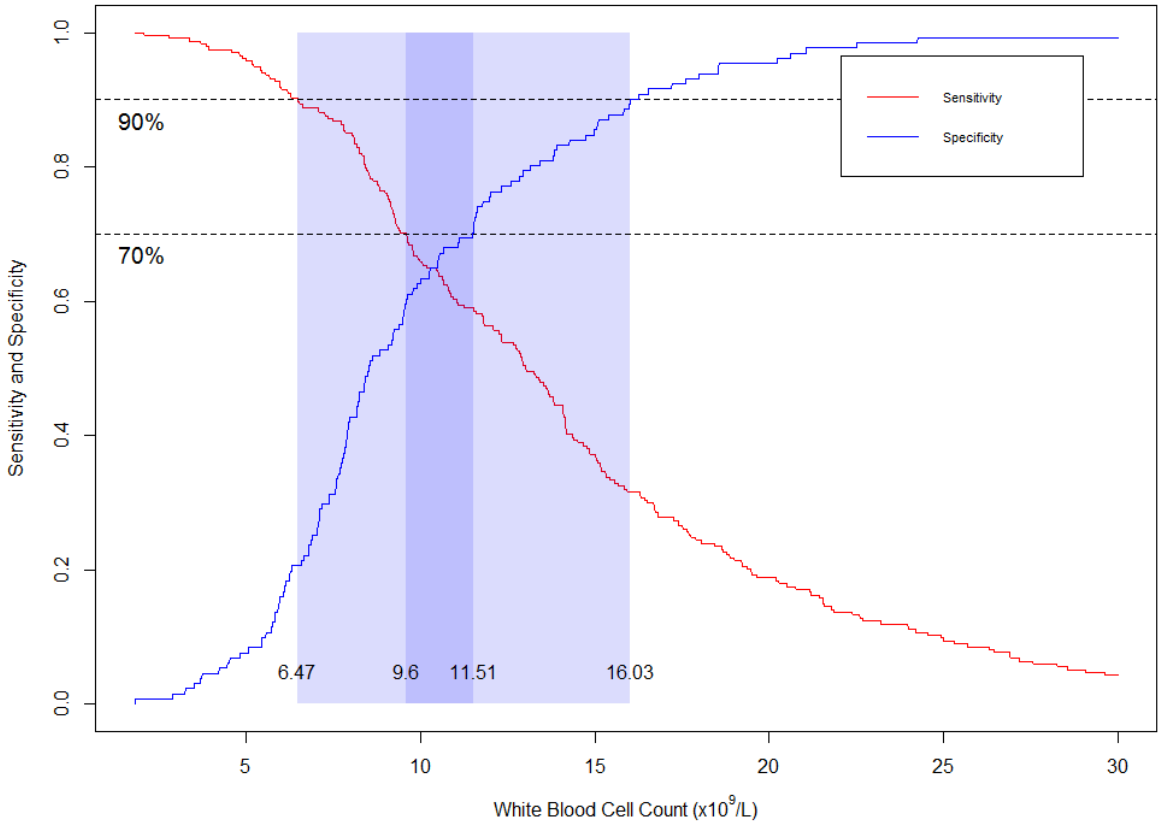


Figure 5.6. TG-ROC results for white blood cell count with additional uninformative range highlighted

Figure 5.7 shows the results for white blood cell count when the uninformative range is added to those of the Grey Zone method. Placing additional thresholds at a positive likelihood ratio of 2 and a negative likelihood ratio of 0.5 is equivalent to selecting an upper threshold at a positive predictive value of 61.8% and a lower threshold at a negative predictive value of 28.8%.

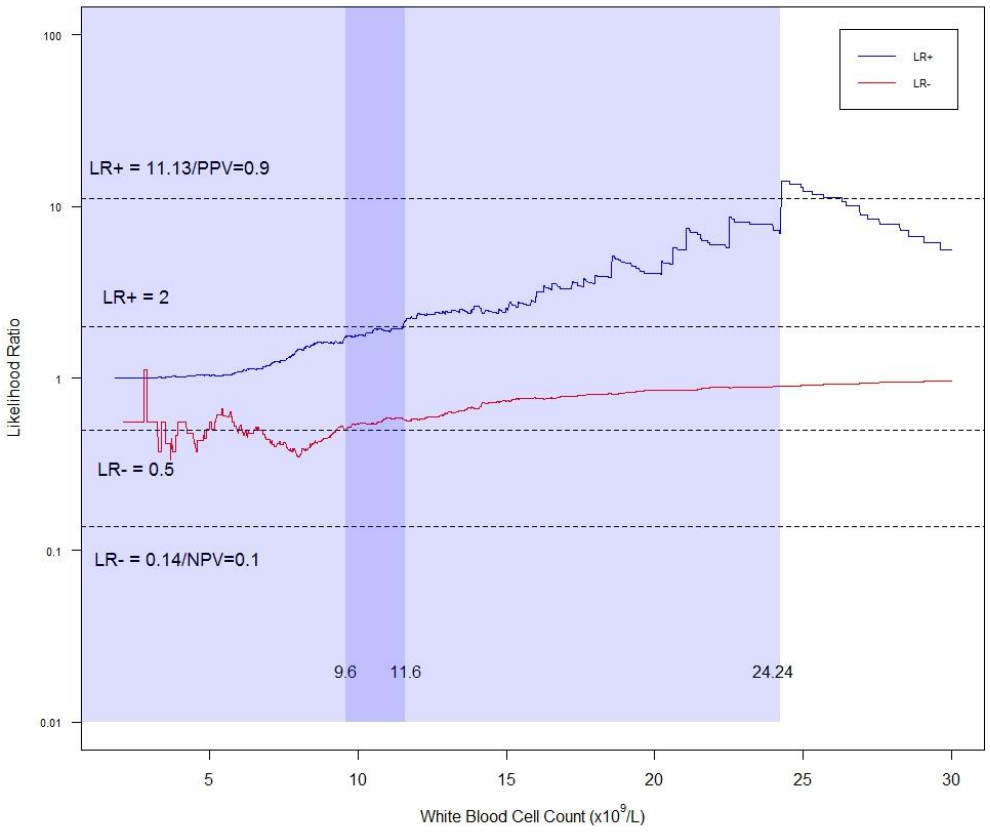


Figure 5.7. The Grey Zone results for white blood cell count with additional uninformative range highlighted

Table 5.5 summarises the multiple categories of test result when an uninformative range is added to the results of the existing methods for identifying an intermediate range for white blood cell count. The results for procalcitonin and C-reactive protein can be found in Appendix 5C. The percentages in each category highlight that the TG-ROC method results in a more even spread of the distribution of patients, whereas a substantial proportion of patients fall in the categories either side of the uninformative range for the Grey Zone method. Category-specific likelihood ratios highlight the poor performance of white blood cell count in this dataset. Other than the upper categories for each method, all of the categories have an average likelihood ratio between 0.5 and

2, with the uninformative range providing an average likelihood ratio close to 1 implying no predictive value.

TG-ROC Method			Grey Zone Method		
Test Ranges	**% in each category**	**LR**	**Test Ranges**	**% in each category**	**LR**
<6.47 (x10^9/L)	14.0%	0.50	**No Solution**	0%	-
6.47 - 9.6 (x10^9/L)	26.8%	0.52	**<9.6 (x10^9/L)**	40.8%	0.51
9.6 – 11.51 (x10^9/L)	10.7%	1.03	**9.6 – 11.51 (x10^9/L)**	10.7%	1.03
11.51 – 16.03 (x10^9/L)	24.7%	1.37	**11.51 – 24.24 (x10^9/L)**	41.1%	1.71
>16.03 (x10^9/L)	23.8%	3.19	**>24.24 (x10^9/L)**	7.4%	7.00

Table 5.5. Adding an uninformative range to the existing methods: proportion in each category and category-specific likelihood ratios for white blood cell count

5.5. Discussion

In this dataset, the intermediate ranges were very wide, in part, due to the poor accuracy of the three tests under evaluation. For both the TG-ROC method and the Grey Zone method, over half of the patients fell in the identified intermediate ranges.

5.5.1. Evaluation of Existing Methods

The Grey Zone method failed to identify lower intermediate range limits for all three tests under evaluation, as none achieved the desired negative predictive value of 0.1. The Grey Zone method has also been applied in a study evaluating the diagnostic value of procalcitonin for the diagnosis of serious bacterial infection in elderly patients (13). The authors set out to identify an intermediate range for a maximal negative likelihood ratio of 0.1 and a minimal positive likelihood ratio of 10, however as the test didn't achieve these accuracy levels, the authors resorted to selecting minimal sensitivities and specificities of 95%, unwittingly implementing the TG-ROC method for delimiting an intermediate range instead.

This is a key limitation of the Grey Zone method; if the 'desired' accuracy levels are beyond the discriminatory capabilities of the test, intermediate range limits cannot be identified. In contrast, an attractive property of the TG-ROC method is that, unless the performance of the test at a single threshold exceeds 90% sensitivity and specificity, it will always identify intermediate range limits. The reason for this is that the selection of a likelihood ratio places demands on both the sensitivity and the specificity of the test, a requirement that the test may never achieve. For the TG-ROC method, it is only making demands on either sensitivity or specificity, and therefore by making sacrifices with one of these parameters, high values for the other parameter can always be achieved.

Another difficulty encountered when applying the Grey Zone method is that the likelihood ratio curves are non-monotonic at both ends of the test scale due to the sparseness of data. This creates

two issues: 1) the 95% confidence intervals for the likelihood ratios are very wide, resulting in a range of possible values for each threshold and 2) the curves may meet the required likelihood ratios at multiple points, making it difficult to determine where the thresholds should be placed. These issues would be particularly exacerbated if applying the Grey Zone method to a small dataset. In contrast, the sensitivity and specificity curves used to create the TG-ROC plots are very smooth and monotonic, making it easy to ascertain the test value at which the required accuracy levels are achieved.

Both methods involve identifying accuracy levels at which to place thresholds, however these accuracy levels will only be achieved for the whole patient population if either 1) none of the patient population falls in the intermediate category or 2) there is an alternative test which is 100% accurate to diagnose those patients who fall in the intermediate category. The proof for this assumption can be found in the Appendix 5D and 5E. Therefore referring to them as 'desired' accuracy levels is slightly misleading since they only apply to the patients who fall in the positive and negative categories, which in our example is less than half of the patient population. It is also important to note that these methods are only applicable to tests where one end of the scale indicates abnormality, rather than a test such as glucose where both extremes of the scale can indicate abnormality.

