Systematic reviews and meta-analyses of test accuracy: developing methods that meet practitioners' needs

by

Dr Clare Frances Davenport

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Department of Public Health, Epidemiology and Biostatistics
School of Health and Population Sciences
College of Medical and Dental Sciences
University of Birmingham
Submitted April 2012

Volume 1 of 2

Main text (the Appendices are in a separate file)

Abstract

Testing can be a substantial driver of health care costs. Increase in test use over recent decades has occurred despite disappointing results from test accuracy evaluations. Poor quality and reporting of primary test accuracy studies and difficulties with understanding and application of test accuracy information are purported to be important contributors to this observed evidence 'gap'.

The objectives of this thesis were to:

- Systematically review evidence concerning the understanding and application of test accuracy metrics.
- Undertake primary research building on the review of understanding and application.
- Assess whether the contribution of test accuracy reviews to the test accuracy evidence base is compromised by deficiencies in their contextual fit, or of included primary studies.

Existing research concerned with understanding and application of test accuracy information is not driven by the needs of decision makers. Contrary to the prevailing view in the literature, findings of original research from this thesis demonstrate that probability revision is not a feature of diagnostic decision making. Choice of test accuracy metric however was shown to have a profound influence on diagnostic decision making. Deficiencies in question formulation and contextualisation of test accuracy reviews are undermining their contribution to the test accuracy evidence base.

Executive Summary

Background

In recent decades the total number of tests ordered by doctors has increased substantially, despite observations that primary studies of test accuracy are characterised by poor reporting, poor quality and lack of contextual fit. A particular feature of this gap between evidence and practice in diagnostic decision making, is that application of test accuracy evidence is purported to be far more problematic than application of evidence about interventions. Improving the accessibility (contextual fit and informed use of outcome measures) of test accuracy evidence has the potential to positively impact on test use. Systematic reviews and meta-analyses of test accuracy are increasing in number and prominence as a resource for diagnostic decision making and offer the opportunity to mitigate some of the current limitations of primary studies. In particular, by considered framing of research questions and by enabling a comparative approach to test evaluation. The potential of systematic reviews to improve evidence based testing has been supported by the relatively recent and rapid development of statistical techniques for undertaking meta-analyses. However the potential for new summary test accuracy measures generated by these statistical developments to exacerbate problems with application of test accuracy evidence has not been investigated.

Aims

The aims of this thesis were to assess the accessibility of existing secondary test accuracy research with respect to the extent to which their conduct and reporting reflects testing context and the extent to which decision makers can interpret and apply test accuracy outcome metrics.

Objectives

Contextual fit of existing test accuracy research:

- An epidemiological mapping exercise of systematic review databases in order to document the volume and characteristics (disease category, review purpose and test application) of existing secondary test accuracy research.
- A methodological review of a representative (in terms of quality) sample of systematic reviews of test accuracy, identified from the mapping exercise, in order to assess the extent to which review authors considered clinical context at each stage of the review process (question formulation; reporting primary study findings; synthesis; making recommendations).

Interpretation and application of test accuracy outcome metrics by decision makers:

- Systematic reviews of qualitative and quantitative research concerned with the understanding and application of test accuracy and risk measures.
- A survey of general practitioners to assess: sources of test accuracy information used for diagnostic decision making; familiarity with a range of test accuracy metrics; ability to apply a range of test accuracy metrics to a hypothetical scenario.

Results

Contextual fit of existing test accuracy research:

Reviews of test accuracy dominate the test evaluation landscape. Within this body of research, tests more commonly applied in secondary care and certain disease topic areas predominate and there is a lack of comparative test accuracy evaluation. Based on the epidemiological characteristics of test evaluations it is unlikely that the existing evidence base reflects the clinical need for evidence.

Assessment of the contextual fit of test accuracy reviews reveals ill-defined objectives which are reflected in question formulation, review synthesis (including investigation of heterogeneity) and reporting of findings. The place of index tests within a testing pathway is mostly not articulated by consideration of test role, (add, replace, triage), healthcare setting, patient presentation, prior tests or current testing practice: Seventy six percent of reviews did not state the setting in which index tests were to be used and only 24% of reviews detailed all of index test application, role and prior tests as part of question formulation. Reporting of study characteristics was poor: setting, participant presentation and age were documented by just over 50% of reviews whilst chronicity and severity of the target disorder were documented by less than one third of reviews. A minority of reviews (between 1% and 8% of reviews depending on characteristic) cited limitations in primary studies as a reason for this poor reporting.

Interpretation and application of test accuracy outcome metrics by decision makers:

The literature reviewed was characterised by well educated and self-selected samples and the UK, policy making and generalist perspectives were under-represented. Evaluation of metrics more common to meta analyses of test accuracy is have received very limited attention in the literature to date. The features of test accuracy measures that are perceived to impact on the extent to which they facilitate formal probability revision include:

-Having the test result (rather than disease status) as the reference class when interpreting conditional probabilities.

-Discrimination between the 2 dimensions of test accuracy and quantification of test errors (ability to rule in or rule out a diagnosis or the value of a positive test result separate to a negative test result).

-Portability of test accuracy metrics across populations.

With the exception of predictive values (self reported use 80%), all other metrics are reported to be used by <4% of clinicians. The utility of different metrics for diagnostic decision making has only been evaluated from the perspective that formal probability revision is a necessary

pre-requisite for informed diagnostic decision making. This is despite evidence that application of Bayes' theorem is not commonplace in practice, estimation of pre-test probability and the accuracy of named tests is inaccurate and highly variable and provision of quantitative test accuracy information does not appear to improve probability revision.

There is no empirical evidence supporting the superiority of a single test accuracy metric for diagnostic decision making. Natural frequency and multiple presentation formats appear to facilitate understanding. No consideration has been given to how metrics may be used in a complimentary way to assist with diagnostic decision making. Although comprehension of test accuracy information by academic clinicians has been shown to be superior to practising physicians, no consistent difference is observed between practising health professionals and non-health professional samples, suggesting that medical education per se may offer no advantage in this respect.

A general practice survey revealed the majority of respondents were familiar with predictive values, sensitivity and specificity and the diagnostic 2x2 table, in contrast to likelihood ratios and metrics more commonly associated with systematic reviews (DOR, AUC, ROC curves). Clinical experience, colleagues and guidelines were reported as sources most commonly used to assist with diagnostic decision making whilst use of quantitative estimates of test accuracy was generally low (55% for sensitivity and specificity; 20% for predictive values; 13% for the diagnostic 2x2 table and less than 2% for likelihood ratios, the DOR, AUC and ROC curve).

Application of test accuracy metrics to a hypothetical scenario resulted in marked variation in responses to both positive and negative test results. Summary measures that separate the two dimensions of test accuracy in the absence of prevalence information (for example sensitivity and specificity) appeared to result in a misplaced emphasis in one or other of false positive or false negative test errors. Presenting test accuracy data using the 2x2 diagnostic table or a pictograph attenuated this effect.

Conclusions

At the current time, inadequacies in question formulation and the subsequent impact on contextualisation of test accuracy review findings may be undermining the potential for statistical and methodological advances in meta-analysis of test accuracy to positively impact on diagnostic decision making.

Choice of test accuracy metric appears to have a profound effect on diagnostic decision making. Understanding, contextual factors and motivational biases are likely to be contributing factors to the observed variability. It is unclear to what extent any advantage of test accuracy metric for informed decision making is based on familiarity as opposed to their intuitive nature. Simultaneous illustration of both dimensions of test accuracy in order to facilitate informed diagnostic decision making requires further exploration.

Dedication

To Caolán, Finnian and Liam.

Acknowledgements

This thesis represents the culmination of some 7 years work. Inevitably over such a time period there have been several key individuals in both my professional and family life who have helped make it happen.

Professor Chris Hyde and Professor Christine MacArthur have been, and continue to be inspirational mentors. Their role as PhD supervisors is just one way in which they have supported my professional development. They both have a profound gift of being able to share their expertise and experience in a non-directive way. I sincerely thank them for their encouragement and their faith in my ability and I look forward to many more years of collaborative working.

I am indebted to Sue Bayliss, information specialist and co-author on completed and planned publications; all four reviews contained in this thesis were challenging and I would like to think we both explored new ground and learned a lot from completing them. I would like to thank Matthew Thompson and Carl Heneghan for advice on the design and distribution of my survey in primary care; their generosity in sharing experience of research in primary care undoubtedly contributed to its success.

Latterly I would like to thank Jon Deeks, Lavinia Ferrante di Ruffano, Susanna Wishniewska and Anne Walker for stimulating discussion, practical and emotional support and encouragement during the final months.

My relief at completing this work is shared by my family and friends. I would like to dedicate this thesis to my three boys (one of whom has not known life without a mummy doing a PhD), who have followed my progress closely and marvelled at the words I have accumulated over the years. I would particularly like to thank my dear friend Dot for her endless patience and support.

VOLUME I

Contents

Figures and tables

Chapter 1: Background: The role of test accuracy information and interpretation in clinical decision making

Chapter 2: Review of literature concerned with the understanding and application of test accuracy and risk measures

Chapter 3: Mapping the epidemiological characteristics of test evaluation systematic reviews

Chapter 4: Methodological Review: An investigation of the extent to which clinical context shapes the conduct and reporting of systematic reviews of test accuracy

Chapter 5: Survey of understanding and application of test accuracy measures

Reference Lists

VOLUME II

Appendices

Appendices to chapter 2:

Appendix 2.1: Search strategies employed for test accuracy and risk communication literature reviews (2010; 2007; 2005)

Appendix 2.2: Characteristics and results of included empirical test accuracy studies

Appendix 2.3: Characteristics and results of included empirical risk communication literature

Appendices to chapter 3:

Appendix 3.1: Pragmatic search filters created for use with HTA, DARE and ARIF databases

Appendix 3.2: Flow of references from the HTA, DARE, Medion, C-EBLM and ARIF databases

Appendix 3.3: Pro-forma for coding review references according to Title and Abstract

Appendix 3.4: Yield from searches of combinations of databases excluding primary research, research not concerned with test accuracy and duplicates

Appendix 3.5: Characteristics of specialist reviews databases

Appendices to chapter 4:

Appendix 4.1: Consistency rules for inclusion of reviews on the basis of test

Appendix 4.2: Demographic details of included reviews

Appendix 4.3: Details of review question formulation

Appendix 4.4: Reporting of review findings

Appendix 4.5: Use of outcome measures in reviews

Appendix 4.6: Detail of contextualisation of review findings

Appendix 4.7: Reporting of study characteristics, consideration of applicability of review findings and methodological quality of reviews reporting the most complete question formulation.

Appendices to chapter 5:

Appendix 5.1: Paper based version of questionnaire used for the survey of understanding and application of test accuracy information

Appendix 5.2: Details of web based resources cited as sources of test accuracy information by survey respondents (accessed 27-07-11)

Systematic reviews and meta-analyses of test accuracy: developing methods that meet practitioners' needs.

Evolution of the Plan of Research

Research rationale

Diagnosis is self-evidently a key clinical activity. Further, from a health policy perspective it can be a substantial driver of health care costs[1]. In recent decades the total number of tests ordered by doctors has increased substantially. This is despite observations that primary studies of test accuracy are characterised by poor reporting, poor quality and lack of contextual fit [2,3] suggesting a gap between diagnostic evidence and clinical diagnostic activity. A particular feature of the test accuracy evidence 'gap' is that application of test accuracy evidence is purported to be far more problematic than application of evidence about interventions [4]. In addition the considerable impact of contextual variables on estimates of test accuracy and on the value placed on testing outcomes has implications for the interpretation and application of test accuracy measures.

Improving the quality and accessibility of test accuracy evaluations has the potential to positively impact on test use. Parallel developments in both primary and secondary research will be required as the value of systematic reviews of test accuracy as a resource for decision making is dependent on the nature of the primary test accuracy evidence base. However systematic reviews of test accuracy offer the opportunity to mitigate some of the current limitations of primary studies, in particular framing of review questions to optimally use primary evidence pertinent to a particular testing context, investigation of reasons for observed variation in test accuracy estimates and consideration of the downstream consequences of test results reflecting the context in which a test is to be used. Systematic reviews and meta-analyses of test accuracy are increasing in number and prominence as a resource for diagnostic decision making. Systematic reviews offer the potential for an

immediate improvement in the contextual fit of evidence and by way of research recommendations, to improvements in the primary test accuracy evidence base. Considerable challenges to the conduct of systematic reviews of test accuracy including identification of primary studies and statistical challenges associated with meta-analysis of test accuracy have fuelled research in this area and in particular statistical techniques have developed rapidly over the last 15 years [5]. However the potential for new summary test accuracy measures generated by these important statistical developments to exacerbate any existing problems with understanding and application of test accuracy evidence has not been investigated.

Research aims

The aim of this thesis is to focus on systematic reviews of test accuracy, as an increasingly prominent resource for decision makers and to redress the current imbalance in methodological developments taking place in systematic reviews of test accuracy by focusing on the accessibility of review findings to decision makers. The accessibility of test accuracy reviews will be assessed with respect to the extent to which their conduct and reporting reflects testing context and the extent to which decision makers can interpret and apply test accuracy outcome metrics.

Original research plan

Assessing the contextual fit of systematic reviews of test accuracy

In order to assess the contextual fit of secondary test accuracy research, a review of existing test accuracy reviews was planned, in order to capture how testing context was being incorporated into systematic review methods and represented in the reporting of review findings.

Accessibility of test accuracy metrics to decision makers

In order to assess the extent to which decision makers are able to interpret and apply test accuracy outcome measures, a review of research concerned with diagnostic decision making was planned. In addition, focus groups were to be used to access the perspectives of decision makers from a range of healthcare settings with respect to barriers and facilitators to the interpretation and application of existing meta-analytic summary measures.

Development of novel test accuracy metrics

Drawing on the findings from the literature reviews and focus groups novel summary measures that better met the needs of decision makers were to be developed.
The original focus of the research was on meta-analytic summary measures of test accuracy on the basis that their derivation would be relatively unfamiliar to decision makers and pose challenges to their interpretation and application.

Emerging findings from the literature reviews of diagnostic decision making

Empirical literature concerning the application and understanding of more established summary test accuracy measures shared by primary and secondary test accuracy research (including sensitivity and specificity, predictive values and likelihood ratios), was considered pertinent because of the popularity of these measures for reporting the findings from narrative and meta-analytic syntheses of systematic reviews of test accuracy [6,7]. At the outset it was anticipated that the volume of existing empirical research concerning the understanding and application of less familiar meta-analytic summary measures (including the Diagnostic Odds Ratio (DOR), the Receiver Operator Characteristic (ROC) space and curves, the Area Under the Curve (AUC), relative measures of test accuracy and forest plots) would be limited compared to more established metrics.

However emerging findings from the empirical literature revealed the volume, quality and applicability of empirical research was limited even for the more established test accuracy measures. Research was characterised by highly selected academic clinical samples and almost exclusively confined to evaluation of sensitivity, specificity and likelihood ratios (LRs). Evaluation of understanding and application of test accuracy measures was largely based on the premise that formal probabilistic reasoning was commonplace in clinical practice and there was no empirical investigation of the extent to which testing context might influence application. Primary care professionals were under-represented and the perspective of policy makers was entirely absent. Understanding of more established test accuracy metrics, even in these highly selected decision makers was poor and self-reported use of test accuracy information raised questions about the extent to which clinicians seek and use quantitative estimates of test accuracy. Additionally an unexpected finding from the literature reviews was a rich qualitative data set of the perspectives of clinical academics concerning facilitators and barriers to understanding of the more established test accuracy metrics.

Modification of the original plan of research concerning the understanding and application of test accuracy metrics

Broadening the scope of the literature review in place of the planned focus group

On the basis of initial findings from the literature review revealing the considerable difficulty decision makers have interpreting and applying even the more established test accuracy metrics and the limitations of the evidence base concerning newer meta-analytic summary measures, the plan of investigation concerning the understanding and application of test accuracy metrics was modified. The original aim to develop novel meta-analytic summary metrics seemed unrealistic and premature. In addition the existence of a relatively rich qualitative dataset in combination with empirical evidence that self-reported familiarity with test accuracy metrics was not necessarily associated with understanding and application,

raised questions concerning the added value of a focus group at this stage of the investigation.

In place of the focus group originally planned, the scope of the literature review concerned with diagnostic decision making was broadened. The perspectives and experience of other disciplines (psychology and education) with respect to the communication of probability and risk was sought. In addition to a synthesis of quantitative empirical literature concerned with test accuracy interpretation, a synthesis of published perspectives, comments and analyses about diagnostic decision making was undertaken.

Primary research to address the limitations of the existing empirical literature concerning the understanding and application of test accuracy measures

Addressing the considerable limitations of the existing empirical evidence base was now considered a priority as part of this programme of research. In particular:

- capturing the perspectives and experience of a representative sample of decision makers
- assessment of the extent to which quantitative estimates of test accuracy are sought and used to assist with diagnostic decision making in practice
- assessment of barriers and facilitators to the use of test accuracy information for diagnostic decision making in practice
- assessment of the impact of quantitative test accuracy estimates on diagnostic decision making

The scope of the primary research possible in the time available was limited and test accuracy metrics, testing context and the decision maker sample were selected to best address limitations of the existing published literature whilst supporting the evolution of the evidence base.

Improving understanding about the accessibility to decision makers of the more established test accuracy metrics was considered a natural first step in this investigation. In addition

established test accuracy metrics (sensitivity and specificity, predictive values (PVs) and LRs) are relevant to the application of results of both primary and secondary test accuracy research. The results of the literature review suggested that global test accuracy metrics were unfamiliar to decision makers and their inclusion in the primary research was not to be at the expense of a more detailed consideration of the more established metrics. Reflecting findings from the review of risk communication, alternative formats to percentage and normalised frequency representations of probabilistic information were also included. Primary care clinicians were chosen over secondary care as the decision maker sample because the primary care perspective was under-represented in the published literature. In addition although the culture of testing would be expected to differ between specialists and generalists, a generalist perspective is likely to encompass a broader mix of testing contexts compared to any particular medical speciality alone. Clinicians were chosen over policy makers because the existing evidence base is concerned with decision making at the bedside and improving understanding of the clinician perspective was seen as a natural progression in this respect.

Thesis outline

- ***Chapter 1*** outlines the potential role of test accuracy information in diagnostic decision making and provides a framework, drawing on behavioural decision theory, within which probabilistic information and professional, patient and contextually dependent utilities might combine to influence testing behaviour.
- ***Chapter 2*** presents a review of the existing empirical and non empirical literature concerning understanding and application of test accuracy measures, complimented by a review of the more established risk communication evidence base.

- ***Chapter 3*** presents an epidemiological mapping exercise of existing test reviews with respect to coverage of disease topic areas, representation of healthcare setting and review purpose.
- ***Chapter 4*** presents a review of the degree to which clinical context shapes the conduct and reporting of existing test accuracy reviews.
- ***Chapter 5*** is a survey of use, understanding and application of test accuracy measures in a sample of primary care clinicians.

List of Abbreviations

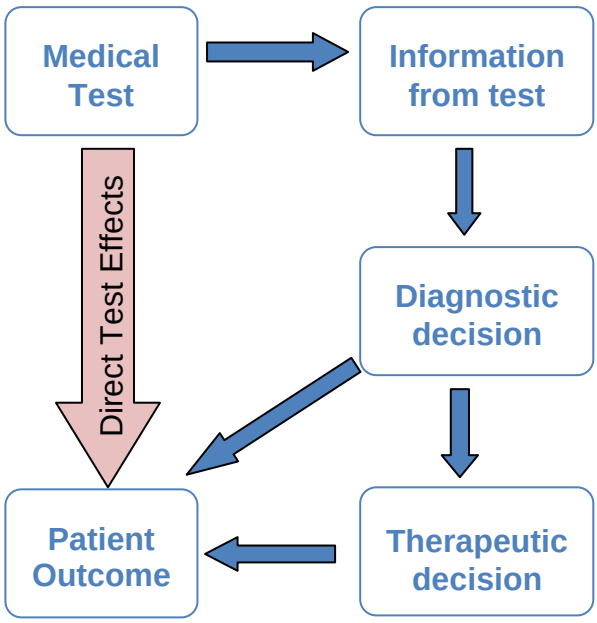
Abbreviation	Explanation
AR	Attributable Risk
ARR	Attributable Risk Reduction
AUC	Area Under the Receiver Operator Characteristic Curve
DOR	Diagnostic Odds Ratio
EBM	Evidence Based Medicine
FN	False Negative
FP	False Positive
GP(s)	General Practitioner(s)
LR-	Negative Likelihood Ratio
LR(s)	Likelihood ratio(s)
LR+	Positive Likelihood Ratio
NNT	Number Needed to Treat
NPV	Negative Predictive Value
PPV	Positive Predictive Value
PV(s)	Predictive Value(s)
Q	Point on the ROC curve where sensitivity=specificity
RCT	Randomised Controlled Trial
rDOR	Relative Diagnostic Odds Ratio
ROC	Receiver Operator Characteristic Curve
RR	Relative Risk
RRR	Relative Risk Reduction
SnNOUT	Sensitivity high, Negative test result rules OUT
SpPIN	Specificity high, Positive test result rules IN
sROC	Summary Receiver Operator Characteristic Curve
TN	True Negative
TNT	Tablets Needed to Take
TP	True Positive

Chapter 1: Background: The role of test accuracy information and interpretation in clinical decision making

1.1 The impact of testing on patient outcomes

Diagnosis is self-evidently a key clinical activity; making a correct diagnosis is a pre-requisite for appropriate management. Testing may result in the correct identification of a greater number of individuals who might benefit from effective treatments as well as the avoidance of unnecessary further interventions in those without disease. In addition acquiring knowledge about diagnosis and prognosis has a value to both clinicians and patients regardless of its impact on eventual health outcomes [8-10]. For example Pauker (1998)[11] observed that 20% of tests ordered in a primary care setting were described by general practitioners (GPs) as tests performed to re-assure patients. Due to the inevitability of test errors, testing will also generate false positives and false negatives with associated negative outcomes for patients. In addition to errors as inherent properties of a test itself, errors can occur as a result of misinterpretation of test results[12]. Tests can also have a direct effect on patients if the test itself causes anxiety or carries risk.

Fig 1.1 The pathway from testing to patient outcomes



1.2 Trends and variations in testing behaviour

In recent decades the total number of diagnostic tests ordered by doctors has increased substantially [13-17]. This increase may in part be due to political and organisational factors. Organisational factors include the shift of patient care from secondary to primary care; technological advances that have vastly expanded the range of investigative technologies available to clinicians and simultaneously created a barrier to evaluating each new test [18] patient preference and demand[18,19], fear of litigation[11,19] and a quest to reduce uncertainty which is argued to be a part of medical culture [18,19]. As well as an increase in the number of tests being ordered, international [20,], regional [21,22] and between-doctor variation in test ordering [15,23] has been shown to be large to a degree that is unlikely to be explained by differences in patient demographics or need.

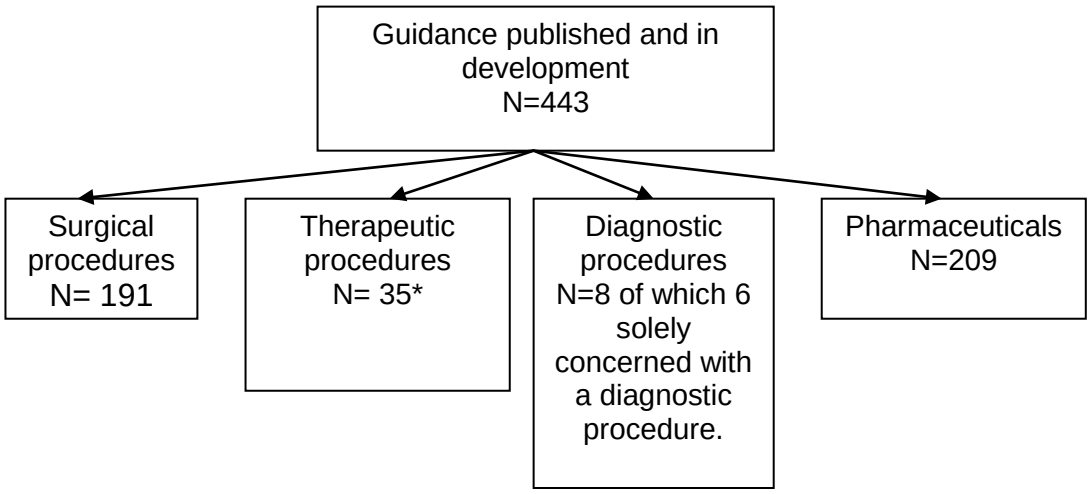
Although it has been suggested that variations in testing are likely to be due to both under and over utilisation based on clinical need [23], the overall rise in test use is generally assumed to be largely due to over utilisation[14,24].

Uncritical adoption of new tests is encouraged by the lax regulatory system for the introduction of new tests. The Medicine and Healthcare products Regulatory Authority (MHRA) is the competent authority responsible for regulating devices in the United Kingdom (UK). There is currently no requirement for a particular standard of evidence or level of accuracy in order for in vitro diagnostic devices to meet the performance standards required for award of a certified (CE) mark that is primarily designed to assure safety rather than improvements in accuracy. In addition MHRA certification is not a well regulated process: devices considered low risk may be self-certified by manufacturers and although those considered to pose a greater risk are required to be assessed by commercial organisations, these organisations are chosen by manufacturers of the test who are at liberty to disregard a negative assessment in preference to a positive alternative (personal communication John Webster, Medical Devices Consultants International Ltd (MDCI) February 2009). A similar

situation exists in the USA equivalent, the Food and Drug Administration (FDA)[25]. Rink (1993)[15] for example demonstrated that the rate of testing with 6 new near-patient tests increased by 16% following their introduction without a concomitant reduction in laboratory based testing.

A survey in 2007 [26] demonstrated that the independent body charged with producing guidance for the National Health Service (NHS) in England and Wales, the National Institute for Health and Clinical Excellence (NICE) health technology appraisal programme, appraised an estimated 9.5% of all pharmaceuticals between 1996 and 2004; the comparable figure for devices (diagnostic and interventions) was 0.1%. There remains an imbalance between national guidance issued concerning interventions and interventional procedures in comparison to diagnostic procedures. In 2009 less than 2% of NICE guidance published and in development involved detailed consideration of the diagnosis of a condition and only 1% was solely concerned with evaluation of a diagnostic procedure, (see figure 1.2)

Fig 1.2: Details of NICE guidance published and in development as at 24-07-09



* This includes 6 interventional procedures listed apparently incorrectly under diagnostic procedures.

Further factors that may fuel inappropriate increases in testing are test errors. Inevitably increases in test use generate increased numbers of false positives and results of uncertain

clinical significance, of which one of the consequences is further testing [16,27]. For example Wennberg (1996) [27] demonstrated significant, positive relationships between tests used early in the diagnostic pathway for coronary vascular disease (stress testing) and tests used later in the diagnostic pathway (coronary angiography) across 72 hospital service areas that were of a magnitude unlikely to be explained by variation in disease prevalence, variation in availability of technology or unmet need. There is a volume of evidence supporting an inappropriate excess of testing which will have direct and indirect adverse outcomes for patients as well as being a substantial driver of health care costs.

1.3 Variables affecting testing behaviour

Two recent systematic reviews [28,29] (see table 1.3) have attempted to summarise variables that affect test ordering by clinicians.

A systematic review of interventions aimed at changing testing behaviour [24] found that enabling interventions aimed at changing testing behaviour directly, such as restrictions on ordering and feedback, were the most potent intervention, followed by reinforcing interventions (such as audit) with educational interventions having the most modest effect. Educational interventions included in the review were poorly described but included guidelines, conferences and lectures covering the clinical utility of testing and cost-effectiveness. Interventions targeting more than one behavioural factor were more successful than those targeting single factors (62% versus 56%). Evidence was generally of poor quality and only 12% (6/49) of included studies were conducted in the primary care setting.

This body of research demonstrates that reasons for test ordering and influences on test ordering are many. The relationship between individual variables is likely to be complicated and in the studies considered here, only a small percentage of the observed variation in test ordering behaviour could be explained by contextual, doctor, or patient variables studied.

Clearly, understanding the causes of variation in testing is important in order to encourage testing that is considered evidence based.

Importantly the reviews above provide no direct evidence concerning the impact of test accuracy evidence on testing behaviour. Some of the variables investigated may be proxies for knowledge of test properties. Involvement in research or guideline development, provision of educational materials and provision of feedback were observed to decrease test use. Clinical experience did not demonstrate a consistent relationship with test use but this variable is likely to reflect the relationship between experience and the process of decision making in a broader sense rather than decisions about test use alone.

Table 1.3: Factors associated with test ordering derived from empirical research

Category	Example	Effect on test ordering (↑ or ↓)
Patient-related factors	Patient preference	Variable depending on test and condition being sought
	Patient acceptability (side effects of test)	Variable depending on test and condition being sought
	Impact of diagnosis or lack of diagnosis	Variable depending on test and condition being sought
	Consequences of test errors	Variable depending on test and condition being sought
	Patient reassurance	Variable depending on test and condition being sought
	Patient demographics	Variable (eg older patients ↑; female patients, ↑)
Doctor-related factors	Confidence in clinical judgement/ clinical experience	↓ with ↑ confidence and ↑experience [28]
		Inconsistent effect of clinical experience [29]
	Knowledge of test properties	No evidence found [28]
	Involvement in research / guideline development	↓
	Attitudes to risk litigation	↓with ↓ fear or risk taking /fear of litigation
	Response to patient requests (appropriate & inappropriate)	↑ or no effect according to individual doctor
	Feedback on test ordering behaviour	↓
	Doctor speciality	Variable depending on speciality [28]
		Specialists ↑ tests from a narrower repertoire compared to generalists [29]
	Doctor demographics (age and sex)	Variable and contradictory [28]
		Female doctors ↑ tests [29]
Policy and organisational-related factors	Time constraints	↑ with time constraints
	Primary care practice size	↓ with ↓ practice size
	Availability of tests	↓ with ↓access
	Method of doctor payment	↓ with payment by salary; ↑with fee for service
	Existence of testing policies / clinical guidelines	↓ with introduction of clinical guidelines and policy recommendations
	Structured test ordering forms	↓ with introduction of structured test ordering forms
	Geographical location	Inconsistent
	Knowledge of test costs	Inconsistent

Categorisation adapted from Whiting et al. *Journal of Clinical Epidemiology 2007* [28]

It has been observed that important barriers to the use of evidence include accessibility (accessing information in a timely manner and possession of skills necessary to interpret information) and the acceptability and applicability of the research evidence (provision of information relevant to the decision context) [30]. In addition clinicians may not be aware of the extent to which their own practice diverges from the evidence base [31]. This has implications for education; not least the motivation of practitioners to question their understanding.

There is a widespread belief that decision makers are less familiar with evidence about test accuracy compared to evidence about effectiveness and have difficulty understanding and applying test accuracy evidence [32,33]. In order to address any gap between the evidence base and testing behaviour, the relationship between understanding and application of test accuracy information and testing behaviour needs to be examined. However to date there has been systematic interrogation of the evidence base to allow quantification or characterisation of the extent of the problem and therefore the impact this might have on testing behaviour.

Certainly the observed increase in test use outlined above is not congruent with observations made about the existing primary test accuracy evidence base. Primary studies of test accuracy are characterised by poor reporting, poor quality and lack of contextual fit [3,34-38]. In addition the existing evidence base is characterised by evaluations of single tests divorced from diagnostic pathways which does not assist or encourage decision makers to consider test replacement or addition based on demonstrable gains in accuracy.

Systematic reviews of test accuracy currently represent a small proportion of all systematic reviews (an estimated 4% in 2004 [39]) but they are increasing in number [40-42] and prominence as a resource for decision makers. Systematic reviews of test accuracy offer the opportunity to mitigate against limitations in the primary evidence base by improved framing of questions, in particular inclusion of a comparative element. However the extent to which systematic reviews of test accuracy are realising their potential in this respect has not been investigated.

Similarly, although limitations of primary studies are proposed to present challenges to review conduct, it is not known the extent to which contextual, quality and reporting limitations documented for primary test accuracy studies are impacting on the quality of systematic reviews of test accuracy.

The considerable statistical challenges associated with meta-analysis of test accuracy data have encouraged a rapid evolution in statistical methods for synthesising data over the last 10 years [5] accompanied by an increasing repertoire of metrics and graphics. Indeed advancement of statistical techniques appears to have occurred without consideration of the accessibility of meta-analytic summaries of test accuracy to decision makers. There is a need to investigate the extent to which existing meta-analytic summaries of test accuracy are understood by decision makers because of the potential to mislead those attempting to apply the results of these reviews in practice [6,43,44 (TAR18)].

1.4 Clinical problem solving

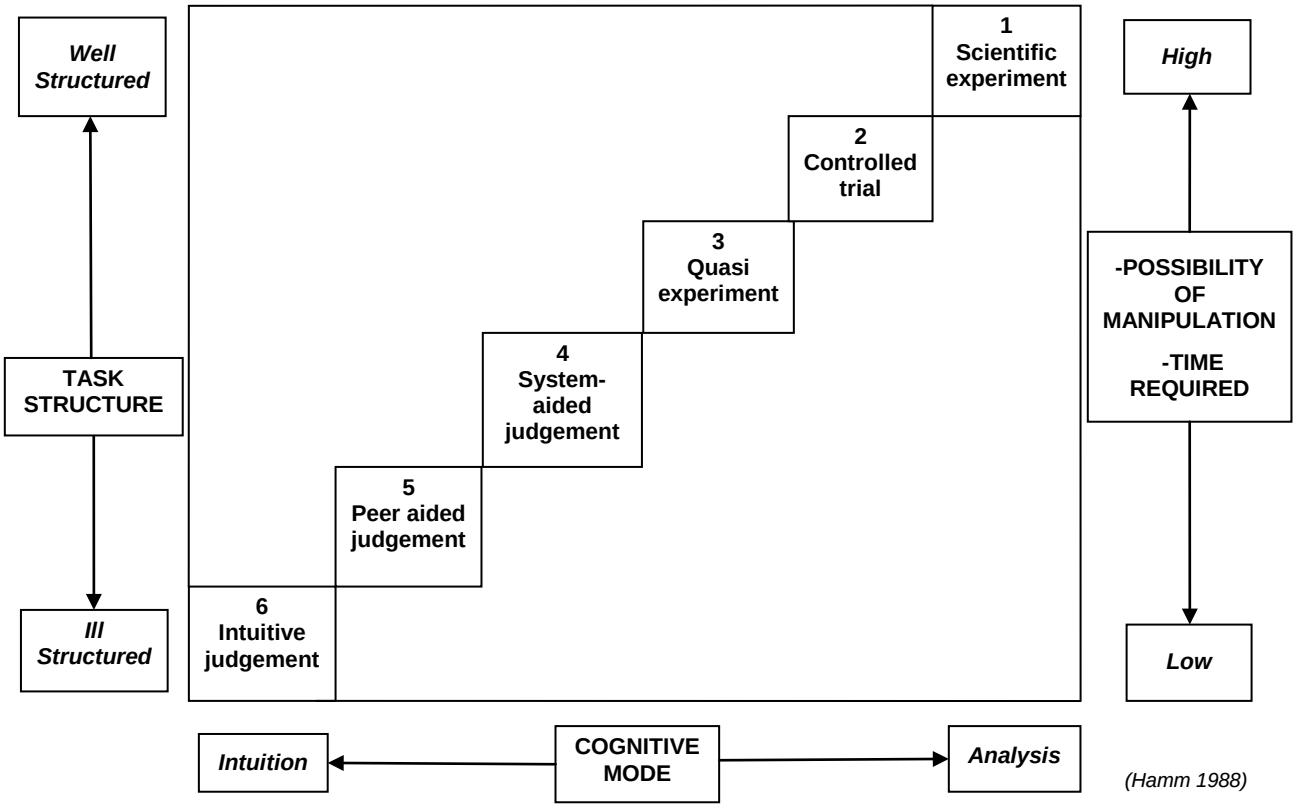
Clinical problem solving and behavioural decision theory are frameworks within which to consider the relationship between accessibility and relevance of test accuracy information, decision making and testing behaviour.

As the potential health benefits, health risks and economic implications associated with investigative techniques have expanded there has been an accompanied increase in research attempting to explain and improve clinicians' decision making, including judgements about the use of diagnostic tests.

Problem solving can be characterised by two distinct modes of cognition: intuitive and analytical. Hamm 1988 [45] proposed a cognitive continuum framework to represent six different approaches to problem solving ranging from intuitive judgement ('un-criticised private judgements') which are rapid, largely unconscious and characterised as inconsistent and moderately accurate to analytical approaches which are slow, conscious, likely to

combine information using organising principles, consistent and more accurate (see figure 1.4). Hamm proposed that the precise nature of the decision making task induced a particular problem solving approach (intuition inducing to analytical inducing) and that accuracy in decision making was determined by the clinician's expertise in choosing a problem solving approach appropriate to each task; ill-structured tasks induce intuition and well-structured tasks an analytical approach.

Fig 1.4: The Cognitive Continuum



Modes 5 and 6 in the cognitive continuum (see figure 1.4) are seen as problem solving approaches adopted in practice settings with minimal direct support from empirical research. Modes 1-4 involve the use of formal decision analytic frameworks including probabilities and utilities. Modes 1-3 would draw on probabilities and utilities estimated from empirical research whereas mode 4 would involve subjective estimates.

Hamm [45] suggests that tasks should not be seen as always externally controlled and that the nature of task features may be open to manipulation by the clinician as problem solver. However manipulation requires clinicians to be able to use the tools and knowledge information systems that are available to them. Institutional and social contextual factors may also impact on the problem solving approach adopted so that for example, staff training, knowledge management and the time available for each patient will impact on whether a clinician can adopt an intuitive or an analytical approach to problem solving and whether they are able to manipulate the task. It is argued that knowledge constraints in the use of modes 1-3 should be viewed as a lost educational opportunity [46].

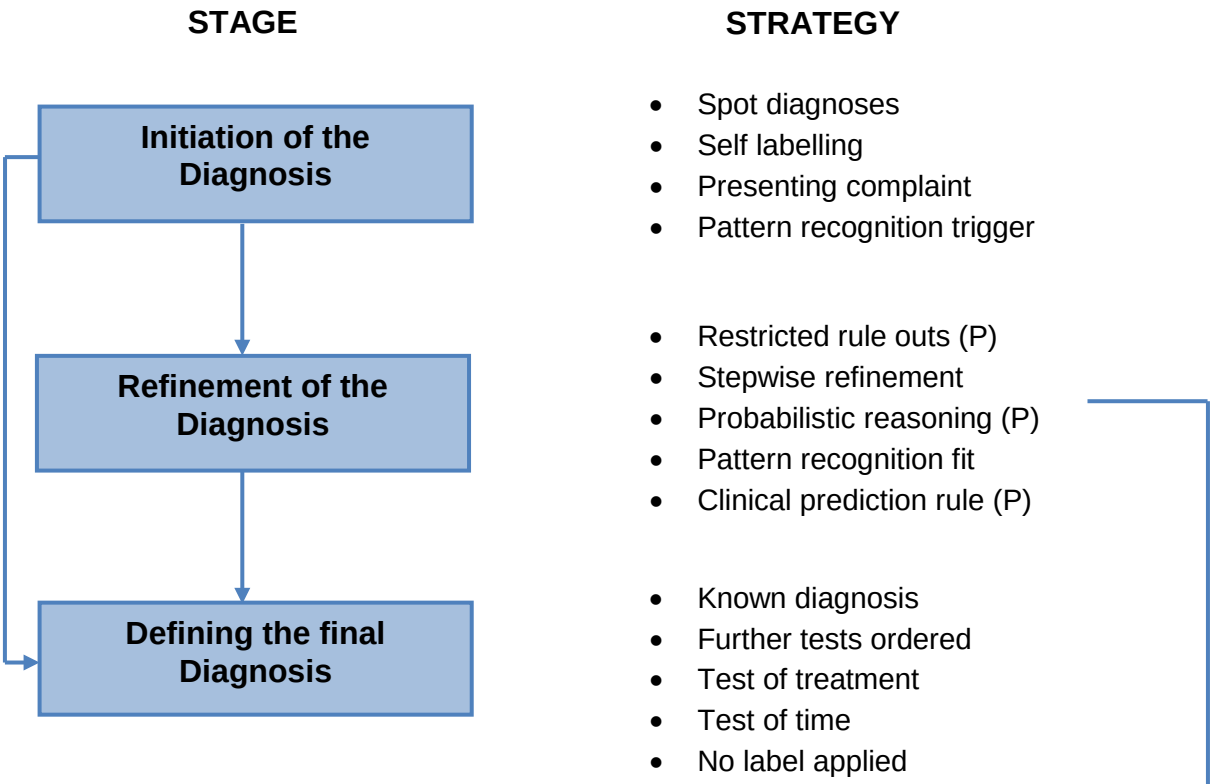
However based on the observation that problem solving strategies differ between novices and experts, alternative theorists have proposed a theory of expert cognition where the problem solving approach is based on expertise rather than task, social or institutional factors. This theory proposes that more experienced clinicians adopt increasingly intuitive approaches whereas novices rely on guiding principles and rules to make sense of clinical presentations and as a consequence are more likely to adopt an analytical approach. Indeed there remain many problem solving strategies, particularly those of experts, which are rapid, automatic and not well understood [47].

Expertise as defined by Kassirer [47] is likely to be underpinned by clinical experience in a speciality, knowledge about the evidence base and sources of information to help with decision making, as well as experience in decision making per se. Indeed if the definition of ‘expert’ includes mastery of a particular domain of knowledge then it would be expected that problem solving strategies may not be generalisable across medical specialities and in particular between specialists and generalists.

Adopting a qualitative approach using verbal rather than mathematical rules to explain behaviour, process tracing research also highlights the importance of subject-specific

expertise. Process tracing approaches have also demonstrated that probabilistic reasoning is a prominent feature of clinical problem solving but suggest that the relative contribution of intuitive and analytic thinking vary at different stages in the decision making process rather than only being dependent on the structure of the decision making environment [48,49] (see figure 1.5).

Fig 1.5: Diagnostic strategies used by 6 General Practitioners across a total of 300 consultations



RULE	*DEFINITION*
Restricted rule outs	*Learning the most common (probable) causes of a presenting problem and a shortlist of serious diagnoses which must be ruled out.*
Stepwise refinement	*Diagnostic refinement based on either anatomical location of the problem or the putative underlying pathological process.*
Probabilistic reasoning	*The specific but probably imperfect use of symptoms, signs, or diagnostic tests to rule in or rule out a diagnosis and requiring knowledge of the degree to which a positive or negative test result adjusts the probability of a given disease.*
Pattern recognition fit	*Symptoms and signs are compared to previous patterns and a disease is recognised when a pattern fits.*
Clinical prediction rule	*A formal, validated version of pattern recognition based on the probability of a diagnosis in the presence of a pattern of symptoms, signs and test results.*

Notes to fig 1.5: (P) Involves manipulation of probabilities either formally or informally

Heneghan, C. et al. BMJ 2009;338:bmj.b946 [48]

It is therefore unlikely that problem solving *ability* will be dependent entirely on experience in problem solving per se but rather mastery of a particular (speciality-specific) domain of knowledge. With increasing experience a less analytical and more intuitive approach might be taken, except in novel clinical situations. In addition the relatively more structured environments of the medical specialities in secondary care may be more suited to an analytic approach than the less structured, diverse and less well-defined clinical decision making environment in generalist, primary care settings. Medical knowledge tends to be organised according to disease rather than individual signs and symptoms, an organisation better suited to specialist settings and requiring knowledge of the relative probabilities of disease for its application [49].

Dowie [50] proposes that differences in the way decision tasks are structured and the cognitive modes employed, as proposed in the cognitive continuum model (see figure 1.4), explain much of the difference in research and practice cultures that hinders implementation of research findings. This is in contrast to the more traditional explanations of weaknesses in research dissemination or practitioner attitudes and motivations. Research activity is described as 'truth driven' and is characterised by well-structured, highly analytic environments, whereas practice based decision making is characterised by complexity and lack of structure which encourages intuition, implicit assumptions about outcomes such as test performance, and integration of 'value judgements'. Dowie [50] suggests that the difference between the task structure of researchers and practitioners produces an evidence gap - researchers are unable to represent their findings in a way that is relevant to practitioners; an observation that has been made by others [51]. Decision analysis (mode 4 figure 1.4) is proposed as a form of system-aided judgement which could act as a potential bridge between the research and practice cultures and assist with the implementation of research findings. Indeed evidence based medicine is viewed as a mechanism for introducing quantification into medical (including diagnostic) decisions whilst leaving a substantial role for clinical judgement [52,53].

The extent to which evidence based medicine has achieved its aims with respect to diagnostic decision making, particularly the extent to which decision makers seek and use quantitative test accuracy information, is unclear. Certainly the relative lack of guidance concerned with the use of tests compared to treatments (see figure 1.2 above) presents a challenge to those attempting to integrate test accuracy evidence into their practice. Probabilistic reasoning will almost certainly play a role in diagnostic decision making although its use is likely to vary depending on the skill and expertise of the decision maker, subject-specific experience, the structure of the problem, availability of evidence and contextual factors that determine utility judgements. However the existence of intuitive judgement as a legitimate problem solving approach calls into question the degree to which formal, quantitative probability revision is a necessary pre-requisite for informed diagnostic decision making.

1.5 Behavioural decision analysis: a framework for considering diagnostic decision making

The introduction of evidence into diagnostic decision making requires knowledge about the way in which clinicians solve problems. As probabilistic reasoning is proposed to be a prominent part of diagnostic decision making and the paradigm on which evidence based medicine is based, behavioural decision analysis provides a useful framework for considering how this might be facilitated. Diagnostic decision making is a term that can be used to encompass the integration and application of test accuracy information and other pertinent contextual considerations, into decisions about the use of tests and the interpretation of test results.

1.5.1 Normative Decision Theory

Decision analysis seeks to determine the best course of action, under conditions of uncertainty and has applications in situations when choices need to be made between clearly defined courses of action. Moving beyond the initial narrow economic concept, optimality is seen as conditional on context [54]. Normative decision theory proposes that when faced with a number of choices, the rational procedure is to place a value on each outcome, multiply this by the probability of the outcome occurring to derive an expected value, and make a choice that will result in the highest total expected value; expected values distinguish the right from the wrong decision.

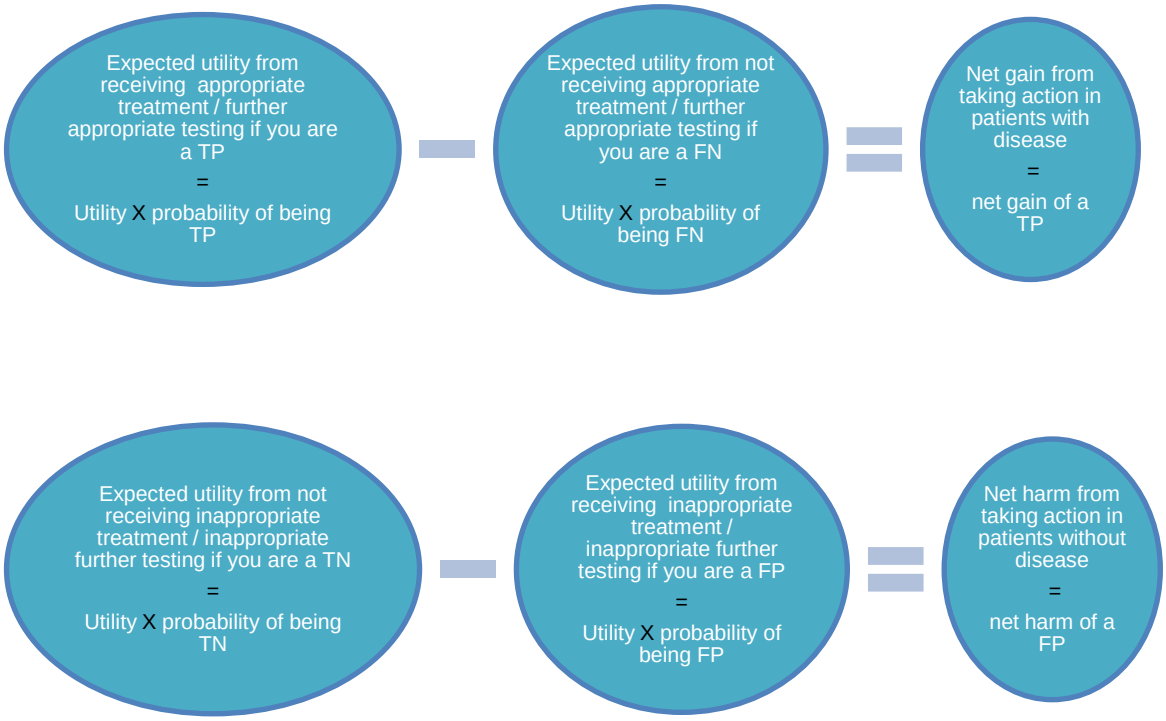
Most tests are imperfect and merely adjust the probability of having a disease rather than confirm its presence or absence. Test accuracy information is an expression of the probability (risk) of one of four possible outcomes of testing. These four outcomes are illustrated by the 2x2 diagnostic contingency table (figure 1.6):

Fig 1.6: The 2x2 Diagnostic contingency table

	DISEASE +VE *(verified by a reference standard test)*	**DISEASE -VE** *(verified by a reference standard test)*
INDEX TEST RESULT +VE	**T**RUE **P**OSITIVE	**F**ALSE **P**OSITIVE
INDEX TEST RESULT -VE	**F**ALSE **N**EGATIVE	**T**RUE **N**EGATIVE

Each possible testing outcome (**T**rue **P**ositive, **T**rue **N**egative, **F**alse **P**ositive, **F**alse **N**egative) has a value attached which will be determined by the context in which a test is being used. For example the relative values associated with true positives and false negatives will increase as the target disease increases in seriousness and those of true negatives and false positives will increase as the toxicity of management of positive test results increases.

Fig 1.7 Diagrammatic representation of a normative guide to decision making applied to testing



(Adapted from Matcher 2007 [55] and Kassirer 1989 [56])

1.5.2 Departures from normative decision making: behavioural decision analysis

According to normative decision theory, an individual is expected to make a decision based on maximising the expected value of the outcomes possible from competing choices. In order to make an optimal decision it is assumed that individuals will be fully informed, fully rational and able to compute accurately. Behavioural decision analysis is based on the notion that due to limitations of working memory in complex decision making environments, (bounded or limited rationality) [57], humans simplify complex problems using heuristics and as a result introduce errors at subsequent stages in the decision making process: hypothesis generation; probability estimation and revision and assessment of utility (for example the consequences of testing itself, correct and incorrect diagnoses). Errors and biases can be divided into those associated with probability estimation and revision (cognitive biases) and

those associated with emotional reactions to uncertainty, the utility associated with testing per se and the consequences of different testing outcomes (motivational biases) [(ETA36);58,59]

1.5.2.1 Fully informed and able to compute accurately: cognitive errors and biases

Fully informed

In order to make appropriate decisions about test use and interpretation of test results, information about the probability of disease prior to testing and information about test accuracy is needed. This information is most commonly communicated quantitatively either as single event probabilities (the probability of the occurrence of a test result or the probability of the presence of a disorder prior to testing (pre-test probability) after testing (post-test or posterior probability)) or communicated as conditional probabilities (the probability of the presence or absence of x disorder given y test result or the probability of having y test result in the presence or absence of x disorder). Verbal expressions of probability require standardisation for application: for example a numerical equivalent of low, medium or high pre-test probability and a numerical definition of what constitutes a poor, moderate or highly accurate test. Problems with standardising the language of risk are well rehearsed [60,61].

The extent to which diagnostic decision makers can be regarded as fully informed will depend on an assessment of their knowledge of the information required to undertake probability revision and sources of that information; accuracy of pre-test probability estimation (in the absence of external evidence) and the extent to which quantitative probabilistic expressions of the uncertainty associated with testing (test accuracy information) is understood.

Only one test accuracy measure, the Likelihood Ratio (LR), has guidelines for interpretation reflecting the magnitude of change in pre to post-test probability it conveys [62] (see table 1.8) However the extent to which these are a reliable guide to the clinical utility of a test across different settings has not been evaluated in practice. Table 1.8 compares features of the

more commonly used test accuracy summary measures with respect to their interpretation and application.

Table 1.8: Characteristics of commonly used test accuracy metrics

TEST ACCURACY METRIC	STATISTICAL MEANING & PROPERTIES	CLINICAL APPLICATION	CONTEXTUAL CONSTRAINTS
DISEASE AS REFERENCE CLASS		ESTIMATION OF THE PROBABILITY OF DISEASE IN INDIVIDUALS	
Sensitivity (true positive fraction)	-The proportion of people with disease who test positive. TP/(TP+FN) -Values closer to 1 (100%) indicate increasing accuracy -Disease as reference class* -Requires discrimination and calibration	SnNOUT (sensitivity high, negative test result, rule out disease) (Sackett 1997 [TTA27] is suggested as a heuristic . Estimation of the post test probability of disease in individuals is possible but requires probability revision using the formula of Bayes' theorem **. Indication of the ability of a test to rule out disease and therefore may help with decisions about test use when downstream consequences of a –ve test result are considered > than the downstream consequences of a +ve test result. No guidelines for translation of quantitative estimates of sensitivity into a measure of their diagnostic utility.	Accuracy varies according to spectrum of patients therefore cannot be assumed to be portable across different populations. The SnNOUT rule may mislead as prevalence increases because the absolute number of FNs becomes large.
Specificity (true negative fraction)	-The proportion of people with disease who test positive. TN/(TN+FP) -Values closer to 1 (100%) indicate increasing accuracy -Disease as reference class* -Requires discrimination and calibration	SpPIN (specificity high, positive test result, rule in disease)(Sackett 1997 [TTA27] is suggested as a heuristic . Estimation of the post test probability of disease in individuals is possible but requires probability revision using the formula of Bayes' theorem **. Indication of the ability of a test to rule in disease and therefore may help with decisions about test use when downstream consequences of a +ve test result are considered > than the downstream consequences of a -ve test result. No guidelines for translation of quantitative estimates of specificity into a measure of their diagnostic utility.	Accuracy varies according to spectrum of patients therefore cannot be assumed to be portable across different populations. The SpPIN rule may mislead as prevalence decreases because the absolute number of FPs becomes large.

Table 1.8 continued

TEST ACCURACY METRIC	STATISTICAL MEANING & PROPERTIES	CLINICAL APPLICATION	CONTEXTUAL CONSTRAINTS
TEST RESULT AS REFERENCE CLASS		ESTIMATION OF THE PROBABILITY OF DISEASE IN INDIVIDUALS	
Positive predictive value (PPV)	The proportion of individuals testing positive who have disease. TP/(TP+FP) -Values closer to 1 (100%) indicate increasing accuracy) -Test result as reference class* Requires discrimination and calibration	Allows estimation of the probability of disease in individuals based on test result* and is therefore viewed as clinically intuitive. Setting specific PPVs negate the need for pre-post test probability revision: the post test probability of disease following a positive test result =PPV. No guidelines for translation of quantitative estimates of PPV into a measure of their diagnostic utility.	Accuracy depends on prevalence of target disease which exacerbates observed variation in estimates of accuracy due to spectrum variation. As prevalence decreases, the PPV decreases.
Negative predictive value (NPV)	The proportion of individuals testing negative who do not have disease. TN/(TN+FN) -Values closer to 1 (100%) indicate increasing accuracy -Test result as reference class* Requires discrimination and calibration	Allows estimation of the probability of disease in individuals based on test result* and is therefore viewed as clinically intuitive. Setting specific NPVs negate the need for pre to post test probability revision: the post test probability of disease following a negative test result is 1-NPV. No guidelines for translation of quantitative estimates of NPV into a measure of their diagnostic utility.	Accuracy depends on prevalence of target disease which exacerbates observed variation in estimates of accuracy due to spectrum variation. As prevalence decreases, the NPV increases.
Likelihood ratio (LR)	-Probability of a test result in diseased individuals divided by probability of same test result in non-diseased individuals LR+ve $\frac{TP/(TP+FN)}{FP(FP+TN)}$ LR-ve $\frac{FN/(TP+FN)}{TN(FP+TN)}$ -Accuracy increases as LR values differ from one (>1 for LR+ and <1 for LR-). -Test result as reference class* -Requires discrimination and calibration	The degree to which the probability of disease changes following a test result*. Multi-level likelihood ratios allow linkage of test accuracy with the degree of abnormality of a test result. Estimation of the post test probability of disease in individuals is possible but requires probability revision using the formula of Bayes' theorem **. Graphical tools can be used to simplify the conversion of odds to probabilities with likelihood ratios (Fagans nomogram (Fagan 1975[63]). The likelihood ratio scale is non linear and therefore is not intuitive to interpret. Clinical utility guide available (Jaeschke 2006[62]).	LR convey magnitude of change in probability therefore in situations where pre-test probability is very low or very high, LRs alone may be misleading with respect to the value of a test. Accuracy varies according to spectrum of patients therefore cannot be assumed to be portable across different populations.

Table 1.8 continued

TEST ACCURACY METRIC	STATISTICAL MEANING & PROPERTIES	CLINICAL APPLICATION	CONTEXTUAL CONSTRAINTS
OVERALL DISCRIMINATION OF A TEST		DOES NOT ALLOW ESTIMATION OF PROBABILITY OF DISEASE IN INDIVIDUALS	
Test accuracy	Proportion of total diagnostic judgements that are correct. (TP+TN)/N -Values closer to 1 (100%) indicate increasing accuracy) -Requires discrimination	Overall discrimination of a dichotomous test. Does not allow estimation of probability of disease in individuals. Useful if comparing 2 tests with similar properties. No guidelines for translation of quantitative estimates of test accuracy into a measure of diagnostic utility.	Accuracy varies with spectrum therefore cannot be assumed to be portable across different populations. Implicitly values FN and FP equally.
Error rate	Proportion of total diagnostic judgements that are test errors. (FN+FP)/N -Values closer to 0 (0%) indicate increasing accuracy -Requires discrimination	Overall discrimination of a dichotomous test. Does not allow estimation of probability of disease in an individual. Useful if comparing 2 tests with similar properties. No guidelines for translation of quantitative estimates of error rates into a measure of diagnostic utility.	Accuracy varies with spectrum and cannot be assumed to be portable across different populations. Implicitly values FN and FP equally.
Diagnostic odds ratio (DOR)	The cross product ratio of the 2x2 diagnostic table. (TPXTN) / (FNxFP). -Accuracy increases the more the DOR increases from one Requires discrimination	Overall discrimination of a dichotomous test. Does not allow estimation of probability of disease in an individual. Useful if comparing 2 tests with similar properties. No guidelines for translation of quantitative estimates of the DOR into a measure of diagnostic utility.	Accuracy varies with spectrum and cannot be assumed to be portable across different populations. Implicitly values FN and FP equally. Relatively constant with changes in test threshold.
Area Under the Curve (AUC)	-Area under the Receiver Operator Characteristic curve (ROC curve) - AUC 0.5: a non-informative test; AUC 1: a perfect test -Requires discrimination	Overall discrimination of a dichotomous test. Does not allow estimation of probability of disease in an individual. Useful if comparing 2 tests with similar properties. No guidelines for translation of quantitative estimates of AUC into a measure of diagnostic utility.	Accuracy varies with spectrum and cannot be assumed to be portable across different populations. Implicitly values FN and FP equally. Relatively constant with changes in test threshold.

Notes to table 1.8: TP: true positive; TN: true negative; FP: false positive; FN: false negative. *Test result as reference class: TP as a proportion of test +ve results and TN as a proportion of test –ve results. Disease as reference class: TP as a proportion of diseased individuals and TN as a proportion of non diseased individuals. **Bayes' theorem : $P(D+|X) = \frac{(P(D+) \times P(X|D+))}{(P(D+) \times P(X|D+) + P(D-) \times P(X|D-))}$

Table 1.8 illustrates that test accuracy metrics can be subdivided into those that allow the estimation of disease probability in individuals and those that provide information only on the overall discriminatory ability of test. The former are more aligned to decision making at the bedside (interpretation of test results) whereas the latter may be more useful for decisions about testing policy, with the caveat that such comparisons of test performance implicitly value false positive and false negative test errors equally.

Of those metrics that allow the estimation of disease probability at the bedside, those that communicate test accuracy with test result as reference class might be considered more clinically intuitive than those with disease status as reference class, on the basis that a diagnosis is made on the basis of a test result; disease status is unknown at the time testing is performed.

Able to compute accurately

The normative rule for updating of opinion (pre-test probability of a disorder) with imperfect information (uncertainty conveyed by measures of test accuracy) is Bayes' theorem (see notes to table 1.8 above). The derived post-test probability becomes the pre-test probability for any subsequent tests. In clinical medicine a final diagnosis is usually reached following a number of tests which may be applied sequentially or simultaneously. This complicates Bayes' theorem and makes its application increasingly impractical, particularly where diagnostic tests are not independent (as is often the case). Linear and logistic regression play a role in deriving clinical prediction rules under such circumstances but they are not exhaustive in their coverage, they are disorder rather than symptom based which may limit their applicability in generalist settings and there are many problems with their derivation, validation and portability across settings and populations [64]. Prediction models are not seen as substitutes for medical decision making by doctors [65].

The ability to compute accurately will therefore depend on the complexity of the problem and the knowledge, skill and preferences of the decision maker [45,48]. Cognitive shortcuts

(heuristics) have been observed to introduce systematic errors during the processing of probabilities. Box 1.9 outlines common biases which are believed to overlap and interact. Understanding the impact of these cognitive biases on diagnostic decision making would be necessary if educational interventions to improve probability revision were to be pursued.

Box 1.9: Cognitive errors identified by the decision making literature [59,66-71]

COGNITIVE ERROR / BIAS	**DESCRIPTION**
Base rate neglect	*The tendency to ignore information about base rate (the pre-test probability) of an event. Base rate neglect has been observed to occur more frequency when dealing with low prevalence conditions / low frequency events. Base rate neglect may underlie the observation that sensitivity and specificity are confused with positive and negative predictive values (for example individuals given information about sensitivity assume this is the post-test probability of having disease thereby effectively ignoring pre-test probability (base rate)).*
Conjunction Fallacy	*The tendency to judge that the conjunction of two events is more probable than one of the events in a direct comparison. The explanation for the existence of this heuristic is purported to be due to judgements of representativeness (similarity to stereotypes).*
Availability heuristic	*The tendency to overestimate the probability of events that come easily to mind, for example that observed recently or that which is memorable.*
Confirmatory bias	*The tendency to look for information that fits pre-existing expectations and to dismiss information that contradicts pre-existing information.*
Hindsight bias	*Assignment of the posterior probability to disease (after testing) to the prior probability resulting in overconfidence in initial diagnoses and a tendency to ignore diagnostic cues.*
The unpacking effect	*An increase in the subjective probability assigned to an event when its description is more detailed.*
Sub-additivity	*Overestimation of probabilities leading to the sum of the probabilities of competing hypotheses exceeding 1.0.*

1.5.2.2 Fully rationale: Motivational errors and biases

Individuals can be considered rational if they base decisions on reason and knowledge rather than emotional response. Objectivity when processing information about the value of different outcomes is compromised by prior beliefs, expectations and the value (utility) attached to risks. Professional utility may be affected by many factors including fear of litigation [11,19], financial motives [19] and risk preference[11].

Estimation of probabilities has been shown to be distorted in risky decision making, depending on the perceived seriousness of the outcome [10]. An individual's risk preferences are therefore unlikely to be fixed, but dependent on context and the time frame being considered.

Different attitudes to gains and losses have also been observed to impact on values attached to risk; a departure from behaviour predicted by normative decision theory. This is outlined by prospect theory [72] which describes how people think of possible outcomes relative to a reference point (the framing effect). Losses from a reference point are perceived as worse than gains and as a result risk attitudes to gains are different to risk attitudes to losses. Framing of outcomes positively (as gains) or negatively (as losses) can therefore affect the values placed on them – individuals are more likely to choose an option framed as a gain from a reference point compared to the same option framed as a loss from the same reference point. Thus option A, with an 80% chance of not having disease X (a gain) would be preferred to option B, with a 20% chance of having disease X (a loss). Prospect theory predicts that when faced with gains individuals tend to be risk averse but when faced with losses they tend to be risk takers.

In healthcare decision making the healthcare professional as agent, acting on behalf of the patient, introduces complexity into the notion of utility, incorporating both patient and professional values [10]. With professional as agent, the extent to which patient utility is

incorporated in decision making will depend on an accurate assessment of patient utility by professionals, itself dependent on effective communication of uncertainty in the consultation.

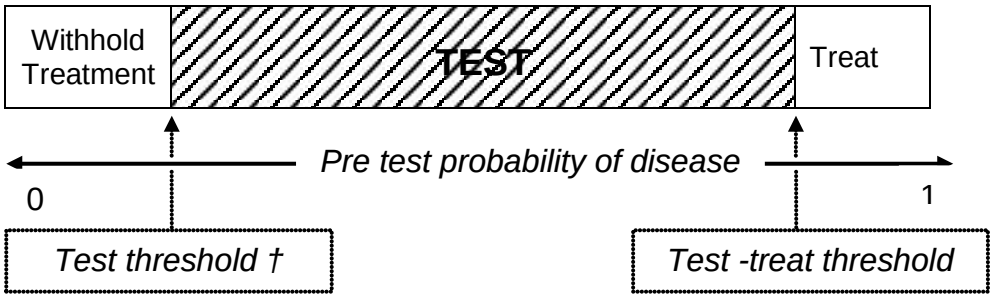
Rational decisions about testing should be concerned with maximising the outcome for the recipient of the test; the patient. However there is a utility associated with testing decisions for professionals as well as patients. For example Hozo (2008)[73] introduced the notion of regret to incorporate professional utility into diagnostic decision making: the difference in utility between the best possible action in retrospect (which may be the action taken) and the utility of the action taken. Relevant to the concept of professional utility are the relative values placed on acts of commission and acts of omission. Traditionally harm resulting from inaction (omission) is viewed as more morally acceptable in comparison to harm resulting from action (commission) as commission carries with it notions of intent [74]. This phenomenon is termed omission bias [75]. Indeed it has been argued that omission bias may be responsible for the majority of medical errors, a large proportion of which are errors concerned with diagnosis [75]. Performing a test could be considered clearly as an act (a commission) and not performing a test an act of omission. However when considering a management decision based on a test result the existence of test errors adds complexity to the assessment of whether an act is one of commission or omission. Initiating treatment or further testing following a positive test result might be considered an act of commission if the positive test result is a false positive but not if it is a true positive. Similarly, not initiating further treatment or testing on the basis of a negative test result might be viewed as an act of omission if the negative test result is a false negative. Assessment by decision makers of the nature of an action (commission or omission) following a test result will therefore be complicated by a requirement for knowledge about test error rates and assessment of the relative utility associated with each type of test error.

In summary the consideration of professional and patient utility in healthcare decision making, particularly where these may be incongruous, introduces complexity into the

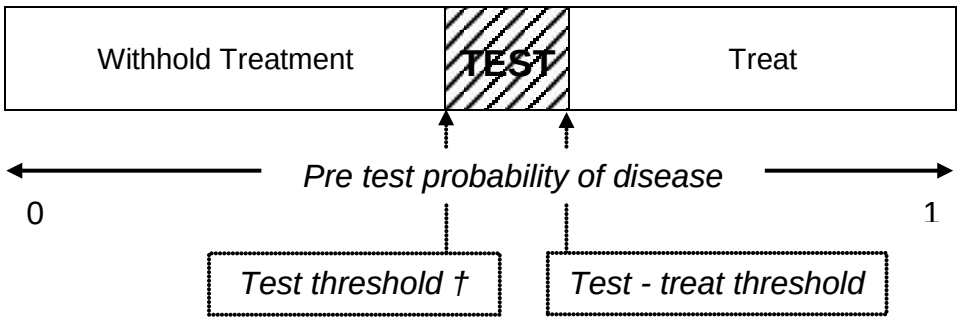
definition of optimality as defined by Einhorn and Hogarth 1981 [54] and may not be well represented by traditional economic behavioural decision theory. The fact that testing is associated with four possible outcomes, each with its own utility, magnifies this complexity. The threshold approach to clinical decision making as proposed by Pauker [76] is a useful model for illustrating the incorporation of utility in decisions about test use (see fig 1.10 below).

Fig 1.10: Diagrammatic illustration of the test and test-treat threshold model (Pauker 1980) [76]

(a)Test with greater accuracy or lower direct risks associated with testing



(b)Test with lower accuracy or greater direct risks associated with testing



Notes to Fig 1.10:
Test threshold: probability of disease at which the utility associated with testing to determine the subsequent treatment decision = utility of withholding treatment without testing first.
Test-treat threshold: probability of disease at which the utility associated with testing to determine the subsequent treatment decision = utility of administering treatment without testing first.

Figure 1.10 illustrates how testing context might impact on decisions about test use and in the application of a given test result. The model assumes that for a particular disease and its associated treatment, there exists a therapeutic probability threshold (the probability of disease at which the utility associated with administering treatment is equal to the utility associated with withholding treatment). The therapeutic (probability) threshold is determined by the risks associated with untreated disease and the risks associated with treatment itself. Pauker [76] extended the treatment threshold concept to incorporate the utility associated with testing, which for a particular test is determined by the risks associated with administering the test and the risk of test errors (false positives and false negatives). The test threshold is defined as the probability of disease at which there is no difference in the utility associated with withholding treatment without testing first or testing first to determine the subsequent treatment decision. The test-treat threshold is defined as the probability of disease at which there is no difference in the utility of administering treatment without testing first, or testing first to determine the subsequent treatment decision. At probabilities of disease higher than the test threshold but lower than the test-treat threshold, the balance of utility is in favour of testing to determine the subsequent treatment decision. At probabilities higher than the test-treat threshold the balance of utility is in favour of treating without testing first. For a given disease and associated treatment, tests with higher accuracy (and therefore a lower risk of test errors) and /or lower risks associated with test administration, result in a decision to test prior to determining a management approach over a larger range of disease probabilities than tests with lower accuracy and/or higher risks associated with their administration. The implication for research is that judgements about the quality of diagnostic decision should incorporate utilities and not rely on demonstration of accurate probability revision alone. Optimising the utility from testing requires an appreciation of the downstream consequences of each of the four possible outcomes of testing and a comparison of their probabilities for each specific testing context.

1.5.3 Application of information about test accuracy in clinical practice

This chapter has outlined the gap between testing practice and that predicted from clinical need and the existing evidence base. Variables that have been investigated as potential modifiers of test ordering behaviour explain a small amount of the observed variation in testing and the potential contribution of informed diagnostic decision making appears to have received little attention in this respect. This is despite the fact that probabilistic reasoning is proposed to be a prominent part of diagnostic decision making and the paradigm on which evidence based medicine is based. Contextual considerations are likely to be particularly important for decisions about test use and in the application of test results. Clinical context encompasses variables that are potential modifiers of test accuracy. These include factors that shape the spectrum of the population to be tested (for example severity of the target disorder, co-morbidities, the stage in a testing pathway that an index test is being used as a proxy for prior tests received and features of the healthcare setting that might impact on test conduct), as well as variation in test technology, application and interpretation that are independent of healthcare setting, prevalence of the target disorder and the intended role and application of the test under evaluation including consideration of the downstream consequences of test results. Knowledge concerning the extent to which test accuracy measures effectively convey the uncertainty associated with testing to decision makers and the degree to which the testing context is represented by test accuracy evidence is important if interventions are to be designed with the aim of reducing the observed gap between evidence and testing practice.

1.5.4 Thesis outline

The following chapters represent an investigation of the extent to which the test accuracy evidence base supports informed diagnostic decision making. The framework for this

investigation will draw on the assumptions of behavioural decision theory: a decision maker who is fully informed, fully rational and able to compute accurately.

- ***Chapter 2*** presents a review of the existing empirical and non empirical literature concerning understanding and application of test accuracy measures by decision makers, complemented by a review of the more established risk communication evidence base. The scope of the review will be comprehensive with respect to use of test accuracy evidence and information about pre-test probability (fully informed); the role and adequacy of probability revision in practice (able to compute accurately) and the influence of utility judgements on decision making (fully rational).
- ***Chapter 3*** presents an epidemiological mapping exercise of existing test reviews with respect to coverage of disease topic areas, representation of healthcare setting and review purpose. The mapping exercise will measure the extent to which secondary test accuracy evidence is fit for purpose to fully inform decision makers across different healthcare settings.
- ***Chapter 4*** presents a review of the degree to which clinical context shapes the conduct and reporting of existing test accuracy reviews. The focus of the review will be to assess the extent to which secondary test accuracy evidence provides information to inform decisions in specific testing contexts (fully informed) (fully rational).
- ***Chapter 5*** is a survey of use, understanding and application of test accuracy measures in a sample of primary care clinicians. The survey aims to assess the degree to which test accuracy evidence is sought and used in practice (fully informed) (able to compute accurately) as well as gaining an insight into the diagnostic decision making process beyond probability revision, (fully rationale) and exploring the purported centrality of probability revision as a pre-requisite for informed decision making.

Chapter 2: Review of literature concerned with the understanding and application of test accuracy and risk measures

2.1 Abstract

Background

The widespread belief that decision makers have difficulty understanding and applying test accuracy information has not been based on a systematic interrogation of the evidence base to allow quantification or characterisation of the extent of the problem.

Aims and objectives:

To comprehensively ascertain literature pertinent to the understanding and application of test accuracy measures in order to identify facilitators and barriers to their use by decision makers.

Methods

Bibliographic searches were conducted in 2003, 2005, 2007 and 2010, across 11 databases representing medicine, psychology and education. Searches were iterative, purposive and supplemented by reference checking included studies and contact with experts. A narrative synthesis of empirical and theoretical test accuracy and risk communication literature was undertaken.

Results

64 test accuracy and 21 risk communication papers were included. Research is characterised by self selected samples, lacks external validity and primary care is under-represented. Ability to define the most commonly used metrics (sensitivity, specificity, predictive values) is poor. Predictive values and test errors are promoted as most intuitive although there is no empirical evidence supporting the superiority of a single test accuracy metric for diagnostic decision making. Natural frequency and multiple presentation formats facilitate understanding. Verbal descriptions and negative test results may be less well understood. Self-reported use of measures varies: predictive values 80%, sensitivity and

specificity 4% and ROC curves and LRs < 1%. Pre-test probability and test accuracy estimation is inaccurate and highly variable which has implications for probability revision.

Conclusions

The emphasis in the literature has been on identifying the best single metric rather than identifying an optimal combination and understanding of meta-analytic summary measures has not been investigated. Investigation of contextual and motivational influences on test and test-treat thresholds is required to identify test accuracy magnitudes that will have most impact on diagnostic and therapeutic yield.

2.2 Review rationale and aims

2.2.1 Rationale

The rationale for the review was to build up a theoretical picture of the proposed strengths and weaknesses of existing measures of test accuracy and to summarise existing empirical literature examining their interpretation and application.

2.2.2 Aims

The aims were to comprehensively capture empirical literature concerned with the understanding and application of test accuracy measures, whilst a review of theoretical perspectives (non-empirical literature) adopted a qualitative approach to searching and synthesis, with the aim of providing a representative map of comment and analysis offering insights and opinions about characteristics of existing or novel accuracy measures that may impact on their interpretation and application.

2.2.3 Objectives

- To assess the extent to which decision makers understand and can apply test accuracy metrics
- To make recommendations for the practice of reporting evaluations of test accuracy, particularly systematic reviews and meta-analyses
- To identify the need for further research.

In anticipation of a paucity of literature specifically concerned with understanding and application of test accuracy measures, a review of literature concerned with the communication of risk more generally in healthcare settings was planned. Drawing on the common theme of communicating uncertainty, the rationale was that a larger body of research concerned with the communication of risk may provide insights into formats for

presenting probabilistic information that were novel for the communication of test accuracy information. Although the focus of this programme of research was the communication of test accuracy in systematic reviews, searches were not restricted to summary measures more typical or unique to reviews. The challenges in the interpretation of summary measures of test accuracy [34,34] are common to both systematic reviews and primary studies as the majority of outcome measures in use are shared by both types of research [6,7].

2.3 Review methods

2.3.1 Methods: Review Search strategy

In order to fulfil the aims of a review it was anticipated that perspectives from a range of disciplines would be relevant so the search strategy included interrogation of the medical, psychological and educational literature. In addition the preliminary results of the literature search were presented at a national conference (Methods for Evaluating Tests and Biomarkers: second international symposium, University of Birmingham, July 2010) as a test of face and content validity.

Bibliographic searches were conducted in 2003, 2005, 2007 and 2010, from inception to date in each database. The searches conducted in 2005, 2007 and 2010 were designed to be iterative and purposive, building on knowledge and sources of relevant literature obtained to date, and in addition an aim of the 2010 search was to update the currency of the previous search strategies. During this period literature obtained passively through discussions with experts and identified opportunistically by information specialist colleagues was also assessed and reference checking of included articles was undertaken.

Over the period of searching the following bibliographic databases were searched using various configurations of text and MESH terms in order to identify published, unpublished and on-going work concerned with understanding and interpreting test accuracy measures (see appendix 2.1 for search strategies employed 2010):

Cochrane methodological register (CMR)

Science Citation Index (Web of Science)

Social Science Citation Index (Web of Science)

Web of Science including Proceedings

Cochrane library (DARE; HTA; CDSR; Central)

MEDION Methods database

MEDLINE

Embase

PsycInfo

ERIC

ISI proceedings

ZETOC

NIHR NETSCC

CADTH NETCC

In addition an author search was conducted for authors who had published any empirical research concerning understanding of test accuracy by health professionals or who had published substantively in related areas.

2.3.2 Methods: Inclusion criteria

Following completion of the test accuracy reviews and a scope of the risk communication literature, a decision was made to restrict the risk review to reviews of empirical studies and primary, empirical research updating these reviews (see table 2.1).

Table 2.1: Inclusion criteria for risk communication literature following scoping searches: Map of risk communication outcome by review date and population (Health and Non-Health (N-H) professionals)

		REVIEW, DATE PUBLISHED AND SEARCH END DATE										
RESEARCH QUESTION		**Edwards 2000 (1996)** (ER5)	**Kuhberger 1998 (1997)** (ER13)	**McGettigan 1999 (?)** (ER16)	**Lipkus 1999 (1998)** (ER14)	**Edwards 2001 (1999)** (ER6)	**Julien-Reynier 2003 (2002)** (ER12)	**Epstein 2004 (2003)** (ER10)	**Ancker 2006 (2005)** (ER2)	**Edwards 2006 (2005)** (ER7)	**Edwards 2008 (2006)** (ER9)	**Albada 2009 (2007)** (ER1)
		N-H	**N-H**	**Health**	**Health and N-H**	**N-H**	**N-H**	**Health and N-H**	**Health and N-H**	**N-H**	**N-H**	**N-H**
	Framing /Risk taking attitudes		X	X	X	X	X		X			
Tailoring	**Tailored content**	X					X			X	X	X
	Tailored presentation	X					X				X	
Graphical presentation					X	X		X	X			
Numerical Presentation	**Numerical presentation (Frequencies vs probabilities)**							X	X		X	
	Numerical presentation (RR vs AR vs NNT)			X		X		X				
	Numerical presentation (probability)			X	X	X	X			X	X	
Verbal presentation				X	X	X	X			X	X	
Other presentation format						X		X			X	

Notes to table 2.1: RR: Relative Risk. AR: Attributable Risk. NNT: Number Needed to Treat. Other presentation format includes: more versus less information; lay versus medical; range versus point estimate; presentation order; manipulation of base rates

Literature concerned exclusively with communication of uncertainty in non-healthcare settings was not included on the basis that contextual features unique to the healthcare setting may substantially affect the interpretation, communication and use of risk and test result information. However perceptions and understanding of both health and non-health professionals was considered. At the outset only non-English language studies concerned with empirical evaluation of understanding of test accuracy measures were to be considered for translation.

2.3.3 Methods: Data extraction, quality assessment and synthesis

Literature concerned with communication of test accuracy was sub-divided into empirical and non-empirical. Empirical literature included any quantitative or qualitative investigation of understanding; application; behaviour (effect on test or treatment use or uptake); use of measures in practice; preference or attitudes to test accuracy or risk information by individuals. Research investigating preferences and attitudes were included as these were considered important variables impacting on use and understanding. Non-empirical literature comprised opinions about the communication of uncertainty in healthcare testing settings. Empirical studies were tabulated, making a distinction between risk and test accuracy research and health professionals and non-health professionals. Information was extracted on study setting, study design, type and presentation format of test accuracy or risk measures and study findings. Methodological quality of papers was not comprehensively assessed using formal checklists due to the diversity of the literature included. For reviews the adequacy of question formulation, search strategy, study flow, quality assessment and presence of double data extraction were assessed in addition to consideration of external validity. For Randomised Controlled Trials (RCTs), adequacy of randomisation, allocation concealment, similarity of groups at baseline, blinding and attrition were assessed. For other study designs assessment of selection and measurement bias was undertaken where

possible. The aim was to provide an indication of the applicability and quality of evidence in the field. Empirical findings were synthesised drawing on the assumptions of normative decision theory: the expectation that (diagnostic and treatment) decisions are made with the aim of maximising the value of the outcomes possible from competing choices. This requires a decision maker who is fully informed, fully rational and able to compute accurately (see 1.5.2). For the purposes of synthesis, these decision maker attributes were mapped to the outcomes outlined in table 2.2. For empirical research, outcomes of interest were defined a priori whereas for non empirical, theoretical papers the outcomes presented in table 2.3 represent the result of thematic coding.