The recognition of an additional, narrower range of 'uninformative' test values appeared to be a useful way of further clarifying the accuracy of the test. For very accurate tests, however, it will not be possible to identify an uninformative range if a single threshold produces both a negative likelihood ratio smaller than 0.5 and a positive likelihood ratio greater than 2. If this is the case, the upper threshold will fall earlier on the test scale than the lower threshold. In this scenario, it may be useful just to identify rule-in and rule-out thresholds or it may be useful to identify the optimal single threshold to further break down the intermediate range.

5.5.2. Limitations of 'Fixed' accuracy levels

The Grey Zone method has been criticised for failing to take into account or recognise the many complex issues which are faced when estimating clinically relevant post-test probabilities at which to place thresholds (14). It has been suggested that the initial step, the estimation of the pre-test probability and the selection of acceptable post-test probabilities, could be avoided by working directly with likelihood ratios. The TG-ROC method also addresses this issue by placing thresholds at fixed levels of sensitivity and specificity. Although appealing in terms of its simplicity, this inflexible approach to identifying intermediate test ranges will not always provide clinically relevant intermediate ranges, as priorities for ruling out and ruling in disease vary across different settings.

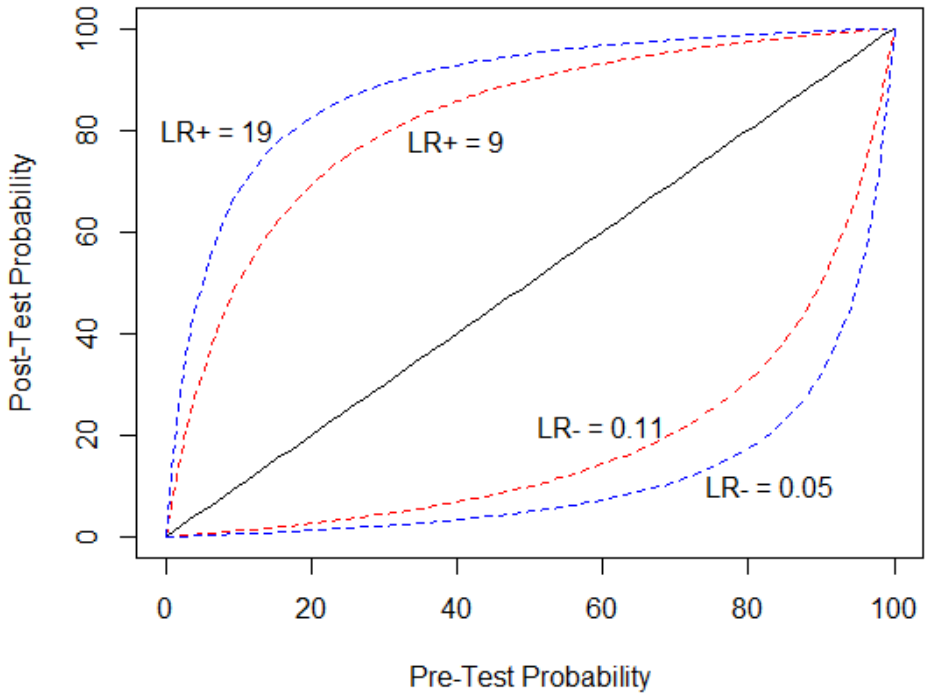


Figure 5.8. The relationship between pre-test probability and post-test probability for fixed sensitivity and specificity: 90% (red) or 95% (blue)

Figure 5.8 shows how the post-test probability varies with pre-test probability for the fixed likelihood ratios that equate to 90% and 95% sensitivities and specificities. Based on our estimated pre-test probability of disease (45%), a sensitivity and specificity of 90% equates to a LR+ of 9 and LR- of 0.11 and a PPV and NPV of 88% and 8%, respectively. These post-test probabilities are not

unreasonable for a paediatric unit, however if we consider an alternative clinical setting such as general practice, where the prevalence of serious bacterial infection is far lower (estimated to be less than 2% in the UK (15)), using these fixed accuracy levels to define an intermediate range produces less useful post-test probabilities. Thus, defining an intermediate range based on fixed levels of specificity and sensitivity overlooks the impact that the pre-test probability of disease has on the accuracy levels demanded of a test.

The Grey Zone method already involves the derivation of scenario-relevant accuracy levels at which to place thresholds, and also accounts for the prevalence of disease in the clinical setting under evaluation. Furthermore, likelihood ratios, in addition to sensitivity and specificity, are often thought of as prevalence-independent summaries of accuracy; however this has been strongly refuted in recent literature where dependence on disease spectrum has been clearly demonstrated (16). This more flexible approach facilitates the selection of more clinically relevant test categories and should also be applied to the TG-ROC method.

As a caveat to this argument, for the purposes of defining an 'uninformative' range of test results (see 5.4.4), using a fixed definition based on weak likelihood ratios (0.5 and 2) worked well. The maximal change in post-test probability that a LR+ of 2 and a LR- of 0.5 can provide is an increase or decrease of 16.6% (only achievable at a pre-test probability of 50%). These test values therefore provide minimal adjustment to the probability of disease across all levels of disease prevalence, supporting the argument that any results within this range are unlikely to be of clinical impact.

5.5.3. Clinical Implications

Serious bacterial infections, such as meningitis or pneumonia, are notoriously difficult to diagnose because many of the presenting symptoms are indicative of several conditions, most of which are self-limiting and do not require intervention. The consequences of missing serious bacterial infection, however, can be life-threatening and the ramifications of missing positive cases outweigh

those of subjecting some patients to unnecessary investigations or antibiotics. In an ideal scenario, we could identify a risk threshold which is low enough to capture the majority of serious bacterial infection, but which also does not overburden the health system and patients with redundant over-testing and treatment (17). However, the NICE clinical guideline for managing feverish children recognises that this is not possible given the current diagnostic technologies available. To allow for the uncertainty often faced in this diagnostic scenario, a traffic light system is described where an amber classification indicates an uncertain diagnosis and the need to provide the patient with safety netting advice (18).

Currently, the NICE recommendations for interpreting biomarkers for serious bacterial infection are based on a single threshold. The recognition of intermediate diagnostic test results on these quantitative test scales would highlight to clinicians when these tests are not providing sufficiently strong evidence to help rule in or rule out a diagnosis. The relatively poor accuracy of the three tests found in this dataset is consistent with results of a recent systematic review evaluating the accuracy of laboratory tests for serious bacterial infection (19).

Before the clinical implications of the current results can be considered, further exploration of how the tests perform for specific types of serious bacterial infection is necessary. Unfortunately, limitations in sample size in the current dataset make this disease-specific evaluation infeasible. A further key limitation of this analysis is that not all of the patients received all three tests. In order to evaluate comparative test performance, a dataset where the same group of patients have received all three tests would be necessary.

5.5.4. Methodological Implications

The real challenge of this task is the selection of clinically relevant accuracy levels at which to place thresholds. These existing methods sidestep this issue by either selecting fixed levels of accuracy

(already discussed as being inappropriate) or by simply stating that 'desirable' accuracy levels should be selected.

Battaglia and Pewsner argue that the thresholds can only be applied to populations similar to the one originally used to identify the threshold, and even then there are varying levels of disease severity and comorbidity within any given population which will affect the accuracy of the test (14). To account for individual variations in pre-test probability, diagnostic clinical prediction rules can be developed which allow the probability of disease to be adjusted for other key factors such as disease severity, comorbidity and treatment history. An additional advantage of these models is that they avoid the need to categorise the test scale at all, retaining all of the original information available from the test. The methods evaluated in this chapter could just as easily be applied to a risk score, as demonstrated by Santos and colleagues when applying the TG-ROC method to the results of CART and logistic regression models for the prediction of pneumonia onset after cardiac surgery (20). However, even when applied to multivariate model predictions, the issue of where to place thresholds on the risk scale so that they are clinically relevant still remains a key challenge.

Ideally, where evidence from randomised controlled trials is available, a full decision analysis can be conducted to quantify the costs associated with each test category outcome and determine the optimal placement of thresholds. Traditionally when identifying a single threshold, the trade-off is between the proportion of false positive and false negative results. However, when an intermediate category of test result is identified, the cost associated with obtaining an intermediate test result e.g. further testing or a delay in diagnosis needs to be considered (21). In our example, patients falling in the intermediate category of test result may be sent home with safety netting advice. The cost in this instance is the possible delay in diagnosing serious bacterial infection, which could impact on the effectiveness of treatment and increase the mortality rate.

The review in Chapter Three highlighted the difficulties faced at the meta-analysis level when accuracy is reported at lots of different thresholds across primary studies. When identifying an

intermediate range, the use of different cost specifications and consequently different thresholds would make it challenging to then pool the results from multiple studies for the purposes of a meta-analysis. The identification of an intermediate range of test values may therefore be more appropriate at the meta-analysis level. Traditional meta-analysis methods do not allow for thresholds to be extracted from the summary ROC curve, however a novel method which overcomes this has been proposed (22). The next chapter will therefore explore the feasibility of applying the Grey Zone method and the TG-ROC method to the summary ROC curve to derive lower and upper intermediate range limits based on evidence from multiple studies.

What this chapter adds:

- This chapter applies and compares two alternative methods for identifying an intermediate test range
- A number of issues were faced with the Grey Zone method, the most notable being that it failed to identify lower limits for all three tests under evaluation
- Neither method accounts for the costs incurred by having patients fall in the intermediate range
- This evaluation highlights the complexity of identifying clinically relevant intermediate ranges and the need for standardised methods

5.6. References

1. Greiner M, Sohr D, Gobel P. A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. J Immunol Methods. 1995 Sep 11;185(1):123-32.

2. Greiner M. Two-graph receiver operating characteristic (TG-ROC): a Microsoft-EXCEL template for the selection of cut-off values in diagnostic tests. J Immunol Methods. 1995 Sep 11;185(1):145-6.

3. Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. Int J Epidemiol. 2003 Apr;32(2):304-13.

4. Jaeschke R, Guyatt GH, Sackett DL, Group tE-BMW, Guyatt G, Bass E, et al. Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients? JAMA. 1994;271(9):703-7.

5. Knottnerus JA. The Evidence Base of Clinical Diagnosis: BMJ Books; 2002.

6. Papadakis MA, McPhee SJ, Rabow MW. Current medical diagnosis & treatment 2013: McGraw-Hill Medical; 2013.

7. Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. Brit Med J. 2012 Oct 24;345(6717).

8. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013 2013-05-16 11:27:50;346:f2778.

9. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making. 1987 Apr-Jun;7(2):107-14.

10. Bossuyt PM, Reitsma JB, Standards for Reporting of Diagnostic A. The STARD initiative. 2003;361(2985213r, l0s, 0053266):71.