The method of qualitative synthesis chosen for literature concerned with theoretical perspectives (non-empirical) was on the basis of the desired synthetic product [77]:

The impact of moderators of understanding and application of test accuracy information in healthcare settings.

By definition theoretical, non empirical literature represents second order interpretations of phenomena based on observation or empirical research conducted by the authors. The aim of this qualitative synthesis was therefore twofold: firstly to translate second order explanations (conceptual themes) across individual articles; secondly to develop a line of argument analysis, taking into consideration the perspectives of different research traditions and differences in theoretical assumptions. This approach borrows from applications of meta-ethnography as applied to non-ethnographic studies [78] and critical interpretive synthesis [79].

Table 2.2: Mapping of normative decision theory assumptions to outcomes considered in the reviews of test accuracy and risk communication

	Test accuracy literature outcomes	General risk communication literature outcomes
Fully informed	**Estimation of pre-test probability or test accuracy**	**Estimation of baseline or intervention/exposure risk**
	Accuracy of pre-test probability or test accuracy estimation as measured by one or more of the following: -Perception (size relative to a reference point) -Quantitative or semi-quantitative estimation -Precursor of change in testing / treatment /other behaviour	Accuracy of risk estimation as measured by one or more of the following: -Perception (size relative to a reference point) -Quantitative or semi-quantitative estimation -Precursor of change in testing / treatment /other behaviour
	Test accuracy metrics	**Risk metrics**
	Familiarity / Understanding of test accuracy measures as indicated by one or more of: -Preference -Use in practice -Comprehension -Appreciation of contextual variation in test accuracy -Consideration of downstream consequences of test results on patient outcomes - Precursor of change in testing / treatment /other behaviour	Familiarity / Understanding of risk measures as indicated by one or more of the following: -Preference -Use in practice -Comprehension -Appreciation of contextual baseline variation in risk -Behaviour change following risk communication -Precursor of change in testing / treatment /other behaviour
Fully rational	Consideration of patient and professional utility in the decision making process (anxiety / attitude / affect / framing effects)	Consideration of patient and professional utility in the decision making process (anxiety / attitude / affect / framing effects)
Able to compute accurately	-Ability to undertake probabilistic reasoning as indicated by one or more of: -Quantitative or semi-quantitative adjustment in disease probability pre to post testing. - Precursor of change in testing / treatment /other behaviour	-Ability to undertake risk manipulation and compare risks as indicated by one or more of: -Quantitative or semi-quantitative manipulation of risk measures -Precursor of change in testing /treatment /other behaviour

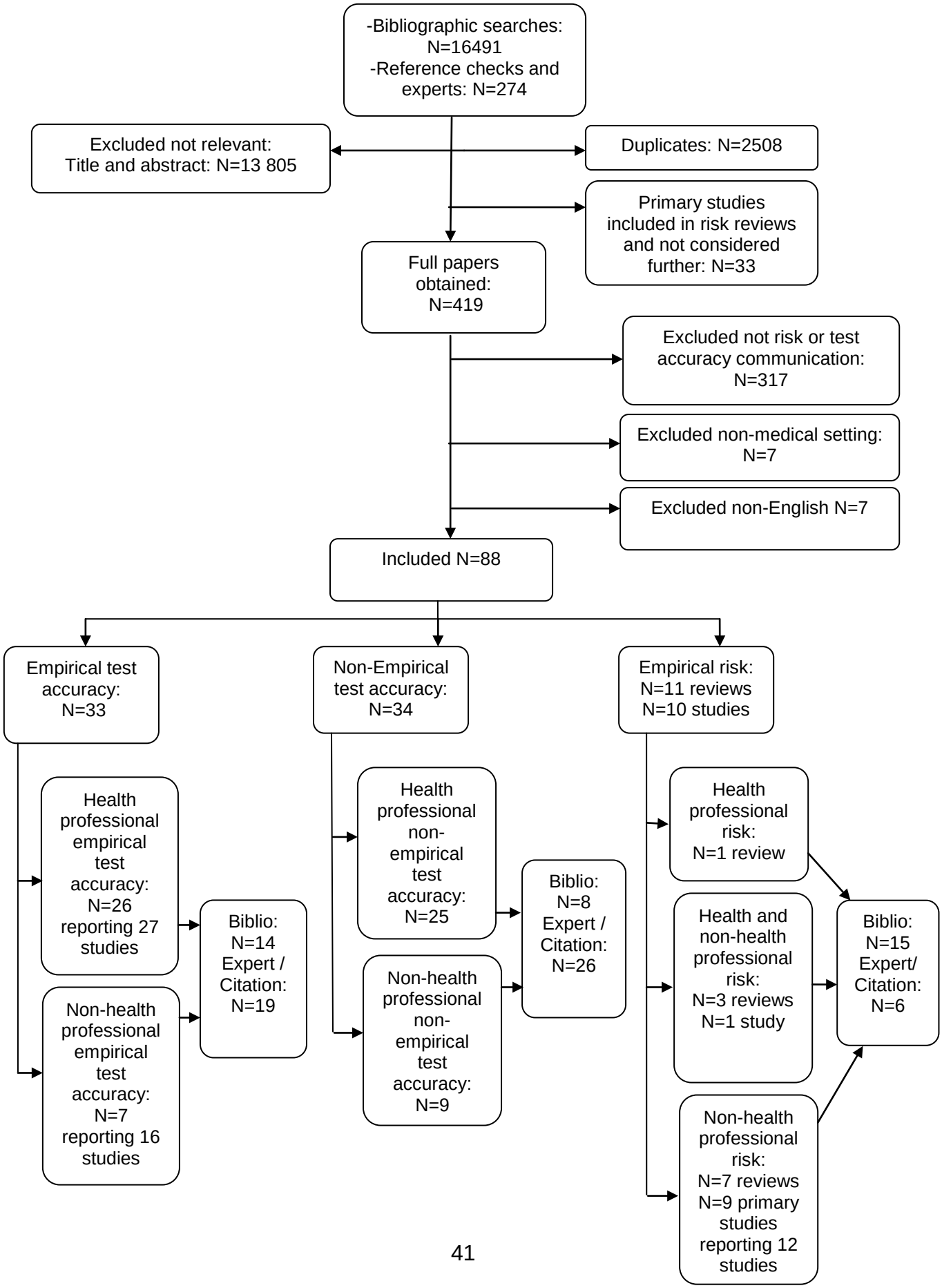
Initial themes for synthesis of non-empirical literature included familiarity, understanding and use of existing test accuracy measures, probabilistic versus frequentist expressions of uncertainty, contextual considerations including knowledge about disease prevalence and attitudes to risk and probabilistic reasoning. These were generated a priori and expanded as additional themes emerged from the literature. In order to assist with the development of a line of argument analysis, literature was considered in date order and according to authors'

profession as far as could be gleaned from published affiliations. Second order interpretation themes from authors were grouped, translated across articles and linked to the three assumptions underlying normative decision theory. A priori themes that were not represented in the literature included familiarity of decision makers with measures of test accuracy and the extent of use of test accuracy measures in practice.

Table 2.3: Organisation and linking of 2nd order interpretation themes

2nd order interpretation themes	3rd order interpretations				
	Collapsing of themes				Assumptions underlying normative decision theory
Likelihood Ratio (LR)	Test accuracy measures				Fully informed
Sensitivity and Specificity					
Global measures of test accuracy					
Predictive values (PVs)					
Graphical methods of expressing test accuracy					
2x2 table					
Test errors		Contextual factors			
Patient and setting-specific factors that might impact on test use/interpretation					Fully rational
Estimation of pre-test disease probability					Fully informed
Decision maker attitudes to risk			Attitudes to risk		Fully rational
Probabilistic versus frequentist expression of uncertainty				Probabilistic reasoning	Able to compute accurately
Pre to post-test probability revision					

Fig 2.4: Review Study flow



2.4 Results: Non-empirical test accuracy literature

Table 2.5 documents the author affiliation, date, publication details and discussion themes of the 34 articles by 30 unique authors. Most articles identified were published in general medical journals by clinicians, 16/25 of whom were affiliated with an academic institution. The majority of literature identified was published in the last two decades (88% after 1990), the rapid increase during the 1990s coinciding with the introduction of the Evidence Based Medicine (EBM) movement [80] which promotes the integration of external evidence with individual expertise for medical decision making [81].

Table 2.5: Non-empirical literature: Date, place of publication and discussion themes

Author, Year & place of publication	Title	Test accuracy measures	Contextual factors including knowledge of pre-test probability	Attitudes to risk	Probabilistic reasoning
(TTA1) Akobeng 2007(a) (C) Acta Paediatrica	Understanding diagnostic tests 1: sensitivity, specificity and predictive values	-Sensitivity, specificity, PVs and 2x2 table -SnNOUT and SpPIN heuristics -Graphical presentation of test accuracy	-Prevalence effects		
(TTA2) Akobeng 2007 (b) (C) Acta Paediatrica	Understanding diagnostic tests 2: likelihood ratios, pre and post-test probabilities and their use in clinical practice	-Advantages of LRs -Disadvantages of sensitivity and specificity			-Graphical tools (nomogram) for probability revision
(TTA3) Benish 2003 (C) Methods of Information in Medicine	Mutual information as an index of diagnostic test performance	-Mutual information as a novel test accuracy measure	-Contextual modifiers of diagnostic information		
(TTA4) Bianchi 2006 (C) BMJ	Evidence based diagnosis: does the language reflect the theory?	-Test accuracy language - 2x2 table, PVs and prevalence -Disadvantages of sensitivity and specificity	-Spectrum effects -Prevalence effects		-Test accuracy language as a barrier to probabilistic reasoning -Confirmatory bias
(TTA5) Daniel 1993 (C) Medical Decision Making	Graphic representation of numerically calculated PVs: an easily comprehended method of evaluating diagnostic tests	-Advantages and disadvantages of sensitivity and specificity -Graphical tools for test comparisons.			- Limitations of existing test accuracy measures for probability revision

Table 2.5 continued

Author, Year & place of publication	Title	Test accuracy measures	Contextual factors including knowledge of pre-test probability	Attitudes to risk	Probabilistic reasoning
(TTA6) Doust 2010 (CA) BMJ	Using probabilistic reasoning	-SnNOUT and SpPIN -Graphical f test accuracy presentation -PVs -Test errors	-Prevalence effects and test errors		-Qualitative and quantitative probabilistic reasoning
(TTA7) Dujardin 1994 (CA) European Journal of Epidemiology	Likelihood ratios: a real improvement for clinical decision making?	-Advantages and disadvantages of LRs, PVs, sensitivity and specificity -Test errors	-Spectrum effects		-Advantages of LRs for probabilistic reasoning
(TTA8) Falk 2009 (C) BMJ	Diagnosis in General Practice: Clinical Prediction Rules	-Sensitivity, specificity and LR interpretation	-Healthcare setting and the two dimensions of test accuracy		-Clinical prediction rules -Sequential testing
(TTA9) Gigerenzer 1995 (A) Psychological Review	How to improve Bayesian reasoning without instruction: frequency formats				- Frequentist expression facilitates probability revision
(TTA10) Gigerenzer 1996 (A) Medical Decision Making	The psychology of good judgment: frequency formats and simple algorithms				-Human cognitive algorithms. - Frequentist expression facilitates probability revision -Satisficing algorithms and heuristic reasoning
(TTA11) Gigerenzer 2003 (A) BMJ	Simple tools for understanding risks: From innumeracy to insight				- Frequentist expression facilitates probability revision -Reference class confusion: sensitivity & PPV

Table 2.5 continued

Author, Year & place of publication	Title	Test accuracy measures	Contextual factors including knowledge of pre-test probability	Attitudes to risk	Probabilistic reasoning
(TTA12) Gill 2005 (A) BMJ	Why clinicians are natural Bayesians	- Advantages of LRs	-Estimation of pre-test probability		- Bayesian reasoning -Frequentist expression of uncertainty -Semi-quantitative vs quantitative probabilistic reasoning -Sequential testing
(TTA13) Gorry 1978 (A) The New England Journal of Medicine	The diagnostic importance of the normal finding				-Semi-quantitative pre to post test probability revision -Frequentist probabilistic reasoning implicit
(TTA14) Grimes 2005 (CA) Lancet	Refining diagnosis with likelihood ratios	- Advantages and disadvantages of LRs, sensitivity and specificity. LR non-linear scale may complicate interpretation			-Facilitation of probability revision with LRs -Graphical tools (nomogram) for probability revision
(TTA15) Halkin 1997 (C) Quarterly Journal of Medicine	Likelihood ratios: getting diagnostic testing into perspective	- Advantages of LRs	-Spectrum effects		-Quantitative versus qualitative probabilistic reasoning.
(TTA16) Henderson 1998 (CA) BMJ	Test accuracy is example of redundant information	-Advantages and disadvantages of global measures of test accuracy -Consideration of two dimensions of test accuracy			

Table 2.5 continued

Author, Year & place of publication	Title	Test accuracy measures	Contextual factors including knowledge of pre-test probability	Attitudes to risk	Probabilistic reasoning
(TTA17) Hoffrage 2002 (A) Cognition	Representation facilities reasoning: what natural frequencies are and what they are not				- Frequentist expression facilitates probability revision -Natural and normalised frequencies
(TTA18) Kassirer 1989 (C) New England Journal of Medicine	Our stubborn quest for diagnostic certainty: a cause of excessive testing		-Erroneous estimation of pre-test probability	-Attitudes to uncertainty. -Attitudes to uncertainty as a modifier of test use	-Categorical / quantitative expressions of uncertainty -Limitations of medical training for probability-orientated thinking
(TTA19) Klein 2005 (A) BMJ	Five pitfalls in decisions about diagnosis and prescribing				-Cognitive biases in diagnostic decision making
(TTA20) Knottnerus 1985 (CA) Journal of the Royal College of General Practitioners	Interpretation of Diagnostic Data - An Unexplored Field in General-Practice		-Healthcare setting as a moderator of pre-test probability and test accuracy -Pre-test probability estimation. - Secondary care focus of medical training		
(TTA21) Loong 2003 (CA) BMJ	Understanding sensitivity and specificity with the right side of the brain	-Graphical presentation of test accuracy			-Graphical presentation of test accuracy facilitates probabilistic reasoning
(TTA22) McCowan 2006 (CA) British Journal of General Practice	Diagnosis and diagnostic testing in primary care		-Contextual difficulties in primary care -Secondary care focus for test accuracy research		

Table 2.5 continued

Author, Year & place of publication	Title	Test accuracy measures	Contextual factors including knowledge of pre-test probability	Attitudes to risk	Probabilistic reasoning
(TTA23) Miettinen 1998 (CA) Journal of Clinical Epidemiology	Evaluation of Diagnostic Imaging Tests: Diagnostic Probability Estimation				-Use of logistic regression in probabilistic reasoning
(TTA24) Moons 2003 (A) Academic Radiology	Sensitivity and Specificity Should be De-emphasized in Diagnostic Accuracy Studies	-Importance of reference class for interpretation of conditional probabilities	-Spectrum effects		
(TTA25) Pewsner 2004 (A) BMJ	Ruling a diagnosis in or out with "SpPIN" and "SnNOUT": a note of caution	-Advantages and disadvantages of sensitivity and specificity -Advantages and disadvantages of LRs			
(TTA26) Richardson 2003 (CA) Journal of General Internal Medicine	Could our pre-test probability estimates become evidence based? A prospective survey of hospital practice		-Pre-test probability estimation -Translation of research to practice		
(TTA27) Sackett 1998 (CA) Evidence Based Medicine	On some clinically useful measures of the accuracy of diagnostic tests	-Importance of reference class for interpretation of conditional probabilities - Advantages of LRs	-Pre-test probability estimation -Contextual modifiers of test and test-treat thresholds		-Graphical tools (nomogram) to facilitate probabilistic reasoning
(TTA28) Sonis 1999 (CA) Family Medicine	How to use and interpret interval likelihood ratios	-Advantages and disadvantage of LRs -ROC curves	-Secondary care focus for research on pre-test probability -Spectrum effects		

Table 2.5 continued

Author, Year & place of publication	Title	Test accuracy measures	Contextual factors including knowledge of pre-test probability	Attitudes to risk	Probabilistic reasoning
[TTA29] Sox 1986 (CA) Annals of Internal Medicine	Probability theory in the use of diagnostic tests: An introduction to critical study of the literature	-SnNOUT and SpPIN heuristics	-Healthcare setting as a moderator of pre-test probability. -Heuristics introduce bias in pre-test probability estimation		-Graphical aids to probability revision -Test-treat thresholds - SnNOUT and SpPIN heuristics
[TTA30] Sox 2006(b) (CA) Annals of Internal Medicine	Better care for patients with suspected pulmonary embolism	-Test accuracy language	-Clinical prediction rules for improving pre-test probability estimates		
[TTA31] Stengel 2003 (C) Journal of Medical Screening	A likelihood ratio approach to meta-analysis of diagnostic studies	-Advantages /disadvantages of sensitivity, specificity &LRs -Graphical test accuracy presentation -Meta-analysis of test accuracy			
[TTA32] Summerton 2008 (CA) British Journal of General Practice	The medical history as a diagnostic technology	-Advantage/ disadvantages of PVs & LRs	-Prevalence effects -Pre-test probability estimation		- Sequential testing -Bayes' theorem -Logistic regression
[TTA33] Van den Ende 2005 (CA) The Lancet	The trouble with likelihood ratios	-Disadvantages of LRs			-LRs and probabilistic reasoning

Table 2.5 continued

Author, Year & place of publication	Title	Test accuracy measures	Contextual factors including knowledge of pre-test probability	Attitudes to risk	Probabilistic reasoning
[TTA34] Zaat 1992 (CA) Medical Care	General practitioners' uncertainty, risk preference and use of laboratory tests			-Contextual, cultural and person-specific modifiers of attitudes to uncertainty -Attitudes to uncertainty modify test use	

Notes to table 2.5: A: author affiliation included an academic institution; C: Author qualifications included MD or DR with affiliation to a clinical placement; LRs: Likelihood ratios; PVs: predictive values; two dimensions of test accuracy performance of test in diseased and non-diseased populations or discriminatory value of a positive or negative test result.

2.4.1 Results: Non-empirical test accuracy literature: fully informed

2.4.1.1 Having test result as reference class for interpretation of conditional probabilities

Having the test result (predictive values (PVs) and likelihood ratios (LRs)) as opposed to the disease state (sensitivity and specificity) as the reference class for interpretation of conditional probabilities was emphasised as intuitive:

"In our view, a single test's sensitivity and specificity are of limited value to clinical practice....They are reverse probabilities with no direct diagnostic meaning. In practice, of course, patients do not enter a physician's examining room asking about their probability of having a particular test result given that they have or do not have a particular disease; rather they ask about their probability of having a particular disease given the test result. The predictive value of test results reflects this probability of disease, which might better be called post-test probability.....In our view these parameters (sensitivity and specificity) are of limited relevance to practice, and their estimation should not necessarily be pursued in diagnostic research." (TTA24)

"As clinicians our interest isn't in the vertical columns (of the 2x2 diagnostic contingency table) ....if we knew what column our patient was in we wouldn't need the diagnostic test. We want to know the horizontal significance of this test result" (TTA27)

"Diagnostic tests are mainly used in clinical medicine to answer the Bayesian question, "What is the probability that the patient has the disease given an abnormal test?" not "What is the probability of an abnormal result given that the patient has disease?"" (TTA12)

"This is because sensitivity and specificity are defined on the basis of people with or without a disease. However because the patient would have presented to you with a set of symptoms rather than a diagnosis, you would not know at the time whether the patient has a disease or not and cannot, therefore, apply these parameters to them." (TTA2)

"A clinician will not start from diseased or not diseased, but from a positive or negative test. Therefore sensitivity and specificity are intuitively not so evident as the likelihood ratio." (TTA7)

2.4.1.2 Limitations introduced by having a fixed threshold of test accuracy

A minority of authors introduced the potential limitation of adopting a single fixed diagnostic threshold and multi-level LRs and the Receiver Operator Characteristic Curve (ROC curve) were discussed with respect to expressing the relationship between diagnostic threshold and test accuracy:

"..a newer way of expressing (a test's) accuracy with likelihood ratios reveals its even greater power and.....shows us how we can be mislead because the old sensitivity-specificity approach restricts us to just 2 levels (positive and negative) of the test result." (TTA27).

"When test results with continuous or ordinal outcomes are dichotomized for calculation of sensitivity and specificity, valuable information is lost, because results that are markedly

abnormal are lumped together with results that are only mildly abnormal. Interval likelihood ratios, however, assign a specific value to each of level of abnormality, and this value can be used to calculate the post-test probability of disease for a given level of a test." (TTA28)

"Collapsing multiple categories into positive and negative loses information. Likelihood ratios enable clinicians to interpret and use the full range of diagnostic test results." (TTA14)

The use of the ROC curve to illustrate the effects of changing thresholds on the two dimensions of test accuracy was also noted:

"In addition, sensitivity and specificity are not fixed values because they can be and should often be, altered by moving the decision threshold. This aspect is best examined by means of analysis of receiver operating curves." (TTA16)

2.4.1.3 Contextual factors influencing the interpretation and application of summary test accuracy metrics

Throughout the period covered by the review of theoretical literature there was an emphasis on the non-portability of PVs as a consequence of their mathematical dependence on prevalence.

"...predictive values are of course useless in other settings where the prevalence rate or pre-test probability is different..... PPV and NPV will be determined by the combination of the sensitivity and the specificity values of a test for a given disease and by disease prevalence." (TTA7)

"Although the positive and negative predictive values are the clinically useful measures, they are not generally reported in studies of the accuracy of diagnostic tests as predictive values vary greatly with changes in pre-test probability." (TTA6)

"Both PPV an NPV vary with changing prevalence of disease. It will therefore be wrong for clinicians to directly apply published predictive values of a test to their own populations when the prevalence of disease in their population is different from the prevalence of disease in the population in which the published study was carried out." (TTA2)

"Considerable confusion has been the consequence (of the belief that there are universal, standard predictive values) and clinicians continue to misunderstand this issue." (TTA7)

"...the positive predictive value... often makes the most intuitive sense to clinicians, and yet it is a constant source of misunderstanding between GPs and our secondary care colleagues. It is imperative to be aware that the predictive value is affected by prevalence." (TTA32)

"...likelihood ratios are portable. By contrast, predictive values of test are driven by the prevalence of the disease in question." (TTA14).

By contrast the non-portability of all test accuracy metrics as a function of population spectrum received relatively less attention and the result was to undermine the value of PVs in this respect:

"It is well known that post-test probabilities depend on disease prevalence and therefore vary across populations and across subgroups within a particular population, whereas sensitivity and specificity do not depend on the prevalence of disease. Accordingly the latter are commonly considered characteristics or constants of a test. Unfortunately, it is often not realised that this is a misconception.... in fact there can be no generally valid estimates of a test's sensitivity, specificity or likelihood ratio that apply to all patients of a particular population, nor should such values be sought." (TTA24)

"Specificity, like sensitivity, is often considered an intrinsic property of a test and therefore independent of the population under study. As specificity is determined by unaffected individuals who have positive results, however, it is in fact dependent on the characteristics of the comparison population." (TTA4)

"Like sensitivity and specificity, likelihood ratios values can change with different settings." (TTA7)

The different approach to diagnosis adopted in generalist and specialist setting was also noted with implications for test application and interpretation:

"In summary- in general practice sick people need to be distinguished from healthy people and in hospital healthy people need to be distinguished from sick people. That the validity of tests and the interpretation of symptoms should be directly related to the populations consulting and to the degree of clinical differentiation of the disease at reporting has only recently received attention in the literature." (TTA20)

"Whether a clinician wishes to rule in or rule out a disorder is likely to be specific to the setting of care and the nature and severity of the target disorder. For instance, clinical prediction rules may be used in primary care to rule out a disorder, provide reassurance or adopt a watchful waiting strategy. ....Ruling in a diagnosis is desirable in a secondary care setting where the emphasis is usually on establishing a firm diagnosis and starting appropriate treatment or conducting more expensive and invasive diagnostic tests." (TTA8)

There was recognition that an acceptable level of test accuracy would be context dependent and this was exclusively articulated using test errors; the relative values placed on false negatives and false positives:

"Positive and negative predictive values may not be equally important to diagnostic test users in individual use situations. The costs of false positives and false negative errors may be very different in given situations." (TTA5)

"Knowledge of both indices (sensitivity and specificity) is required to appraise test precision fully. However one might think of clinical situations in which only one of these characteristics is of real interest." (TTA31)

"What is the optimal cut off point for a test? The answer depends on the subjective values attached to false positive and false negative results." (TTA20)

" ...the feelings of uncertainty regarding medical problems can differ depending on the situation, not only because one physician may be faced with more complicated puzzles than the other but also, and primarily because the consequences of a vague and uncertain diagnosis may vary in each situation." (TTA34)

The view that without quantification of test errors the SpPIN (high specificity, low number of false positives, positive test result rules in a disorder) and SnNOUT (high sensitivity, low number of false negatives, negative test result rules out a disorder) mnemonics (TTA27) may mislead as a guide to the usefulness of a test:

"We make errors by believing false positive and false negative test results....When the prevalence or pre-test probability is low, the probability that a positive test result is a false positive becomes quite high. This is often the case in general practice." (TTA6)

"...one might be tempted to choose test A as a better confirmatory test because of its higher specificity. However test B yields the greater number of correct diagnoses in a population where the pre-test likelihood of disease is high." (TTA5).

There was a suggestion that the 2x2 diagnostic contingency table may mislead with respect to the portability of test accuracy metrics because it allows the derivation of both PVs and metrics not mathematically dependent on prevalence:

"The use of the 2x2 table to teach these concepts (sensitivity, specificity, PPV and NPV) also frequently creates the erroneous impression that the positive and NPVs calculated from such tables could be generalised to other populations without regard being paid to different disease prevalence." (TTA2)

"One of the potentially confusing aspects of predictive values is that it seems to be determined by simple calculations with the 2x2 box, similar to sensitivity and specificity and therefore it may be misconstrued as a characteristic of the test itself." (TTA4)

The only novel measure of test accuracy presented in the theoretical literature: the "Mutual index as an Index of Diagnostic Test Performance", was in the context of quantifying the setting-specific information provided by diagnostic test:

"Because I (D;R) (mutual information) is dependent on pre-test probabilities, knowledge of the setting in which a diagnostic test is employed is a necessary condition for quantifying the amount of information it provides." (TTA3)

2.4.1.4 The language of test accuracy

The non linearity of the likelihood ratio scale was suggested as a potential source of confusion about the magnitude of change in probability they convey:

"Likelihood ratios are not linear... This has important implications. Intuitively the clinician rates the discriminative strength of the LR+ of 100 as ten times that of an LR+ or 10, which is an overestimation" (TTA7)

" ..the counterintuitive scale of likelihood ratios. Why is a test with a likelihood ratio of 100 not 10 times more powerful than a test with a likelihood ratio of 10" (TTA33)

The observation that test accuracy descriptor terms often do not represent how test accuracy measures are to be applied was suggested as a source of confusion concerning their interpretation and application:

"Referring to the 'predictive value of a test' gives the false impression that a test's predictive power stands alone and can be applied to any patient... the predictive value is a characteristic of a test result in a specific patient, not of the test result in general, nor of the test itself." (TTA4)

"... the absence of an appropriate language for clinical logic. Instead of indicating what it means for clinicians, the word 'likelihood ratio' states where it comes from... For years we have tried in vain to introduce (likelihood ratios) in clinical teaching in four continents... Never in 20 years of teaching clinical logic, have we found a clinician who used the word "positive likelihood ratio"." (TTA33)

" Although interest in evidence based medicine has increased in recent years, and is taught in most medical schools, evidence based strategies have been adopted inconsistently into routine care.... it is worth considering whether the manner in which we verbally communicate these ideas (about diagnostic testing) may represent a fundamental (yet reparable) hindrance to diagnostic reasoning." (TTA4).

In addition there was considerable discussion about the importance of using test accuracy measures that encompass all information about the discriminative power of a test to inform diagnostic decision making:

"LRs summarise the information of both sensitivity and specificity and give the discriminative power of a test." (TTA7)

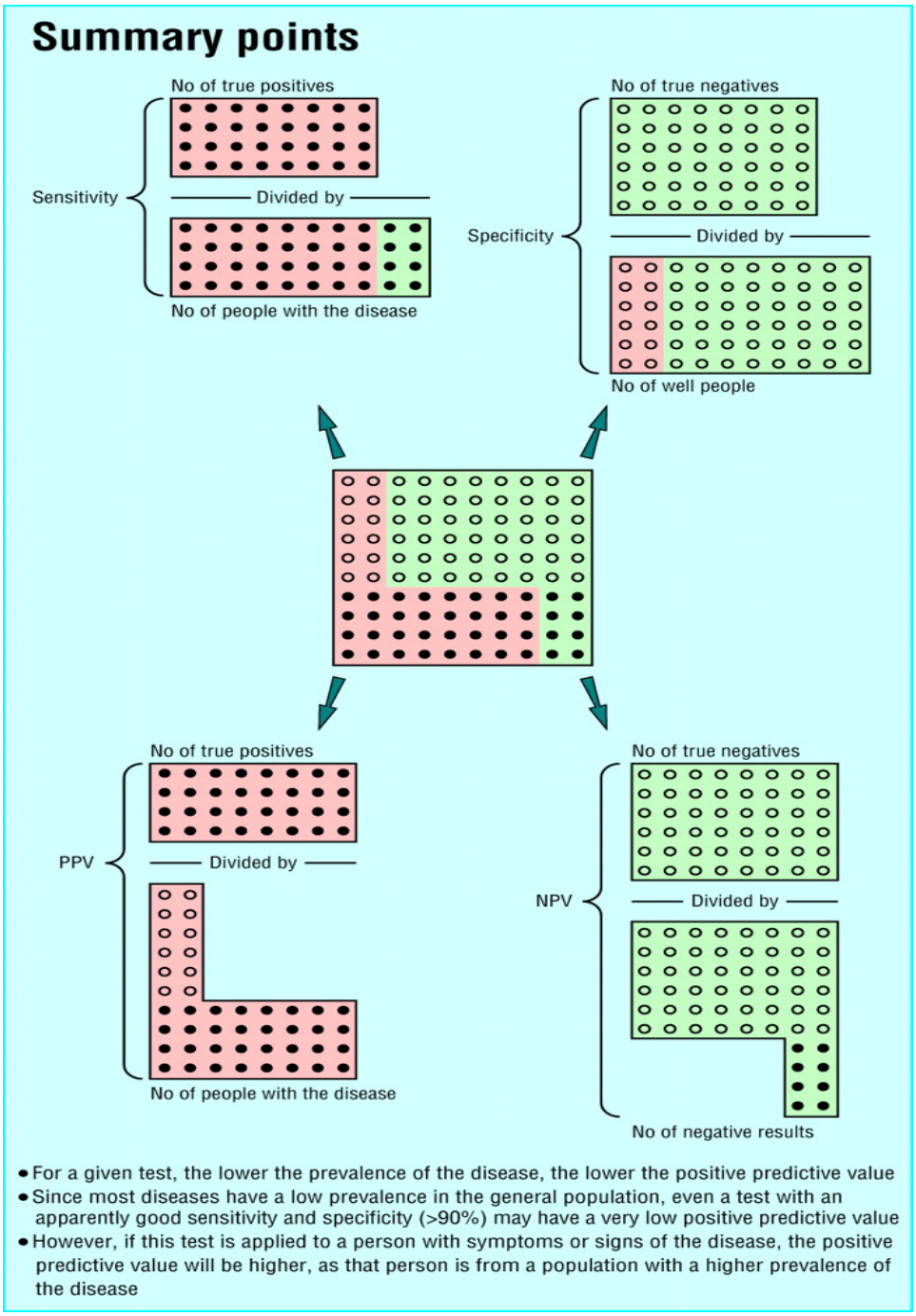
"...(it is) a common misperception that sensitivity and specificity can be considered in isolation." (TTA4)

"Reliance on sensitivity and specificity frequently leads to exaggeration of the benefits of tests." (TTA14)

Methods for simultaneously presenting both dimensions of test accuracy were suggested to assist with the interpretation and application of test accuracy information. Dot graphics were presented as examples to illustrate the relationship between the two dimensions of test accuracy conveyed by sensitivity and specificity, positive predictive value (PPV) and negative predictive value (NPV) and positive likelihood ratio (LR+) and negative likelihood ratio (LR-).

"What follows are diagrams that were useful for me in attempting to visualise sensitivity, specificity and their cousins' positive predictive value and negative predictive value." [TTA21] (figure 2.6)

Fig 2.6: Dot graphic illustrating the two dimensions of test accuracy

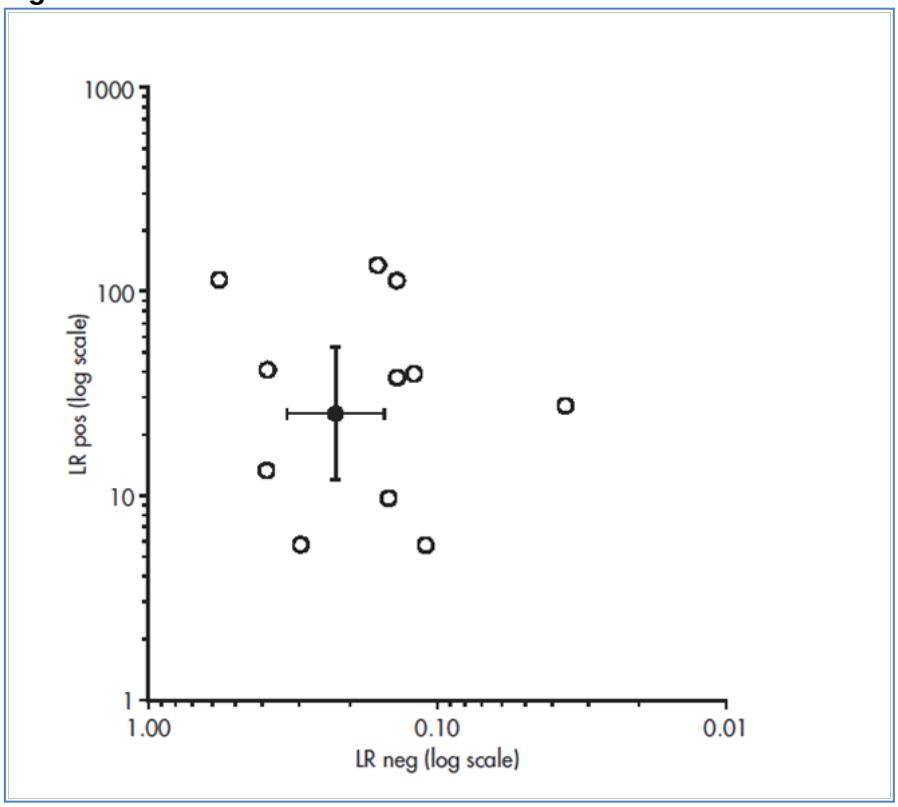


[TTA21] *(Loong, BMJ 2003; 327:716-718)*

Similarly a likelihood scatter plot was proposed as a method for representing both dimensions of test accuracy:

"Our objectives were to develop a clearly arranged graphical presentation of the results from individual diagnostic studies... We hypothesised that this method of graphical presentation could be easily interpreted, especially by readers already used to the "classic" forest plots of therapeutic meta-analyses.
The matrix presentation enables a quick visual impression of the strengths and the weaknesses of a diagnostic test in either direction." (figure 2.7) [(TTA31)]

Fig 2.7: Likelihood Ratio Scatter Plot



Likelihood ratio scatterplot matrix meta-analysis. Unfilled circles represent individual studies. The filled circle shows the weighted summary likelihood ratios (random-effects model). Error bars represent 95% confidence intervals.

[(TTA31)]*(Stengel et al, J Med Screen 2003;10:47–51)*

2.4.1.5 Facilitating decisions about test use and testing policy

Discussion of the strengths and weaknesses of test accuracy measures in the theoretical literature reviewed was almost exclusively from the perspective of the bedside rather than the perspective of the development of testing policy.

There was recognition that the prevailing architecture of test accuracy research was concerned with evaluation of tests in isolation and not their role or contribution to testing pathways:

"Test research merely quantifies the 'characteristics' of a test rather than the test's contribution to estimate the diagnostic probability of disease presence or absence. By 'diagnostic' research we refer to studies that aim to quantify a test's added contribution beyond the test results readily available....Moreover the focus (of test research) is on the value of a single test rather than on the value of that test in combination with other, previous tests." (TTA24).

The existence of a scale with which to judge the clinical utility of a test's accuracy was viewed as an advantage of LRs:

"By convention, marked changes in prior disease probability can be assumed in positive likelihood ratios exceeding 10.0 and negative likelihood ratios below 0.1....since no threshold values of sensitivity or specificity are available that would allow either the adoption or the rejection of the routine application of a diagnostic procedure likelihood ratios appear as preferable indices of test performance, at least in the setting of clinical decision making." (TTA31)

Global test accuracy measures were suggested to facilitate comparisons between tests to inform decisions about test use:

"To evaluate the performance of a diagnostic test...Our goal is to find a number that summarises the performance of the diagnostic test." (TTA3)

"This aspect (decision thresholds) is best examined by means of analysis of receiver operating curves,,,such analyses are provided in reports on diagnostic tests and allow the comparisons of one test with another." (TTA16)

However there was also recognition that consideration of both dimensions of test accuracy was important for decisions about test use. Test accuracy measures that did not communicate information separately on the two dimensions of test accuracy were seen as inadequate in isolation for diagnostic decision making:

"Although (LRs) together contain all the information given by sensitivity and specificity and are sufficient for most clinical decisions, sensitivity and specificity are still necessary when false positives or false negatives have to be avoided as much as possible. The same LR+ can be the result of the combination of very different values for sensitivity and specificity." (TTA7)

Similarly, several authors noted that global measures do not communicate information about the two dimensions of test accuracy:

"The problem that occurs in a meta-analysis of diagnostic studies is the multi-directional performance of the diagnostic instrument regarding its ability to detect (specificity) or exclude (sensitivity) the characteristic of interest. Multi-dimensional outcomes cannot be summarised well by a single estimate." (TTA31)

"(Test accuracy) condenses two fundamental test variables- sensitivity and specificity-which apply to diseased and non diseased populations, respectively. So what is the point of merging these populations when all our efforts are directed at distinguishing between them?" (TTA16)

2.4.1.6 Pre-test probability estimation

Pre-test probability was either generically conceptualised as the prevalence of disease reflected by healthcare setting or as a combination of prevalence estimated form healthcare setting and the results of clinical history and examination: the point at which tests, other than history and clinical examination, were being considered.

It was suggested that considering the results of clinical history and examination as contributing to pre-test probability estimation rather than considering the accuracy of individual components might reduce inappropriate testing, particularly in primary care where disease prevalence is low.

"Dismissing (a test) for its low likelihood ratio risks setting clinicians on a slippery slope towards clinical impotence. If we pursued this reasoning...many, perhaps most, other questions of examinations might also prove minimally useful. But this conclusion follows only by considering each test in isolation. Instead, suppose we applied the arbitrary minimally useful positive likelihood ratio of 2 to each of the above 16 tests. If all returned positive the aggregate likelihood ratio could reach 65 356." (TTA12)

"*Unfortunately for the primary care clinician the reduced magnitude of the prior odds is compounded by the finding that many items in the medical history have positive likelihood ratios that are too small to be clinically useful. For a number of inexperienced doctors this may perhaps account for some of their tendency to order an increasing number of diagnostic tests."* (TTA32)

"*There are particular diagnostic challenges for GPs in primary care: the pre-test probability of disease is lower."* (TTA22)

For this reason the potentially important contribution of clinical history and examination was emphasised:

"When diagnosis is viewed as a processing pathway founded on a robust medical history, it becomes clear that in some situations investigations may become unnecessary." (TTA32)

The importance of pre-test probability estimation for diagnostic decision making and an appreciation of its variation with healthcare setting were highlighted:

"One of the specific skills of a general practitioner is to understand the pre-test probabilities of disease in his or her clinical setting...The difference in pre-test probabilities between primary and secondary care is one reason why clinicians find it difficult to move between test settings." (TTA6).

"Pre-test probabilities for the same target disorder can vary widely among and within countries and among primary, secondary and tertiary care settings." (TTA27)

The impact of pre-test probability on test error rates was also highlighted:

"When the prevalence of disease is low, the probability that a positive test is a false positive becomes quite high. This is often the case in general practice..." (TTA6).