11. Greiner M. Two-graph receiver operating characteristic (TG-ROC): Update version supports optimisation of cut-off values that minimise overall misclassification costs. Journal of Immunological Methods. 1996 May 10;191(1):93-4.

12. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, Vienna, Austria; 2012.

13. Steichen O, Bouvard E, Grateau G, Bailleul S, Capeau J, Lefevre G. Diagnostic value of procalcitonin in acutely hospitalized elderly patients. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology. 2009 Dec;28(12):1471-6.

14. Battaglia M, Pewsner D. Commentary: black and white or shades of grey? Int J Epidemiol. 2003 Apr;32(2):314-5.

15. Van den Bruel A, Bartholomeeusen S, Aertgeerts B, Truyers C, Buntinx F. Serious infections in children: an incidence study in family practice. BMC Fam Pract. 2006 March 28;7:23.

16. Moons KGM, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Academic Radiology. 2003 Jun;10(6):670-2.

17. Buntinx F, Mant D, Van den Bruel A, Donner-Banzhof N, Dinant GJ. Dealing with low-incidence serious diseases in general practice. British Journal of General Practice. 2011 Jan;61(582):43-6.

18. Richardson M, Lakhanpaul M, Guideline Development G, the Technical T. Assessment and initial management of feverish illness in children younger than 5 years: summary of NICE guidance. BMJ. 2007 Jun 2;334(7604):1163-4.

19. Van den Bruel A, Thompson MJ, Haj-Hassan T, Stevens R, Moll H, Lakhanpaul M, et al. Diagnostic value of laboratory tests in identifying serious infections in febrile children: systematic review. BMJ. 2011;342(7810):d3082.

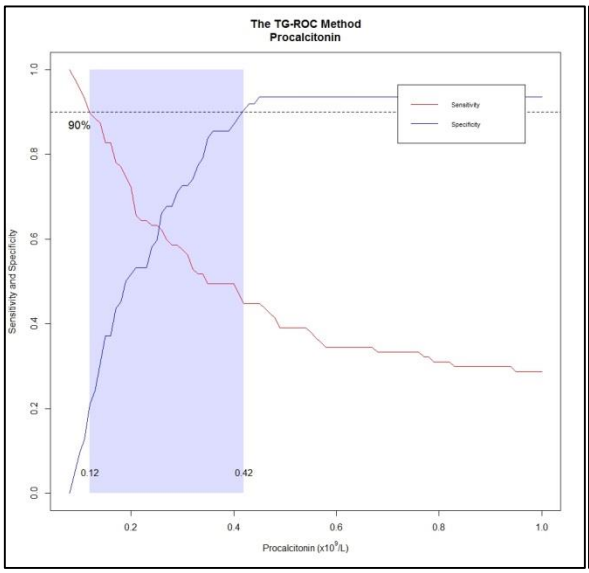
20. Santos M, Braga JU, Gomes RV, Werneck GL. Predictive factors for pneumonia onset after cardiac surgery in Rio de Janeiro, Brazil. Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America. 2007 Apr;28(4):382-8.

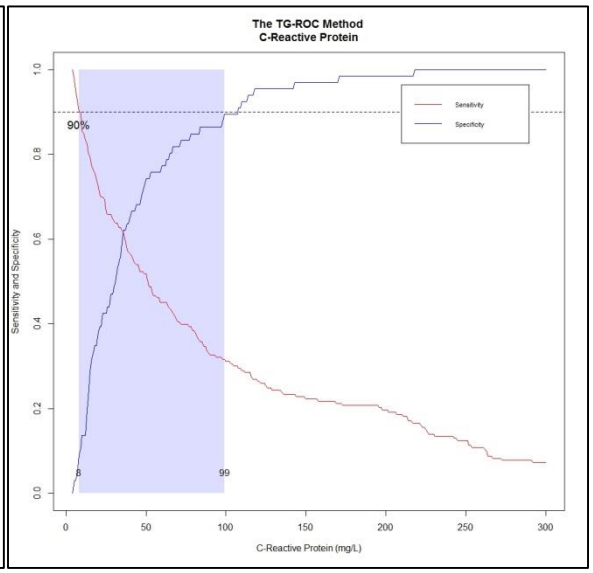
21. Tortorella F. A ROC-based reject rule for dichotomizers. Pattern Recogn Lett. 2005 Jan 15;26(2):167-80.

22. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003 Dec;59(4):936-46.

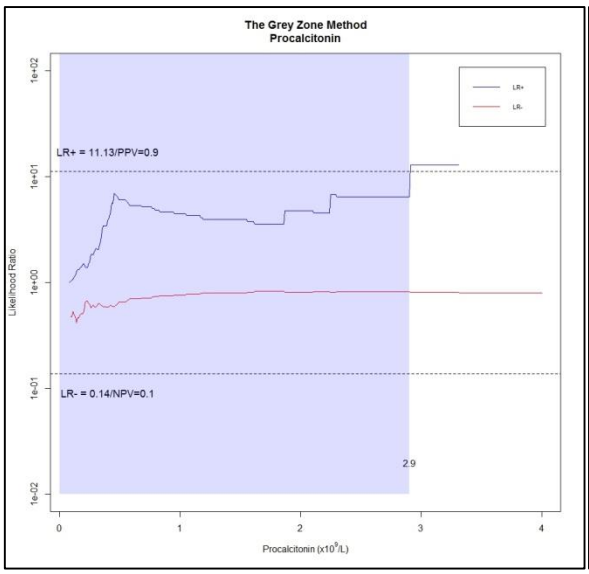
Chapter 5 Appendix

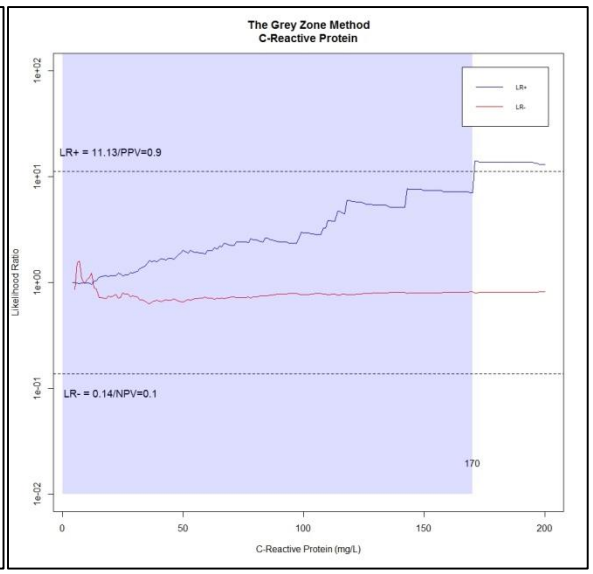
5A. TG-ROC Plots



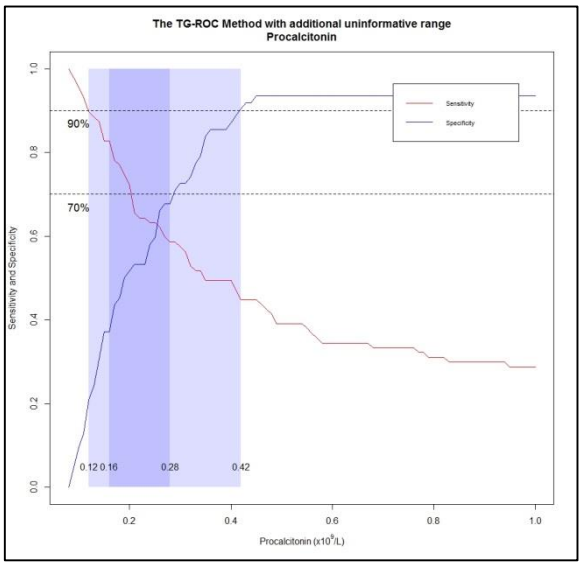


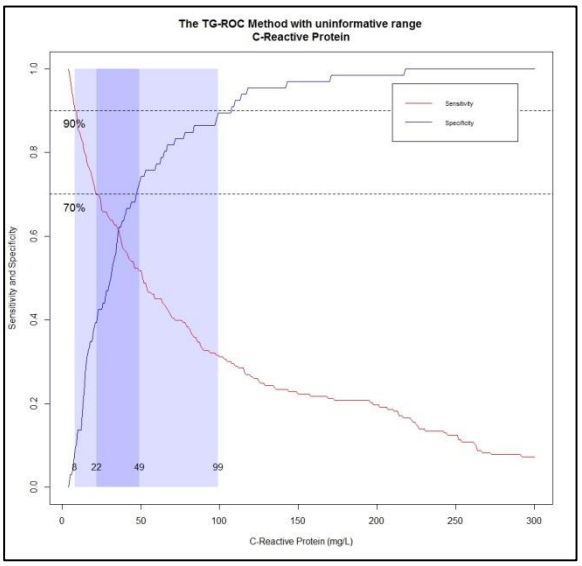
5B. Grey Zone Plots

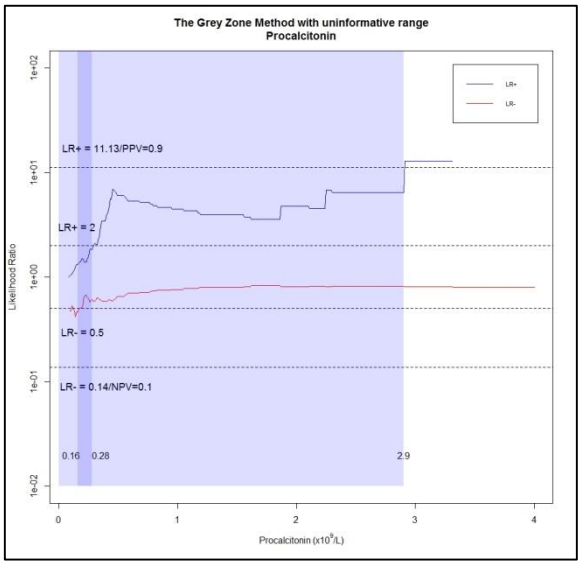


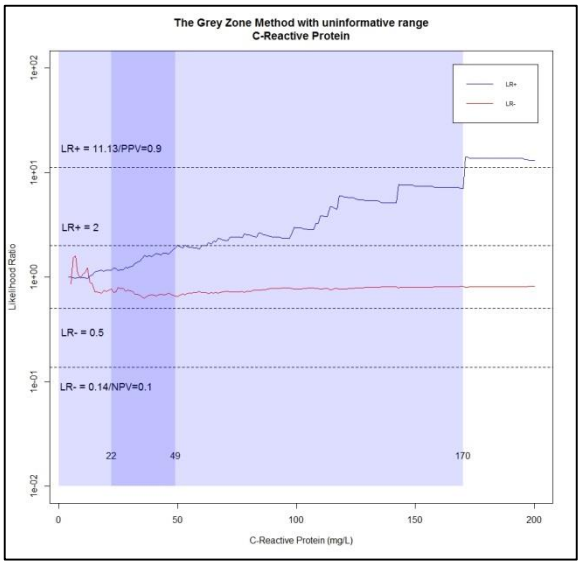


5C. Uninformative Range Plots









5D. Proof for Specificity

Theorem 1: If $SP_{t2} = 1$ when $a \le x \le b$
Then $P_o(T^- | D^-) = P_{t1}(x < b \mid D^-)$
Theorem 2: If $SP_{t2} < 1$ when $a \le x \le b$
Then $P_o(T^- | D^-) < P_{t1}(x < b \mid D^-)$

Where:
a, b are lower and upper limits of the IR
x = test value
P_o = Overall probability
P_{tn} = Probability for test n
SP_{tn} = Specificity of test n
SP_n = Specificity for specified range of values n
T^- = Test Negative
D^- = Disease Negative
$\%_k$ = Percentage of individuals in specified distribution range k

Proof

$$P_o(T^-|D^-) = (SP_{t1-(a\le x\le b)} \cdot \%_{1-(a\le x\le b)}) + (SP_{(a\le x\le b)} \cdot \%_{(a\le x\le b)})$$

$$P_o(T^-|D^-) = \left(\frac{P_{t1}(x<a|D^-)}{Pt_1(x<a|D^-)+ P_{t1}(x>b|D^-)} \cdot (P_{t1}(x>b|D^-) + P_{t1}(x<a|D^-))\right) + (P_{t2,a\le x\le b}(T^-|D^-) \cdot P_{t1}(a\le x\le b|D^-))$$

$$P_o(T^-|D^-) = P_{t1}(x<a|D^-) + (P_{t2,a\le x\le b}(T^-|D^-) \cdot P_{t1}(a\le x\le b|D^-))$$

Let $p = P_{t2,a\le x\le b}(T^-|D^-)$

$$P_o(T^-|D^-) = P_{t1}(x<a|D^-) + p\,P_{t1}(a\le x\le b|D^-)$$

If $p = 1$, $P_o(T^-|D^-) = P_{t1}(x<b|D^-)$ (3)

If $p < 1$, $P_o(T^-|D^-) < P_{t1}(x<b|D^-)$ (4)

5E. Proof for Sensitivity

Theorem 1: If $SE_{t2} = 1$ when $a \le x \le b$
Then $P_o(T^+ \mid D^+) = P_{t1}(x > a \mid D^+)$
Theorem 2: If $SE_{t2} < 1$ when $a \le x \le b$
Then $P_o(T^+ \mid D^+) < P_{t1}(x > a \mid D^+)$

Where:
a, b are lower and upper limits of the IR
x = test value
P_o = Overall probability
P_n = Probability for test n
SE_{t2} = Sensitivity of test n
SE_n = Sensitivity for specified range of values n
T^+ = Test Positive
D^+ = Disease Positive
$\%_k$ = Percentage of individuals in specified distribution range k

Proof

$$P_o(T^+|D^+) = (SE_{1-(a\le x\le b)} \cdot \%_{1-(a\le x\le b)}) + (SE_{(a\le x\le b)} \cdot \%_{(a\le x\le b)})$$

$$P_o(T^+|D^+) = \left(\frac{P_{t1}(x>b|D^+)}{P_{t1}(x>b|D^+)+P_{t1}(x<a|D^+)} \cdot (P_{t1}(x>b|D^+) + P_{t1}(x<a|D^+))\right) + (P_{t2,a\le x\le b}(T^+|D^+) \cdot P_{t1}(a\le x\le b|D^+))$$

$$P_o(T^+|D^+) = P_{t1}(x>b|D^+) + (P_{t2,a\le x\le b}(T^+|D^+) \cdot P_{t1}(a\le x\le b|D^+))$$

Let $p = P_{t2,a\le x\le b}(T^+|D^+)$

$$P_o(T^+|D^+) = P_{t1}(x>b|D^+) + p\,P_{t1}(a\le x\le b|D^+)$$

If $p = 1$, $P_o(T^+|D^+) = P_{t1}(x>a|D^+)$ (1)

If $p < 1$, $P_o(T^+|D^+) < P_{t1}(x>a|D^+)$ (2)

Chapter Six

Deriving an intermediate range of values from the results of a diagnostic accuracy meta-analysis

6.1. Overview

Research Objective: To explore the feasibility of applying two existing methods for identifying an intermediate range of test values to the results of a diagnostic accuracy meta-analysis.

Methods: Methods for defining an intermediate range of test values were applied to the results of a previously published systematic review which explored the accuracy of two commonly used inflammatory markers for detecting serious bacterial infection in children: procalcitonin and C-reactive protein. Summary ROC curves were generated using a hierarchical model proposed by Dukic and Gatsonis (1), from which intermediate range thresholds were extracted according to each method. Point-wise confidence intervals were produced for each threshold estimate using simulation.

Results: Problems were encountered when extracting thresholds from the summary ROC curve as it requires extrapolation beyond the primary study data, resulting in the recognition of infeasible (negative) test values. If restricted to the range of reported primary data, the limits of the intermediate range fall outside the scope of the summary ROC curve in some cases, and therefore cannot be estimated.

Conclusion: Although Dukic and Gatsonis' method is advantageous in that it allows accuracy results for multiple thresholds to be incorporated per study, thus minimising information loss, it is still restricted by limited reporting in the primary studies. The need for complex models to reconstruct study-specific ROC curves can easily be circumvented if accuracy across all thresholds is reported in an extractable way at the primary study level.

6.2. Introduction

In Chapter Three it was shown that the accuracy of diagnostic tests on a quantitative scale is typically evaluated at a single or multiple binary thresholds in primary research. The dependence of test performance on the diagnostic threshold selected is typically conveyed using ROC curves, which describe accuracy across the whole test scale. For the purposes of a meta-analysis though, the reporting of this plot alone is not helpful as the underlying data cannot be extracted. This information loss is problematic as it then hampers the possibility of exploring optimal decision threshold(s) in meta-analyses and economic evaluations further down the research pathway.

Traditional methods for diagnostic accuracy meta-analysis are based on the pooling of a single 2 x 2 table from each primary study. A new method has been proposed for quantitative tests which reduces information loss by allowing more than one 2 x 2 table per primary study to be incorporated into the meta-analysis (1). Instead of pooling the performance of the test at a single threshold, study-specific ROC curves are estimated from the available data and then these are pooled to form a 'summary ROC curve'. Unlike the summary ROC curves produced using standard methods which depict between-study heterogeneity in accuracy; the summary ROC curve produced by this method is an estimate of the pooled accuracy across the whole test scale. This has the potential to overcome the issue of information loss and provide sufficient evidence to allow threshold exploration at the meta-analysis level.

In the previous chapter, methods for identifying an intermediate range of test values were evaluated. Reporting accuracy at just two thresholds in primary studies, however, will still result in significant information loss at the meta-analysis level. Furthermore, these data-driven methods will also perpetuate the problem of pooling accuracy estimates at inconsistent thresholds. Dukic and Gatsonis' method facilitates the extraction of pooled sensitivity and specificity estimates at all thresholds from the summary ROC, which means it may be possible to apply these methods at the meta-analysis level.

In this chapter, I explore the feasibility of applying two existing methods for identifying an intermediate range of test values to the results of a diagnostic accuracy meta-analysis.

6.3. Methods

A systematic review evaluating the accuracy of laboratory tests for the diagnosis of serious bacterial infection in children has been previously published (2). In this chapter, data from this review is re-analysed. The review consisted of an extensive literature search and a full quality assessment of the relevant primary studies. The authors of the review found the between-study heterogeneity to be acceptable for procalcitonin and C-reactive protein, and therefore the data from these primary studies were extracted for inclusion in a meta-analysis. For both tests, there were primary studies that reported accuracy at multiple binary thresholds. Consequently the Dukic and Gatsonis method was applied to allow all of the available data to be included in the meta-analysis (1).

The authors of the review kindly provided the primary study data included in their meta-analyses for these two tests. The Dukic and Gatsonis model was programmed and applied using the statistical software R (R core team, 2013).

6.3.1. Statistical Analyses

Estimation of the summary ROC model

This section will describe the meta-analysis model proposed by Dukic and Gatsonis, adopting their notation[1] (1). For further depth (for example, details on how to incorporate more than one covariate or the random-effects extension model), readers are advised to refer back to their original methods paper.

At the individual study level, the ordinal categorical test results (reported in *J* ordered categories) can be modelled using a simple ordinal linear regression model [eq. 6.1]. The only covariate in the

[1] Please note that the formula in the original paper contained an error (no intercept parameter) but the correct formula (as agreed via personal correspondence with the authors) is presented here.

model is the disease status of the patient (D_i = 1 if disease is present and D_i = 0 is disease is absent). Estimation of the intercept and location (α_1 and α_2) and scale (β) parameters of the ordinal regression model allows derivation of the false-positive and true-positive rates for any threshold θ. By applying the monotone logit function (g) and assuming that Y_i (the estimated test result for the i-th patient) comes from a latent continuous variable, the linear model is transformed to produce a smooth, study-specific, ROC curve.

$$g\{P(Y_i \leq j|D_i)\} = \frac{\theta_j - \beta D_i}{exp(\alpha_1 + \alpha_2 D_i)}$$

[eq. 6.1]

These study-specfic ordinal regression models are used as the basis of a fixed-effects model for estimating a summary ROC curve at the meta-analysis level (1). A common underlying ROC curve across studies is assumed i.e. each study has the same scale and location ROC curve parameters, but with independent sets of thresholds for each study. By only having one covariate in the model (disease status D_i), all between-study differences are assumed implicitly to be a result of threshold differences.

Expanding on eq. 6.1 to accommodate data from K studies, the intercept, location and scale parameters of the summary ROC curve can be estimated via numerical maximisation of the function described in eq. 6.2.

$$L\,(\Theta|\mathrm{Y}) = \prod_{k=1}^{K} \prod_{j=1}^{J_k} \prod_{i \in G_{jk}} \left[g^{-1}\left(\frac{(\theta_{j,k} - \beta D_{i,k})}{exp(\alpha_1 + \alpha_2 D_{i,k})}\right) - g^{-1}\left(\frac{(\theta_{j-1,k} - \beta D_{i,k})}{exp(\alpha_1 + \alpha_2 D_{i,k})}\right)\right]^{I(y_{ik}=j)}$$

[eq. 6.2]

where Θ is a vector consisting of the thresholds from every study, the location and scale parameters (θ, α, β). G_{jk} is the subset of patients from study k who has a test result that falls in the j-th category and $I(\cdot)$ denotes the indicator function.