The literature acknowledged that clinicians are inaccurate in their pre-test probability estimation:

"Some say clinicians can generate post-test probabilities on the basis of clinical experience...Yet research has shown that that clinicians' estimates of probability vary widely and are often inaccurate....by itself, clinical experience appears insufficient to guide accurate probability estimation." (TTA26)

Also that pre-test probability estimation was often qualitative rather than quantitative:

"We are not arguing that the Bayesian approach is a perfect means of reaching a diagnosis. Admittedly, the definition of pre-test odds of a disease for a given patient is inherently subjective." (TTA12)

Concern was expressed that undergraduate medical training and test accuracy research has a secondary and tertiary care rather than a primary care focus and this was suggested to contribute to difficulty in pre-test probability estimation and the application of test accuracy estimates in generalist settings*:*

"...most of the existing data on pre-test probabilities were obtained from tertiary care populations and may not be generalisable to primary care populations.... This highlights the need for research on pre-test probabilities in primary care settings." (TTA28)

"Attention must be paid to the differences in clinical stages encountered in general practice and specialist practice. It is concluded that a large part of the diagnostic field of general practice has still to be discovered and developed." (TTA20)

"Frontline clinicians are gaining increasing access to high quality evidence about diagnostic tests....Using this evidence requires more than knowing a test's discriminatory power. Clinicians also need to estimate pre-test probabilities for the disorders being considered. But where do these pre-test probabilities come from?" (TTA26)

Several approaches were proposed to assist clinicians with pre-test probability estimation including compilation of a catalogue of setting-specific LRs for individual components of the clinical history and examination and computer based decision support:

"The medical history is more than a nostalgic relic of little relevance to modern practice. However there is now an urgent requirement for careful consideration to be given to how, where and for whom this diagnostic technology is being used. A key component of this initiative would be the development of a library of setting-specific likelihood ratios and kappa for individual feature and combinations of items of medical history and information." (TTA32)

"Computer based decision support can be linked to the electronic health record...such a system could also tell the physician how to estimate the pre-test probability." (TTA30)

"...the accurate recall and implementation of (decision) rules can be facilitated by computer based decision clinical decision support systems that quantify diagnostic and prognostic information and provide clinicians with patient specific recommendations." (TTA8)

2.4.2 Results: Non-empirical test accuracy literature: Fully rational

The impact of motivational biases on testing behaviour received limited attention in the literature identified here. Testing was portrayed as a risk averse behaviour and it is proposed that health professionals may not be aware of the limitations of testing in reducing uncertainty or in estimating the degree to which testing reduces uncertainty.

It was suggested that the observed increase in testing may be due to a 'stubborn quest for diagnostic certainty' and a manifestation of risk aversive behaviour:

"...some physicians order all the tests that may be even remotely applicable in a given clinical situation. Such a practice may comfort the patient and enhance the physician's belief that all diagnostic avenues have been pursued, but more tests do not necessarily produce more

certainty...Despite the limitations of our diagnostic procedures we continue to test excessively, partly because of our discomfort with uncertainty." (TTA18)

Further it was suggested that risk averse behaviour manifest by excessive testing may be a symptom of difficulties with probabilistic reasoning:

"...why are clinicians uneasy with uncertainty? (Clinicians) have been taught to think categorically. When (clinicians) try to think in terms of probabilities, (they) often falter. (Clinicians) disregard uncertainty or behave as if it doesn't exist; use inexact expressions such as 'probable, 'occasional' and 'likely'". (TTA18)

Healthcare setting, the utility associated with different test outcomes and variation between individuals was proposed to be an important modifier of attitudes to uncertainty. Attitudes to risk and the size and significance of uncertainty were suggested to be context-dependent:

"... feelings of uncertainty regarding medical problems can differ depending on the situation, not only because one physician may be faced with more complicated diagnostic puzzles than the other, but also, and primarily because the consequences of a vague and uncertain diagnosis may vary in each situation." (TTA34)

"(No-treatment- no test and test-treatment) thresholds vary amongst diseases and individual patients." (TTA27)

It is suggested that uncertainty may be larger in generalist settings, whilst the consequences of an uncertain diagnosis may be greater in specialist settings:

"...further research should take into account the fundamental differences between the significance of uncertainty and risk taking in the minds of GPs and specialists." (TTA34).

"GPs generally deal with far greater diagnostic uncertainty than their hospital based colleagues, this being part of the gate keeping role of primary care." (TTA22)

Further personality and cultural differences were suggested as important modifiers of attitudes to risk:

"Some physicians find it difficult to tolerate any diagnostic uncertainty...The cultural differences between physicians in American hospitals and Dutch GPs are considerable...there is a difference in opinion regarding the use of the laboratory between American and European physicians." (TTA34)

2.4.3 Results: Non-empirical test accuracy literature: able to compute accurately

The emphasis in the literature was that diagnostic problem solving should be based on quantitative probability revision rather than being a more qualitative process. However there was an appreciation of the complexity of probability revision and a suggestion that the process required for evidence based diagnostic decision making was less familiar to clinicians compared to other aspects of evidence based medicine:

"We (clinicians) use probabilistic reasoning intuitively whenever we consider the likelihood of a patient having a disease in the light of new information." (TTA6)

"...clinicians apply Bayesian reasoning in framing and revising different diagnoses without necessarily undergoing, or requiring, any formal training in Bayesian statistics... Bayesian reasoning is a natural part of clinical decision making... Bayesian approaches are a powerful and intuitive approach to the differential diagnosis" (TTA12).

"Many trainees appreciate the concepts of sensitivity and specificity and learn how to combine the "art" of the history and physical examination with the "science" of diagnostic testing without explicit use of quantitative probability theory. Nevertheless it seems that quantitative reasoning is neither intuitive nor well understood." (TTA4)

"Choosing the appropriate test.... for a particular diagnostic setting is often difficult for medical decision makers... Various schemes including nomograms based on Bayes' theorem, probability ratios, receiver operating curves and formal decision analysis have been used to compare the performances of various tests available in a given setting. These methods are often cumbersome, limited to a single disease prevalence and not intuitive to interpret." (TTA13)

"Despite general awareness of the other concepts of evidence based medicine, the estimation of pre-test probability and adjustment of disease probability in the setting of thresholds for testing and treating is not commonplace." (TTA4)

"...(in undertaking probability revision using Bayes' theorem) the physician violates the statistical requirement that the tests operate independently....his reflects the reality that there is some redundancy in our clinical evaluations." (TTA12)

It was suggested that improving the use of quantitative probabilistic reasoning ability would improve diagnostic decision making:

"We (clinicians) should not be satisfied with descriptions of probabilities that are vague, subject to varying interpretations and not adaptable to calculations. Instead we should be more quantitative and teach how to combine numerical representations of probabilities and risks....We (clinicians) shun probability-orientated thinking (and are) taught to think categorically...(the consequences of which are that clinicians) judge the likelihood of diseases erroneously and combine data on probabilities inaccurately...Our shunning of probability-orientated thinking is reflected in our textbooks, which are rife with absolutes." (TTA18).

2.4.3.1 Facilitation of the process of probabilistic reasoning

Several examples of strategies to assist with the quantitative probabilistic reasoning process were suggested including the use of clinical prediction models for integrating the results of multiple tests:

"...a Bayesian approach to diagnosis: estimating a clinically likely pre-test probability for a target disorder, then applying a likelihood ratio derived from the presence or absence of the clinical features of the rule (similar to applying a test result), which in turn enables a revised estimate of clinical probability." (TTA8)

LRs were suggested to simplify the conversion from pre to post-test probability in comparison to the use of sensitivity and specificity, mostly with reference to the use of Fagan's nomogram (figure 2.8), a graphic tool to simplify this conversion:

"Bayesian approaches are a powerful and intuitive approach to the differential diagnosis...the pre-test odds of a hypothesis being true multiplied by the weight of the new evidence (likelihood ratio) generates post test odds of the hypothesis being true." (TTA12)

"The probability of disease given a positive or negative test result (post-test probability) is usually obtained by calculating the likelihood ratio of the test result and using formulas based on Bayes' theorem or a nomogram, to convert the estimated (pre-test probability) into a post-test probability which takes the (test) result into account." (TTA25)

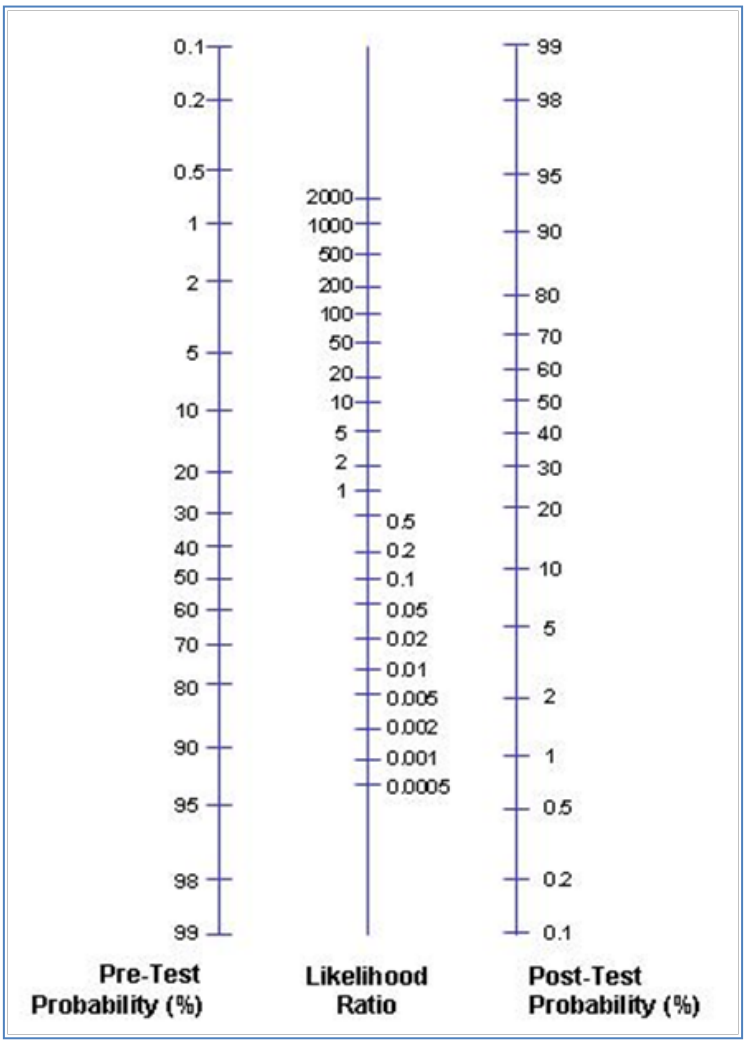
"*Calculation of positive predictive value and negative predictive value with sensitivity and specificity is quite time consuming...Likelihood ratios...are intuitive; they simplify the predictive value calculation and the overall evaluation of sequential testing."* (TTA7)

"If a disease's pre-test probability is known or can be estimated, likelihood ratios allow for direct calculation of post-test probability using a formula that can easily be derived from Bayes' theorem. This is the major advantage of likelihood ratios, and gives it superiority over predictive values for given prevalence, which have to be calculated by the rather complicated Bayes' theorem." (TTA7)

"The Fagan's nomogram is a useful and convenient graphical tool that allows likelihood ratios to be used in conjunction with a patient's pre-test probability of disease to estimate the post-test probability of disease." (TTA1)

"...there's an easier way to manipulate all these probabilities↔odds calculations and a nomogram for doing so." (figure 2.8) (TTA27)

Fig 2.8: Likelihood ratio nomogram



A minority of authors did not agree that pre to post-test probability revision was simplified using LRs:

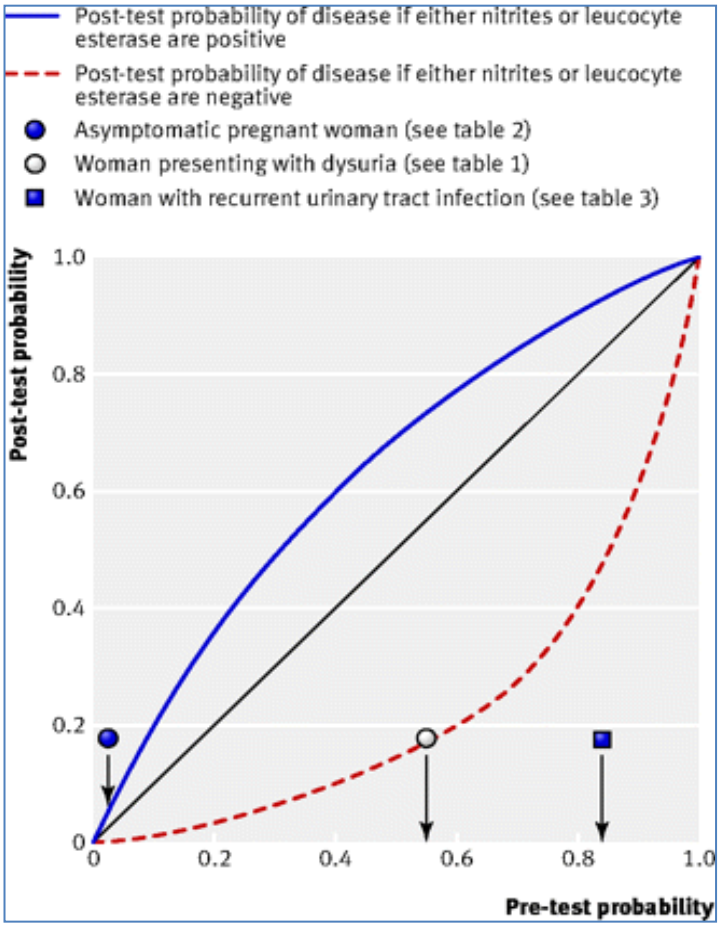
"the complex chain of calculations involved. Clinicians should transform probabilities into odds, multiply by a series of likelihood ratios, and finally reconvert odds to probabilities." (TTA33)

"The need to convert back and forth between pre-test probability/pretest odds and post-test odds/predictive value may be confusing, but fortunately Fagan's nomogram obviates all the calculations." (TTA7)

A graphical illustration of post-test probabilities derived using a single estimate of test accuracy over a range of pre-test probabilities was promoted as a means of facilitating decisions about test use in different populations, defined according to the prevalence of disease (figure 2.9) (TTA5;TTA6;TTA29).

"The graphic analysis of numerically calculated predictive values that we describe...provides a simple, reliable method for comparing the predictive values of available test options at disease prevalence relevant to the use of the proposed test." (TTA5)

Fig 2.9: Graphical illustration of pre to post-test probability



(TTA6) *Doust J, BMJ 2009; 339:1080-1083*

2.4.3.2 Test accuracy presentation format as a means of facilitating probabilistic reasoning

The contribution of non-health professionals to the theoretical literature reviewed was concerned with representation of test accuracy information.

Rather than the belief that errors in probability revision result from inherent limitations in our *"cognitive processes"* (ability to perform the arithmetic required for probability revision) or motivational biases *"passion and desire",* (TTA9; TTA17) frequentists propose that errors in probability revision arise as a result of how probabilistic information is presented:

"In the 1990s, intuitive Bayesian reasoning began to be seen in a new light, that is, from an ecological angle....one can facilitate reasoning from the outside by changing the external representation from probabilities and relative or normalised frequencies, to natural frequencies" (TTA17)

The premise of the frequentist philosophy is that natural frequency representations facilitate probabilistic reasoning by mimicking natural sampling, removing reference class confusion as a result of use of a single reference class and removing the need to incorporate base rates (pre-test probability) in calculations.

Natural sampling

Natural frequency representations of probabilistic information (see table 2.10) facilitate probabilistic reasoning by mimicking natural sampling (acquisition of data by direct experience):

"...organisms did not acquire information in terms of probabilities and percentages until very recently. We assume that as humans evolved, the natural format was frequencies as actually experienced in a series of events...the sequential acquisition of information by updating frequencies." (TTA9)

"...because information was experienced during most of the existence of Homo sapiens in terms of discrete cases, for example three out of 20 cases rather than 15%." (TTA10)

Table 2.10: Comparison between natural frequency, normalised frequency and probabilistic expression and equivalent test accuracy expression (adapted from Gigerenzer 1995;1996[TTA9; TTA10])

Natural Frequency Expression	Normalised (relative) frequency expression	Probabilistic expression (% / decimal)	Test accuracy expression
In a population of 100, 10 individuals will have disease X and 90 will be unaffected by disease.	In a population of 100, 10 individuals will have disease X and 90 will be unaffected by disease.	The prevalence of disease is 10% (0.1).	Pre-test probability.
Of the 10 individuals with disease, 8 will test positive with test A.	Of every 100 individuals with disease 80 will test positive with test A.	The probability of testing positive with test A if you have disease X is 80% (0.8).	The true positive rate (sensitivity).
Of the 90 individuals without disease, 80 will test negative with test A but 10 will test positive.	Of every 100 individuals without disease, 89 will test negative. *AND*	The probability of testing negative with test A if you do not have disease X is 89% (0.89). *AND*	The true negative rate (specificity).
	Of every 100 individuals without disease 11 will test positive.	The probability of testing positive with test A even if you do not have disease is 11% (0.11).	The false positive rate (1-sensitivity).
How many patients who test positive will have disease?	How many patients who test positive will have disease?	What is the probability of having disease X if you test positive with test A?	Positive predictive value or post test probability given a +ve test result.
Answer: 8/ (8+10) = 8/18.	Answer: $\frac{(80/100) \times (10/100)}{((80/100) \times (10/100)) + ((11/100) \times 90/100))}$	Answer: $\frac{(0.8) \times (0.1)}{(0.8 \times 0.1)+(0.11 \times 0.9)}$	

Notes to table 2.10: Normalised frequencies: expression in relation to a constant (normalised) denominator and do not carry inherent information about the base rate (prevalence) in contrast to natural frequencies.

Reference to a single reference class

In a natural frequency format, all frequencies explicitly refer to the same reference class and the base rate (disease prevalence) can be ignored, whereas conditional probabilities refer to more than one reference class which may cause confusion:

"Conditional probabilities such as sensitivity and specificity refer to different classes (the class of people with and without illness respectively), which makes their mental combination difficult." (TTA11)

One suggested manifestation of difficulties in the manipulation of more than one reference class was confusion between the concepts of sensitivity and PPV and specificity and NPV:

"'If a woman has breast cancer the probability that she will have a positive result on mammography is 90%'. This statement is often confused with: 'If a woman has a positive result on mammography the probability that she has breast cancer is 90%'". (TTA11).

2.4.3.3 Systematic errors in probability revision

There was recognition that systematic errors are introduced into probability revision by heuristics (cognitive short-cuts used for complex problems) (also see box 1.9):

"Psychologists have shown that rapid decision making is aided by heuristics – strategies that provide shortcuts to quick decisions- but they have also noted that these heuristics frequently mislead us. Good decision making is further impeded by the fact that we often fall prey to various cognitive biases...Even worse it is common for people who are particularly prone to cognitive biases to believe that they are good decision makers. The greatest obstacle to making correct decisions is seldom insufficient time but distortions and biases in the way information is gathered and assimilated." (TTA19).

"To use Bayes' theorem wisely, one must be aware of pitfalls in estimating probability." (TTA29)

"...clinicians' memories are fallible and their thinking is prone to numerous biases." (TTA26)

Heuristics that were described include base rate (pre-test probability) neglect:

"...it is important to be aware of base rates of the occurrence of a particular condition and to avoid giving too much weigh to one piece of information." (TTA19)

Placing undue weight on diagnoses that come easily to mind, are significant or unusual (the availability heuristic):

"They (doctors) should ask if their decision is influenced by any salient pieces of information and, if so, whether these pieces of information are truly representative or simply reflect recent or otherwise memorable experiences." (TTA19)

The tendency to pursue, and take notice of information that fits with a pre-existing expectation about the correct diagnosis (confirmatory bias or anchoring):

"In taking medical histories, doctors often ask questions that solicit information confirming early judgements. Even worse, they may stop asking questions because they reach an early conclusion, thus failing to unearth key data." (TTA19)

"Many experiments have shown that clinicians do not adjust their initial estimate enough to take account of new information." (TTA29)

In addition it was suggested that clinicians often have a misplaced confidence in their diagnostic ability:

"Research has shown that almost all of us are more confident about our judgements than we should be. Since medical diagnoses typically involve some uncertainty, we know that almost all doctors make more mistakes in diagnosis than they think they do." (TTA19)

2.4.3.4 The place of probability revision in clinical practice

Accompanying the debate concerning the intuitive nature and complexity of probabilistic reasoning there was an appreciation that formal probability revision may not be commonplace in clinical practice:

"Despite their usefulness in interpretation of clinical findings, laboratory tests and imaging studies, likelihood ratios are little used. Most doctors are unfamiliar with such ratios and few use them in practice." (TTA14)

"Simplifying aids such as the Fagan nomogram are rarely used, should be done for every test and need published likelihood ratios." (TTA33).

Although the emphasis of the majority of literature was quantitative probabilistic reasoning, informal, semi-quantitative estimation was proposed as an acceptable alternative to formal quantitative probabilistic reasoning by several authors:

"(Clinicians) need to have a sense of both the pre-test probability of disease and the diagnostic accuracy of test results. We do not need to be able to do these calculations exactly." (TTA6).

"Whilst we rarely know what the sensitivities, specificities or likelihood ratios are for these tests. At best clinicians carry a general impression about their usefulness." (TTA12)

Similarly Gorry (1978) (TTA13) suggests ranking the relative probability of false negative test results for each competing diagnosis to arrive at the most probable diagnosis:

"...if properly interpreted, the normal value may help to differentiate among diagnoses that yield normal results with different frequencies." (TTA13)

Gigerenzer (1996) (TTA10) proposes the use of satisficing algorithms as an alternative to formal probability revision in situations where application of Bayes' theorem and logistic regression for the integration of information about sequential, often interdependent tests becomes *"mathematically complex and computationally intractable – at least for the human mind."*:

"...simple algorithms that exploit the structure of information to make good inferences under constraints of limited time and knowledge...non linear, non compensatory and work with the principle we call 'one good reasoning', that is, they base inference on only one predictor (the first identified as discriminating between two or more differential diagnoses) as opposed to an integration of several (predictors)." (TTA10)

2.5 Results: Empirical test accuracy literature

For results tabulated by study see appendix 2.2

Characteristics of included studies

Twenty six papers reporting 27 studies were concerned with the understanding and application of test accuracy in health professionals or medical students whilst seven papers reporting 16 studies were concerned with understanding and application in non-health professionals. The literature spanned 1978 to 2010, although 2/3rds of papers were published after 1995. This may be a reflection of the emergence of the evidence based medicine movement in the early 1990's and the promotion of the integration of quantitative external evidence into medical decision making [80].

The majority of health professional samples were self-selected, convenience samples from medical education courses or sampling methods were unclear or not reported. The exceptions were one study based on a random sample from a professional register, obtaining a 91% response rate [(ETA29)] and one study carried out as part of a medical undergraduate OSCE examination [(ETA31)]. It is therefore likely that the review findings represent more motivated practitioners. Eleven studies comprised secondary care clinician samples, four comprised primary care clinician samples (one conducted in the setting of an army clinic), five comprised a mixture of primary and secondary care clinicians, two were restricted to medical undergraduates and two study samples were a mix of health and non-health professionals. In one study the setting was unclear. Of the studies conducted on primary and / or secondary care clinician samples seven also included medical undergraduates and five also included professions allied to medicine.

Most studies of non-health professionals were undertaken on students (13/16), one on a patient sample [(ETA17)], one on women attending for screening [(ETA1)] and one on females from the general population [(ETA10)]. One study on women eligible for screening achieved an 85%

response rate [ETA1]. Other samples were either self-selected, incentivised or the method of sampling was unclear.

Three health professional studies were conducted in the UK [ETA4; ETA18; ETA30] and six in the rest of Europe. The majority of research was undertaken outside of Europe: USA (N=16), Japan (N=1), Israel (N=1) where medical education may differ. In addition due to differences in healthcare organisation, particularly the absence of a primary care gate-keeping system, the spectrum of patients, prevalence of disorders and testing culture [TTA34] will not be generalisable to the UK or the rest of Europe. Of the 16 non-health care professional studies, ten were undertaken in the USA, one in Australia four were undertaken in the UK and one in France.

Only six of twenty seven of the health professional studies evaluated self reported use of test accuracy information in practice [ETA27; ETA7; ETA29; ETA21; ETA34; ETA36], three undertaken in the USA, one in Belgium and two in the Netherlands. All other studies investigated the application of test accuracy metrics to hypothetical scenarios which were self-administered in the form of questionnaires. All non-health professional studies assessed the use of test accuracy using hypothetical scenarios.

Study Designs and Quality

Five of the health professional studies were randomised controlled trials (RCTs), three of which were considered to be of good quality [ETA4; ETA28; ETA32] although likely to be underpowered. Four health professional studies were controlled trials [ETA3; ETA9; ETA12; ETA27]. Only one within-subject comparison with N of 6 [ETA12] was reported adequately enough to allow an assessment of moderate quality. Of the remaining 18 health professional studies, 15 were modifications of cross sectional design of which two were supplemented by qualitative interviews [ETA29; ETA30]. One study was a cohort study [ETA21] and one study employed both cross sectional and cohort study designs [TTA34]. One study was qualitative using covert observation [ETA14].

One non-health professional study was an RCT [ETA1] which was judged to be of good quality. Eight studies were controlled trials [ETA8] for which poor reporting precluded assessment of study quality. One study was a face-face qualitative interview study of moderate quality [ETA10]. The remaining five non-health professional studies were cross sectional in design.

2.5.1 Results: Empirical test accuracy literature: Fully informed

2.5.1.1 Test accuracy information

Thirteen studies were concerned with test accuracy measures; 10 in health professionals [ETA2; ETA9; ETA14; ETA18; ETA22; ETA25; ETA29; ETA30; ETA33; ETA34] and three in non-health professionals [ETA1; ETA10; ETA17]. Three studies were concerned with the relationship between sensitivity, specificity and false negatives and false positive rates; seven with sensitivity and specificity; two with predictive values (PVs); three with test errors, two with likelihood ratios (LRs) and one with Receiver Operator Characteristic (ROC) curves.

Comprehension: definition of test accuracy metrics

Whilst some studies reported the majority of respondents were familiar with definitions of sensitivity and specificity (75-98%) [ETA9; ETA14; ETA33] and PVs (61%) [ETA33] other studies provided evidence of confusion of sensitivity with positive predictive value (PPV), specificity with negative predictive value (NPV), false positive rate with 1- PPV and false negative rate with NPV [ETA2; ETA29]. Similarly, although Gigerenzer (1998) [ETA14] observed a high level of recognition of the link between sensitivity and false negative test results (75% of respondents), only 25% of respondents recognised the relationship between specificity and false positives.

One study compared respondents' familiarity with diagnostic and effectiveness metrics and found performance was worse for diagnostic metrics [ETA2].

Comprehension: estimating the accuracy of named tests

Those authors attempting to elicit estimates of the sensitivity and specificity of named tests generally observed high levels of error [ETA25; ETA22; ETA17]. Accuracy was neither consistently over nor underestimated. Interestingly Noguchi (2002) [ETA25] observed a large discrepancy between estimates of a test's sensitivity (3% underestimate) compared to its specificity (21% underestimate). Reid (1998) [ETA29] demonstrated that estimates of test accuracy were based largely on clinicians' own clinical experience of test use, rather than published estimates. The one study eliciting non-health professional estimates of test accuracy for named tests [ETA17] demonstrated clustering of estimates and suggested this was a reflection of a lack of understanding of test accuracy properties and their application.

Preference: test accuracy metric

Four authors investigated reported health professionals' use of test accuracy measures in practice. ROC curves and LRs were reported to be used by <1% of clinicians and sensitivity and specificity by < 4% [ETA29; ETA34]. By contrast PVs were reported to be used by 80% of clinicians in one study, partly on the basis that these measures were intuitive for quantifying test errors and provided a direct estimation of post-test probability without the need for complicated calculations [ETA29] (see 2.5.3.2 below). However, it is important to note that in this study, respondents' self reported use referred to the way they conceptualised the performance of a test based on their own experience of using it rather than use of published estimates of PVs to inform their testing practice. In fact these respondents confused the definition of sensitivity and PPV and specificity and NPV [ETA29] suggesting that they did not rely on published test accuracy estimates. Other authors have demonstrated that information about test errors are a prominent part of decisions concerning test use [ETA30; ETA18].

Preference: presentation format

One non-health professional study investigated the preference of a female population sample for information about post-test probability when receiving information about the results of mammography (ETA10). Information presented as normalised frequencies or percentages was perceived as being about 'other' people. There was no clear preference for verbal or graphic representations. Verbal descriptions of post-test probability were suggested as helpful accompaniments to numerical representations, although wide variability (10-90%) in numerical definitions of verbal probabilities was observed.

Behaviour

The one study investigating the use of test accuracy information in non health professionals demonstrated that providing information about test errors and mortality and morbidity risk resulted in a reduction in the number of women intending to attend for screening (ETA1) although it not possible to distinguish the effects of test accuracy and risk information on intended behaviour.

2.5.1.2 Pre-test probability

Twelve health professional studies (ETA5; ETA11; ETA12 ; ETA14; ETA18; ETA21; ETA22; ETA25-ETA28;ETA34;) and two non-health professional studies (ETA17; ETA10) were concerned with estimation or use of pre-test probability.

Comprehension: defining pre-test probability

Two health professional studies (ETA12; ETA14) conceptualised pre-test probability as the prevalence of disease in the presenting population, four studies as the probability of disease following clinical history (ETA11; ETA18; ETA25; ETA33) and six studies as the probability of disease following clinical history and examination (ETA5; ETA21; ETA22; ETA26;-ETA28). In one study (ETA34) it was unclear what information was used to estimate pre-test probability.

Comprehension: pre-test probability estimation

The majority of studies concerned with pre-test probability (n=11) were concerned with the accuracy of pre-test probability estimation by health professionals for a range of diseases. Between-person variation in quantitative, pre-test probability estimation for a common scenario was considerable (75-100%) in three studies (ETA11; ETA14; ETA18) and 20-25% in two studies (ETA26; ETA28). Overestimation rather than underestimation was a feature of studies (ETA5; ETA25; ETA27; ETA18) with overestimation of atypical or severe and less probable diagnoses (the availability heuristic) reported by three studies (ETA5; ETA11; ETA25). An educational intervention designed to improve the accuracy of pre-test probability estimation was effective in reducing overestimation but had no effect on subsequent test use (ETA27).

One non-health professional study demonstrated a clustering of quantitative estimates (37-50%), regardless of disease (ETA17).

Three studies concerned with contextual modifiers of pre-test probability estimation all demonstrated appropriate directional adjustment by clinicians (ETA21; ETA22; ETA33).

Preference

One health professional study investigated the preferred presentation format for pre-test probability information (ETA34). The majority of health professionals studied used verbal, categorical descriptions of pre-test probability (52%), followed by frequentist and percentage expressions. However 76% stated that they did not find pre-test probability estimation useful for diagnostic decision making.

One qualitative study in non-health professionals did not demonstrate a preference for the presentation of pre-test probability in educational materials to encourage patient involvement in decisions about screening (ETA10).

2.5.2 Results: Empirical test accuracy literature: Fully rational

Only two studies were concerned with the investigation of motivational biases on testing behaviour [ETA21; ETA36]. Both studies were observations of actual practice in primary care settings.

In the context of investigating primary care clinicians' approaches to test use and interpretation in low prevalence settings, Houben (2010) [ETA21] observed that the emphasis of testing in primary care was to rule out disease and that this was most often done to reassure the clinician (62% of tests ordered), followed by reassurance for the patient (20% of tests ordered). Only 19% of tests were performed to confirm suspected disease. Only 9% of abnormal test results were pursued. It is unclear whether this represents an appreciation of the magnitude of test errors in low prevalence populations or confirmatory bias.

In the context of investigating reasons for variation in test use in primary care, Zaat (1992) [ETA36] did not find an association between individual attitudes to risk and self-reproach and laboratory test use.

2.5.3 Empirical test accuracy literature: Able to compute accurately

The majority of empirical studies (32/33 papers, 40/43 studies) included an examination of the ability of respondents to manipulate information to derive the probability of disease following testing; 16 studies (17 papers) in health professionals [ETA2; ETA6; ETA7; ETA9; ETA11; ETA14; ETA20-ETA23; ETA25; ETA28; ETA29; ETA32-ETA34],14 studies (6 papers) in non-health professionals [ETA8; ETA10; ETA13 (x3); ETA15 (x7); ETA17; ETA19] and two in a mixed health and non-health professional sample [ETA3;ETA4].

Most of these papers (25) required respondents to estimate post-test probability either quantitatively or semi-quantitatively. Six of 25 studies required respondents to estimate post-test probability on the basis of pre-test probability and test accuracy represented as sensitivity and specificity [ETA3; ETA4; ETA23; ETA32-ETA34]. Eight papers (16 studies) required

respondents to estimate post-test probability on the basis of pre-test probability and one or more of false positive rate, false negative rate, true positive rate (sensitivity) and true negative rate (specificity) (ETA 2; ETA6; ETA9; ETA8; ETA13 (x3); ETA15 (x7); ETA20; ETA32). Five studies required respondents to estimate post test probability on the basis of clinical experience of disease prevalence and test accuracy (ETA14; ETA17; ETA19; ETA22; ETA25). Two studies compared the utility of different test accuracy metrics and graphics for probability revision (sensitivity and specificity (%), a plain language explanation of LRs and a graphical representation of test accuracy (ETA28; ETA33). In addition, six of these 25 papers included a comparison of test accuracy presented as one or more of natural frequencies, normalised frequencies or % (ETA4; ETA20; ETA32; ETA8; ETA13; ETA15).

2.5.3.1 Comprehension: ability to undertake probability revision

Ability to derive post-test probability was poor, (average of < 46% across studies, range 0% - 73%) with correct estimates above 33% achieved only by academic clinicians. It is unlikely that these studies were adequately powered to detect differences by medical speciality or between health professionals and non-health professionals. With the exception of the above average performance of academic clinicians there were no consistent patterns observed across participant groups.

2.5.3.2 Comprehension: the effect of presentation format on probability revision

More correct responses were obtained when test accuracy was presented as natural frequencies compared to normalised frequencies or percentages. The exceptions to this observation were studies that employed partitioning or explication of subsets of information (ETA8; ETA13; ETA15). In studies employing partitioning or explication of subsets the difference between natural frequencies, normalised frequencies and percentage presentation was attenuated. Features of incorrect responses when information was presented as normalised

frequencies or percentages were base rate neglect (ETA2-ETA4; ETA6; ETA8; ETA9; ETA13; ETA20; ETA22; ETA33) and reference class confusion (sensitivity is confused with PPV and specificity with NPV) (ETA7; ETA14; ETA15; ETA29). A feature of incorrect responses when information was presented as natural frequencies was neglect of test accuracy information (ETA4; ETA8; ETA13; ETA20). Two studies observed that base rate neglect was inversely associated with pre-test probability and concluded this was a result of respondents' difficulty handling very small percentages or proportions (ETA19; ETA23).

2.5.3.3 Comprehension: The different effect of positive and negative test results on probability revision

The majority of studies concerned with probability revision were restricted to probability revision following a positive test result. Only seven of 32 studies included an investigation of disease probability estimation following both positive and negative test results (ETA7; ETA11; ETA21-ETA23; ETA25; ETA28), all in health professionals samples. The findings of three of the seven studies suggested that respondents had relatively more difficulty deriving the probability of disease after a negative test result (1-NPV) compared to a positive test result (ETA7; ETA22; ETA23). The results of an additional three of the seven studies suggested confirmatory bias, whereby a test result had an impact on estimates of post-test probability, only if it concurred with pre-test probability estimates, (box 1.9), was operating (ETA11; ETA21; ETA25).

2.5.3.4 Comprehension: The effect of test accuracy metric for probability revision

The two studies comparing the utility of different test accuracy measures rather than different presentation formats to facilitate probability revision compared sensitivity and specificity as percentages and a verbal description of LRs; one study also included a graphical presentation of test accuracy. Stuerer (2002) (ETA33) demonstrated that a verbal description of LRs reduced error in estimating post-test probability compared to sensitivity and specificity

whereas Puhan (2005) [ETA28] found no difference in the accuracy of post-test probability for different test accuracy measures.

2.5.3.5 Comprehension: Clinical experience alone as a basis for post-test probability estimation

Seven health professional studies [ETA11; ETA14; ETA21; ETA25; ETA22; ETA32; ETA33] and one non-health professional study [ETA17] investigated the ability of respondents to estimate post-test probability based on clinical experience alone without provision of pre-test probability or test accuracy information.

Only one health professional study demonstrated a majority of respondents were able to adjust pre to post-test probability, in the correct direction, on the basis of clinical experience alone, following a positive test result [ETA22] this finding was not replicable for negative test results (2.5.3.3 above). Two studies demonstrated a minority of health professionals were able to accurately estimate post-test probability following a positive test result on the basis of clinical experience [ETA14; ETA33]. In two health professional studies there was evidence of confirmatory bias [ETA11; ETA25], whereby a test result (positive or negative) had an impact on estimates of post-test probability, only if it concurred with pre-test probability estimates (box 1.9).

One non-health professional study [ETA17] demonstrated clustering and overestimation of post-test probability estimates across a variety of diseases and test results and concluded this reflected a lack of appreciation of the use of test results.

2.5.3.6 Preference: use of probability revision in practice

One health professional study [ETA29] surveyed clinicians about their use of Bayes theorem for probabilistic reasoning in practice. Only 3% of respondents reported using Bayes theorem for probability revision whilst 80% of respondents reported using predictive values as the basis for estimating post-test probability.

2.5.3.7 Behaviour: impact of probability revision on practice

In one health professional study relying on respondents' own estimates of test accuracy, confirmatory bias was observed to influence test ordering behaviour as well as comprehension (ETA21). One study demonstrated no difference in patient management when respondents were given test accuracy information or relied on their own test accuracy estimates for post-test probability estimation (ETA32).

2.6 Results: Empirical risk literature

For results tabulated by study see appendix 2.3.

Characteristics of included studies

Eleven reviews and 10 papers reporting 13 studies were concerned with the understanding and communication of medical risks. Table 2.1 (2.3.2) illustrates how primary studies updating the review of reviews were chosen for inclusion on the basis of review date, health or non-health professional samples and research question addressed. One review was concerned with health professionals' understanding, seven reviews and 12 primary studies with non-health professionals' understanding and three reviews and one study with health and non-health professionals understanding. The literature spanned 1996 to 2009 and as for the test accuracy literature, this is likely to reflect the promotion of the integration of quantitative external evidence into medical decision making as part of the evidence based medicine movement [80].

The country of origin of included studies was not reported for 7/11 reviews. For the remainder of reviews, 50% to 98% of included studies originated from the USA [(ER1; ER6; ER7; ER9)]. Six of the ten primary studies represented respondents from the USA, two Europe, one Norway and North America and one Australia.

Eight of the 13 primary studies were conducted using face to face or self-completed paper questionnaires, four on line and one by telephone. Ten studies were conducted in general adult populations and three in medical settings or in individuals at high risk of disease. In 10/13 included studies, greater than 50% of the sample had had higher education.

Included studies in 4/10 reviews and 11/13 primary studies were concerned exclusively with hypothetical presentation of risks, two reviews and two primary studies were concerned with presentation of risks in actual practice (ecological), three reviews were concerned with presentation of risks in both hypothetical and practice settings and in one review this information was not reported.

Types of risk portrayed by included studies

Literature was concerned with risks associated with a variety of healthcare decisions. Included studies in 2/10 reviews and 2/13 primary studies were concerned with presentation of testing risks alone (for example survival 'risk' associated with uptake of screening), one review and nine primary studies with intervention risks alone (for example risk of morbidity, mortality and adverse effects), and six reviews and one primary study a combination of two or more of testing risks, intervention risks and population risks (for example the population risk of developing a disorder). Two reviews included presentation of medical and non medical risks [ER2; ER13]. In six of the 11 reviews no information was provided about the population characteristics of included studies. Two reviews were concerned with screening populations (low risk, high risk, workplace and self selected [ER1; ER7]. Two reviews were concerned with risk communication in the context of cancer genetics. One third of the studies included in the single review concerned with health professionals' understanding of risks [ER16] were conducted at educational events and are therefore likely to represent highly selected samples.

Study Quality: Health professional samples

One review concerned exclusively with undergraduate medical students and health professionals from a range of health care settings [ER16] was of moderate quality, although included studies were described as generally being of poor quality.

Study Quality: Non-health professional samples

Five of the seven reviews of non-health professionals were judged to be of high quality [ER1; ER5-ER7; ER9] although only one reported the quality of included studies as good [ER7].Two of seven reviews of non-health professionals were of low quality [ER12; ER13]. Of the 12 primary studies of non-health professionals reported in nine papers, two were RCTs judged to be of

good quality (ER8; ER18), five were RCTs judged to be of poor quality or poor reporting precluded quality assessment (ER3; ER4; ER11; ER19 (x2)) and five studies were cross sectional in design (ER15; ER20 (x3); ER21).