The summary ROC curve can then be constructed by plotting the estimated false-positive and true-positive rate pairs across all thresholds θ using the estimated intercept ($\hat{\alpha}_1$), location ($\hat{\alpha}_2$) and scale ($\hat{\beta}$) parameters [eq. 6.3].

$$\widehat{TPR}(\theta) = g^{-1}\left(\frac{\left(\theta_{j,k} - \hat{\beta}D_{i,k}\right)}{exp\left(\hat{\alpha}_1 + \hat{\alpha}_2 D_{i,k}\right)}\right)$$

$$\widehat{FPR}(\theta) = g^{-1}(\theta)$$

[eq. 6.3]

Applying the intermediate range methods to the estimated summary ROC curves

The Grey Zone method and the TG-ROC method for identifying a range of intermediate values were applied to the estimated summary ROC curves for procalcitonin and C-reactive protein.

A full description of these two methods can be found in **Chapter Five**. To apply the TG-ROC method, the intermediate range was placed at the thresholds at which each test achieved an estimated pooled sensitivity and specificity of 90%. To apply the Grey Zone method, positive and negative likelihood ratios were derived across all thresholds from the sensitivity and specificity summary ROC estimates using standard formulae (see Eq. 1.1). So that results are comparable with those in **Chapter Five**, the same pre-test probability of 44.7% is assumed, and results for positive and negative post-test probabilities of 0.9 and 0.1 are reported.

All analyses and plots presented in this chapter were produced using R (3).

6.4. Results

6.4.1. Primary Study Data

Data were available from three primary studies for procalcitonin as a predictor of serious bacterial infection (Table 6.1). Accuracy was reported at multiple thresholds in two of the studies, and three distinct thresholds have been used across all three studies.

Study	Threshold (ng/mL)	Sample Size	True Positives	False Positives	False Negatives	True Negatives
Andreola 2007	0.5	408	69	74	25	240
Andreola 2007	1.0	408	60	32	34	282
Andreola 2007	2.0	408	45	11	49	303
Galetto-Lacour 2008	0.5	202	51	47	3	101
Thayyil 2005	0.5	72	7	32	1	32
Thayyil 2005	2.0	72	4	9	4	55

Table 6.1. Procalcitonin results for the diagnosis of serious bacterial infection reported in primary studies

For C-reactive protein, data were available from five studies. Accuracy has been reported for a wide range of thresholds (Table 6.2), and two of the studies reported accuracy at multiple thresholds.

Study	Threshold (ng/mL)	Sample Size	True Positives	False Positives	False Negatives	True Negatives
Andreola (2007)	20	408	83	123	11	191
Andreola (2007)	40	408	67	59	27	255
Andreola (2007)	80	408	43	17	51	297
Berger (1996)	20	127	25	32	5	65
Galetto-Lacour (2008)	40	202	44	36	10	112
Hsiao (2006)	2	387	41	246	0	100
Hsiao (2006)	5.2	387	35	158	6	188
Hsiao (2006)	9.8	387	21	68	20	278
Thayyil (2005)	50	72	6	20	2	44

Table 6.2. C-reactive protein for the diagnosis of serious bacterial infection reported in primary studies

6.4.2. Producing the Summary ROC Curves

The estimated summary ROC curve for procalcitonin can be found in Figure 6.1. The model estimated the scale (β) and location (α_2) parameters as 0.57 (95% CI: 0.23 to 0.91) and 2.10 (95% CI: 1.66 to 2.55) respectively, with a model intercept (α_1) of -0.47 (95% CI: -0.61 to -0.32). The summary ROC curve appears to be largely driven by the data provided in Andreola (2007) and this is mainly due to the larger comparative sample size in this study.

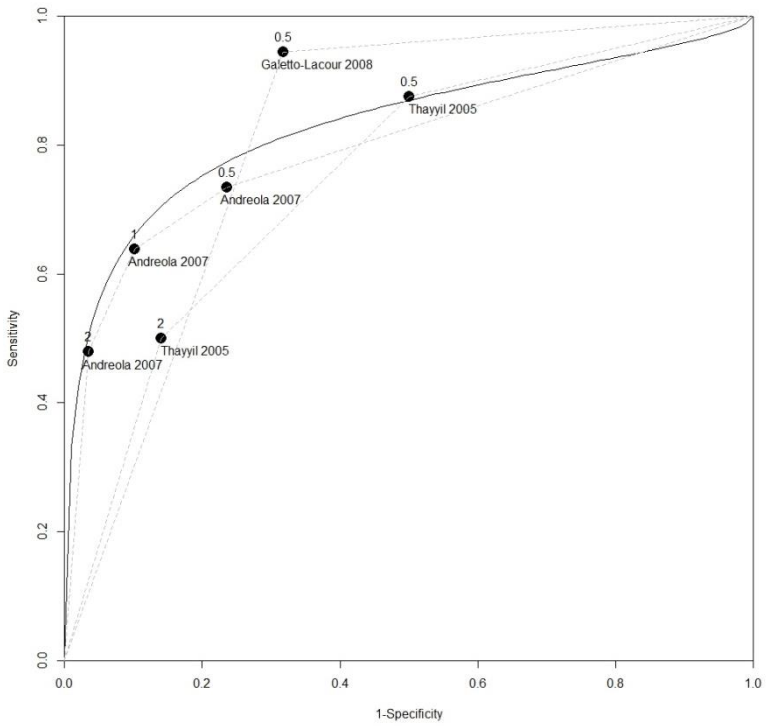


Figure 6.1. Summary ROC curve for procalcitonin: data points from primary studies and thresholds are highlighted with dashed lines denoting study-specific empirical ROC curves

The estimated summary ROC curve for C-reactive protein can be found in Figure 6.2. The estimated location and scale parameters were 0.40 (95% CI: 0.18 to 0.62) and 66.63 (95% CI: 57.53 to 75.73), respectively, with an intercept of 2.98 (95% CI: 2.90 to 3.07). The summary ROC curve does not appear to be a particularly good fit; it fails to run through the core cluster of study results. This is likely to be driven by the data-point from the Hsiao study estimating 100% sensitivity at a threshold of 2ng/mL. This study has a relatively large sample size and therefore will have significant weight in the overall model, thus increasing the sensitivity of the whole summary ROC curve.

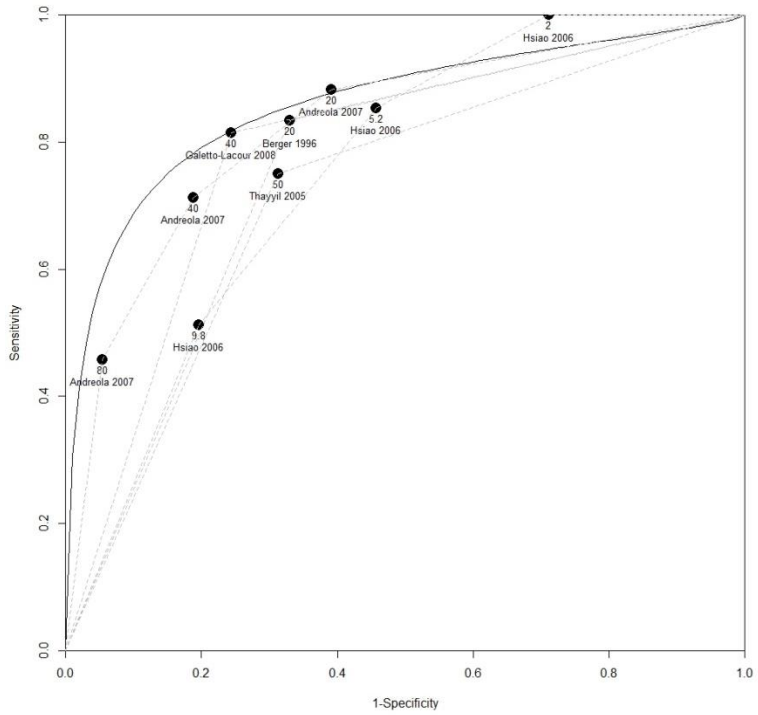


Figure 6.2. Summary ROC curve for C-reactive protein: data points from primary studies and thresholds are highlighted with dashed lines denoting study-specific empirical ROC curves

6.4.3. The TG-ROC Method

The thresholds at which 90% sensitivity and 90% specificity were achieved on the estimated summary ROC curve (based on the Dukic and Gatsonis method) for C-reactive protein and procalcitonin were identified.

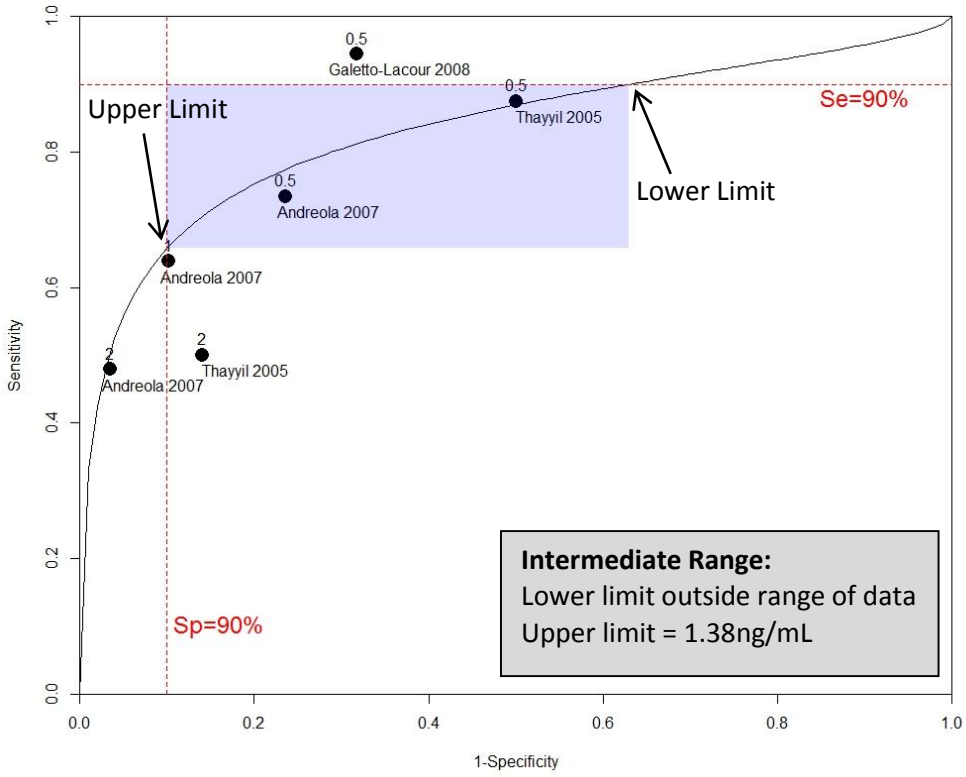


Figure 6.3. Summary ROC curve for procalcitonin with TG-ROC intermediate range highlighted in grey. Studies and thresholds are also plotted

The TG-ROC intermediate range plotted on the summary ROC curve for procalcitonin can be found in Figure 6.3. The upper limit of the intermediate range was 1.38ng/mL, however an implausible (negative) test value was estimated for the lower limit. The maximum sensitivity achieved in the primary studies was 87.5% at a threshold of 0.5ng/mL, and therefore the lower limit of the intermediate range fell outside the range of the primary study data and is based on extrapolation. This can be seen clearer on the TG-ROC plot (Figure 6.4), where the test scale has been limited to the range of thresholds at which accuracy has been reported in the primary studies.

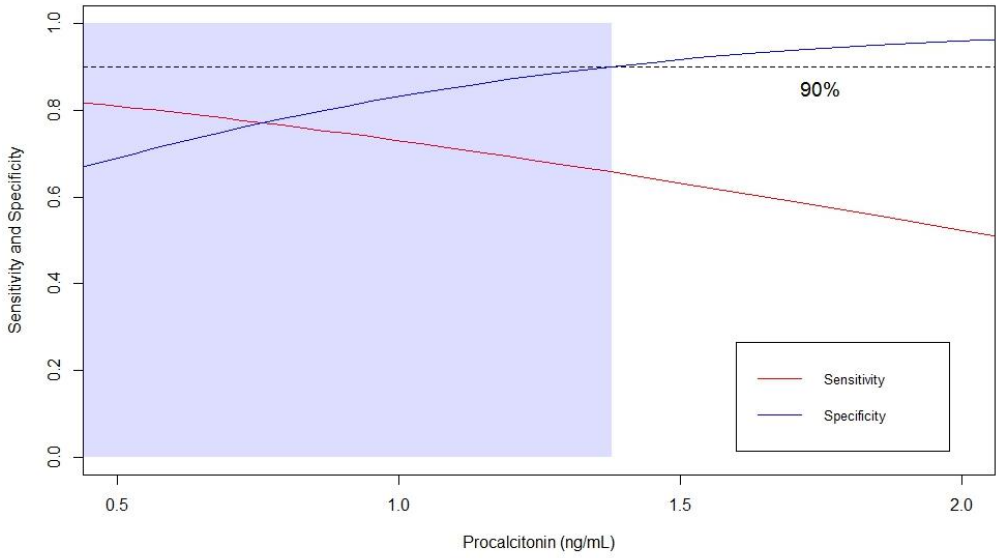


Figure 6.4. TG-ROC curve extracted from the Dukic and Gatsonis summary ROC curve, restricted to the range of thresholds reported in the primary studies

The TG-ROC intermediate range for C-reactive protein can be found in Figure 6.5. In this instance, the intermediate range fell within the range of the primary study data and the model therefore produced conceivable lower and upper intermediate range limits of 1.6 mg/L and 43.4mg/L.

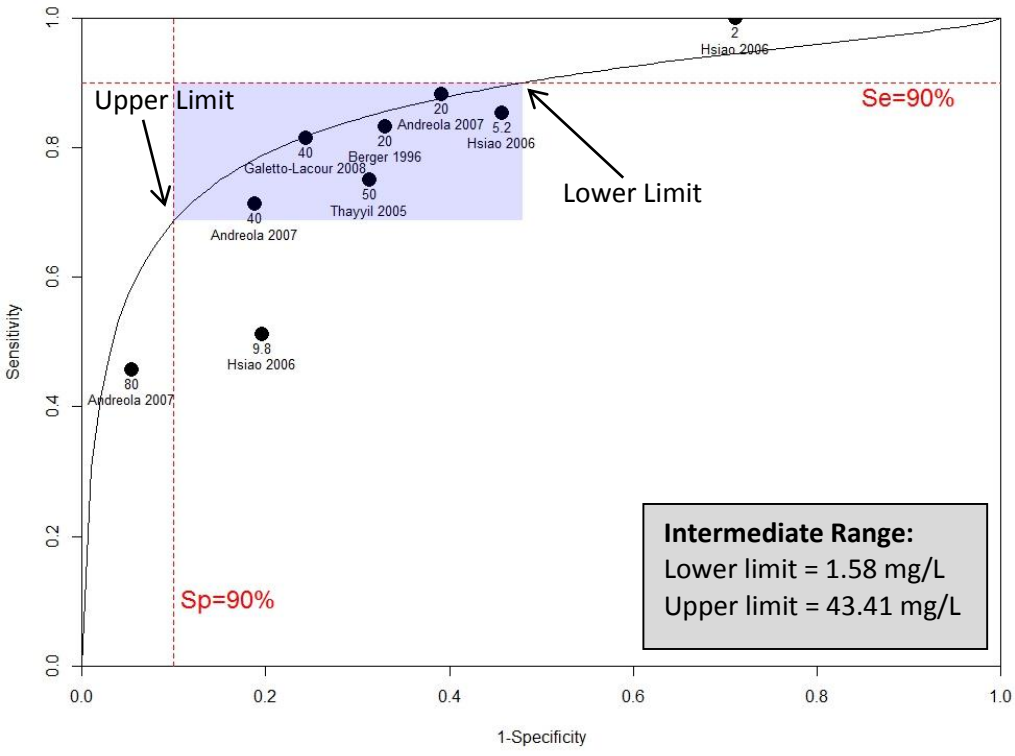


Figure 6.5. Summary ROC curve for C-reactive protein with TG-ROC intermediate range highlighted in grey. Studies and thresholds are also plotted.

6.4.4. The Grey Zone Method

The prevalence of serious bacterial infection in the primary studies was highly variable, making it difficult to identify the most relevant pre-test probability. For a pre-test probability of 44.7% and positive and negative post-test probabilities of 0.9 and 0.1, respectively, the upper intermediate range limit falls at a positive likelihood ratio of 11.1 and the lower limit at a negative likelihood ratio of 0.14.

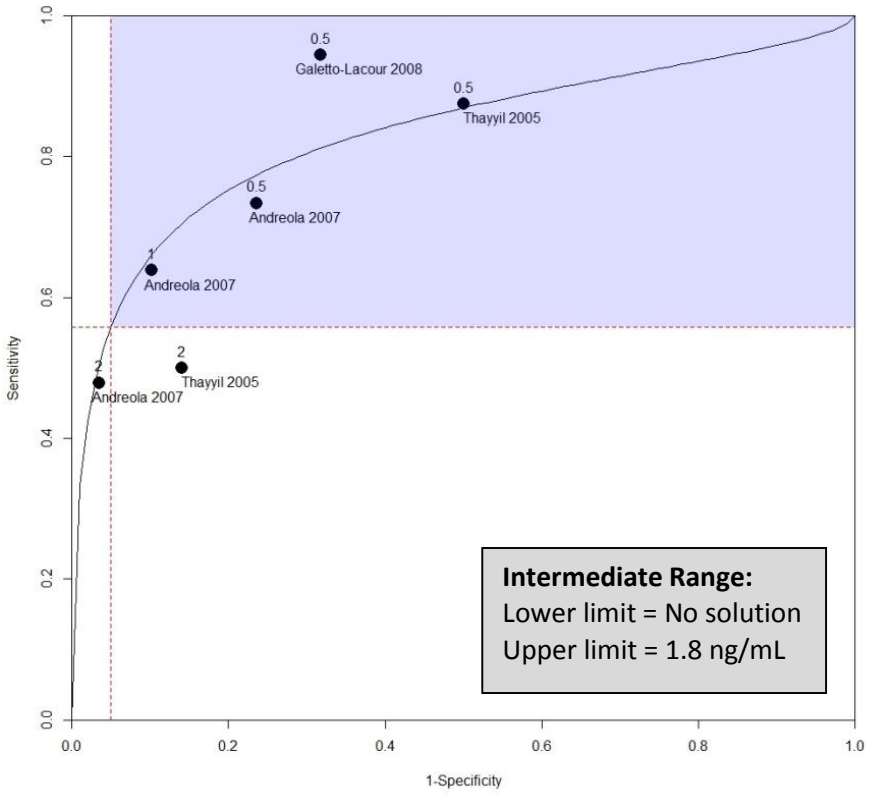


Figure 6.6. Summary ROC curve for procalcitonin with Grey Zone intermediate range highlighted in grey. Studies and thresholds are also plotted

Figure 6.6 shows the Grey Zone intermediate range plotted on the Dukic and Gatsonis sROC curve for procalcitonin. The negative likelihood ratio curve does not achieve an NPV of 0.1 and therefore a lower limit cannot be identified. In contrast to the previous TG-ROC results for procalcitonin where it was possible to extract a lower, albeit implausible, threshold from an extrapolated section of the model, in this instance the summary ROC model predicted that the best negative likelihood

ratio procalcitonin could ever achieve is 0.26. The estimated upper threshold of the intermediate range was 1.8ng/mL.

Figure 6.7 shows the equivalent summary ROC curve and likelihood ratio curves for C-reactive protein with the Grey Zone intermediate range highlighted in grey. The same issue that the test never achieving the required negative likelihood ratio to identify a lower intermediate range limit was encountered. An upper threshold of 58.2mg/L was identified.