Study Quality: Mixed health and non-health professional studies

Three reviews of mixed patient and undergraduate medical student samples (ER2; ER10; ER14) were of poor quality and for the one cross sectional study including patients and physicians poor reporting precluded quality assessment (ER17).

2.6.1 Results: Empirical risk literature: Fully informed (comprehension, accuracy of perception, preference, behaviour change)

Several indicators of the extent to which individuals are fully informed about risk have been utilised in the risk communication literature.

Comprehension and accuracy of perception

Although a distinction is made in the literature between comprehension and perception, measures of comprehension were often not clearly reported, were variable and were often semi-quantitative. In addition risk perception was almost exclusively informed by risk information provided to respondents rather than based on participants' own experience. There is therefore likely to be a significant overlap between comprehension and perception as reported in the literature here.

Preference

Interestingly of the minority of studies investigating the relationship between preference and comprehension, no association was observed for patients (ER15) or health professionals (ER2). This was suggested to be due, at least in part, to respondents applying heuristics selectively to presentation formats most familiar to them (ER2).

Behaviour change

With the exception of three reviews (ER1; ER6; ER7) investigation of behaviour change in the risk communication literature has been restricted to measurement of intended rather than actual behaviour. Evidence on intended behaviour would be expected to more closely reflect comprehension than actual behaviour; the latter would be expected to vary according to contextual factors, including those acting as motivational biases. These relationships were not examined formally in the literature reviewed.

2.6.1.1 Numerical versus verbal presentation of risk

Comprehension and perception

Three reviews included an investigation of the effects on comprehension of numerical versus verbal presentations of risk [ER6; ER9; ER16]. McGettingan (1999) [ER16] demonstrated increased consistency of ratings of effectiveness in health professional samples with numerical compared to verbal presentations of risk. Similarly one review of health professionals [ER16], two reviews of non-health professionals [ER7; ER12] and one primary study of non-health professionals [ER20] demonstrated greater consistency in behaviour for numerical compared to verbal risks. This observation may be a reflection of the difficulties in standardising verbal descriptions of risk magnitude [60,61]. Numerical presentations of risk are observed to result in greater comprehension compared to verbal presentations in non-health professional samples [ER6; ER9]. Verbal presentations of risk have been observed to improve accuracy of perception compared to numerical risk presentation formats [ER20] and relative risk presentations to result in overestimation of risk [ER16; ER21].

Behaviour

Evidence concerning the direction of effect of verbal versus numerical presentation of risks on behaviour is inconsistent. Two reviews of non-health professional studies observed an increase in behavioural uptake with verbal compared to numerical presentation of risks of harm (developing disease); one for preventive behaviour uptake not otherwise specified [ER12] and one for screening uptake [ER7]. One primary study of non-health professionals [ER20] observed a decrease in treatment acceptance when risk of harms (treatment side effects) were presented verbally compared to numerically whereas one review of non-health professional studies [ER6] demonstrated a decrease in treatment acceptance with numerical presentation of risks of harms (treatment side effects) compared to verbal presentation. This apparent inconsistency may be due to contextual features of the risk scenarios including

perception of the nature of the risk or differences between respondents in attitudes to risk and comprehension.

2.6.1.2 Graphical presentations of risk

Comprehension

Two reviews including studies of both health professionals and non-health professionals [ER2; ER14] and two non-health-professional studies [ER8; ER21] included an investigation of graphical risk communication on comprehension. One study of non-health professionals demonstrated improved comprehension with pictographs compared to normalised frequency or percentage representations of risk [ER21] whilst two reviews of health professional and non-health professional samples did not find evidence for an improvement in comprehension with the addition of graphics to numerical presentations of risk [ER8; ER14]. Features of graphical presentations of risk that have been shown to improve comprehension in both health professional and non-health professional samples include part-whole representations compared to non part-whole representations [ER2], block versus random icon displays in pictographs [ER2] and for comparison of risks, risk ladders [ER14].

Perception

There is some evidence from health professional and non-heath professional studies that provision of graphical, numerical and verbal information about risk improves accuracy of perception over either presentation format alone [ER14; ER19; ER12]. However, graphical presentations did not result in improved accuracy of perception over numerical presentation formats in non-health professional samples [ER6; ER21]. One review of health professional and non-health professional studies suggested that pictographs resulted in overestimation of risk compared to other graphical presentations [ER10].

Behaviour

There has been limited investigation of the effects of graphical presentation of risks on behaviour in health professional and non-health professionals [(ER2)]. Presentation of harms (risk of developing disease) has been observed to increase preventive behaviour uptake when presented as pictographs or bar charts compared to numerical presentation. Graphical part-whole relationships have not been shown to have an effect compared to non- part-whole graphical representations of harm on preventive behaviour uptake.

2.6.1.3 Frequentist versus probabilistic presentation of risk

Comprehension and perception

A high level of comprehension (70%) was observed in one motivated non-health professional sample when risk was presented both in normalised frequency and percentage format [(ER15)]. Although normalised frequencies (constant denominator) were observed to improve comprehension compared to frequencies presented with a constant numerator (1/n) in one review of health professionals and non-health professionals [(ER2)] a later study of non-health professionals did not observe a similar effect on accuracy of perception [(ER19)].

Natural frequencies were observed to improve comprehension in comparison to normalised frequencies in both non-health professional and health professional samples [(ER17; ER3)] and in comparison to comparative measures (Relative Risk Reduction (RRR); Number Needed to Treat (NNT); Attributable Risk (AR); Tablets Needed to Take (TNT)) in another non-health professional sample [(ER3)]. Frequencies (not otherwise specified) were observed to improve comprehension compared to probabilistic representations in a review of non-health professional studies [(ER12)].

One review of non-health professionals [(ER6)] observed base rate neglect with manipulation of risk denominators.

A review of non-health professionals and health professionals suggests that frequentist representations of risk are perceived as pertaining to self and probabilistic representations to others [ER10].

Although one review and one primary study representing health professionals and non-health professionals observed a decrease in comprehension with lower compared to larger magnitudes of probability [ER21; ER14] this observation was not replicated for accuracy of perception [ER21].

2.6.1.4 Comparative risk measures (Relative Risk (RR); Relative Risk Reduction (RRR); Attributable Risk (AR); Attributable Risk Reduction (ARR); Number Needed To Treat (NNT); Tablets Needed To Take (TNT))

One review of health professional and non health professional studies [ER10] and three non health professional samples [ER8; ER18; ER21] were concerned with comparisons of comparative risk measures.

Comprehension and perception

Overall, 44% of a non-health professional sample were able to identify the more effective treatment when risk was presented either as RR, AR or NNT [ER18].

An improvement in comprehension with absolute risk measures compared to relative risk measures was observed in medical students but the finding was not replicable in patients [ER10]. Use of absolute measures of risk (AR, NNT) was observed to lessen comprehension compared to relative risk in two non-health professional samples [ER18; ER21].

One review of health professional studies [ER16] and one primary study of non-health professionals [ER21] demonstrated a magnification of perception of risk with relative risk compared to absolute risk (AR and NNT) measures.

Behaviour

One review of health professionals [ER16] and three primary studies in non-health professionals [ER3; ER6; ER11] included a comparison of comparative risk measures on behaviour. All report greater uptake of screening or acceptance of treatment with relative compared to absolute risk measures.

2.6.1.5 More versus less information

Comprehension

One primary study of non-health professionals observed that presentation of multiple numerical risk metrics was perceived as unhelpful although this did not result in a detectable difference in comprehension or accuracy of perception [ER8].

Behaviour

One review of non-health professional studies observed an increase in treatment uptake with increasing explanation of data concerning risks of treatment benefit but no effect concerning risks of treatment harm [ER6].

2.6.1.6 Tailored versus non-tailored presentation

Content tailoring and presentation tailoring

There was a lack of clarity in many reviews about the exact nature of tailoring of information that was the focus of investigation in primary studies. Content-tailored and non-content-tailored risk information is closely aligned to presentation of absolute rather than relative risk measures whereas tailoring of risk presentation is a more heterogeneous concept that encompasses presentation format, respondent preference and factors that are perceived to affect motivational biases.

Comprehension and perception

Five reviews of non-health professionals included an investigation of the effects of tailoring information on comprehension.

Three reviews demonstrated an improvement in comprehension with content-tailored compared to non-content-tailored information [ER1; ER5; ER7] and one review an improvement in comprehension with tailored compared to non-tailored-information (content or presentation not clearly specified) [ER9]. One review did not find any effects of tailoring (not clearly specified) on comprehension [ER12].

The effect of tailoring information on accuracy of perception is observed to be mixed, with some studies reporting an improvement in perception with content-tailored versus non-content-tailored information [ER5; ER7] and some studies reporting inconsistent effects with content-tailored information [ER1] or any type of tailoring (content or presentation) [ER12].

Behaviour

One review [ER1] suggested that tailoring risk information had inconsistent effects on screening behaviour when communicating risks of harms (developing disease), although these conclusions were based on studies heterogeneous for the type of tailoring (content only; content and presentation; presentation only). Three reviews [ER6; ER7; ER12] observed content-tailored risk information to increase uptake of screening compared to non-content-tailored information when communicating risks of harms (developing disease).

2.6.1.7 Anchoring to familiar risks / lay versus medical terminology

Comprehension and perception

One primary study demonstrated that anchoring of health risks to familiar non-health risks [ER8] resulted in improvements in comprehension whilst one review demonstrated improved comprehension with the use of lay compared to medical terminology [ER6].

Behaviour

Use of lay terminology has been observed to increase treatment uptake compared to use of medical terminology when presenting treatment harms [ER6].

2.6.2 Results: Empirical risk literature: Fully rational

One review of health professionals [ER16], four reviews of non-health professionals [ER5; ER6; ER12; ER13], four primary studies of non-health professionals [ER3; ER19; ER20; ER21] and one review including health professional and non-health professional studies [ER2] were concerned with the effects of risk presentation on anxiety and affect and the effects of framing of risk information and attitudes to risk on comprehension, perception, preference and behaviour.

2.6.2.1 Numerical versus verbal presentation

Anxiety

Numerical presentation of risks appears to decrease anxiety compared to verbal presentation in health professionals and non-health professionals and across a variety of health care decisions [ER6; ER20].

2.6.2.2 Comparative risk measures (RR, RRR, AR, ARR, NNT, TNT)

Anxiety

One study of non-health professionals demonstrated that presentation of Absolute Risk (AR) metrics resulted in less anxiety than presentation of Relative Risk (RR) [ER21] which is coherent with the finding that perception of risk is magnified when presented in relative rather than absolute terms (see above) [ER16; ER21].

2.6.2.3 Lay versus medical terminology

Anxiety

One review of non-health professional studies demonstrated that use of lay terminology concerning potential side effects of a drug resulted in increased anxiety compared to medical terminology [(ER6)].

2.6.2.4 Vivid (personalised) versus abstract (population) descriptions of risk

Anxiety

One review of non-health professional studies demonstrated no difference in anxiety for vivid compared to generic based descriptions of risk [(ER6)].

2.6.2.5 Graphical presentation

Affect

One study of non-health professionals demonstrated negative affect was significantly higher with the use of a graphic to present risk information (the Paling scale) (appendix 2.3) compared to frequencies with a constant numerator (1/n) followed by a pictograph graphic or normalised frequencies (with a constant denominator) [(ER20)].

2.6.2.6 Framing (loss versus gain and positive versus negative frames)

The effects of positive framing (communicating effects in positive terms, for example survival) and negative framing (communicating effects in negative terms, for example mortality were investigated in both health professional [(ER16)] and non-health professional studies [(ER3; ER6; ER12; ER2; ETA10; ER13)]. In addition one review of non-health professionals [(ER6)] made a distinction between positive and negative framing and loss and gain framing; the latter defined as emphasising benefits over losses or losses over benefits respectively.

Comprehension and perception

McGetiggan (1999) [ER16] demonstrated an increase in the perception of treatment benefit by health professionals with positive framing of risk of benefit.

Behaviour

Loss framing and negative framing have both been observed to increase uptake of screening by patients [ER6; ER12] whilst positive framing has been observed to increase treatment use by professionals [ER16]. Temporal considerations may also add complexity to interpretation of framing effects on behaviour. Negative framing (mortality) was observed to result in risk aversion in the short-term (avoidance of potentially toxic treatment) whereas positive framing (survival) resulted in risk taking in the short-term in the context of survival curve interpretation [ER2]. These apparently contradictory observations may be a reflection of the complexity of defining optimality when considering both professional and patient utility: differences in the definition of loss and gain, risk and certainty. Positive and negative framing effects appear to have a greater effect on intended behaviour for business and gambling domains compared to health and social domains [ER13] making the application of prospect theory (1.5.2.2) to healthcare settings problematic. As an illustration, testing was conceptualised as risk aversive behaviour by many authors contributing to the non-empirical test accuracy literature, on the basis that it results in greater certainty about the presence or absence of disease (2.4.2). In the risk literature, uptake of screening was considered risk taking in the short-term on the basis that screening may reveal the presence of disease that would otherwise not be apparent to an individual. The fact that a review of the effects of intervention risks and testing risks on behaviour found no consistent effect of positive or negative framing effects [ER6] is likely to be a reflection of the complexity associated with defining optimality in healthcare rather than the absence of a framing effect.

2.6.2.7 Attitudes to risk

Comprehension, perception, anxiety and behaviour

Different types of healthcare decision and individual variation in attitude to risks may modify the effects of risk presentation. For example in a review of non-health professional studies comparing the effects of different types of medical decisions, intervention risks had greater effects on comprehension, perception, anxiety and behaviour compared to testing risks [(ER5)]. Further, individual variation in attitudes to risk type, risk magnitude, type of outcome and associated costs have been observed to modify the effects of risk presentation and the effects of framing on intended behaviour [(ER16; ER3; ER13)].

2.6.3 Results: Empirical Risk literature: Able to compute accurately (manipulation of risks; comparison ≥ 2 risks)

Three studies in non-health professionals [(ER4; ER11; ER18)] and one review [(ER2)] and one study in health and non-health professionals [(ER17)] examined the ability of respondents to quantitatively or semi-quantitatively manipulate risk measures.

2.6.3.1 Frequentist versus probabilistic presentation

Comprehension and perception

In a study of non-health professionals, 57% overall were able to correctly mathematically manipulate risks although percentage presentation of risk resulted in the largest number of correct responses followed by normalised frequency presentation and least for frequencies with a constant numerator (1/n) [(ER4)].

2.6.3.2 Comparative risk measures (RR, RRR, AR, ARR, NNT, TNT)

Comprehension

Although only 13% of a non-health professional sample were able to correctly manipulate comparative risk measures, RRR was demonstrated to result in a larger number of correct

responses followed by ARR, followed by a combination of measures and least for NNT [ER18]. The authors suggest that this counterintuitive finding (ARR are easier to manipulate mathematically) may be due to the fact that RRR are familiar representation of probabilities to non-health professionals – for example they are encountered when adjusting retail prices during sales.

Correct responses of between 52% and 87% were observed in a highly selected sample of health and non-health professionals for manipulations of RRR, ARR, RR and baseline risk to derive treatment effects. Overall health professionals achieved more correct responses compared to non-health professionals [ER17].

Behaviour

Manipulation of RRs was observed to result in more risk aversive behaviour compared to manipulation of ARs in one non-health professional sample [ER11]. This is consistent with the observation of magnification of perception of risk with RR compared to AR [ER16; ER21].

2.6.3.3 Graphical

Comprehension

One review including studies of health and non-health professional demonstrated an improvement in correct responses for probability problems when information was presented as pictographs (part-whole information) compared to numerical representation (probabilities or percentages [ER2].

2.7 Strengths and limitations: Literature reviews

The breadth and iterative nature of the search strategy is likely to have captured the key areas that have been discussed and researched to date. The relatively recent development of test accuracy research methods and application is reflected in the literature identified. In

addition, continued opportunistic literature acquisition since completion of the formal searches and a check for face and content validity at a recent international diagnostic test symposium, (Methods for Evaluating Tests and Biomarkers: second international symposium. University of Birmingham. July 2010) provides some reassurance that key evidence has not been missed and that for the review of theoretical literature, saturation had been reached.

However given the challenges of searching the test accuracy literature [82-84] and the breadth of disciplines covered it is possible that relevant studies have been missed despite the comprehensiveness of the literature searches.

2.7.1 Non Empirical test accuracy literature

It is inevitable that exclusive use of the published literature and the relatively large proportion of articles accessed as a result of reference checking and experts may have limited the perspectives represented by this review, despite the breadth of the bibliographic database search strategy. Reliance on the published literature may also have resulted in under-representation of the perspectives of practising clinicians as it is unclear the time which clinician authors included in these reviews spend in clinical practice or the degree to which their opinions are informed by the perspectives of clinical colleagues, particularly those affiliated with academic institutions. Further, conclusions regarding the strength, order and discipline-specific nature of the line of argument presented, depend on the assumption that the literature identified is representative. However, despite these potential limitations this qualitative synthesis provides a point of reference to appraise the extent to which empirical investigation of the understanding and application of test accuracy measures reflects and reinforces the issues raised.

2.7.2 Empirical test accuracy literature

The empirical test accuracy literature identified is limited in volume and quality and is heterogeneous; many of the observations reported are based on the findings of a very small number of studies and are not supported by consistency. Although the findings of older studies may not reflect current knowledge, particularly given the impact of the evidence based medicine movement and the fact that diagnostic research is likely to be less familiar to clinicians compared to effectiveness research [(ETA2)], 60% of empirical test accuracy studies identified were conducted in the last two decades.

2.7.3 Empirical risk literature

The review of risk communication relied heavily on existing reviews. Relying on reviews rather than primary studies restricts consideration of outcomes to those considered by review authors which may not coincide with the themes raised by the test accuracy literature. However, the range of outcomes considered across the numerous reviews undertaken in the risk communication literature suggests that they are a comprehensive reflection of issues investigated in primary studies.

Although risk reviews were varied with respect to the outcomes they considered, it is likely that some degree of duplication in inclusion of primary studies occurred. It is considered that this is unlikely to have an impact on the conclusions drawn from this review, providing included primary studies were representative of the evidence available.

Variability in presentation formats across studies limited comparability. For example, for the investigation of the effect of presentation of risks as frequencies, there was a lack of a consistent distinction between natural frequencies and normalised frequencies which alone has been shown to result in differences in comprehension of treatment risks, testing risks and understanding of test accuracy [85]. The use of reviews may therefore obscure

associations if effect modifiers are not taken into consideration, as was observed for a review of framing effects across a variety of medical decision making contexts [ER6].

2.8 Quality and applicability of included literature

2.8.1 Quality

Included reviews and primary empirical studies were of variable quality. For empirical risk reviews considered high quality only one reported the quality of included studies as mostly good [ER7]. The majority of empirical test accuracy studies were cross sectional in design and study reporting precluded quality assessment in a large proportion of included studies.

2.8.2 Applicability

The empirical test accuracy research is largely congruent with the theoretical literature. However this may be a reflection of the similar and highly selected nature of both samples. As would be expected, comprehension, accuracy of perception and ability to manipulate risks were associated with numeracy and education [ER4; ER15; ER10; ER18]. Similarly, empirical test accuracy studies attempting to distinguish between the ability of academic and practising clinicians demonstrate large differences in ability [ETA2; ETA3; ETA18].

The review of non-empirical test accuracy literature is therefore likely to represent the perspectives of experts rather than practising clinicians and the reviews of empirical studies to overestimate comprehension, accuracy of perception and ability to manipulate risks, with less clear impact on preference for metric and presentation format.

Findings are almost exclusively based on hypothetical scenarios and self-reported practice. The generalist perspective was under-represented in both the non-empirical and empirical test accuracy evidence base. Unique aspects and challenges posed by the early stages of the diagnostic work up in primary care settings, such as the different emphasis of test use (an emphasis on ruling out disease rather than reaching a definitive diagnosis) and symptom

rather than disease based investigation is not represented. The observations that test accuracy terms concerned with the absence of disease (specificity) and negative test results are less well understood and that manipulation of small probabilities cause difficulty may therefore reflect the restricted testing context represented by the literature rather than a generic problem. Research concerning the use, understanding and application of test accuracy information should be undertaken mindful of the potential differences in patient spectrum, testing culture and types of challenges encountered in generalist compared to specialist settings. Representation of both settings is required in order to address needs specific to either group as well as to facilitate evidence based testing across entire care pathways.

In addition to a specialist contextual focus, UK practice was under-represented in the empirical test accuracy and risk literature. The majority of evidence originated from the USA where medical education differs and healthcare organisation and cultural differences may limit generalisability, particularly with respect to behaviour.

The non-empirical test accuracy literature was almost exclusively concerned with the use of test accuracy measures for decision making at the bedside and an assumption that probability revision is a pre-requisite for informed diagnostic decision making with limited consideration of their utility to guide testing policy. The requirements of policy makers and interpretation of meta-analyses of test accuracy were not well represented and there was relatively little attention given to the use of test accuracy measures to facilitate test comparisons or evaluation of multiple tests in a testing pathway. Comparisons between tests and testing policy are more likely to be decided on the basis of reviews, where global measures of test accuracy have the potential for greater application.

2.9 Conclusions

Whilst there is widespread belief that clinicians have difficulty applying test accuracy information, this has not been based on a systematic interrogation of the evidence base. As a result it has not been possible to date to quantify or characterise the extent of the problem in order to identify characteristics of existing test accuracy measures or expressions of probability more generally that might facilitate their understanding and application. This review represents the first attempt to bring together evidence pertinent to the facilitation of evidence based diagnosis and the findings provide a framework for further research.

2.9.1 A decision maker who is fully informed?

2.9.1.1 Desirable properties of test accuracy metrics: sensitivity, specificity and predictive values

The majority of non-empirical test accuracy articles were concerned with one or more of sensitivity, specificity and PVs with frequent comparison to LRs. The features of test accuracy measures that are perceived to impact on the ease and appropriateness with which they are interpreted and applied include:

- Having the test result (rather than disease status) as the reference class for interpretation of conditional probabilities.
- Discrimination between the two dimensions of test accuracy and quantification of test errors (ability to rule in or rule out a diagnosis or the value of a positive test result separate to a negative test result)
- Portability across populations

Predictive values are repeatedly described as the more intuitive of the test accuracy metrics on the basis that they have a test result rather than disease status as reference class. However they are subsequently dismissed on the basis that they are mathematically

dependent on prevalence. This dismissal of PVs on the basis of their dependence on prevalence has been fuelled by a relative neglect of the effects of population spectrum as a modifier of all test accuracy metrics. The result has been a lack of empirical investigation of how PVs impact on understanding and application of test accuracy information.

Familiarity with metrics as measured by the empirical test accuracy literature was not a good indicator of understanding and available research suggests that sensitivity and specificity, although predominant metrics in test accuracy research, are not well understood and their practical application is difficult. One empirical study investigating the use of test accuracy metrics in practice reported that sensitivity and specificity were used by 4% of respondents compared to 80% respondents reporting to use PVs [(ETA29)].

There are important parallels to be drawn between the development of outcome reporting for primary test accuracy research and those for meta-analyses of test accuracy. Sensitivity and specificity have been shown to be the most commonly used test accuracy metric in meta-analyses of test accuracy [6] . As for primary test accuracy studies this is likely in part to have been based on a misperception that sensitivity and specificity are fixed properties of tests and that their use will reduce heterogeneity. However recent research suggests that directly deriving PVs from meta-analyses produces similar estimates to PVs derived indirectly from summary estimates of sensitivity and specificity [86] . In addition the use of PVs may mitigate against partial and differential verification bias and have advantages in situations when it would be unethical or impractical to verify index test negatives, such as the application of tests for screening. The complex relationship between prevalence, spectrum and heterogeneity requires further research [51,87,88]. It is possible that metrics that are not mathematically dependent on prevalence (sensitivity, specificity and LRs) may offer no great advantage in this respect.

2.9.1.2 Likelihood ratios, ROC curves and test accuracy metrics common to systematic reviews

Investigation of the ability of respondents to define and interpret more recently introduced test accuracy metrics such as LRs and those more common to systematic reviews of test accuracy (for example ROC curves, AUC and Forest plots) were almost entirely absent from the empirical and non-empirical test accuracy literature and the reported use of these metrics is < 1% [(ETA29)]. The relative lack of consideration of global measures and development of testing policy may reflect a lack of familiarity with these measures, as a result of the relatively recent increase in volume of test accuracy reviews [39,42]. The relatively recent emergence of systematic reviews of test accuracy may also explain the emphasis in the literature on the use of test accuracy information for diagnostic decision making at the bedside rather than to support testing policy.

2.9.1.3 Complimentary use of test accuracy metrics

Lacking from the literature identified was discussion of how test accuracy measures might be used in a complimentary way to assist with diagnostic decision making. Comparison of test accuracy measures was approached with the aim of advocating a single, preferred metric rather than identifying a suite of metrics that would be complimentary in terms of presentation format and conveying different aspects of test accuracy (for example the two separate dimensions of test accuracy; the overall discrimination of a test; the relationship between pre-test probability and the clinical utility of a test). Similarly, there was very limited discussion of the links between different test accuracy measures (for example the similar information provided by sensitivity, NPV and LR- or specificity, PPV and LR+) that might help decision makers make sense of the multiple outcome measures in use [43]. The potential role of the 2x2 diagnostic table as a test accuracy presentation format that could be used to illustrate the relationship between summary test accuracy metrics as well as an explicit representation of test errors in a natural frequency format appears to have been overlooked.

2.9.1.4 Estimation of pre-test probability and test accuracy

The literature mostly conceptualises the clinical history and examination as characteristics of patients contributing to 'pre-test' probability. This may be a feature of the secondary care focus of the literature reviewed (see 2.8.2 above) which may serve to undermine the contribution of the clinical history and examination as diagnostic tests in their own right. There is a need for greater consistency in the use of the term 'pre-test probability' in order to provide contextual clarity. As a concept, pre-test probability needs to reflect spectrum, including specification of the point in the diagnostic pathway that the test is to be used. Findings from the empirical test accuracy literature do suggest that clinicians are aware of contextual modifiers of pre-test probability although accuracy of pre-test probability estimation and knowledge of the accuracy of tests used in practice appears poor with wide variability and a tendency to overestimation.

The importance of accurate pre-test probability and test accuracy estimation will depend on the extent to which formal probabilistic reasoning takes place as part of the diagnostic decision making process (see 2.9.3.1 below).

2.9.2 A decision maker who is fully rational?

There was very limited consideration of motivational biases in diagnostic decision making in both the non-empirical and empirical test accuracy literature. Discussion in the non-empirical literature included consideration of individual and contextual variation in attitudes to risk as a modifier of decision making behaviour. The empirical test accuracy literature was restricted to two studies in generalist settings and no association was found between individual practitioners' attitudes to risk and test ordering although it is unclear whether this is due to limitations of the measurement tools used or confounding. Patient and practitioner motivation were observed to be important modifiers of test use in these studies and patient and practitioner reassurance were viewed as legitimate reasons for testing.

Testing was portrayed as a risk aversive behaviour in the non empirical test accuracy literature and one contributory factor to the observed increase in testing (2.4.2). Explicit quantification of test errors is one method of conveying the degree to which a test reduces uncertainty but is not a feature of any existing summary test accuracy metrics and may serve to obscure the uncertainty associated with the testing process. Consideration of factors that impact on test and test treatment thresholds is an important aspect of the evaluation of the proposed role of new tests.

2.9.3 A decision maker who is able to compute accurately

2.9.3.1 Probability revision in practice

Both the non-empirical and empirical test accuracy literature is dominated by consideration of methods for simplifying probability revision. Although difficulties with undertaking quantitative probability revision were discussed, this was from the perspective that this was the problem solving approach to be aspired to, both by clinical and non-clinical authors. Approaches proposed by clinicians were grounded in probabilistic expression of uncertainty and probability revision using Bayes theorem whereas psychologists proposed a frequentist approach to probability revision. The emphasis on probability revision is likely to be a reflection of the promotion of the integration of quantitative evidence into clinical decision making endorsed by the Evidence Based Medicine Movement [81]. With the exception of academic respondents, quantitative probability revision appears poor, even in these predominantly highly selected samples.

Existing efforts to integrate test accuracy information into the diagnostic decision making process have been based on an assumption that quantitative test accuracy information is sought but not understood. This review raises questions about the extent to which clinicians seek quantitative estimates of test accuracy and pre-test probability and the extent to which formal probability revision is used in practice. For example the absence of effect on test use

of an intervention to improve the accuracy of probability revision [ETA27] and reported use of Bayes' theorem in practice by respondents in one study of 3% [ETA29] suggest formal probability revision may not be commonplace in diagnostic decision making. The impact that quantitative estimates of test accuracy might have on diagnostic and therapeutic yield requires consideration of not only the extent to which test accuracy information is understood but also the perceived added value of the information over clinical experience alone. A single study found no difference in patient management between clinicians provided with information about pre-test probability and test accuracy and those who were expected to rely on their own experience [ETA32] suggesting that clinical experience of test use rather than use of test accuracy estimates from the published literature are used in practice. Although evidence concerning the effect of clinical experience, as measured by years since completion of training, on the ability to undertake probabilistic reasoning is limited and conflicting [ETA2; ETA6; ETA11], the impact of individual and setting-specific variations in test and test-treatment thresholds has not received attention in this respect.

2.9.3.2 The impact of test accuracy metric and presentation formats on probability revision

On the basis of two studies in the empirical test accuracy literature, sensitivity and specificity, LRs and a graphical representation of test accuracy could not be distinguished with respect to their ability to facilitate probability revision.

There was a large body of evidence supporting the effect of presentation format on the probabilistic reasoning ability. The advantages of natural frequency presentation format is proposed to be as a consequence of their natural separation of reference classes (thereby avoiding reference class confusion) and simplification of probability revision by negating the need to incorporate base rates. The empirical literature here suggests that clear definition and partitioning of reference class may be the more important characteristic, which has implications for the use of sensitivity and specificity for probability revision.

Reflecting the importance of reference class, most studies from the psychological literature avoided the use of summary measures such as sensitivity and specificity; true positive rates (sensitivity) and true negative rates (specificity) rates were described as 'defective partitioning' (ETA15) due to the fact they refer to the disease as reference class. Instead false negative rates (1-sensitivity) and false positive rates (1-specificity) were commonly used reflecting test result as reference class. Indeed despite the finding by one study (ETA29) that practising clinicians used PPVs (PPV and 1- NPV) for information on the post-test probability of disease, the potential value of PVs as a summary test accuracy metric that avoids the requirement for probability revision has not been discussed or addressed by the empirical or non-empirical literature.

2.9.3.3 The impact of negative test results on probability revision

Three of the five studies in the empirical test accuracy literature investigating respondents' ability to interpret and use negative test results found respondents' had problems interpreting and using negative results. It has been suggested that clinicians' 'insensitivity' to negative test results may reflect problems processing absent problems, epidemiological terminology (negative predictive value) linking negative findings to the absence rather than the presence of disease (ETA7) and due to an emphasis on the 'abnormal' by patients. However, this finding should be interpreted with caution as all of these studies were conducted solely with secondary care clinicians where the emphasis of testing is proposed to be ruling in disease rather than ruling out disease (TTA8).

2.9.4 Contribution of the empirical risk literature

2.9.4.1 Consistencies with the test accuracy literature

Findings consistent in both the test accuracy and general risk literature include:

- difficulty with the comprehension of low probabilities (2.5.3.3)

- presentation of frequencies result in greater comprehension than percentages and natural frequencies appear to have a more marked effect on comprehension compared to normalised frequencies (2.5.3.3)
- frequentist representations of risk are perceived as pertaining to self whereas probabilistic representations are perceived as pertaining to others (2.5.1.1)

In addition, in the test accuracy literature, an important feature of natural frequencies believed to contribute to their accessibility was the fact they represented sequential acquisition of information based on direct experience (2.4.3.2). This may have parallels with attempts to facilitate understanding in the risk communication literature by anchoring unfamiliar risks to familiar risks [(ER8)].

2.9.4.2 Inconsistencies with the test accuracy literature

One striking difference between the body of literature concerned with the understanding and application of test accuracy measures and that concerned with the understandings and application of risk measures is the use of comparative metrics. The risk literature is almost entirely concerned with comparison of risks, whereas comparative test accuracy evaluation is almost entirely absent from the test accuracy literature. One explanation for this observation may be the delay in the development of methods for test evaluation relative to evaluations of interventions including the relatively more recent emergence of test accuracy reviews [39,42] (2.9.1.2 above). However in the absence of any risk literature concerned with understanding and application of single intervention risks, it is likely that this observed difference between the test and risk literature, at least in part, reflects barriers to rigorous test evaluation, such as the relatively more rapid pace of technological advancements in testing compared to drugs [18] and the relatively less lax regulatory system for the introduction of tests compared to drugs which does not encourage comparative evaluation [25].

2.9.4.3 Additional insights provided by the empirical risk literature

Characteristics of metrics that facilitate understanding

Overall comprehension of metrics and the ability to manipulate probabilities was superior in the risk communication literature compared to the test accuracy literature. This is supported by the one study in the empirical test accuracy literature that compared the ability of health professionals and non-health professionals to define and manipulate measures of effect and measures of test accuracy [(ETA2)] (appendix 2.2). Both the test accuracy and risk literature were characterised by educated and highly selected samples suggesting selection bias as an unlikely explanation for this observed difference.

Familiarity and understanding

The observed difference may be a reflection of the relative lack of familiarity with test accuracy metrics and less advanced understanding of the challenges posed by the use and application of test accuracy information compared to information about intervention risks. Indeed, the observation that empirical test accuracy studies were almost exclusively concerned with health professionals whereas empirical risk studies had a larger proportion of studies concerned with non-health professionals might, by itself, suggest that the evidence base concerning understanding and application of risk metrics is more advanced than that for test accuracy metrics.

However, if familiarity and the state of evolution of the evidence base were the sole explanations for the differences in understanding observed between the test accuracy and risk literature, it might be expected that health professionals would be superior to non health professionals on the basis of the advantages of medical training. There were no consistent differences in comprehension or ability to manipulate probabilities observed between health and non-health professional samples in the test accuracy literature and no consistent differences in the risk literature with the exception of a single study, where health professionals only were selected on the basis of 'strong' critical reading skills [(ER17)]. In

addition familiarity did not appear to be related to ability to define or use metrics in the test accuracy literature (2.9.1.1 above). In conclusion therefore, it is likely that differences in the characteristics of test accuracy and intervention risk information are contributing to observed differences in comprehension.

The considerable body of literature concerned with probability revision in the test accuracy literature reflected the need to derive the probability of disease consequent on a test result: the probability of having disease following a positive test result (equivalent to the PPV) or the probability of having disease following a negative test result (equivalent to 1- NPV). For this reason conditional probability summary test accuracy metrics with (antecedent) test result as reference class (PVs) were emphasised as intuitive for decision making in contrast to summary test accuracy metrics with disease class as test result (sensitivity and specificity) (2.4.1.1). The observation that comprehension and the ability to manipulate metrics was superior in the general risk compared to the test accuracy literature may in part be explained by the fact that all summary risk measures share the property of conveying the probability of having a condition *following* exposure to a risk or preventative factor rather than the probability of being exposed if you have or do not have a condition; in other words having the antecedent event as reference class for conditional probability measures may be a key characteristic facilitating understanding.

Comparative metrics

The body of literature pertaining to communication of risk demonstrated consistent overestimation of magnitude of effect with RR representation compared to AR representation (ER16; ER6;ER3; ER11; ER21) In addition content-tailored risk information had beneficial effects on comprehension and perception (ER1; ER5; ER7) and larger effects on behaviour (ER6; ER7; ER12) compared to non-content-tailored information.

Content-tailored information can be conceptualised as similar to absolute risk information by taking into account information on a baseline risk at the point of exposure whereas non-

content-tailored information is similar to relative risk information. Although comparable metrics are currently not widespread in the test accuracy literature, the potential for similar misinterpretation can be anticipated.

Graphics as an aid to comprehension

Use of graphics received limited attention in the non-empirical test accuracy literature. Although graphical presentation was proposed as a method for facilitating understanding (2.4.1.4) and application (2.4.3.1) of test accuracy information no empirical evidence was found to support this; one study of non-health professionals did not identify a preference for numeric or graphical presentations of test accuracy information and existing graphics, such as ROC curves, do not feature as an aid to decision making for health professionals (2.5.1.1). In contrast a considerable body of literature exists investigating the potential for graphics to aid risk communication. Inconsistent evidence for improvement in comprehension or accuracy of perception with graphical compared to numerical presentations of risk exists (ER2; ER8; ER14; ER21). In one non-health professional study multiple numeric metrics were perceived as unhelpful although no effect on comprehension was observed (ER8). There is however some evidence that provision of graphical, numerical and verbal information about risk improves accuracy of perception over either presentation format alone (ER12; ER14; ER19). This latter observation may be a function of maximising accessibility by including a variety of presentation formats. Graphic aids have appeal as a medium for the simultaneous presentation of the two dimensions of test accuracy and their interdependence and warrant investigation for this purpose. However it is clearly important to distinguish between complimentary presentation formats and indiscriminate presentation of multiple numerical metrics (see also 2.6.1.2; 2.6.1.5). The use of multiple numerical metrics in test accuracy evaluations is commonplace [6] although graphics are not prominent in either primary test accuracy studies or systematic reviews of test accuracy [89].

Contextual and motivational biases

The evidence available on risk communication includes different types of healthcare risk and emphasises risk comparisons. This offers the potential to investigate contextual modifiers of understanding and application of risk measures. Indeed this was raised as an important modifier of behaviour in the review of non-empirical test accuracy literature (2.4.2). There is evidence of contextual and temporal modification to attitudes to risk and risk taking behaviour (ER2; ER16; ER13, ER5) reflecting the notion of optimality as conditional on context (1.5.1) but not well predicted by behavioural decision theory which is largely based on decision making in the financial domain. Individual and contextual modifiers of motivation have implications for the potential of evidence based decision making to improve practice. This finding also suggests that comprehension or the ability to manipulate probabilistic information may not be a good predictor of behaviour.

Chapter 3: Mapping the epidemiological characteristics of test evaluation systematic reviews

3.1 Abstract

Background

There has been a growth in the volume of primary research concerned with testing over recent years. Specialist review databases represent a potential complimentary method for accessing test evaluation research given the well documented problems with searching for test accuracy studies in bibliographic databases. Characterising the test evaluation content of existing databases of systematic reviews may help those looking for specific types of test evaluation as well as identifying areas where test evaluation research is relatively sparse.

Methods

Five specialist review databases (York CRD's DARE, CDSR and HTA databases, the University of Maastricht's MEDION database and ARIF's in-house database at the University of Birmingham) were interrogated with respect to the proportion of included test accuracy reviews, quality assurance, ease of use and currency of databases and the epidemiological characteristics of included test accuracy reviews. Interrogation of databases comprised contact with database owners and application of an in-house search strategy for diagnostic studies. These complementary methods allowed for validation of the in-house search strategy for a proportion of review databases.

Results

Review databases varied significantly with respect to their currency. Difficulties identifying test accuracy reviews in bibliographic databases are mirrored in the review databases interrogated; tagging of test accuracy reviews is currently only conducted in one database (ARIF). A combination of 3 databases would be required to achieve an estimated 76% of available test accuracy reviews. Medion, HTA and C-EBLM databases were characterised by a relatively high proportion of particular disease topic areas. Overall, across databases, tests

applied in secondary care settings, (overall only 4% of reviews evaluated tests for use in primary care); certain disease topic areas (gastrointestinal, cardiovascular and obstetrics and gynaecology) and evaluations of single tests rather than test comparisons, predominate.

Conclusions

Issues pertaining to the identification of primary test accuracy research appear to be pertinent to identification of test accuracy reviews in general review repositories and the considerable ambiguity conveyed by review titles in this investigation also has implications for searching. Based on the epidemiological characteristics of test evaluations it is unlikely that the existing evidence base reflects the clinical need for evidence.

3.2 Rationale

There has been a growth in the volume of primary research concerned with testing over recent years. Evaluation of test accuracy (distinct from test effectiveness or cost-effectiveness) is almost certainly responsible for the majority of this increase. Trials of test and treat combinations have been shown to be rare with an estimated 37 test and treat randomised controlled trials published between 2004 and 2007 [90]. Decision models combining estimates of test accuracy with estimates of treatment effectiveness represent a practical alternative to trials of test and treat combinations, owing to the methodological and practical complexities as well as sample size demands of trials in this area [91]. The increase in test evaluations is reflected by the increasing number of systematic reviews in the area. Systematic reviews are an important resource for summarising existing knowledge and underpin guideline development and needs assessment for research activity.

It is well-documented that using methodological search filters with general bibliographic databases to locate studies of test accuracy is at best unreliable [84,92]. Specialist review databases may represent a complementary and possibly more efficient method for accessing test evaluation research. In addition databases of systematic reviews are an important and efficient resource to support methodological research. Characterising the test evaluation content of existing databases of systematic reviews may help those looking for specific types of test evaluation as well as identifying areas where test evaluation research is relatively sparse.