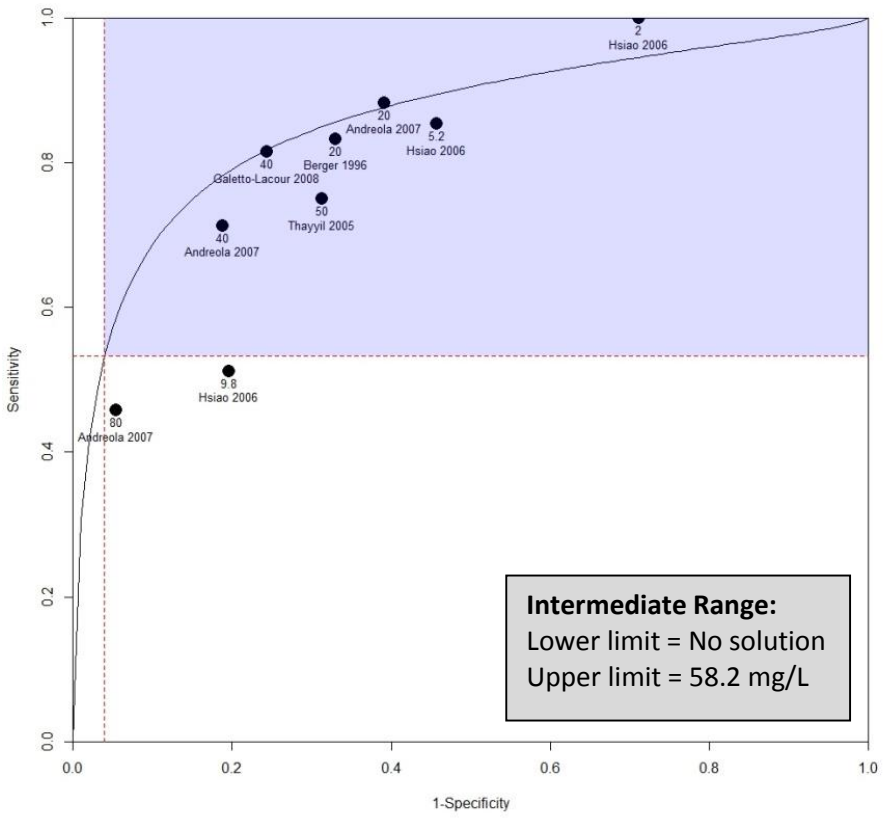


Figure 6.7. Summary ROC curve for C-reactive protein with Grey Zone intermediate range highlighted in grey. Studies and thresholds are also plotted

6.5. Discussion

When applying the TG-ROC method to Dukic and Gatsonis' summary ROC curve for procalcitonin, the estimated threshold at which the test achieved 90% sensitivity was negative and therefore implausible. This value fell was beyond the range of thresholds at which accuracy had been reported in the primary studies and therefore was based on extrapolation. It is therefore possible that the method may fail to produce a solution if the limits of the intermediate range fall outside the range of thresholds at which accuracy is reported in the primary studies.

For procalcitonin, this is a very restrictive limitation as only a very narrow range of thresholds were reported in the primary studies identified. In Chapter Three it was seen that once one threshold has been adopted in a primary study for a given test, then the same threshold is often adopted in subsequent studies. For example, a threshold of 35μ/ml was used in the majority of primary studies evaluating the accuracy CA-125 for the detection of ovarian cancer (4). This practice would greatly limit the potential for exploring optimal thresholds at the meta-analysis level.

Despite this limitation, Dukic and Gatsonis' method offers many benefits compared to traditional methods for meta-analysis. Firstly, it allows the incorporation of accuracy data at multiple thresholds within a single study, regardless of whether they are consistent with the thresholds reported in other primary studies. This allows estimation of a summary ROC curve which actually depicts accuracy across the test scale, rather than the traditional summary ROC curve which just shows the heterogeneity in pairs of sensitivity and specificity between studies.

There are other meta-analysis methods which also accommodate accuracy data if it has been reported at multiple thresholds (5, 6). Putter et al.'s approach is based on methods developed for survival data, but requires accuracy to be reported at the same multiple thresholds across studies (5). The primary data available in the present review reported accuracy at inconsistent thresholds and, as seen in Chapter Three, this is typical of most reviews. Hamza et al.'s multivariate random-

effects model overcomes this issue by allowing for missing data and a possible extension of this chapter would be to apply the intermediate methods to the results of this method.

A key limitation of this method, however, is that there is no restriction on the range of test values for which accuracy is estimated. Both of the summary ROC curves modelled in this analysis had substantial areas of the curve based on negative test thresholds which is clearly unfeasible. Strong caution should therefore be taken when interpreting areas of the estimated summary ROC curve extrapolated beyond the range of the primary data (5, 7). A possible adaptation which would overcome this issue is to limit the summary ROC curve to the range of thresholds reported in the primary studies.

Although Dukic and Gatsonis' model overcomes some of the issues of limited reporting of test accuracy at the primary study level for quantitative diagnostic tests, the analyses presented in this chapter highlight that there remains significant problems when extracting accuracy at specific thresholds. Further methodological work needs to be carried out to explore whether the model can be restricted to clinically plausible values and therefore better represent the distribution of test values. If this is possible, then this model could be a viable method for meta-analyses of historical existing primary studies.

An extension of this analysis could be to apply the random effects equivalent of this model proposed by Dukic and Gatsonis (1). The random effects model allows the within- and between-study variation to be explicitly modelled, which was not a primary objective of the present analysis but may be preferable if applying these methods to future datasets. The authors found that using a random effects model instead of the fixed effects model produces wider point-wise confidence intervals. As this has little effect on the point estimates themselves, there is no reason to assume that the results reported in the current analyses would be substantially different if a random effects model was implemented instead.

For future research however, the loss of information by reporting accuracy at one or a handful of selected binary thresholds could be easily overcome by reporting accuracy across the whole of the test scale in a way that can be extracted for meta-analysis (7, 8). This would circumvent the need for complicated models that attempt to 'recreate' study-specific ROC curves, which are likely to be subject to significant error if based on just a few data points. We therefore need to identify methods for reporting accuracy across the whole test scale in a way that is extractable for the purposes of meta-analysis.

What this chapter adds:

- This chapter explores the feasibility of extracting intermediate range limits from a novel meta-analysis method for producing summary ROC curves
- Although Dukic and Gatsonis' model allows for the inclusion of accuracy data at multiple thresholds, the fit of the model is still heavily compromised by limited reporting in primary research
- More complete data reporting in primary studies would however overcome this issue and allow for threshold selection at the meta-analysis level

6.6. References

1. Dukic V, Gatsonis C. Meta-analysis of diagnostic test accuracy assessment studies with varying number of thresholds. Biometrics. 2003 Dec;59(4):936-46.

2. Van den Bruel A, Thompson MJ, Haj-Hassan T, Stevens R, Moll H, Lakhanpaul M, et al. Diagnostic value of laboratory tests in identifying serious infections in febrile children: systematic review. BMJ. 2011;342(7810):d3082.

3. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, Vienna, Austria; 2012.

4. Medeiros LR, Rosa DD, da Rosa MI, Bozzetti MC. Accuracy of CA 125 in the diagnosis of ovarian tumors: a quantitative systematic review. Eur J Obstet Gynecol Reprod Biol. 2009 Feb;142(2):99-105.

5. Putter H, Fiocco M, Stijnen T. Meta-Analysis of Diagnostic Test Accuracy Studies with Multiple Thresholds using Survival Methods. Biometrical Journal. 2010 Feb;52(1):95-110.

6. Hamza TH, Arends LR, van Houwelingen HC, Stijnen T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. BMC Med Res Methodol. 2009;9:73.

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

8. Trikalinos TA, Siebert U, Lau J. Decision-Analytic Modeling to Evaluate Benefits and Harms of Medical Tests: Uses and Limitations. Medical Decision Making. 2009;29(5):22.

Chapter Seven

Discussion

7.1. Summary of Research Findings

Quantitative clinical tests often struggle to discriminate between patients with and without a given disease, where certainty about disease status cannot be established for a subset of the patients due to the significant overlap between the distributions of test results. In these cases, the current practice of selecting a single test value to treatment decisions can result in a considerable number of potentially costly errors, thus failing to maximise the clinical utility of the test.

The literature review in **Chapter Two** revealed the shortcomings of dichotomising quantitative diagnostic test scales to analyse accuracy and interpret test results. A number of alternative methods that facilitate a richer interpretation of quantitative test results were identified in the literature; however there was a significant lack of evidence regarding the extent to which these methods are currently used in research and whether clinicians would find these methods more helpful when interpreting test results in practice.

The objective of the systematic review of diagnostic accuracy research on cancer biomarkers presented in Chapter Three was to ascertain which methods are being used in contemporary research to analyse accuracy at both the primary study and meta-analysis level. The results of this review indicated that dichotomising the test scale remains the most commonly employed method in primary diagnostic accuracy research; however it was also evident that this practice leads to notable information loss and significantly limits the analyses possible at the meta-analysis level.

Chapter Four presents the results of a survey of GPs exploring preferences for threshold guidance when interpreting quantitative diagnostic test results. The majority of general practitioners reported that they would find a greater depth of information helpful in their clinical decision making than is provided by a single threshold. Rule-in and rule-out thresholds were considered the most helpful, but multiple categories (i.e. more than 3) and result-specific interpretations were also popular. These findings raise the question of why the accuracy of quantitative tests is currently assessed at a single threshold in research when this not the 'preferred' method for interpreting these test results in clinical practice.

Given that there are currently no standardised methods for identifying thresholds for an intermediate range of test results (or a greater number of categories), an evaluation of two existing methods is presented in **Chapter Five**. Both methods were applied to a dataset of test results from biomarkers for detecting serious bacterial infection in children. A number of problems were faced when applying the Grey Zone method to real clinical data, whereas the TG-ROC method was very straightforward to both understand and implement. Recognition that differences in pre-test probability and priorities for ruling in and ruling out disease will affect the selection of thresholds is imperative however, and this was a key shortcoming of the TG-ROC method. Neither method attempted to resolve the key challenge of how to resolve this latter trade-off of costs.

In **Chapter Six**, the Grey Zone and TG-ROC methods are applied to the results of a meta-analysis using a model proposed by Dukic and Gatsonis for estimating a summary ROC curve. The results demonstrated that the application of these methods at the meta-analysis level is feasible so long as thresholds are being placed at accuracy levels for which there is primary data available. This issue could be overcome by reporting accuracy across the whole of the test scale in primary research in a way that can be extracted for meta-analysis. This would remove the need to 'recreate' study-specific ROC curves and facilitate the direct calculation of pooled accuracy statistics across all test values.

7.2. Implications of Research Findings and Future Work

7.2.1. Recommendations for Reporting and Analysing Inconclusive Results

The findings in the initial literature review facilitated the development of recommendations for the reporting and analysis of diagnostic accuracy studies when inconclusive results are present (1). These recommendations distinguish between 'valid' and 'invalid' inconclusive results. Invalid inconclusive results are those where the key diagnostic feature is either uninterpretable or the actual result is missing. Valid inconclusive results are those where an adequate test result has been obtained, but the result is not clearly positive or negative. This distinction affects how inconclusive test results should be reported and analysed.