3.3 Aims and objectives

Aims

The aim of the interrogation of systematic review databases was to establish a repository of systematic reviews of test accuracy and compile a representative sample of test accuracy reviews for a methodological review reported in chapter 4. The process of generating a

representative sample of reviews offered the opportunity to examine the databases from which reviews were sourced in detail and to describe the epidemiology of reviews contained in these databases.

Objectives

- To characterise existing systematic review databases with respect to the number of systematic reviews of test evaluations they contain and to assess the overlap between databases.
- To assess existing systematic review databases with respect to their currency, quality assurance and ease of retrieval of test evaluation reviews.
- To map the following epidemiological characteristics of systematic reviews of test evaluations: disease category, review purpose and test application.

3.4 Methods

Databases making a claim to contain systematic reviews as opposed to narrative reviews and commentaries were included for consideration. The extent to which reviews contained in a database met the definition of a systematic review was not assessed.

The following five databases were included:

- Health Technology Assessment (HTA) database via the Cochrane Library (1998) (http://www.thecochranelibrary.com.ezproxyd.bham.ac.uk/view/0/index.html)
- Database of Abstracts of Reviews of Effects (DARE) via the Cochrane Library (1994) (http://www.thecochranelibrary.com.ezproxyd.bham.ac.uk/view/0/index.html)
- Medion database of diagnostic reviews (University of Maastricht)(1994) (http://www.mediondatabase.nl/)
- International Federation of Clinical Chemists Committee of Evidence Based Laboratory Medicine (IFCC C-EBLM) reviews database (established 1996 as

personal website of Wytze Oosterhuis, publicly available on the IFCC Web site in 2004)

- Aggressive Research Intelligence Facility (ARIF) in house database (University of Birmingham 1996). (http://www.arif.bham.ac.uk/databases.shtml).

Of these specialist reviews databases, Medion and C-EBLM are devoted solely to systematic reviews of tests. Inclusion of the ARIF in-house database was on the basis of plans to make the database publicly accessible and in addition the author's familiarity with the database facilitated its use as a point of reference for evaluation of the other databases. For those databases containing both systematic reviews of interventions and systematic reviews concerned with test accuracy (HTA, DARE, ARIF), a strategy for comprehensively capturing the test accuracy content was devised. The ARIF database tags diagnostic and screening reviews as such on inclusion. At the time of conducting the research test accuracy reviews were not denoted by any special indexing in the DARE and HTA databases. A pragmatic filter was therefore created in order to retrieve as many test accuracy reviews as possible whilst maximising specificity in the absence of reliable methodological search filters [84,92,93]. Searches of HTA and DARE were limited to MeSH index terms to make them as specific as possible. The choice of MeSH terms were based on an analysis of the performance of 12 validated diagnostic search filters [93]. The most frequently used MeSH term used by 11 of the 12 filters: 'Sensitivity and Specificity' (exp) (92% of filters) was combined with the term 'Mass Screening' in order to capture a variety of testing applications. The MeSH term 'Diagnosis' (exp) or text word 'diagnostic' greatly reduced the specificity of the searches and so were not used. The performance of the filter was verified as far as possible with the help of in-house searches of DARE and HTA performed by database producers CRD (Centre for Reviews and Dissemination, University of York) using their preferred search terms. Diagnostic reviews in DARE are coded in-house although at the time of conducting the research this facility was not available on the public database interface. The ARIF database was searched using the

tag *diagnosis* as well as the text word *screening* and false positive hits were identified by scrutiny of retrieved records (see appendix 3.1 for search terms used).

Searches for all systematic reviews of test accuracy in each database were carried out in January 2007 for the period 1996-2006. Scrutiny of retrieved records for false positive hits (reviews not concerned with test accuracy) also allowed investigation of the specificity of the filter (see appendix 3.2 for flow of references). All records for the relevant period in the specialist diagnostic reviews databases, Medion and C-EBLM, were included. Reference Manager v 11 for Windows was used to store downloaded records from DARE, HTA and ARIF, whilst C-EBLM and Medion were added manually as these databases were not compatible with reference management software.

3.4.1 Inclusion criteria

The focus for the methodological review reported in chapter 4 is systematic reviews of test accuracy, either test accuracy reviews conducted in isolation or systematic reviews of test accuracy undertaken as part of a broader evaluation of a test's effectiveness and cost-effectiveness. Thus the mapping exercise sought to map the epidemiological characteristics of reviews of all aspects of test evaluation.

3.4.2 Coding included references

References were tagged according to their database source. In addition epidemiological characteristics of test evaluation reviews were noted based on review title. To ensure consistency a pro-forma was used as in some instances a review could be placed in more than one category. Appendix 3.3 details the criteria used to code references.

3.4.2.1 Disease

The disease topic area or areas the review was addressing were recorded. Classification of disease was pragmatic and not based on a specific disease classification system.

3.4.2.2 Review purpose

The purpose of retrieved reviews was coded as ‘test accuracy’ only, ‘costs’ of testing, ‘effectiveness’ of testing, ‘cost-effectiveness’ of testing, ‘methodological’ test reviews or ‘other’, (reviews concerned with test acceptability; methods of test execution ; early test development for example promising disease markers, testing strategies; organisation of testing programmes; morphological studies). Test accuracy reviews were further sub-divided into those concerned with estimation of the accuracy of single test or with estimation of accuracy of more than one test.

3.4.2.3 Clinical setting

The clinical setting in which tests were being evaluated was noted. Test setting was defined as the likely origin of patients to be tested and not the setting in which the test was to be applied. Thus for example ultrasound examination and X-rays are likely to take place in a secondary care setting although these tests could be initiated and acted on in primary care. Reviews were coded as being concerned with tests to be used in a screening context (encompassing population based and targeted screening programmes), over the counter, in the community, primary care, secondary care or for use in multiple settings.

The search facility in Reference Manager was used to identify yield of references by single database and database combinations and to map epidemiological characteristics of test reviews contained in the databases.

3.5 Results

3.5.1 Performance of pragmatic search filter in general specialist review databases

Appendix 3.2 illustrates the performance of the pragmatic filter which performed variably for detection of reviews concerned with testing in databases with MESH search facilities (DARE and HTA). There were 89 false positive hits for DARE (19% of DARE hits) and 9 in the HTA database (3% of HTA hits). The number of false positives generated by searching the ARIF in-house database using the terms diagnosis and screening was low; n=13 (3%). In the HTA database, only 16 (5%) of hits were not reviews. In the ARIF database 2, (<1%) of hits had been wrongly added to the database as reviews when in fact they were primary research (mostly case series). One record from Medion (a letter) had been erroneously included in the database. All of the records in the C-EBLM databases were reviews. Both the Medion database and the C-EBLM database contained references not concerned with evaluation of tests (1% of Medion records and 9% of C-EBLM records), all concerned with describing putative causal associations between laboratory based markers and disease.

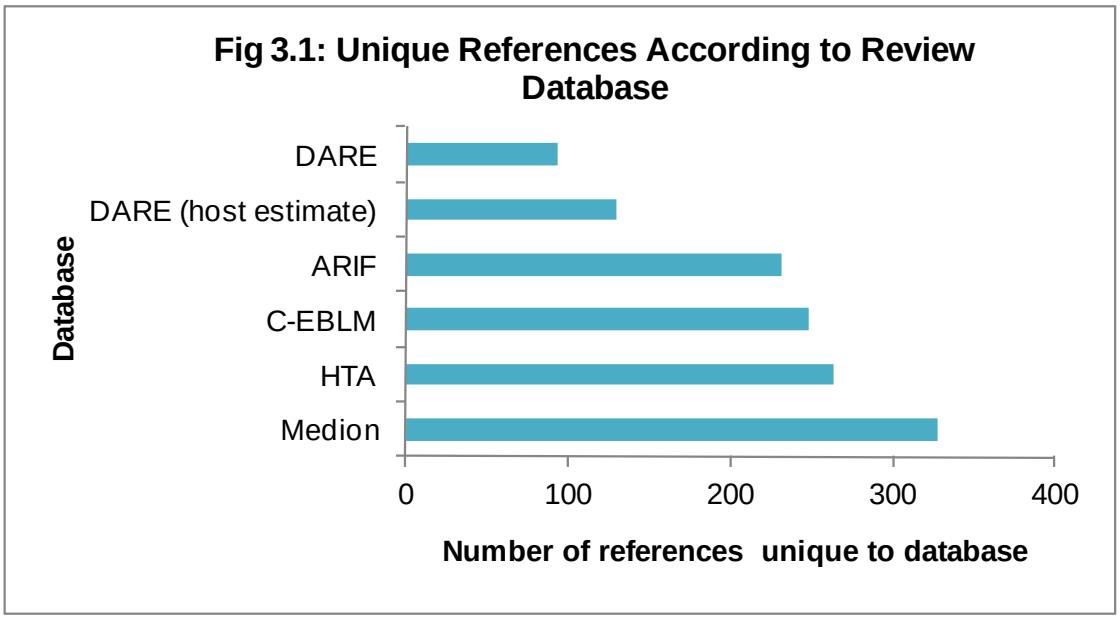
The pragmatic filter identified 383 (72%) of the 542 tagged test evaluation reviews identified by DARE producers for the period 1996-2006. Both estimates are important; the pragmatic filter estimate of 383 is likely to approximate to the yield from a search of DARE on the public interface whilst the database producer estimate is a more valid representation of the number of test evaluation reviews within DARE. Reviews identified by in-house producer searching of the HTA database yielded fewer hits (n=172) compared to the pragmatic filter (n=333) for the period 1996 and 2006. For the purposes of calculating yield of relevant references for single databases and across multiple databases an estimate of 333 test evaluation reviews for the HTA database was used.

3.5.2 Yield of test accuracy reviews by single databases

The yield of test evaluation reviews identified from searching the public interface of a single database would be 664 for the Medion database, 401 for the C-EBLM database, 383 for the DARE database and 333 for the HTA database and 490 for the ARIF database. Using the DARE database producer estimate increased the number of reviews that would be identified by DARE from 383 to 542 (an additional 159 reviews). (Appendix 3.2)

3.5.3 Duplication across databases

After removing reviews not concerned with diagnosis and primary research papers Medion had the most unique test accuracy review references (references not contained in any other database) (n=328) followed by the HTA database (n=264), C-EBLM database(n= 248) and the ARIF database (n=232) (see figure 3.1). DARE had the least number of unique test accuracy review references when using the pragmatic search filter (24% of 383: n=93). The DARE database producer tagged search yielded 542 references for the same period. Extrapolating from the results of searching DARE using the pragmatic filter, an estimated 24% (130) additional references would be unique to DARE. However this does not change the DARE database's rank order for contribution of unique references (figure 3.1).



Appendix 3.4 documents the yield of resources for combinations of 2 and 3 databases. A combination of three publicly available databases (C-EBLM, Medion and HTA) yields 1232 unique references (76% of the total). A combination of Medion and the HTA database or Medion and C-EBLM yielded 948 and 952 unique references respectively (~59% of the total). The lowest yield of references was obtained by a combination of the DARE and HTA databases (561 (35%) using the pragmatic filter on the public interface of DARE or 720 (40%) based on the DARE database producer tagged estimate). However it must be noted that this low yield may be explained by the fact that DARE is a selective, quality assured resource.

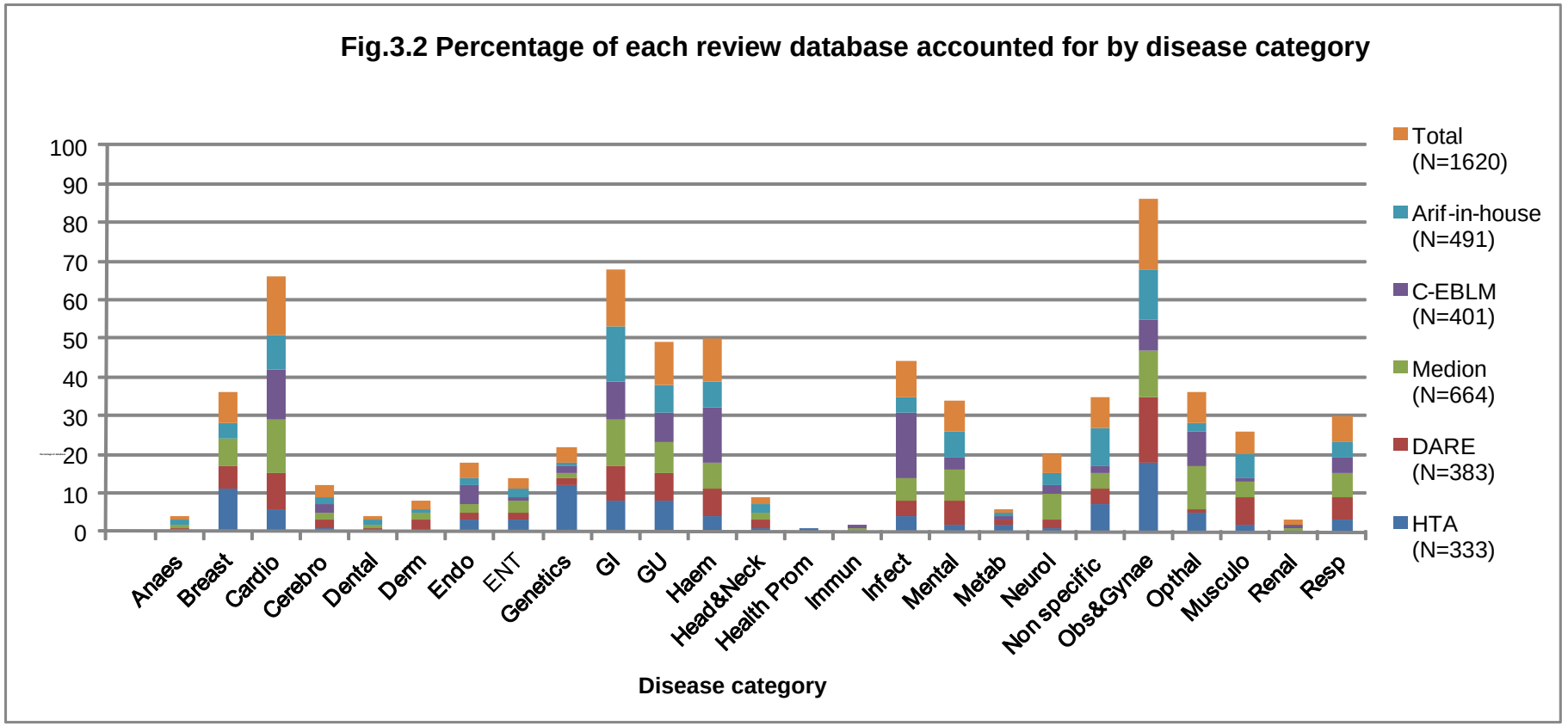
3.5.4 Characteristics of indexed test accuracy reviews

This analysis is based on the content of the 1620 test evaluation reviews identified from a combination of the pragmatic search filter in DARE and HTA; reviews tagged as 'diagnostic' in combination with the use of screening as a text word in the ARIF database; all reviews contained in Medion and all reviews contained in C-EBLM for the period 1996-2006. Additional references identified by the DARE database producer tagged searches (n=130)

were not available for scrutiny within the time available. Description of review characteristics was based on review title and where available review abstract (see appendix 3.5).

3.5.4.1 Disease topic area

Figure 3.2 illustrates a breakdown of reviews according to disease category. Obstetrics and gynaecology accounted for between 8% and 18% of reviews across databases, median 13% of reviews (18% overall). Cardiovascular disease ('cardio') and gastro-intestinal disease ('GI') accounted for between 9% and 15% of reviews, (15% overall). Ophthalmology was prominent in the Medion database (11% citations). The high proportion of reviews concerned with infectious disease ('Infec') and haematology ('Haem') in the C-EBLM database is probably a reflection of the laboratory emphasis of this database. The relatively high proportion of reviews concerned with genetic testing in the HTA database (12%) may be a reflection of this topic area as an emerging health technology [94]. It should be noted that the separate section of the Medion database devoted solely to reviews concerned with genetic testing was not included in this analysis; the number of reviews in the Medion genetics section over our period of study was 119 which would increase the proportion of genetics reviews contributed by Medion to close to 20% of the total across databases, compared to the 1% indicated in figure 3.2.



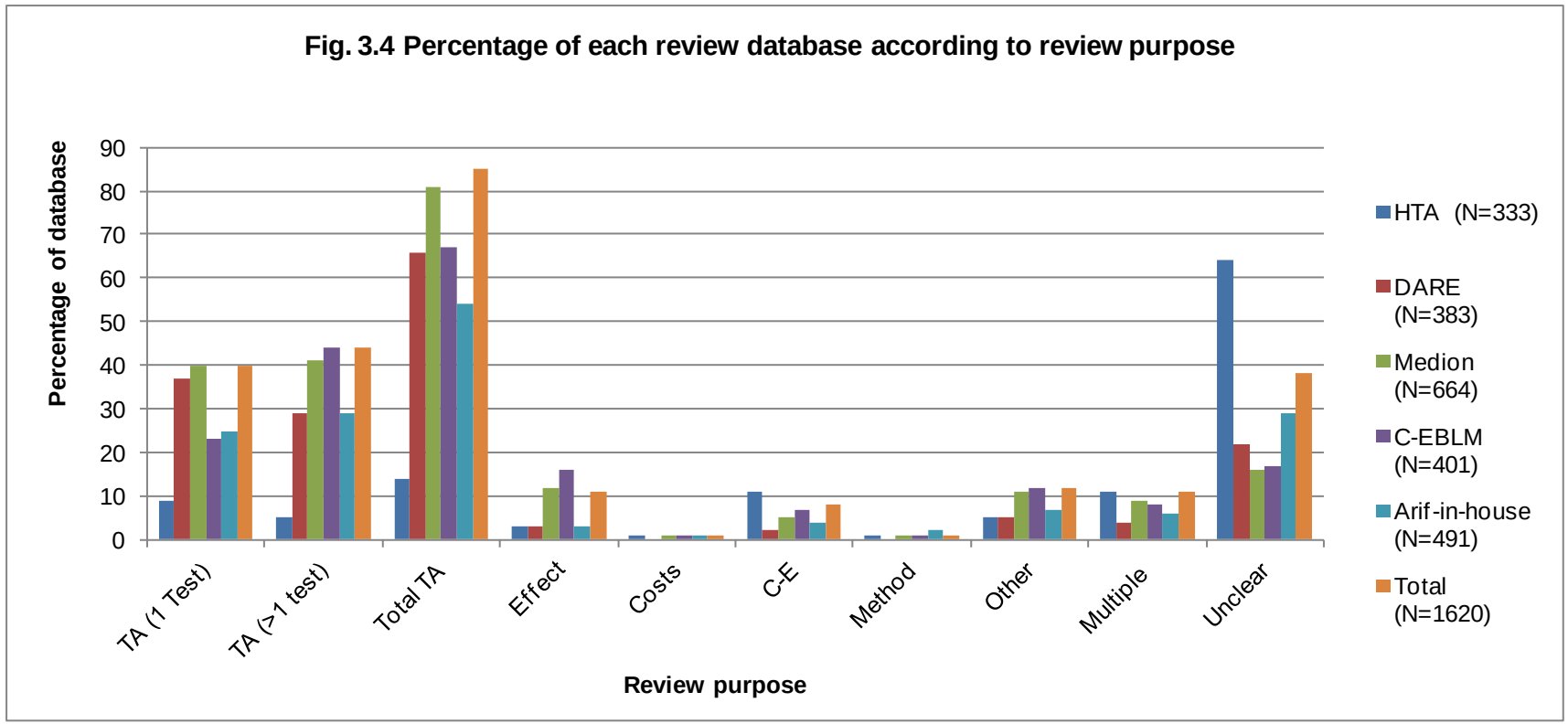
Notes to Fig 3.2: Anaes: Anaesthetics; Cardio: Cardiovascular; Cerebro: Cerebrovascular; Derm: Dermatology; Endo: Endocrinology; ENT: Ear, Nose and Throat; GI: Gastrointestinal;GU: Genito-Urinary; Haem: Haematology; Immun: Immunology; Infect: Infectious Diseases; Neurol: Neurology; Non-Specific (symptoms); Opthal:Opthalmology; Musculo: Musculoskeletal; Resp: Repiratory.

3.5.4.2 Review purpose

Table 3.3 and figure 3.4 illustrate that most test evaluation reviews were concerned solely with the estimation of test accuracy (46%-81% across databases; 85% of all citations). Relatively few test evaluation reviews were concerned with evaluation of effectiveness or costs alone, or were solely methodological in approach. The HTA database contained the highest proportion of test evaluation reviews concerned with cost-effectiveness (36%). Sixteen percent, 12% and 11% of test evaluation reviews contained in the C-EBLM, Medion and the HTA databases respectively were concerned with effectiveness. The ARIF database contained the highest proportion of methodological reviews (24%). The proportion of reviews for which the purpose was unclear was high (16-64% across five databases and 38% overall) and it is unclear what impact the accurate coding of this subset would have on the distribution of review purpose across databases.

Table 3.3: Content of review databases according to review purpose

Review purpose	HTA (N=333)	DARE (N=383)	Medion (N=664)	C-EBLM (N=401)	ARIF (N=491)	Total (N=1620)
Test Accuracy (1 Test)	29 (9%)	142 (37%)	267 (40%)	94 (23%)	121 (25%)	653 (40%)
Test Accuracy (>1 test)	17 (5%)	112 (29%)	271 (41%)	175 (44%)	143 (29%)	718 (44%)
Total Test Accuracy	46 (14%)	254 (66%)	538 (81%)	269 (67%)	264 (54%)	1371 (85%)
Effectiveness	11 (3%)	13 (3%)	81 (12%)	65 (16%)	14 (3%)	184 (11%)
Costs	4 (1%)	1 (0%)	3 (0.5%)	4 (1%)	3 (1%)	15 (1%)
Cost-Effectiveness	36 (11%)	6 (2%)	35 (5%)	30 (7%)	21 (4%)	128 (8%)
Methodological	2 (0.6%)	0 (0%)	8 (1%)	4 (1%)	10 (2%)	24 (1%)
Other	18 (5%)	18 (5%)	70 (11%)	49 (12%)	32 (7%)	187 (12%)
Multiple	36 (11%)	15 (4%)	62 (9%)	33 (8%)	30 (6%)	176 (11%)
Unclear	213 (64%)	85 (22%)	104 (16%)	68 (17%)	142 (29%)	612 (38%)

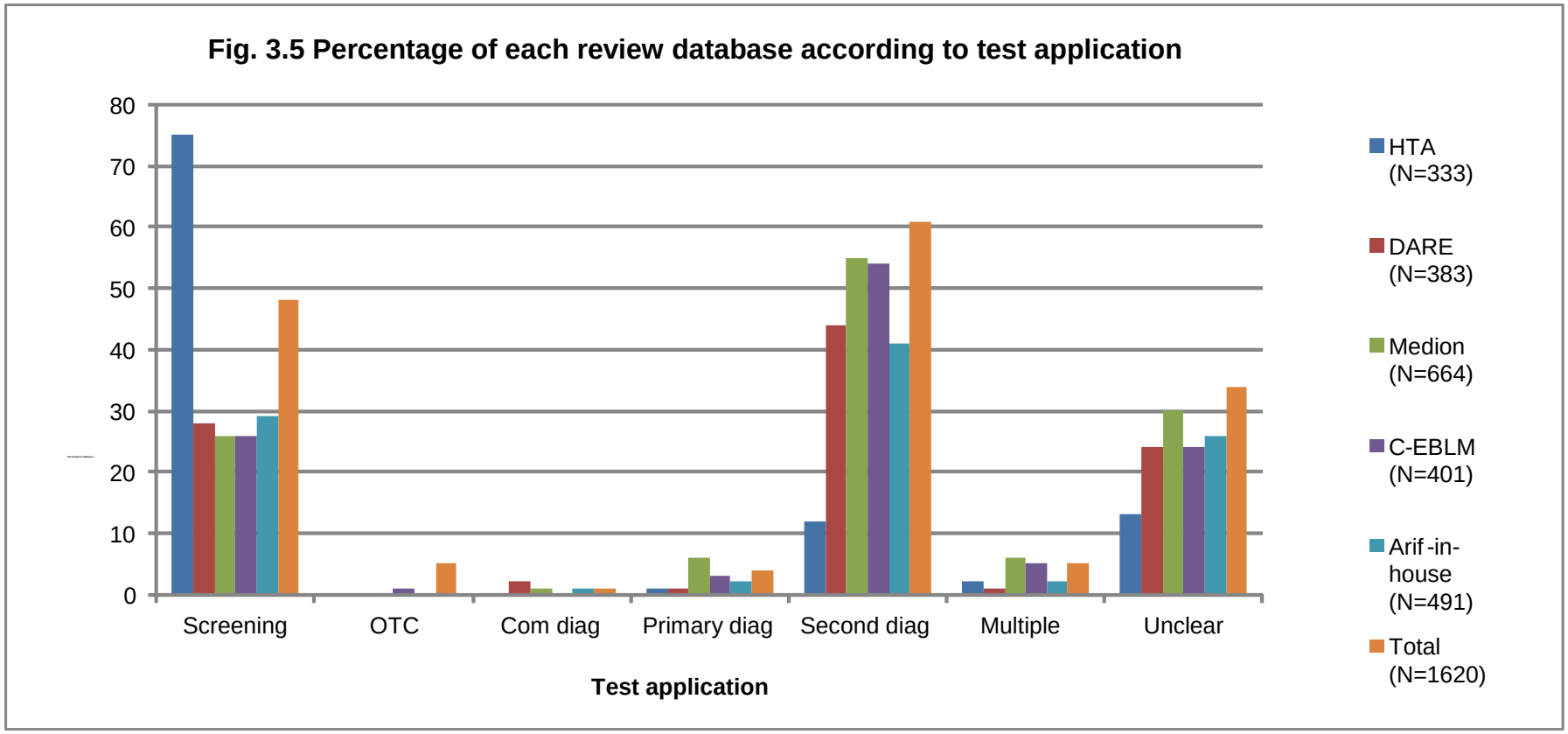


Notes to fig 3.4: TA: Test Accuracy; Effect: Effectiveness; C-E: Cost-Effectiveness; Method: Methodological review; Multiple: Review with multiple purposes

The classification scheme used to describe ‘review purpose’ did not discriminate between evaluations of tests at different stages of development. However it was evident from scrutiny of titles and abstracts that the C-EBLM database contained a larger proportion of reviews concerned with early test development, for example test accuracy employing a case control design. This was in contrast to other databases where the predominant type of test accuracy evaluation was conducted in a clinical setting, (screening, diagnosis, prognosis or disease monitoring).

3.5.4.3 Clinical setting in which tests are applied

Figure 3.5 illustrates that there was a striking preponderance of tests evaluated in secondary care and screening contexts across all databases. Overall only 4% of reviews evaluated tests for use in primary care (1%-6% across individual databases). Secondary care and screening would still dominate as research settings even if all of the reviews coded as ‘unclear setting’ were in fact evaluations of tests in primary care.



Notes to Fig 3.5: OTC: Over The Counter; Com diag: Community diagnosis; Second diag: Diagnosis in secondary care; Multiple: Multiple settings explicitly specified; Unclear: Clinical setting unclear

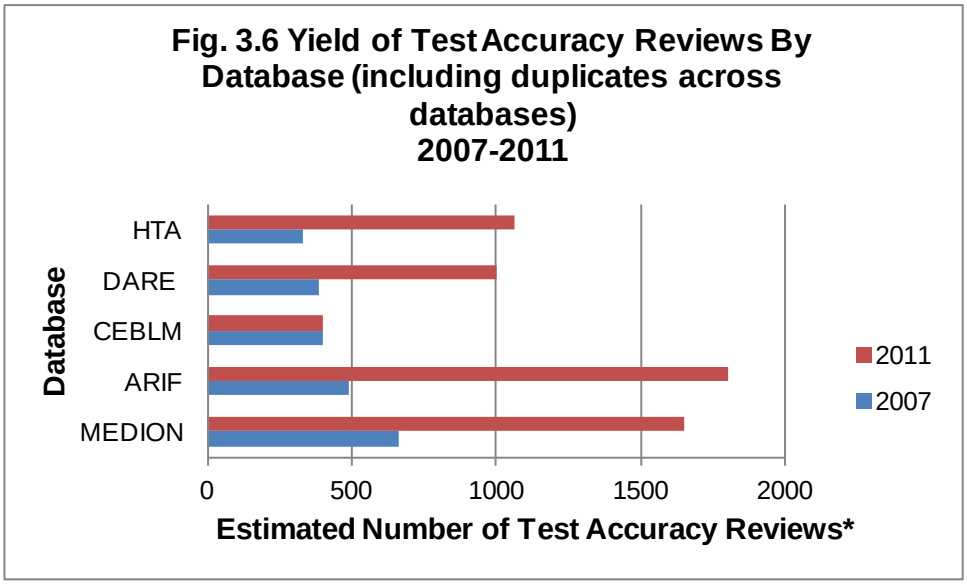
3.5.5 Retrieving Test Accuracy Reviews from Review Databases

Figure 3.6 and appendix 3.5 illustrate the yield of test accuracy reviews and the features of the five review databases at the time of conducting the research (2007) and at the time of writing (2011).

With the exception of the IFCC's C- EBLM database, which is no longer publically available, the number of test accuracy reviews has more than tripled in each database over the intervening 4 year period. In the absence of sensitive methodological search filters for use in general bibliographic databases for the location of studies of test accuracy, review repositories represent an efficient resource for researchers and decision makers.

The ARIF database remains the most up-to-date of the four remaining review repositories included in this investigation and it is now publicly available (http://www.arif.bham.ac.uk/databases.shtml).

The DARE and HTA databases continue to offer relatively sophisticated search and retrieval features but this is offset by the fact that test accuracy reviews are not tagged for public use as they are in the ARIF database. DARE is the only database to contain abstracts of reviews that have been quality-assessed, containing a summary of the review together with a critical commentary about its overall quality and as a result of this is a selective rather than a comprehensive resource. However quality assurance is likely to be an important feature for those undertaking methodological research. Medion is the longest established specialist database devoted solely to test accuracy reviews and has separate smaller databases devoted solely to reviews of genetic tests and methodology. However as at October 2011 no additions had been made to the Medion database since 2010. The IFCC's C- EBLM database with its emphasis on laboratory based tests and contribution of a large number of unique references is no longer available.



Notes to Fig.3.6: The CEBLM database is no longer publically available therefore the estimate of test accuracy reviews has been left unchanged from 2007. Estimated number of test accuracy reviews at both time points is based on the performance of our pragmatic search filter in the HTA database and searches undertaken by hosts of the ARIF and DARE databases. All records contained in the MEDION database are claimed to be test accuracy reviews. In 2007 7% of records retrieved from the HTA database, 3% of references provided by the ARIF host and 1% of MEDION records were not reviews of test accuracy.

3.6 Strengths and Limitations: Epidemiological mapping of test accuracy review characteristics

The use of pragmatic filters may have missed relevant citations in databases where the content was not solely concerned with testing, particularly for the DARE and HTA databases where reviews are not tagged according to type of review question. The impact of any omissions would be to underestimate the contribution of these databases in terms of yield in the analysis, although the possibility that a basic search would skew results towards identifying references with particular epidemiological characteristics cannot be ruled out. The pragmatic filter appears to have performed well in the HTA database although further research would be needed to verify its performance compared to other search strategies. The pragmatic filter did not perform so well in the DARE database and until access to tagged test evaluation reviews is made possible on the public interface of DARE, a more sophisticated search strategy than the one adopted here should probably be advocated.

For pragmatic reasons epidemiological characteristics of reviews were coded based only on review title and where available review abstract. As a result errors in classification may have occurred and in particular reviews that were coded as having an unclear setting or review purpose may have altered the pattern of review characteristics described here. Further, the pattern of review characteristics may have changed in the intervening 4 year period since conducting the original searches. However in the absence of initiatives to selectively encourage test evaluation in relatively neglected topic areas this is not considered likely.

The purpose of this research was to identify reviews concerned in whole or in part with test accuracy. It is likely that reviews concerned with any type of test evaluation would include the terms sensitivity, specificity or screening [82-84]. However the search strategy may have missed reviews where the focus was on test costs, test effectiveness and test cost-effectiveness.

The analysis did not include the NHS EED database or the CDSR database. CDSR does not claim to include systematic reviews of test accuracy studies although our search filter identified 16 hits from CDSR between 1996 and 2006 concerned with various aspects of

screening; it is likely that these reviews are primarily concerned with effectiveness and are unlikely to include a review devoted to accuracy alone. Using our filter in NHS EED (http://nhscrd.york.ac.uk/; accessed 30-11-11) between 1996 and 2006 identified in excess of 800 hits. Without further research it is not possible to comment on the relevance of the NHS EED citations or their content. However, as for CDSR, it is likely that the primary objective of the NHS EED citations will be assessment of costs and cost-effectiveness rather than a review devoted solely to test accuracy evaluation.

3.7 Conclusions: Epidemiological mapping of test evaluation reviews

Recent initiatives encouraging a more critical adoption of new tests and scrutiny of existing tests and testing pathways [95,96] suggest that test evaluation reviews represent a crucial contribution to the evidence base. There are an increasing number of test evaluation reviews and this research suggests that the majority of these are evaluations of test accuracy. Given the widely held concern that applying methodological search filters to capture test accuracy research does not provide adequate sensitivity for systematic review purposes, specialist review databases are an important resource for identifying relevant research. In addition review databases represent an efficient resource to support methodological research, although the unique characteristics of individual review databases and the fact that each database contributed unique references to the repository should be considered when making a choice about which resource or resources to use.

Issues pertaining to the identification of primary test accuracy research [82,83,84] appear to be pertinent to identification of test accuracy reviews in general review repositories and the considerable ambiguity conveyed by review titles in this investigation also has implications for searching. Important characteristics across all review repositories include a predominance of evaluations of single tests rather than test comparisons and a predominance of reviews concerned with application of tests in secondary care or for

screening. The paucity of research concerned with the application of tests in primary care and in community settings has recently been demonstrated to remain current [97]. This imbalance needs to be considered, mindful of the fact that the majority of testing occurs in primary care and the consequences of test errors on the number of tests performed subsequently is likely to be greater in the early stages of the diagnostic workup. Resources such as bibliographic databases can change rapidly and so a watching brief is recommended. At the time of conducting this research the first Cochrane Diagnostic Test Accuracy Review (DTAR) [98] was not published. In the intervening years, five full DTARs and 33 protocols have been published in the Cochrane Database of Systematic Reviews (CDSR) (as at 30-09-11). However currently there is no method for ascertaining a full list of Cochrane DTARs from the Cochrane library website although a visual 'diagnostic' flag appears next to DTARs identified as a result of an author, title, abstract, text or keyword search. The existence of a Cochrane handbook for systematic reviews of diagnostic test accuracy [96] and training of Cochrane review groups by the Cochrane Diagnostic Test Working Group (http://srdta.cochrane.org/welcome; accessed 30-09-11) will encourage improved review methodology although the visibility of the group and its resources in the Cochrane Library (http://www.thecochranelibrary.com/view/0/index.html; accessed 03-10-11) is currently poor. Despite increasing numbers, systematic reviews of test evaluations currently represent a small proportion of all types of systematic review [39]. This favours the timing of an initiative to develop an overarching repository of systematic reviews of test accuracy. Such a resource would be invaluable in a research field still relatively in its infancy. An alternative, more pragmatic approach would be to encourage existing primary study and reviews database producers and publishers to tag studies concerned with the evaluation of tests. Indeed such an initiative led by the members of an expert group advising on the development of the Cochrane Diagnostic Test Accuracy register of primary studies has recently successfully submitted a proposal for a specific indexing term "diagnostic test accuracy study" to Elsevier, as publishers of EMBASE. The indexing term was prospectively introduced in December

2010 [99] and the same advisory group are drafting a formal submission to the National Library of Medicine, as publishers of MEDLINE (personal communication).

Chapter 4: Methodological Review: An investigation of the extent to which clinical context shapes the conduct and reporting of systematic reviews of test accuracy

4.1 Abstract

Background

Statistical and methodological issues have, until relatively recently, dominated the test accuracy research landscape. Consideration of clinical context, (the intended setting application and role of tests, the downstream consequences of test results and the use of test accuracy measures to convey contextual information) has only relatively recently received attention. Systematic reviews and meta-analyses of test accuracy are increasing in number and prominence as a resource for diagnostic decision making and offer the opportunity to mitigate some of the current limitations of primary studies. In particular, by considered framing of research questions and by enabling a comparative approach to test evaluation they offer the opportunity to improve the contextual fit of evidence. Consideration of test accuracy in the absence of contextual information may mislead when making recommendations about test use.

Objectives

To investigate the extent to which a representative sample of systematic reviews of test accuracy represent the clinical context in which index tests are to be used when formulating a review question, deciding on synthesis methods, reporting results and making recommendations. On the basis of results to provide recommendations for how the reporting of contextual aspects of systematic reviews of test accuracy could be improved.

Methods

Published and unpublished reviews were sought by interrogation of the DARE database, the Cochrane Database of Systematic Reviews, the database of systematic reviews hosted by the Aggressive Research Intelligence Facility, University of Birmingham, the UK NHS National Research Register and contact with the Cochrane Diagnostic Test Accuracy

Working group. A final random sample of 100 reviews was included from 271 reviews eligible on the basis of title and abstract.

Results

100 reviews representing 17 disease topic areas and including between 1 and 50 index tests were included. Scrutiny of included reviews reveals ill-defined objectives which are reflected in question formulation, review synthesis (including investigation of heterogeneity) and reporting of findings. The place of index tests within a testing pathway is mostly not articulated by consideration of test role, (add, replace, triage), healthcare setting, patient presentation, prior tests or current testing practice: Seventy six percent of reviews did not state the setting in which index test were to be used and only 24% of reviews detailed all of index test application, role and prior tests as part of question formulation. Reporting of study characteristics was poor: setting, participant presentation and age were documented by just over 50% of reviews whilst chronicity and severity of the target disorder were documented by less than 1/3 of reviews. Poor reporting of primary studies was cited as a reason for this poor reporting by between 1% and 8% of reviews (depending on characteristic).

Conclusions

The findings of this review have implications for the development of standards for reporting of test accuracy reviews. There appears to be no relationship between review quality and review reporting, consideration of applicability of included studies, or completeness of review question formulation. Assessment of the internal validity of systematic reviews according to existing guidance does not appear to be a good reflection of the degree to which review authors have considered the external validity of their findings. There is a need for the development of reporting guidelines specific to systematic reviews of test accuracy; this might be achieved by an annex to the existing PRISMA reporting guidelines.

4.2 Background

Systematic reviews have the potential to offer efficient access to medical knowledge for practising clinicians and policy makers. Although systematic reviews of test evaluations represent a small proportion of reviews overall (approximately 8%) [39], their number has increased substantially over the last decade [40-42]. Systematic reviews of test accuracy are responsible for the majority of this increase and are likely to become an increasingly important source of test accuracy evidence (see chapter 3).

4.2.1 Clinical context and test accuracy

4.2.1.1 Heterogeneity

A particular challenge associated with systematic reviews of test accuracy is a consequence of the fact that test accuracy is not a fixed property and is specific to the circumstances under which a test is being applied. Contextual variation is proposed to play a greater role in the estimation of test accuracy compared to the estimation of effectiveness. Contextual variables that are potential modifiers of test accuracy encompass:

- Features of the population to be tested (population spectrum: for example age, sex, presence or absence of symptoms, disease severity, disease chronicity, prior tests).
- Features of tests being evaluated: for example technical variation of the test itself, operating threshold, the skill and experience of those operating and interpreting the test and the operating environment, for example laboratory or bedside.

Variation in prevalence of the target disorder in study populations is often a proxy for variation in population spectrum as are variations in the intended test application (screening, diagnosis, prognosis, monitoring). In addition proposed test role (add to existing tests, replace existing tests or triage for further testing) will determine population spectrum, as test role determines at what point in a care pathway a test is being evaluated (prior tests that will have been performed) as well as identifying comparator tests that should be considered. The

use of healthcare setting as a measure of clinical context is likely to be fairly congruent with characteristics of index tests being evaluated but a crude measure of variation introduced as a result of population spectrum.

4.2.1.2 The downstream consequences of test results

The application and role of tests is also an important determinant of the relative value placed on the two dimensions of test accuracy – the degree to which a positive test result increases the probability of the target disorder and the degree to which a negative test result decreases the probability of the target disorder [28]. The absolute and relative value placed on erroneous test results (false negatives and false positives) will be contextually dependent. For example the ability of a test to decrease the probability of disease is usually relatively more important when tests are applied for screening purposes where test positives receive further testing providing the opportunity for false positives to be identified but test negatives receive no further testing. Similarly the ability of a test to decrease the probability of disease is usually relatively more important early on in the diagnostic work-up. It has been suggested that sensitivity and negative likelihood ratios (LR-) may be the more important dimension of test accuracy in generalist, primary care settings. General Practitioners, in their role as 'gate-keepers' to secondary care, use tests to rule out serious disease, provide reassurance or adopt a safe watchful waiting approach rather than pursuing a precise positive diagnosis for conditions that have a high probability of being self-limiting [48,100]. Conversely the importance of specificity and the positive likelihood ratio (LR+) increases when there are severe consequences attached to false positive test results, typically at later stages in the diagnostic process where the consequences of a positive test result may be stigmatising, (for example the diagnosis of sexually transmitted diseases) and may result in initiation of treatments that may be lifelong and/or toxic. Added complexity is introduced as the two dimensions of test accuracy are affected to differing and largely unpredictable degrees by clinical context [101]. The relationship between healthcare setting, pre-test probability and the utility of test

accuracy measures has also been highlighted [102]; negative predictive values (NPVs) of comparable magnitude are likely to be of more use in generalist, lower pre-test probability settings compared to specialist, higher prevalence settings. Similarly positive predictive values (PPVs) of comparable magnitude will be of more use in specialist, higher pre test probability settings compared to generalist, lower pre-test probability settings.

4.2.1.3 Translating test accuracy to test effectiveness

Comparable estimates of test accuracy may have different policy implications in different clinical contexts according to those factors that are associated with the translation from test accuracy to test effectiveness (the impact of testing on patient outcomes): ease of access to tests, acceptability of tests to professionals and patients, training implications of introducing a new test, the availability of effective treatments, cost and the clinical and economic burden of the condition for which the test is to be used. Comparable or superior test accuracy does not equate with comparable or superior test effectiveness [91].

4.2.1.4 Incorporating context in evaluations of test accuracy

It is argued that the nature of contextual modification of test accuracy and the implications of this for the application of test accuracy estimates has to date been overlooked relative to addressing the complex statistical issues associated with meta-analysis of test accuracy [2]. Indeed the challenges of identifying and synthesising test accuracy literature have dominated guidelines for systematic reviews and meta-analysis of test accuracy [5,36,40,103,104] until relatively recently [34,101]. Research has demonstrated that lack of contextual information relevant to decision making represents an important barrier to use of evidence [30] and there has been a call for greater clarity about the intended application of tests for those attempting to use evidence about test accuracy and in particular when considering the potential impact of testing on patient management [105].

The potential to use particular properties of different test accuracy metrics to reinforce contextual considerations has also not received attention to date. Although the indiscriminate use of test accuracy metrics is suggested as a potential source of confusion [6,43], there has been no attempt to delineate single or combinations of metrics that might be more useful to convey information in specific testing contexts. For example a feature of the majority of existing summary measures of test accuracy such as sensitivity, specificity, predictive values (PVs), and the area under the curve (AUC), is that these metrics are explicit with respect to correct disease classification whilst test errors are communicated implicitly. The common practice of communicating test accuracy probabilistically may also mislead with respect to the consequences of test errors in different testing contexts: a test with a specificity of 90% and a false positive rate of 10% will result in ten times the number of false positives when pre-test probability is 1% compared to when pre-test probability is 10%.

4.2.2 The potential contribution of systematic reviews for improving the contextual fit of test accuracy evidence

Although dependent on the quality of the primary evidence base, systematic reviews offer the opportunity to improve the contextual fit of test accuracy evidence by synthesising evidence according to the intended application and role of the test under evaluation, by investigation of contextually dependent modifiers of test accuracy and by highlighting deficiencies in the evidence base. Question formulation is crucial to this process and dependent on consideration of the place of the index test in the testing pathway for the target disorder.

4.2.3 Existing research

A recent review of epidemiological characteristics of systematic reviews concluded that reporting of systematic reviews generally was inconsistent [39]. Only a minority of the reviews were concerned with diagnosis or prognosis (23/300 (8%)) and findings were not presented

separately for this subset. Although the poor reporting, quality and contextual fit of primary test accuracy studies have been well documented [2,3,34-38], the extent to which this is true of systematic reviews of test accuracy is unclear.

Recent empirical research on the reporting of systematic reviews of test accuracy confined to the cancer literature [106] did conclude that reporting of reviews of test accuracy in the cancer literature was poor and in particular the clinical setting of studies was reported in only 17% of reviews and details of included patients in only 45% of reviews. Information on disease severity was reported in a minority of studies. However in addition to being limited by topic, this research did not attempt to address the extent to which contextual factors influenced review question formulation, analysis and interpretation or the degree to which inadequacies in primary test accuracy studies contribute to inadequacies in the conduct and reporting of systematic reviews of test accuracy.

4.3 Aims and Objectives

Aims

The aim of this review is to investigate the degree to which clinical context shapes the conduct and reporting of systematic reviews of test accuracy. Clinical context encompasses contextual variables that are potential modifiers of test accuracy, prevalence of the target disorder in study populations and the intended role and application of the test under evaluation including downstream consequences of test results.

Objectives

- Identify a sample of systematic reviews of the accuracy of tests applicable to the primary healthcare setting, representative in terms of quality and target disorder.
- Assess the extent to which reviewer authors have considered clinical context at each stage of the review process:

  - formulation of review question (background, inclusion and exclusion)

  -synthesis, including investigation of heterogeneity (methods; results)

  -summarising and discussing results (results; discussion)

  -making recommendations (discussion; recommendations)
- Assess the extent to which the quality and reporting of primary studies of test accuracy impact on the contextual fit of systematic reviews of test accuracy.

4.4 Methods

4.4.1 Search strategy

Published reviews were sought by interrogation of the DARE (Database of Abstracts of Reviews of Effectiveness, Centre for Reviews and Dissemination, University of York) [107] database, the Cochrane database of systematic reviews via the Cochrane library 2006 Issue 3 [108] and the database of systematic reviews hosted by the Aggressive Research

Intelligence Facility ('ARIF') at the University of Birmingham (West Midlands Commissioning Support Unit 2011 [109]. Searches were carried out in January 2007 and limited to the period 1996-2006. Given the number of references likely to be retrieved and the difficulties caused by poor indexing of systematic reviews of tests already described, searches of DARE and the Cochrane database of systematic reviews were limited to MeSH index terms to make them as specific as possible. The text word 'mass screening' was added to the MeSH term 'Sensitivity and Specificity' (exp) [93] in order to capture a variety of testing applications. The MeSH term 'Diagnosis' (exp) or text word 'diagnostic' greatly reduced the specificity of the searches and so these terms were not used. The ARIF database has been running since 1996 and relies mainly on the weekly alerting services of ZETOC (British Library), Science Direct and PUBCrawler (PubMed). The ARIF database does not have a controlled vocabulary and was searched on the subset *diagnosis* (keyword) as well as the text word *screening*. (See chapter 3 and appendix 3.1 for further details of the search strategy).
The rationale for the selection of these 3 databases followed an interrogation of 5 specialist review databases (see chapter 3) and is as follows:

- Searching these specialist databases was an efficient way to achieve a representative sample of reviews of test accuracy. The search strategy was not designed to be comprehensive of all reviews of test accuracy.
- DARE has a comprehensive strategy for identifying reviews. However inclusion in the database is dependent on reviews meeting 3 of a possible 4 criteria encompassing inclusion of primary studies following the PICO framework, an adequate search strategy, consideration of study quality and presentation of sufficient detail about included studies [107]. The ARIF database has a less comprehensive capture strategy but does not restrict the type of reviews it contains, therefore ensuring a broader representation of reviews.
- The Cochrane Database of systematic reviews was interrogated as although this is primarily a database of effectiveness reviews, a small number of reviews primarily

concerned with screening are contained in the database that might not be captured by searches of the other databases interrogated for this review.

The Cochrane methods database was not considered an important source of test accuracy reviews at the time of searching.

Unpublished reviews were sought by interrogating the National Research Register http://www.nrr.nhs.uk/search.htm 2006 issue 3 and by contacting the Cochrane Collaboration Diagnostic Test Accuracy Working Group.

All searches stopped September 2006.

4.4.2 Inclusion / Exclusion:

To be included reviews had to be concerned in whole or in part with estimation of test accuracy. It was recognised that this strategy might miss reviews primarily concerned with the impact of tests on treatment effectiveness and cost-effectiveness.

As methodological quality was not the focus of this review, reviews not adhering fully to accepted systematic review methods were not excluded but documentation of key aspects of methodological quality were noted. Nine items taken from those used in the AMSTAR checklist [110] and the original QUOROM checklist [111] (current at the time of undertaking this review), were used to score included reviews as it was hypothesised that quality may have an impact on review conduct. Reviews not using a recognised reference standard were not excluded; use of a recognised reference standard is not always possible or appropriate under certain clinical circumstances and is not a pre-requisite for consideration of the clinical context in which a test is to be used.

Generalist settings represent an important part of the diagnostic work up process where the cumulative volume of test errors and in particular their contribution to further testing will be substantial. A mapping exercise of the epidemiology of existing systematic reviews of test accuracy suggested a paucity of test evaluations in the primary care setting, (Chp. 3) [112]. The

practicalities of recruiting sufficient numbers of eligible participants from a broad spectrum of patients [113,114], access to reference standard tests and the necessity for multiple reference standards are possible explanations for this under-representation. In addition patients presenting in primary care are at the beginning of a diagnostic work-up with a greater range of differential diagnoses and testing is often symptom rather than target disease based. By contrast, testing in secondary care is characterised by a narrower disease spectrum following a referral process and the emphasis is on deriving a definitive diagnosis, often relying on a smaller repertoire of tests. In order to capture as diverse a spectrum of disease and test type as possible and to ensure that specific challenges associated with evaluation of tests in generalist settings were represented, reviews were only included if they were concerned with the evaluation of tests that were considered accessible (directly, without the need for consultation with a specialist) to primary care professionals. Few (if any) tests that can be accessed by primary care professionals are not also available in secondary care settings.

Papers were initially screened on the basis of title alone to determine whether they included a review of test accuracy. Potentially relevant reviews were categorised, on the basis of the clinical experience of the author, according to whether the test under evaluation could be applied:

- in the primary care setting or available to the primary care physician via referral but where the primary care physician would normally be responsible for any changes in patient management following a test result (included)
- in the secondary care setting only (excluded)
- for screening as part of national screening programmes. It was considered that the contextual considerations associated with tests at this stage of evaluation would have been well rehearsed as part of the criteria for evaluating screening programmes (National Screening Committee 2009)[115] (excluded)

- in non medical settings such as dentistry (excluded)
- for screening in the primary care setting outside of a UK national screening programme (for example alcohol abuse) (included)

In some instances it was not clear where the main responsibility for performing and interpreting a test lay. In such instances a decision to include was based on the probability that primary care professionals were likely to take responsibility in some testing situations. Whilst recognising that this categorisation would exclude some tests where primary care physicians may be requested by patients to perform and / or interpret test results the process was pragmatic and designed to provide an unambiguous and representative rather than comprehensive sample of reviews of test accuracy relevant to the primary care setting. Reviews were further categorised according to the target disorder being tested for. Appendix 4.1 illustrates the pro-forma used to ensure consistency of inclusion decisions.

A random 150 reviews were initially sampled from the 271 included on the basis of title and abstract, with the aim of achieving a final sample of 100 reviews. Depending on the number of exclusions at full text stage further random samples were to be taken until a minimum 100 reviews had been included.

After obtaining full copies, reviews not concerned in whole or part with estimation of test accuracy or with tests that directly accessible to primary care professionals were excluded. Exclusion decisions at full copy stage were performed in duplicate.

4.4.3 Data extraction

Data extraction was undertaken using a pre-piloted electronic ACCESS data extraction form by a single reviewer. Information was collected on the index test(s), reference standard(s) used, number of included studies, search strategy, and whether testing context was considered at each point in the review methods (question formulation; inclusion and

exclusion process; data synthesis including investigation of heterogeneity; presentation of results; discussion and recommendations).

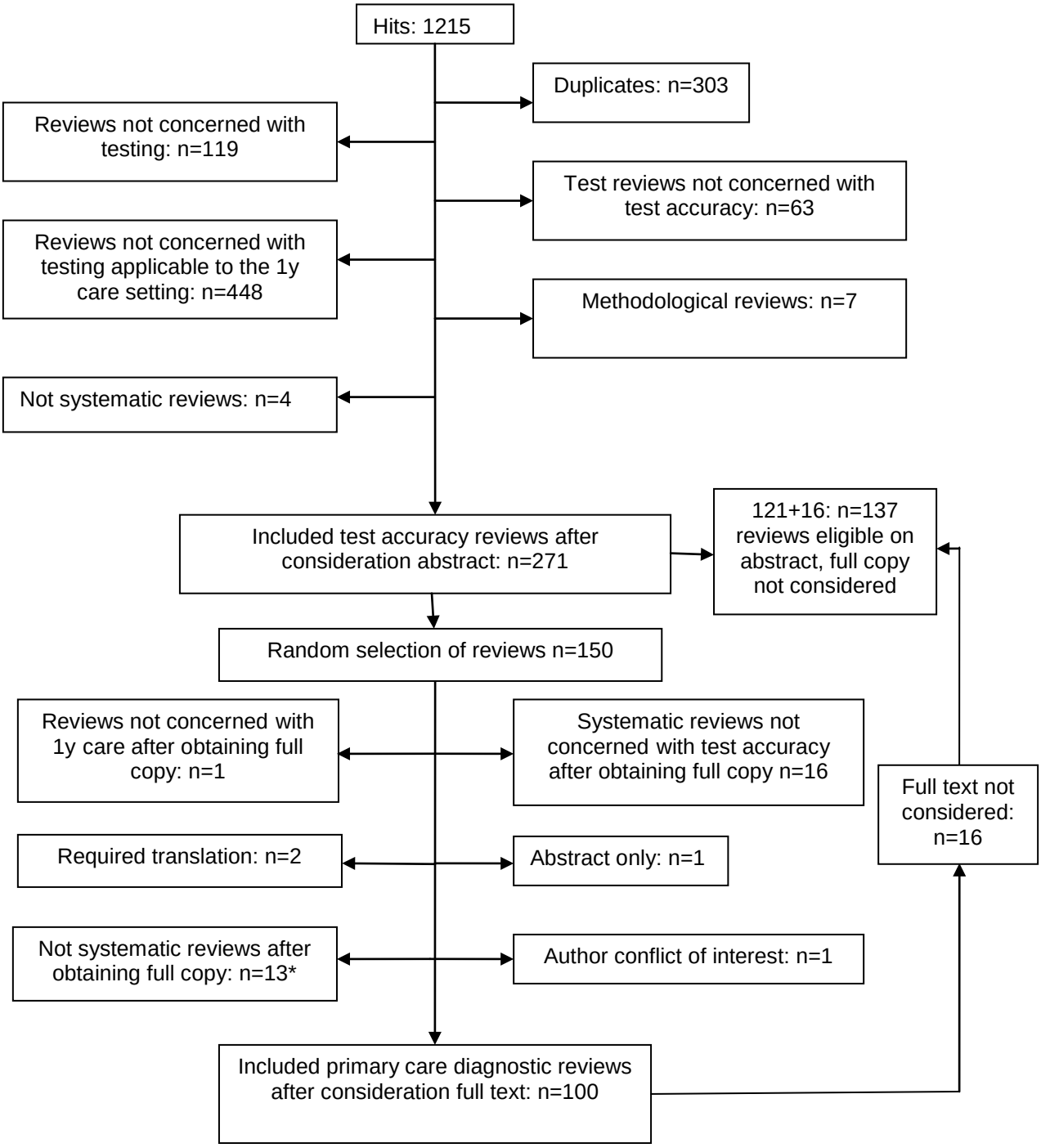
If publications referred to additional electronic files or previous publications these were consulted for information on review conduct. However, with the exception of determining whether a clinician was included on a review team, contact with authors was not considered appropriate as this could have introduced bias given the subjective nature of much of the data being extracted.

4.4.4 Synthesis

Synthesis was narrative. Findings were discussed under the following headings: demographic details of included reviews; question formulation; reporting of study characteristics; outcome reporting; contextualisation of review findings (including investigation of heterogeneity).

4.5 Results

Fig 4.1: Study Flow



* **Notes to fig 4.1:** This includes modelling studies where systematic searches performed to populate the model identified a systematic review which was used to estimate test accuracy; modelling studies where a single study was chosen to populate the model on the basis of quality or relevance; modelling studies where searches used to populate the model were not systematic.

4.5.1 Study Flow

Figure 4.1 documents the volume of literature encountered at successive stages of the inclusion process. No relevant research was identified from the National Research Register and the Cochrane Collaboration Diagnostic Test Methods group did not provide any unpublished accuracy reviews for consideration. A total of 18 potentially relevant reviews were identified in the Cochrane database of Systematic Reviews, 522 from the DARE database and 675 from the ARIF database.

4.5.2 Characteristics of included reviews

For characteristics of included studies tabulated by review see appendix 4.2.

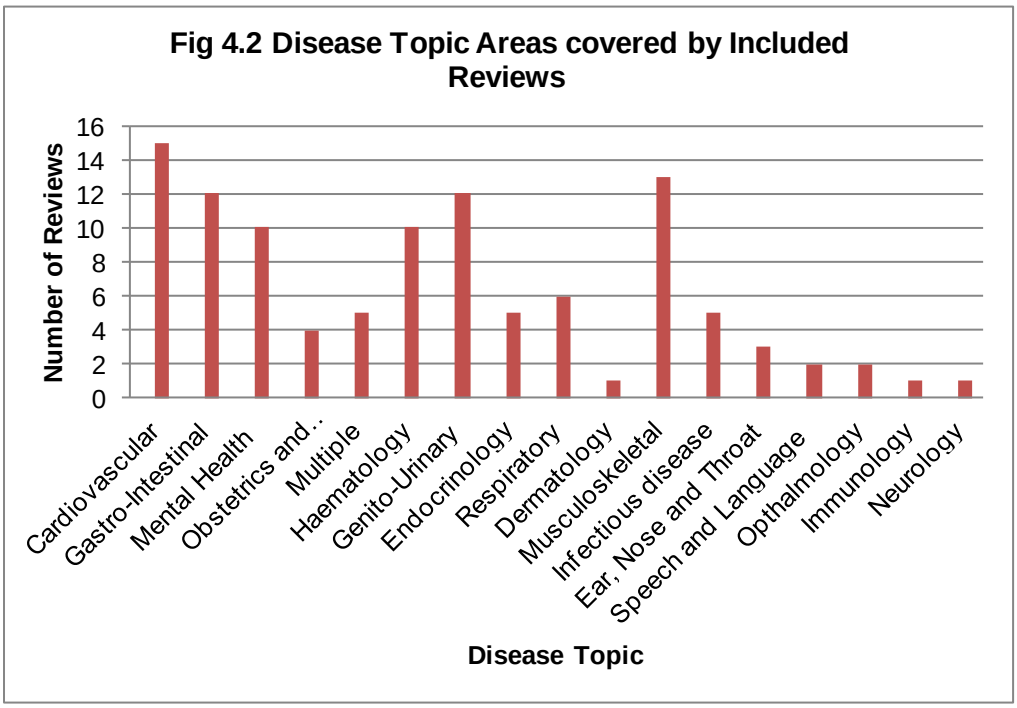
4.5.2.1 Authorship, date, type and place of publication of included reviews

The date of publication of the 100 reviews spanned 1990 to 2006; 23% of reviews before 2000 and 73% on or after 2000. Eleven reviews were undertaken as part of a health technology assessment. The majority of reviews (43/100) were conducted in the USA, 23 in the UK, 12 in the Netherlands and eight in the rest of Europe, six in Australia, four in Canada, two in Peru and one each in Columbia and China. A clinician was not represented in the author contact details of one included review and in a further seven reviews it was unclear whether a clinician contributed. In all remaining reviews (94/100) there was representation from at least one clinician.

4.5.2.2 Disease topic areas covered by included reviews

A total of 16 disease topic areas were represented (see figure 4.2). The distribution of disease topic areas differs from that observed in a concurrent sample of reviews compiled without the ‘accessible to primary care professionals’ restriction applied to selection (3.5.4.1). The greatest difference observed in this selected sample of reviews is the markedly fewer

number of reviews concerned with obstetrics and gynaecology; a finding that might be expected. In addition relatively fewer reviews in the sample selected for this methodological review were concerned with infectious disease and a relatively greater proportion with musculoskeletal disorders.



4.5.2.3 Healthcare setting represented by included reviews

Despite an attempt to include reviews of tests directly accessible to primary care, only a minority of reviews (20/100) were explicitly concerned in whole or part with evaluation of the clinical history and examination (TAR2; TAR5; TAR9; TAR11-TAR13; TAR16; TAR25; TAR29; TAR38; TAR47; TAR49; TAR50; TAR55; TAR61; TAR75; TAR77; TAR79; TAR85; TAR89; TAR93). In addition only five reviews were conducted from the perspective of evaluating the utility of symptoms and signs for a variety of target disorders (TAR5; TAR38; TAR47; TAR55; TAR93). Two reviews successfully approached this task by restricting consideration of target disorders to those representing important rule out diagnoses for low back pain in generalist settings (vertebral cancer, spinal infections, inflammatory spondyloarthropathies, compression fractures, herniated discs and spinal

stenosis [TAR38; TAR93] and restricting inclusion to primary studies concerned with only one of these target disorders. However three reviews were compromised by inclusion of primary studies that failed to follow up negative test results or where positive test results were not confirmed by a reference standard due to the possibility of multiple target disorders [TAR5; TAR47; TAR55]. These three reviews reported the (true positive (TP) + false positive (FP)) / all tested (termed diagnostic yield or detection rate) as an outcome when it was not possible to derive test accuracy. Two further reviews using this outcome measure [TAR24; TAR34] did so alongside test accuracy in the context of discussing the consequences of positive test results (see also 4.5.5.3).

4.5.2.4 Index tests and number of included primary studies

Between one and 50 index tests were evaluated by a single review (median 3) (see figure 4.3). In five reviews the number of included studies was not stated and the number of studies included was unclear in a further two reviews. In 22 reviews the number of participants was not stated. The number of included studies reported by 93 reviews ranged from 0 to 213 (median 24; inter-quartile range 13-47) (see figure 4.4) and the total number of included participants reported by 78 reviews ranged from 0 to 211369 (median 5620; inter-quartile range 2328-15020).

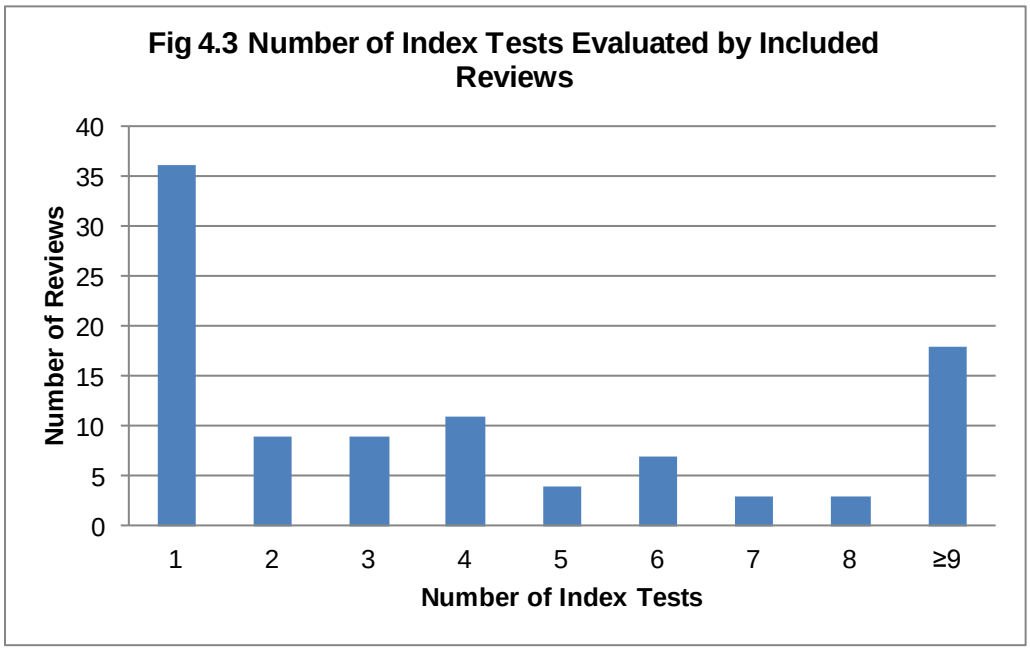


Fig 4.3 Number of Index Tests Evaluated by Included Reviews
Number of Reviews
40
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7
8
≥9
Number of Index Tests

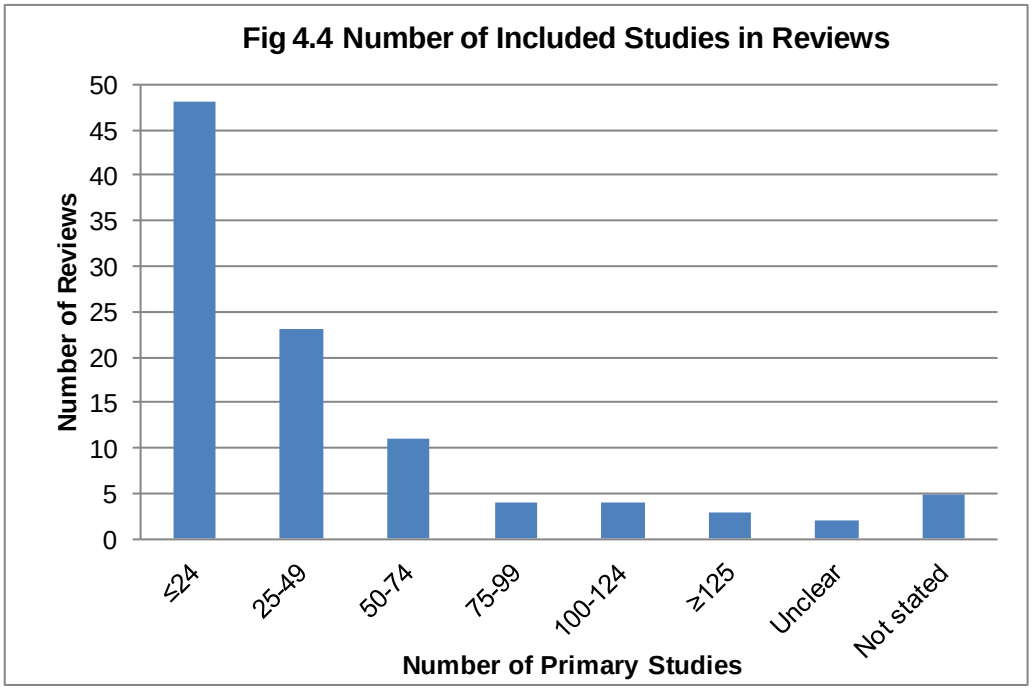
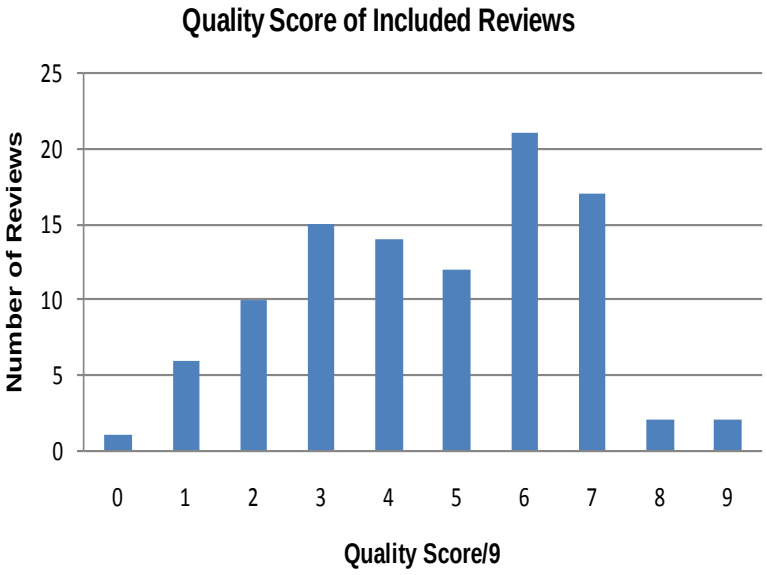


Fig 4.4 Number of Included Studies in Reviews
Number of Reviews
50
45
40
35
30
25
20
15
10
5
0
≤24
25-49
50-74
75-99
100-124
≥125
Unclear
Not stated
Number of Primary Studies

4.5.2.5 Quality of included reviews

Quality of included reviews was assessed using nine criteria taken from QUOROM [111] and AMSTAR [110] checklists (see figure 4.5 below) which were current at the time the research was conducted. The quality score assigned to the reviews ranged from the 0-9/9 (median 4.6; inter-quartile range 3 to 6).

Fig 4.5 Quality of included reviews

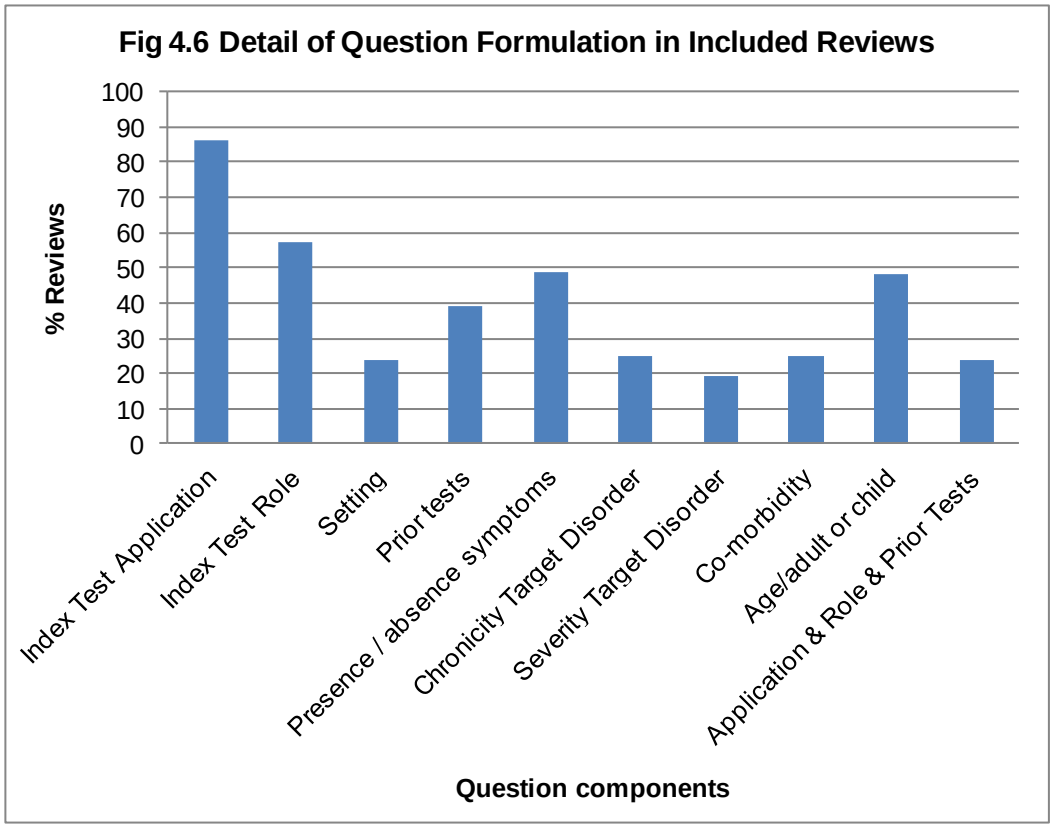


Quality Criteria taken from QUOROM (Moher 1999[111]) & AMSTAR (Shea 2007[110])
Use of >1 bibliographic database
One or more of hand searching; contact with experts / authors; reference checking
Inclusion / Exclusion criteria explicitly stated
Data extraction involves more than one person
Quality assessment undertaken
Level of agreement inclusion
Level of agreement quality assessment
Study flow reported
Potential limitations of review process discussed

4.5.3 Quality of Question Formulation

For quality of question formulation tabulated by included review see appendix 4.3

In the majority of reviews judgements about the clarity of question formulation were based on information from both background and methods sections. Figure 4.6 summarises the level of detail included in included reviews as part of question formulation.



4.5.3.1 Test application

A total of 86 reviews included detail about index test application. Fifty nine of 100 reviews stated they were evaluating tests to be used in the diagnosis of a target disorder or disorders, 16 for screening and one for prognosis. In eight reviews more than one application was stated and it was mostly unclear whether this represented lack of clarity about the intended application of the test or inconsistent use of terminology. For example the concepts of diagnosis and prognosis overlap, particularly when reference standards are applied distant

to index test diagnosis as part of clinical follow up. In six reviews the intended application of the test was not specified.

4.5.3.2 Test role

Assessment of test role was mostly based on implicit information provided in the review and was not explicitly addressed by review authors. Fifty seven of 100 reviews included detail about index test role; 26 reviews were concerned with evaluating a test as a replacement, 21 evaluating a test as an addition, eight evaluating a test for triage and in two reviews multiple included index tests had different roles.

In 35/100 reviews the intended role of the test was unclear and in 8/100 reviews was not specified. In over half of reviews where test role was not clearly detailed, information about tests usually performed prior to the index test was also unclear or not specified (25/43). Although lack of clarity of test role may be part of a review question, this was not clearly articulated for any of the included reviews at question formulation stage.

4.5.3.3 Prior tests

In 39 reviews prior tests were clearly detailed as part of question formulation. Twenty four of 100 included reviews did not specify tests usually performed prior to the index test and in 37/100 reviews other tests used in the diagnostic work up of a condition were mentioned but the testing pathway was unclear.

4.5.3.4 Setting

The proposed setting in which an index test is to be used is a crucial and basic element of question formulation as even within healthcare settings other patient and test characteristics can vary considerably. Setting was the least well articulated component of review questions. Of the 24 reviews specifying settings 10 were to be used in primary care, two in secondary

care, three in the community and in nine reviews more than one setting was specified. Twenty nine of 100 reviews did not specify a setting and in 47/100 reviews the proposed setting was unclear.

4.5.3.5 Spectrum

As healthcare setting can conceal important variation in spectrum, review inclusion criteria were also interrogated for specification of more detailed spectrum characteristics. Twenty five of 100 reviews specified chronicity as part of inclusion criteria, 49/100 the presence or absence of symptoms, 48/100 age, 19/100 target disorder severity and 25/100 presence or absence of co-morbidity. No reviews specified a prevalence range as part of inclusion criteria which may reflect lack of clarity concerning setting, an appreciation of the range of prevalence rates typically encountered in primary test accuracy studies, or the limitations of prevalence as a measure of spectrum (see 4.5.4 below).

4.5.3.6 Inclusion of key components of question formulation: test application; test role and prior tests

In summary only 24/100 reviews clearly specified all of test application, test role and prior tests as part of question formulation; 26% of the 73 reviews published on or after 2000 (TAR2; TAR3; TAR9-TAR11; TAR15; TAR21; TAR23-TAR25; TAR32; TAR33; TAR52; TAR56; TAR60-TAR62; TAR65; TAR78) and 22% of the 23 reviews published before 2000 (TAR5; TAR26; TAR35; TAR77; TAR93).

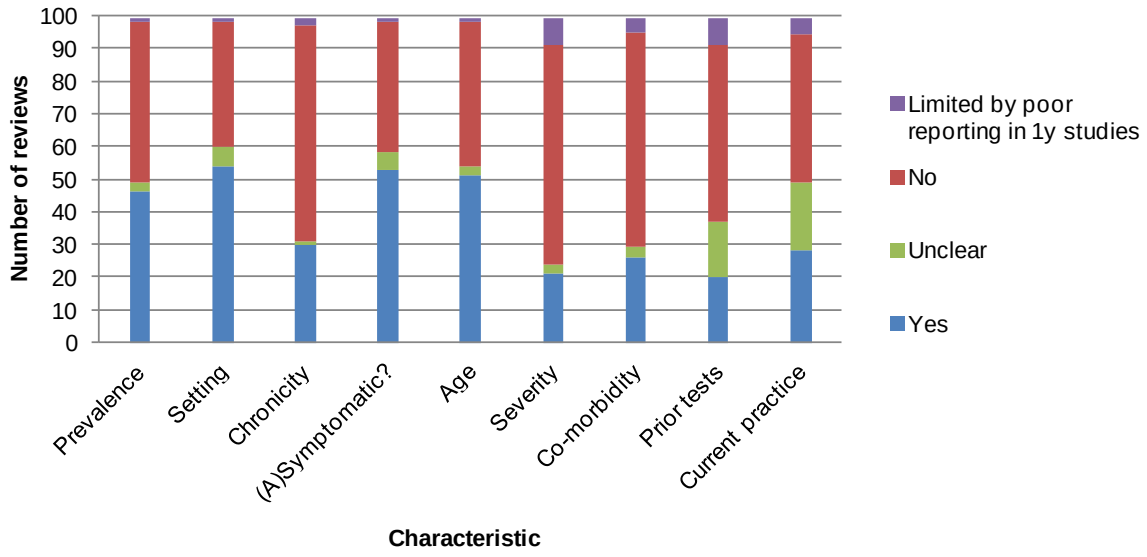
4.5.4 Reporting of primary study characteristics

For reporting of primary study characteristics tabulated by included review see appendix 4.4. One included review (TAR75) reported finding one poor quality relevant study for which no results were reported and therefore review reporting of study characteristics (4.5.4), use of outcome measures (4.5.5) and discussions concerning the contextualisation of review

synthesis and applicability of review findings (4.5.6) is based on a denominator of 99 test accuracy reviews.

Figure 4.7 provides a summary of reporting of study characteristics either in tabular form or discussed in the text for the 99 reviews. Setting, participant presentation and age were documented by just over 50% of reviews. Chronicity and severity of the target disorder, participant co-morbidity and tests performed prior to the index test were documented by less than a third of reviews. Failure to document study characteristics due to limitations in reporting by primary studies was cited by a very small number of reviews (1-8 reviews per characteristic).

Fig 4.7 Reporting of study characteristics in test accuracy reviews

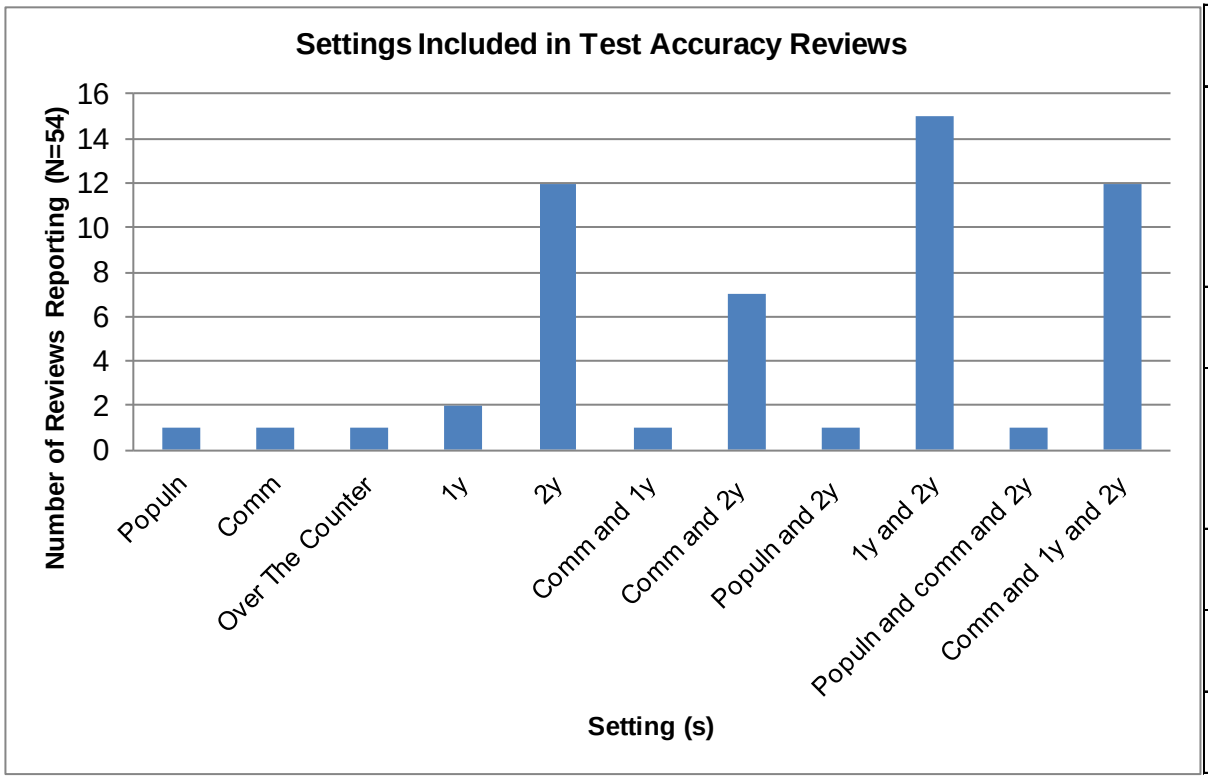


Forty three of 99 reviews commented on the quality of reporting in primary, included studies. Five of 43 reviews commented that primary study reporting was moderate to good, one review commented that primary study reporting was variable and 37/43 reviews commented that primary study reporting was poor: 4/37 reviews did not give further details; 10/37 reviews commented on poor reporting of study methods; 22/37 commented on poor reporting of

aspects of spectrum; 12/37 reviews commented on poor reporting of index test details and 5/37 reviews commented on poor reporting of reference standard details.