Recommendations for Reporting Inconclusive Results

The STARD statement recommends that a flowchart of participants at each stage of the study is reported. This flowchart should be used to facilitate the transparent reporting of all inconclusive results. The number of participants meeting the eligibility criteria (intention to diagnose) should feature in the flow diagram.

The ways in which valid inconclusive results should be analysed depends on the measurement scale of the test: results can be continuous, ordinal, or categorical in nature. For tests on a continuous measurement scale, two or more thresholds can be selected (depending on the richness of interpretation required), leaving a range or ranges of valid inconclusive test values. The number of patients with positive and negative test results for a disease in each category should be cross tabulated, with extra rows to account for any additional categories. A full description of how and when thresholds have been selected should be included in the methods section. In addition to the classification table, it is essential to show the distribution of the raw test results, stratified by disease status (determined by the reference standard). Possible graphical options include paired histograms, dot plots, or cumulative distribution graphs.

For categorical and ordinal index tests, the number of patients in each category should be reported and broken down by presence or absence of the target condition. For ordinal tests with a large number of categories (such as questionnaire scores), it may be sensible to group some of the categories in the classification table. However, it is still important to report a cross tabulation or plot of the frequencies in each original category and explain how the category groupings were determined.

Invalid results should be reported separately from the cross tabulation of valid results by disease status, in addition to any known underlying causes, so that the reader can assess whether they hold any diagnostic value. Clear reporting and discussion of whether these results are related to the patient's disease status, the presence of an alternative target condition, or assumed to be unrelated to patient health enables transparency in how these results should be handled.

Recommendations for Analysing Inconclusive Results

There is no single 'optimal' approach to analysing inconclusive results; diagnostic accuracy should always be analysed in line with how the test will be used in clinical practice.

There are few instances where the exclusion of valid inconclusive results from the analysis can be justified. This approach can lead to overstated summary statistics and the promotion of suboptimal test strategies. One possible approach is to exclude valid inconclusive results from binary statistics but report an additional summary statistic that accounts for them. The risk of simply providing an additional statistic to account for inconclusive results is that readers might struggle to interpret such unfamiliar statistics and interpret only the more popular accuracy measures, such as sensitivity and specificity. Furthermore, these additional statistics are not typically included in meta-analyses, where usually only the sensitivity and specificity are analysed. One way of overcoming the issue of analysing valid inconclusive results is to group them with either the positive or negative results, depending on how these patients would be treated in the clinical context.

Carrying out a sensitivity analysis can help to demonstrate how the grouping decision has impacted on the reported accuracy results.

These reporting and analysis recommendations should enable readers to fully understand if and how inconclusive results have been handled in analyses and provide them with sufficient information to recalculate key statistics if they disagree with the approach adopted by the author. The key next step in this research is to meet with the committee that is currently updating the STARD statement, and encourage these new recommendations to be incorporated.

7.2.2. Clinician Preferences for Interpreting Quantitative Tests

Ascertaining the most helpful way in which quantitative results can be presented to clinicians is another methodological issue that has so far been overlooked in the literature.

The results of the survey exploring preferences for threshold guidance provided overwhelming evidence that GPs would find a greater depth of information helpful in their clinical decision making than is provided by a single threshold. GPs also reported that the seriousness of the disease, its prevalence and the accuracy of the test may also impact their preferences. It seems therefore that there may not be a 'one-fits-all' threshold model across clinical scenarios, and further research is required to identify what factors influence threshold preferences.

For example, the clinical setting may be a key consideration; those who deal with acute conditions where action must be taken immediately may value the simplicity of the dichotomous presentation whereas GPs who typically have more time to carry out gather further evidence to refine their diagnosis may benefit from a greater level of detail. Furthermore, the predictive value of diagnostic tests is notably limited in general practice by low disease prevalence and early patient presentation (2). A test that is capable of producing adequate accuracy in the secondary care setting will not have as strong predictive value in general practice, and therefore it may be helpful to explicitly highlight that only extreme test values are going to be useful for ruling in or ruling out disease.

7.2.3. Rethinking test accuracy methods for quantitative tests

The application of two threshold selection methods in Chapter Four and Chapter Five highlight that accuracy information alone is insufficient to determine clinically-relevant thresholds for a particular scenario. In addition to considering accuracy, both the health and economic costs relating to each possible patient outcome need to be considered. This information is not available from a standard diagnostic accuracy study and therefore the issue of threshold selection is best tackled at the final modelling stages of the research pathway, when patient outcome and economic data should be available.

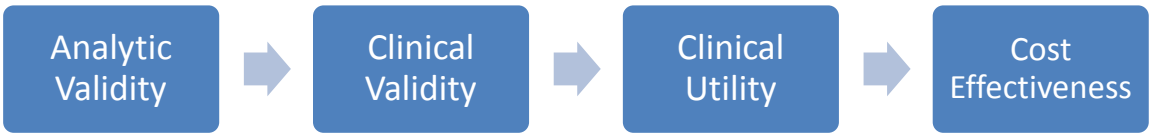


Figure 7.1. Phases of evidence collection necessary to advance a test from research to practice

Information about how accuracy varies across the spectrum of the test scale must therefore be easily extractable from primary research so that it can be later combined in a meta-analysis, the results of which will feed directly into a decision model at the cost-effective stage (see Figure 7.1). The findings of the systematic review in Chapter Three and the analyses in Chapter Six demonstrate that the methods currently used to evaluate quantitative tests in primary research do not facilitate this. A single threshold is typically selected to dichotomise quantitative test results so that they fit into the traditional 2 x 2 diagnostic framework. Reporting accuracy at just one threshold causes notable information loss as key details about how accuracy varies across the test scale are lost.

New methods which allow the extraction of test accuracy across all thresholds are consequently needed for primary research. One possible and very simple solution would be to adapt the existing

ROC plot, perhaps by adding a third axis so that it allows the extraction of sensitivity and specificity across the test scale. Alternatively, a model such as that proposed by Irwig and colleagues (3) may better allow for the adjustment of other covariates. At the same time, methods also need to be developed that accommodate and utilise this additional information at the meta-analysis level. Methods for pooling continuous test accuracy data do exist (4, 5), although they have rarely been applied and fail to account for heterogeneity issues such as the impact of disease spectrum, study design or patient sampling technique.

7.2.4. One size does not fit all: thinking beyond the 2 x 2 diagnostic framework

The methodological adjustments for quantitative tests suggested in the previous section would only solve the problem of threshold selection. In the very early stages of this research, I assessed how different types of inconclusive test results should be reported and analysed (1). This demonstrated that the tendency to mould all test accuracy evaluations to fit the 2 x 2 diagnostic framework can be problematic even for tests that do naturally produce dichotomous results. Results that fall outside of the 'positive' and 'negative' categories can still be produced; for example, test results can go missing, be uninterpretable, or fail to meet the definitions of a positive or negative classification (6).

Schuetz and colleagues carried out a meta-analysis of studies evaluating the accuracy of coronary computed tomography (CT) angiography (7). Inconclusive test results were present in 109 of the 120 (91%) eligible studies, but the reporting of these results was so limited that 3 x 2 tables could only be constructed for 26 of these studies (24%). Furthermore, there were a number of different analytic approaches to handling these results across studies. In the meta-analysis of those adequately reported, alternative analyses were trialled and it was demonstrated that significantly different results could be obtained.

There are many other limitations of diagnostic accuracy research exacerbated by poor reporting (8). For example, very few of the meta-analyses reviewed in Chapter Three fully explored how other factors influence test accuracy because this information was not consistently available in the primary studies. In addition, a recent update of the QUADAS framework highlighted that one of the main barriers to assessing the quality of a diagnostic study is also poor reporting (9).

7.2.5. Making individual patient data available: the solution to (nearly all) of our problems?

The findings of this thesis highlight a more general issue: we are still reporting and analysing primary research as though clinical decisions are going to be made based on the results of a single study. This is completely at odds with current practice; guideline developers now demand high-quality evidence synthesis across the whole research pathway to inform clinical recommendations. By considering diagnostic research from this broader perspective, a key methodological challenge naturally arises: how can we communicate *all* of the relevant results of primary diagnostic accuracy research in a way that can be easily extracted and synthesised with other evidence?

The challenge lies in how we publish primary research that is engaging to the reader and attractive to journal editors, while at the same time containing sufficient detail for successful meta-analysis. This is a tall order and it is often difficult for researchers to anticipate the information required for meta-analysis, especially if the research area has developed notably since completion of the study. Increasing the availability of individual patient data alongside comprehensive study protocols is a viable solution. In the UK at least, most academic research is publicly funded and therefore academic institutions should be mandated to share this publicly-owned data for the purpose of evidence synthesis. Enforcing open data would notably reduce the current waste of evidence that plagues medical research (10).

7.3. Summary

The key implications of the findings presented in this thesis are 1) that methods which convey a greater amount of information than that provided by a single threshold interpretation may be more helpful to decision making, 2) that the 2 x 2 diagnostic framework frequently fails to capture many of the realities and complexities of diagnostic accuracy research, and 3) that the availability of individual patient data would facilitate evidence synthesis across the research pathway.

7.4. References

1. Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. BMJ. 2013 2013-05-16 11:27:50;346:f2778.

2. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care--a systematic review. Fam Pract. 2008 Dec;25(6):400-13.

3. Irwig L. Modelling result-specific likelihood ratios. J Clin Epidemiol. 1992 Nov;45(11):1335-8.

4. Heffner JE, Highland K, Brown LK. A meta-analysis derivation of continuous likelihood ratios for diagnosing pleural fluid exudates. Am J Respir Crit Care Med. 2003 Jun 15;167(12):1591-9.

5. Heffner JE, Heffner JN, Brown LK. Multilevel and continuous pleural fluid pH likelihood ratios for evaluating malignant pleural effusions. 2003;123((Heffner) Div. of Pulmon./Critical Care Med., Med. University of South Carolina, Charleston, NC, United States):1887-94.

6. Knottnerus JA. The Evidence Base of Clinical Diagnosis: BMJ Books; 2002.

7. Schuetz GM, Schlattmann P, Dewey M. Use of 3x2 tables with an intention to diagnose approach to assess clinical performance of diagnostic tests: meta-analytical evaluation of coronary CT angiography studies. Brit Med J. 2012 Oct 24;345(6717).

8. Reid MC, Lachs MS, Feinstein AR. Use of Methodological Standards in Diagnostic Test Research: Getting Better but Still Not Good. JAMA. 1995;274(8):645-51.

9. Whiting PF, Bossuyt PMM, Sterne JAC, Deeks JJ, Reitsma H, Leeflang M, et al. Updating QUADAS: Evidence to inform the development of QUADAS-2. 2011.

10. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009 Jul 4;374(9683):86-9.