Fifty four of 99 studies reported the healthcare setting of included studies and for the majority of these (38/54) more than one healthcare setting was included. Health care settings included population level application of tests, for example screening; tests administered by health professionals in the community (for example testing for sexually acquired infections); tests obtained over the counter (for example pregnancy testing); tests administered following contact with a primary healthcare professional and tests administered in secondary care settings. Figure 4.8 illustrates a breakdown of settings represented in these 54 reviews. Tests restricted to use in the community, primary care and ‘over the counter’ were represented by only 5/54 reviews and tests restricted to use at population level by 1/54 reviews. Forty eight of 54 reviews reporting setting included secondary care populations.

Fig 4.8: Settings Included in Test Accuracy Reviews



Setting combination	Example	Study
Populn and 2y	Abdominal symptoms and gallstones (screening and diagnosis)	Berger 2000
Comm and 1y	Screening for type II diabetes	Harris 2003
Comm and 2y	Whispered voice test in the detection of hearing impairment	Pirozzo 2003
1y and 2y	Assessment of shoulder pain	Dinnes 2003
Comm and 1y and 2y	Detection of domestic violence	Ramsay 2002
Populn and comm and 2y	Detection of dementia	Peterson 2001

Notes to Fig 4.8: Populn: population; Comm: community; 1y: primary care; 2y: secondary care.

Prevalence of the target disorder varied widely across included studies, including the minority of reviews (6/55) including only one healthcare setting. The six reviews reporting prevalence within a single setting were restricted to secondary care and variation in prevalence of the target disorder across included studies ranged from 33% to 76% (TAR2; TAR14; TAR30; TAR33; TAR79; TAR96). This variation of prevalence, even in reviews restricting themselves to single settings, highlights the importance of detailing the characteristics of participants, as opposed to healthcare setting only, in order to convey the spectrum variation.

4.5.4.1 Summary: reporting of primary study characteristics by review authors

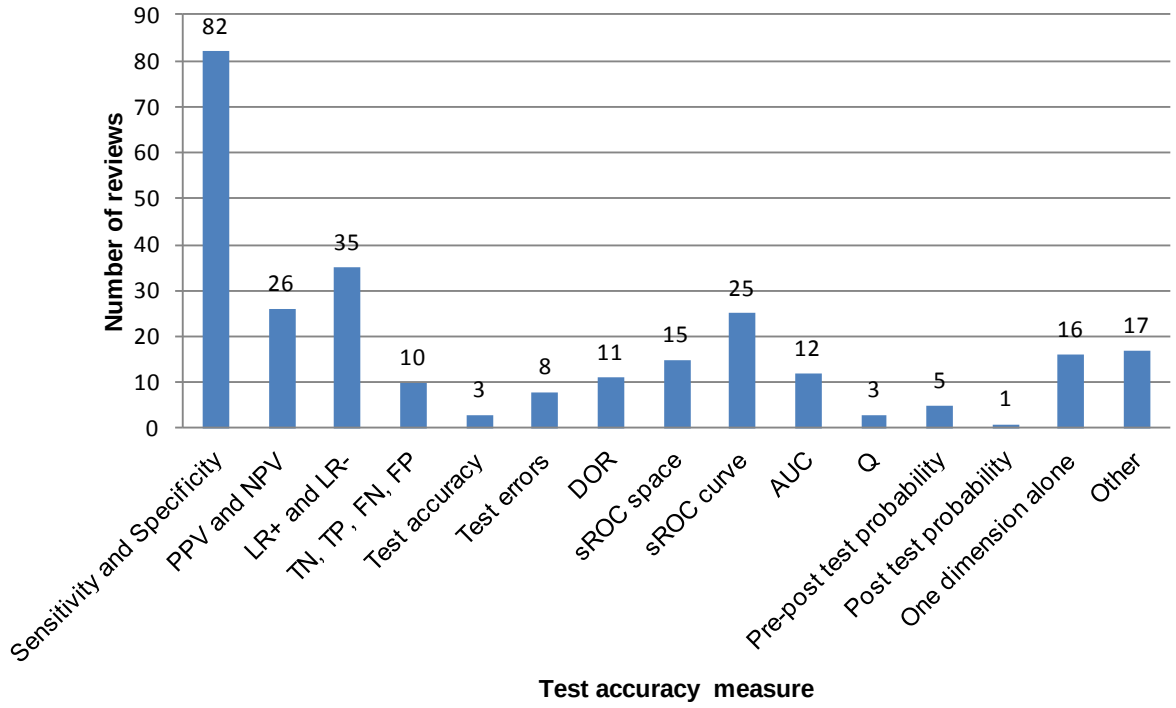
To some extent the importance of recording of individual study characteristics for an assessment of applicability of review findings will vary by review topic. However presentation (symptomatic, asymptomatic or both), healthcare setting and tests performed prior to the index test in included studies could be considered key characteristics. Only 9/99 reviews recorded all of these details (TAR9; TAR15; TAR22; TAR33; TAR40; TAR67; TAR73; TAR78; TAR94) and a further three reviews explicitly stated that poor reporting in primary studies prevented them from doing this (TAR18; TAR50; TAR76).

4.5.5 Use of outcome measures

For use of outcome measures tabulated by included review see appendix 4.5.

One of the 100 included reviews reported finding one poor quality relevant study for which no results were reported [TAR75]. Use of outcome measures, reporting of study characteristics and contextualisation of review synthesis is therefore presented for 99 reviews. The frequency of use of individual outcome measures by the 99 reviews is illustrated in figure 4.9.

Fig 4.9: Test accuracy measures used in 99/100 included reviews



Notes to Fig. 4.9: PPV: positive predictive value; NPV: negative predictive value; LR: likelihood ratio; TN: true negative; TP: true positive; FN: false negative; FP: false positive; DOR: diagnostic odds ratio; ROC: receiver operator characteristic; AUC: area under the curve

4.5.5.1 Synthesis and use of outcome measures

Just over half of the reviews with included studies (60/99) proceeded to meta-analysis. Only 60/99 reviews included confidence intervals for some or all outcome measures. A minority of reviews, (3/99) compared tests using relative pooled accuracy measures: relative Diagnostic Odds ratio (rDOR) [TAR33; TAR59] or difference in pooled estimates of sensitivity and specificity [TAR48].

The majority of reviews (83%) reported sensitivity and specificity, followed by likelihood ratios (LRs) (35%), predictive values (PVs) (26%) and the summary Receiver Operator Characteristic (sROC) curve (25%). Outcome measures illustrating both dimensions of test accuracy (sensitivity and specificity, LRs, PVs, the constituents of the 2x2 table and sROC curves and plots) were preferred over global measures of test accuracy (Diagnostic Odds Ratio (DOR), Area Under the Curve (AUC), Q) although only 41/99 reviews explicitly distinguished between these two dimensions of accuracy for the intended application and role of index tests. A minority of reviews explicitly reported test errors (15%). Only 6% of reviews illustrated the change in disease probability pre to post index test result.

A single dimension of test accuracy (a measure of the degree to which a positive test result increases disease probability or a measure of the degree to which a negative test result decreases disease probability) was reported on 17 occasions by 16 reviews. In five reviews concerned with screening or use of an index test in a triage role the reason for reporting a single dimension of accuracy (PPV in the absence of NPV and sensitivity in the absence of specificity) was due to some or all included studies provided information only on test positives (TAR24; TAR32; TAR67; TAR78; TAR81).

4.5.5.2 Consideration of the downstream consequences of test results

Less than half of reviews made an attempt to link test accuracy to clinical decision making. Only 41/ 99 reviews with included studies made a clear distinction between the ability of a test to increase the probability (rule in) the condition being tested for and the ability of a test to decrease the probability (rule out) the condition being tested for. Forty four of 99 reviews discussed the consequences of test results, (some or all of true positives, false positives, true negatives, false negatives). Where the reason for not distinguishing between the two dimensions of accuracy was due to limitations in primary studies (5 reviews), this was either where tests were being used in a screening context and index test negatives were not verified (TAR24), where the reference standard was invasive and test negatives were not

verified [TAR78] or where a test was being used to detect multiple underlying target disorders, requiring multiple reference standards to comprehensively verify index test results [TAR5; TAR55; TAR47].

In three reviews authors clearly articulated a preference for one or other dimension of accuracy. For example LR- alone was reported by one review, (combined use of d-dimer testing and estimation of clinical probability in the diagnosis of deep vein thrombosis) [TAR21], and sensitivity alone [TAR24] or false negative (FN) rate alone [TAR4] were reported by two reviews concerned with screening, where the ability to rule out the target disorder was clearly articulated as the more important dimension of test accuracy. LR+ alone was reported by one review, (a meta-analysis of the performance characteristics of the free prostate-specific antigen test) ([TAR45]) where review authors commented that false positives (FP) are the more important test error.

In one review, (a meta-analysis of the papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis) authors illustrated variation in PPV with prevalence to discuss when further testing with wet mount following a positive papanicolaou smear might be considered appropriate [TAR99].

In five reviews [TAR44; TAR51; TAR55; TAR58; TAR73] it was unclear why a single dimension of accuracy was reported when data were available to derive both dimensions

4.5.5.3 Less common outcome measures used by review authors

Seventeen reviews used a total of seven outcomes 'other' than those specified in figure 4.9 and these are detailed in table 4.10.

Six of the 17 reviews reporting outcome measures other than those in figure 4.9 were method comparison studies where both index and reference standard tests were on the same continuous scale [TAR3; TAR6; TAR34; TAR39; TAR53; TAR54]. Three of these reviews [TAR53; TAR34; TAR3] reported the limits of agreement [116] in addition to correlation coefficients as provided in primary studies. In one review concerned with early test development, a statistical

comparison of mean index test scores in diseased and non-diseased individuals was reported in addition to correlation coefficients between index and reference tests [TAR54]. Two reviews [TAR6; TAR39] expressed accuracy in terms of the standardised mean difference between index test results in diseased and non diseased individuals, also termed the effectiveness score [117]. The effectiveness score is a simple re-expression of the DOR.

In addition to test accuracy, two reviews provided detail on inter and intra-observer variability [TAR53; TAR54].

Table 4.10: 'Other' outcome measures used by a total of 17 reviews

Outcome measure	Review Topic
Correlation coefficients	Ambulatory blood pressure monitoring and blood pressure self-management in the diagnosis and management of hypertension [TAR3]
	A review of near patient testing in primary care [TAR34]
	Reliability of reporting left ventricular systolic dysfunction by echocardiography: a systematic review of 3 methods [TAR53]
	Application of surface electromyography in the assessment of low back pain [TAR54]
Limits of agreement (Altman 1991)[116]	Reliability of reporting left ventricular systolic dysfunction by echocardiography: a systematic review of 3 methods [TAR53]
	Ambulatory blood pressure monitoring and blood pressure self-management in the diagnosis and management of hypertension [TAR3]
	A review of near patient testing in primary care [TAR34]
Effectiveness score (standardised mean difference) (Hasselblad 1995)[117]	Diagnostic efficacy of home pregnancy test kits [TAR6]
	Methods of Screening for Dementia: A meta-analysis of studies comparing an informant questionnaire with a brief cognitive test [TAR39]
Average, unweighted sensitivity and specificity	Evidence for the Diagnosis and Treatment of Acute Uncomplicated Sinusitis in Children [TAR37]
	Meta-analysis of exercise testing to detect coronary artery disease in women [TAR42]
Inter and Intra observer variability	Reliability of reporting left ventricular systolic dysfunction by echocardiography: a systematic review of 3 methods [TAR53]
	Application of surface electromyography in the assessment of low back pain [TAR54]
(TP+FP)÷ all tested (termed detection rate; yield) TP÷ all tested (termed test positive rate)	Antenatal screening for postnatal depression: a systematic review [TAR4]
	Systematic review of the school entry medical examination [TAR5]
	WHO Systematic Review of Screening Tests for Pre-Eclampsia [TAR14]
	Exercise tolerance testing to screen for coronary heart disease [TAR24]
	A review of near patient testing in primary care [TAR34]
	Diagnosing syncope: Value of history, physical examination and electrocardiography [TAR47]
	A comprehensive Evidence-Based approach to fever of unknown origin [TAR55]
	Screening for depression in adults [TAR67]
	Should health professionals screen women for domestic violence? Systematic review [TAR70]
	Diagnosis, management and screening of early, localised prostate cancer [TAR83]

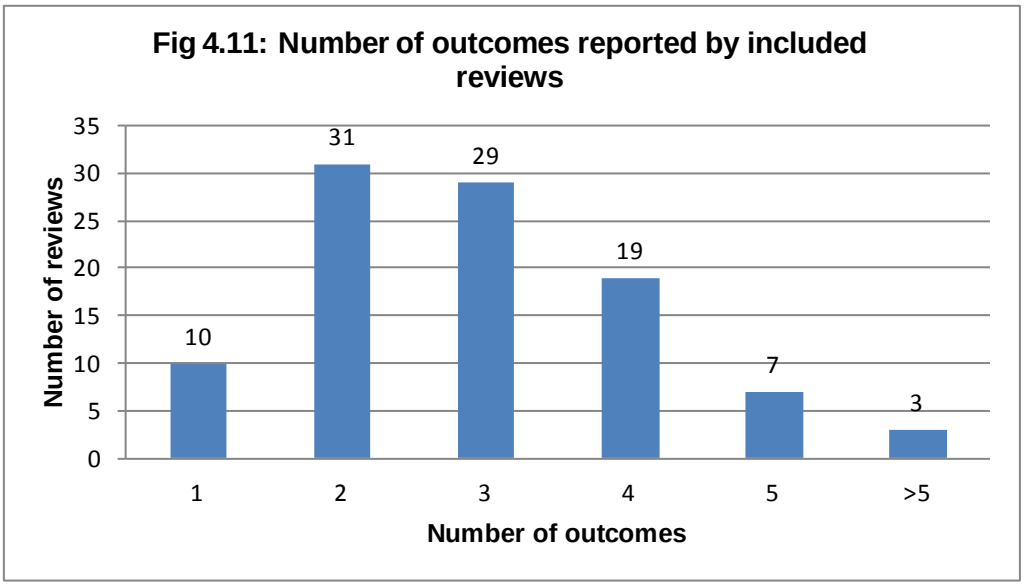
Notes to Table 4.10: TP: true positives; FP: false positives.

Where tests were being used to detect multiple target disorders the outcome measure used was diagnostic yield which was variably defined: “true positives ÷ all tested” where verification of all test positives was possible or “(true positives (TP) + false positives (FP)) ÷ all tested” where verification of all test positives was not possible. Seven of these reviews were concerned with screening tests where test negatives did not receive verification (TAR4; TAR5; TAR14; TAR24; TAR67; TAR70; TAR83) and four were reviews concerned with tests for multiple underlying disorders that would require multiple reference standards in order to comprehensively verify test positives (TAR5; TAR34; TAR47; TAR55).

4.5.5.4 Number of outcome measures used by review authors

The number of outcome measures reported by the 99 reviews with included studies is illustrated in figure 4.11.

Fig 4.11: Number of outcomes reported by included reviews



Notes to Fig 4.11: Sensitivity and specificity reported together considered one outcome. Similarly LR+ and LR-; PPV and NPV; FN, FP, TN, TP.

The majority of reviews reported between two and three outcomes although a substantial minority (26/99) reported four or five outcomes. Three reviews reported a total of six outcomes (TAR34; TAR60; TAR98) although one of these reviews was evaluating a large number of

index tests and review authors commented on the lack of comparable outcomes across included studies [TAR34].

4.5.6 Contextualisation of review synthesis and consideration of applicability of review findings

For details of contextualisation of review synthesis and consideration of applicability of findings tabulated by included review see appendix 4.6

4.5.6.1 Contextualisation of review synthesis

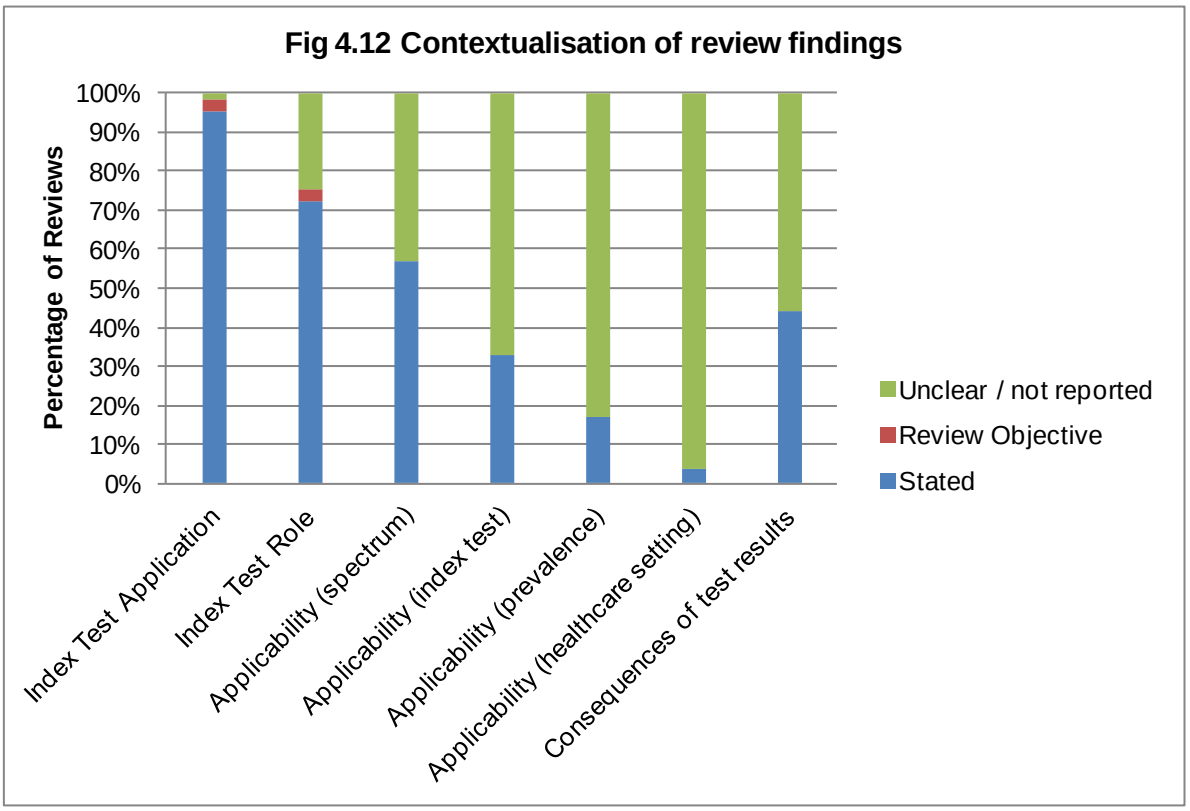
Contextualisation of review findings encompasses consideration of the proposed application, role and intended setting in which index tests are to be used when planning the analysis strategy including investigation of heterogeneity. For example the proposed application of a test will affect which dimension of test accuracy is more important and therefore choice of outcome measures, the intended role of a test will determine the type of synthesis that is undertaken (for example whether a comparative evaluation of accuracy is required) and the intended setting in which a test is to be used should guide any investigation of heterogeneity.

Figure 4.12 illustrates the percentage of reviews that contextualised findings by defining test application, defining test role, discussing the applicability of review findings and considering the downstream consequences of test results.

Index test application

At the point of synthesis of review findings the proposed application of the index test had been defined by 95/99 reviews, in two reviews the proposed test application was unclear [TAR85; TAR36] whilst the objectives of three reviews included determination of the optimal application based on index test properties [TAR45; TAR66; TAR69].

Of the 95 reviews defining test application 74/95 included diagnosis, 32/95 included screening, 7/95 included prognosis and 3/95 included monitoring. Eighteen of 95 reviews stated more than one potential application and of these, 12/18 distinguished between different applications in the review synthesis. In seven reviews where the application of the test was not specified at formulation stage the application was clarified as part of the review synthesis (TAR39; TAR45; TAR54; TAR66; TAR69; TAR76; TAR77).



Index test role

In 15 reviews where the role of the test was not specified or unclear at formulation stage the role was clarified as part of the review synthesis (TAR18; TAR20; TAR27; TAR47; TAR58; TAR66; TAR68; TAR69; TAR71; TAR72; TAR79; TAR81; TAR84; TAR90; TAR94).

Seventy two reviews had defined test role at the point of review synthesis. Twenty eight reviews were concerned with evaluation of tests as replacements to existing tests, 30 /99 reviews with evaluation of index tests as a potential addition to testing practice, 9/99 reviews

with evaluation of index tests in a triage role and 3/99 reviews with multiple concurrent roles. In 3/99 reviews test role was explicitly stated as a review objective [TAR45; TAR79; TAR97] although this had not been clearly articulated at question formulation stage (see 4.5.3). One review was clearly concerned with early test evaluation [TAR54].

In 23/99 reviews the proposed role of the index test was unclear.

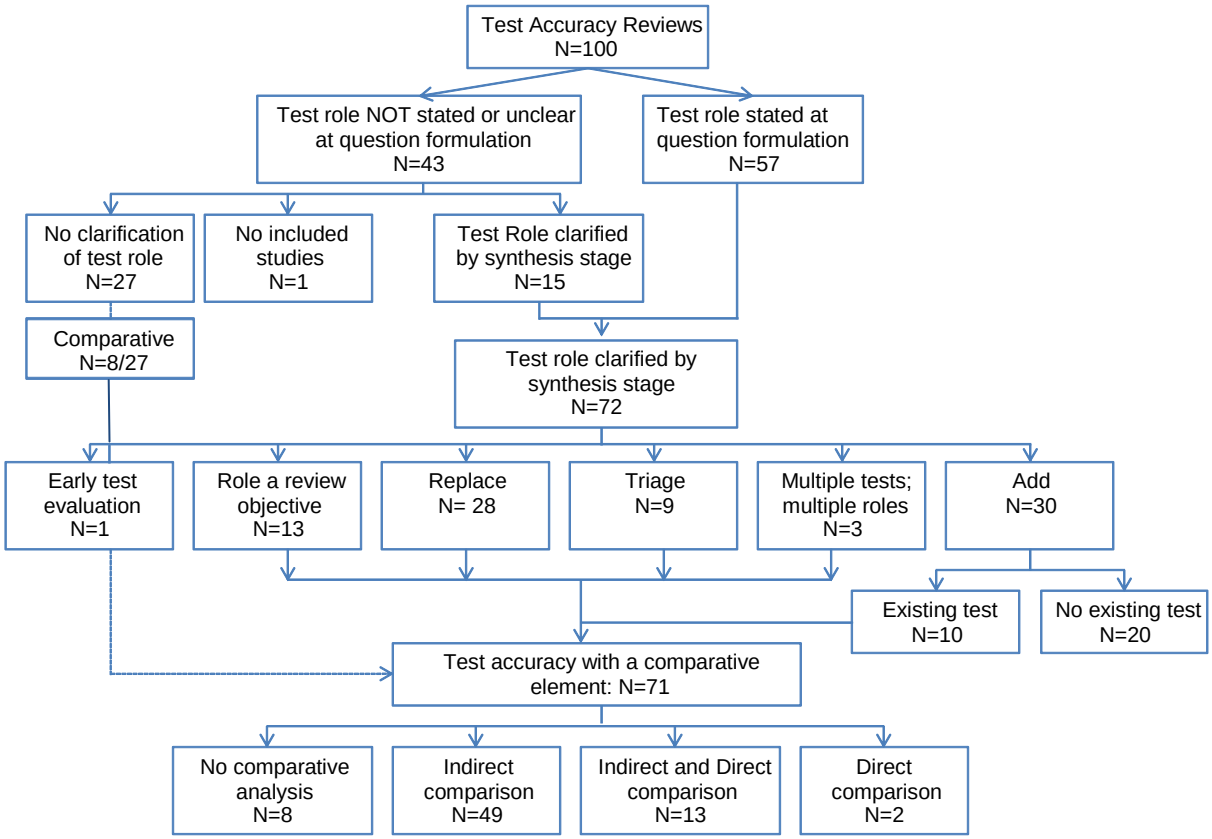
Test comparisons (see figure 4.13)

A total of 64 reviews (both reviews that had defined test role and those who had not) undertook a comparison of tests (comparison of two index tests or replacement of current practice with the index test) or testing strategies (additive or triage roles). Of these 64 reviews, 49 were restricted to indirect comparisons, 13 undertook indirect and direct comparisons and two were restricted to direct comparisons (see figure 4.13). It was not clear the extent to which the comparative approach adopted by reviews (direct or indirect) reflected the study design of included primary studies. Despite the absence of methods for pooling direct comparisons without access to individual patient data, direct comparisons are still considered more valid as they reduce the effects of study level confounding by population spectrum.

Twenty four of 64 reviews undertaking comparisons presented confidence intervals to assist with interpretation, although few reviews explicitly used these when discussing results and 6/64 reported the results of significance testing [TAR2; TAR31; TAR48; TAR52; TAR80; TAR82].

Ten of the 30 reviews concerned with the addition of the index test to existing testing strategy assessed incremental accuracy. Eight of ten reviews did this by means of indirect comparisons of primary studies evaluating the accuracy of existing tests and primary studies evaluating the accuracy of the index test in addition to existing tests [TAR18; TAR25; TAR42; TAR47; TAR62; TAR90; TAR93; TAR98]. One review assessing incremental accuracy undertook Bayesian updating of the accuracy of existing tests with the addition of the index test [TAR51] and one review used both an indirect comparative and Bayesian updating approach [TAR40].

Fig 4.13: Test Role: Flow of Studies



Investigation of heterogeneity

Sixty seven of 99 reviews with included studies considered the potential effects of heterogeneity on test accuracy. Studies adopted one or more of five approaches: restricting inclusion, sub-grouping findings, meta-regression, illustrating the effect of prevalence on post-test probability or discussing effects narratively with the aid of graphics such as Forest plots and sROC space plots. The majority of reviews investigating heterogeneity used sub-grouping of studies (37/67) of which 31/37 stated a priori potential modifiers of heterogeneity. Twenty of 67 reviews used meta-regression, all of which stated a priori potential modifiers of heterogeneity to be investigated. Nineteen of 67 reviews illustrated the effects of variation of prevalence on post-test probability, 14/67 reviews restricted themselves to a narrative

discussion of studies grouped according to characteristics and 5/67 reviews restricted inclusion criteria as a means of achieving a homogeneous sample.

All reviews undertaking an investigation of heterogeneity (67/99) investigated the effects of spectrum variables or prevalence, 48% investigated the effects of index test variation (including threshold) and 38% study quality. In only two reviews the quality or reporting of primary studies was stated as a limitation to the accommodation of spectrum or prevalence as part of the review synthesis [TAR12; TAR16].

4.5.6.2 Assessment of applicability of review findings

Sixty nine of the 99 reviews with included studies considered whether characteristics of included studies were applicable to a specified testing context. Fifty six of the 69 reviews considered the applicability of the spectrum of the tested population whilst applicability of threshold and / or technical aspects of the operation and interpretation of index tests was considered by 35/69. The applicability of the prevalence of the target disorder in included studies was considered by 17/69 reviews and the applicability of the healthcare system in included studies by 4/69 reviews.

4.6 Strengths and limitations: Methodological review

Much of the assessment of included reviews relied on subjective interpretation and the clinical and methodological experience of the person extracting data may have resulted in an overoptimistic representation. For example it was common for information to be available in reviews to form a judgement in the absence of explicit discussion by review authors. In addition a broad framework for assessing the degree to which review authors had considered contextual factors when conducting and reporting reviews was generous and may not be optimal for the topic of any single review.

A potential limitation of this review is that it may not be a current reflection of the conduct and reporting of systematic reviews of test accuracy; searches for the review stopped in September 2006. However guidance pertinent to the contextualisation of test accuracy review questions has been limited to date and largely limited to recent initiatives within the Cochrane Collaboration including the publication of the Cochrane Handbook for Diagnostic Test Accuracy Reviews [96]. The Cochrane handbook remains incomplete as at 2011. Chapters pertinent to question formulation and contextualisation of review findings include Chapter 4: Guide to the content of a Cochrane review and protocol for diagnostic test accuracy: published 2009; Chapter 6: Developing criteria for including studies: published 2008; Chapter 8: Selecting studies and collecting data: not published at the time of writing; Chapter 11: Interpretation of results: not published at the time of writing. http://srdta.cochrane.org/handbook-dta-reviews. Personal experience of training within the Collaboration suggests that review authors, including clinicians, find question formulation in this area difficult.

Other guidance that may have impacted on those aspects of test accuracy review conduct relating to contextualisation between 2006 and the time of writing include the development of the first version of the QUADAS instrument for the quality assessment of primary studies to be included in reviews and the STARD initiative concerned with reporting of primary studies of test accuracy. The original QUADAS item was developed in 2003 [118] and the impact of this tool might be only partially captured by the reviews included in this investigation (see 4.7.1.2 below). However the first version of QUADAS is primarily concerned with assessment of internal validity and a criticism of the original instrument is that it does not make a clear distinction between internal validity and external validity. It is therefore unlikely that QUADAS will have encouraged review authors to consider the applicability of the test accuracy evidence available to them. The STARD initiative, culminating in a checklist for the reporting of primary studies of test accuracy [38] aims to improve accuracy and completeness of

reporting of internal and external validity. The checklist includes items concerned with population spectrum (setting, presenting symptoms, target condition severity, co-morbidity, prior tests received) and technical specification of test execution (including threshold and expertise of the test operator). The STARD checklist may have influenced whether and how review authors considered applicability, both indirectly as a result of improvements in reporting of primary studies and directly by raising awareness of the components of an assessment of external validity for test accuracy research in the research community. However an evaluation of the impact of STARD on reporting of diagnostic accuracy studies up to 2005 reported no observable difference at that time [119] In addition a recent methodological review [97] observed a demonstrable uptake of the QUADAS instrument only three years after its publication in 2003. It is therefore unlikely that the initiatives outlined above will have had a significant impact on the conduct and reporting of test accuracy reviews after completion of searches (2006). Indeed marked variation in uptake of methodological developments for conducting test accuracy reviews of between three and 10 years has been observed [97]. This variation is suggested to be due to variation in the level of complexity and technical barriers to uptake of developments; thus it might be expected that three years is an ambitious lag time for appropriate contextualisation of review questions, review methods and reporting of outcomes, particularly in the light of the personal experience of the author in training authors of test accuracy reviews.

4.7 Discussion

4.7.1 Applicability of findings to reviews of test accuracy

4.7.1.1 Setting

This review sampled three databases including one with no restriction on inclusion (the ARIF database) in an attempt to compile a representative sample of test accuracy reviews. In addition the selective inclusion of tests that could be applied in the primary care setting was

an attempt to mitigate against the preponderance of test accuracy reviews concerned with the secondary care setting [97,112].

Setting was only reported by 54% of reviews and it is therefore difficult to ascertain whether the review sampling strategy was successful. However tests restricted to use in the community, primary care and 'over the counter' were represented by only a minority of reviews, (5/54 reviews reporting setting), whilst 48/54 reviews included secondary care populations. A mix of settings was reported for 38/54 reviews and it is possible that the sampling strategy resulted in greater heterogeneity rather than a greater number of reviews restricted to generalist settings.

Potential barriers to the evaluation of tests in generalist settings include enrolment of sufficient participants when pre-test probability is low, the ability to access reference standards in the primary care setting and the position of primary care early in the diagnostic work up, where the value of tests lies in their ability to identify and distinguish between multiple potential target disorders with an emphasis on ruling out serious conditions. Challenges associated with the simultaneous evaluation of test accuracy for multiple target disorders include the necessity for multiple reference standards which magnifies the problem of access to reference standards in primary care and the practicality of following up index test negatives with potentially invasive further testing early on in the diagnostic work-up where pre-test probability is low.

The ability of this methodological review to identify barriers to evaluating tests in primary care was limited by included reviews, of which only a minority were concerned with generalist settings, with the evaluation of history and examination or with the evaluation of the utility of tests for multiple potential target disorders. Reviews that attempted to evaluate the accuracy of a test for detection of multiple disorders using the traditional test accuracy evaluation framework were successful if they restricted inclusion of individual primary studies to those evaluating only one of a multiplicity of target disorders and therefore a requirement for a single reference standard. This restriction is likely to increase heterogeneity and atypical

presentations in the review sample therefore reducing applicability but without it reviews were limited to reporting the number of test positives generated which provides no information on false positives or the ability of a test to rule out potentially serious disease; the emphasis in primary care settings (TTA8). There is currently no guidance applicable to the evaluation of test accuracy for multiple target disorders simultaneously. This is likely to be a reflection of the nature of medical knowledge, which tends to be organised according to disease rather than individual signs and symptoms; a framework which may be better suited to specialist rather than generalist, primary care settings [49] (see 1.4) and may be an important factor contributing to the under-representation of test accuracy evaluations in primary care and of the clinical history and examination.

4.7.1.2 Quality of included reviews

The median quality score of included reviews was 4.6/9 with 46% of reviews scoring less than 4/9 and 21% of reviews scoring greater than 6/9. A review of 189 diagnostic test accuracy reviews in the Database of Abstracts of Reviews of Effectiveness (DARE) database up to 2002 by Dinnes [7] demonstrated that 48% of included reviews searched more than one database (MEDLINE) and 69% undertook quality assessment. In this review 43% of reviews searched more than one database and 78% undertook quality assessment. The DARE database only contains reviews meeting a minimum quality standard [107] and the expectation would therefore be that the Dinnes' 2002 review would contain a larger proportion of higher quality reviews. The greater number of reviews undertaking quality assessment in this review is likely to be a reflection of the publication of the QUADAS quality assessment tool in 2003, after completion of searches in the Dinnes' review. This is supported by the findings of a recent methodological review that demonstrated widespread uptake of the QUADAS tool approximately 3 years after its publication [97].

4.7.2 Adequacy of question formulation

The clarity of question formulation was generally poor. Only 24% of included reviews detailed all of index test application, role and prior tests as part of question formulation. In addition 51% of reviews did not distinguish symptomatic from asymptomatic presentation, 75% reviews did not specify details about each of chronicity or severity of the target disorder and co-morbidities. Reviews evaluated between one and 50 index tests (median 3) and 76% of reviews did not state the setting in which index tests were to be used. The inclusion of multiple settings and multiple tests in test accuracy reviews has been noted in other work [97]. Possible explanations for the observed inadequacy in question formulation are the limitations imposed on review reporting by publication in journals. However this hypothesis was not supported by improvements in clarity as reviews progressed and only a minority of reviews explicitly acknowledged that lack of clarity was to be addressed as a review objective. Clarity of review question formulation did not appear to change over the time period covered by this review.

4.7.3 Contextualisation of review findings

Reporting of study characteristics was poor in this sample of reviews with no study characteristic being clearly reported by more than 54/99 reviews. Reporting of primary study characteristics is particularly important to assist decision makers with the applicability of review findings given the considerable potential for heterogeneity in evaluations of test accuracy. In addition inadequacies at question formulation stage that result in broad inclusion criteria are magnified if details of primary studies are not reported. Even when question formulation is adequate there are circumstances when inclusion criteria need to be modified in the light of literature searches. For example in this sample of reviews only 9 reviews specified more than one testing setting at question formulation stage and at least 38 included more than one setting. Failure to document study characteristics due to limitations in

reporting by primary studies was cited by a very small number of reviews (one to eight reviews per characteristic). Given the poor quality of review question formulation it appears that inadequacies in reporting review findings may be as much a reflection of poor review methodology as limitations in primary study quality and reporting.

4.7.3.1 Index test role and application

Although the majority of reviews articulated the proposed application of index tests, in one quarter of included reviews the proposed role of the index test was unclear. Accommodation of test role at synthesis stage was variable with limited use of basic statistical techniques (for example use of confidence intervals) to assess the uncertainty associated with any observed difference for test comparisons (30 of 64 reviews undertaking comparisons). Further only one third of reviews evaluating tests in an additive role attempted to quantify incremental accuracy. This is additional evidence to suggest that the observed inadequacies in question formulation and reporting of study characteristics reflect a deficiency in methodological approach rather than poor reporting of methods (see 4.7.2; 4.7.3).

4.7.3.2 Spectrum

Only 9/99 reviews reporting all of population presentation, healthcare setting and tests performed prior to the index test.

Between one half and two thirds of reviews did not report each of setting, details of patient spectrum and prevalence. Only 22% of reviews commented that spectrum was poorly reported by primary studies and less than 10% stated that poor reporting in primary studies was a review limitation. This finding suggests any inadequacies in primary studies may be exacerbated by a lack of appreciation of spectrum effects by review authors. Similarly one third of included reviews did not consider whether the characteristics of included studies

were applicable to a target testing context. This again suggests inadequacies in question formulation rather than poor reporting of methods.

4.7.3.3 Investigation of heterogeneity

The potential to investigate and accommodate heterogeneity has existed for some time and certainly over the time period represented by this review. The definition of investigation of heterogeneity adopted by this review was broad and extended to sub-grouping graphically or narratively in the absence of statistical tests. Using this broad definition only 68% of reviews explored one or more of the potential modifying effects of spectrum, index test variation and methodological quality of included studies. By contrast, using a stricter definition of exploration of heterogeneity, an earlier methodological review sampling the quality assured DARE database demonstrated that 83% of reviews explored heterogeneity potentially introduced by clinical, test or study quality variables [7]. The discrepancy between the two reviews may be a reflection of differences in the quality of included reviews, although this is not obviously apparent from the limited information available for comparison (see 4.7.1.2 above).

4.7.3.4 Use of outcome measures

The results of two reviews conducted in the DARE database in 2000 (Honest 2002 [6]) and 2002 (Dinnes 2005 [7]) are compared to the findings of this review in table 4.14:

Table 4.14: Comparison of outcome measures used in systematic reviews of test accuracy, 2000 (Honest 2002[6]); 2002 (Dinnes 2005[7]); 2006.

Outcome measure	DARE 2000 (Honest 2002[6]) (N=90)		DARE 2002 (Dinnes 2005[7]) (N=189)		This Review 2006 (N=189)	
	Meta-Analysis 60 (67%)	Narrative 30 (33%)	Meta-Analysis 133 (70%)	Narrative 56 (30%)	Meta-Analysis 60 (61%)	Narrative 39 (39%)
ROC curve	44%	NR	64 (48%)	NR	24(40%)	5 (13%)
Sensitivity and specificity	58%	NR	117 (88%)	NR	33 (55%)	28 (72%)
Predictive values	18%	NR	11 (8%)	NR	9** (15%)	14 (36%)
Likelihood ratios	22%	NR	26 (20%)	NR	20 (33%)	11 (28%)
DOR	8%	NR	14 (11%)	NR	13 (22%)	0 (0%)
Effectiveness score	NR	NR	8 (6%)	NR	2 (3%)	0 (0%)
'Q'	NR	NR	18 (14%)	NR	4 (7%)	0 (0%)
Test accuracy	NR	NR	5 (4%)	NR	0%	3 (8%)
AUC	NR	NR	13 (10%)	NR	11 (18%)	1 (3%)
Pre-post-test probability	NR	NR	NR	NR	5 (8%)	0%
Comparative measures (Relative or absolute)	NR	NR	NR	NR	3 (5%)	0%
Test errors	NR	NR	NR	NR	2† (3%)	14 (36%)
TP, TN, FP, FN	NR	NR	NR	NR	0%	10 (26%)

Notes to table 4.14:
TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; ROC: Receiver Operator Characteristic; DOR: Diagnostic Odds Ratio; AUC: Area Under the Curve.
* One included review did not report outcomes
**One review derived PVs from sROC average sensitivity and specificity
† Both reviews derived test errors from 'Q'

The proportion of reviews proceeding to meta-analysis is greater in earlier reviews (67% and 70%) compared to this review (61%). This discrepancy may have been expected if the sampling strategy of this review (to include tests that were available to generalists) had the effect of increasing the mix of settings of included studies and therefore heterogeneity in included reviews. However investigation of heterogeneity in this review was also less than earlier reviews (see section 4.7.3.3 above).

Sensitivity and specificity remain the most commonly used outcome measures although the frequency of their use in meta-analyses is variable. Use of ROC curves in meta-analyses appears comparable across the time period covered by the three reviews whilst the use of LRs, AUC and the DOR in meta-analysis has increased. However the increase in the use of the DOR is less marked if considered in combination with the effectiveness score, which is a simple re-expression of this metric. The use of Q has decreased between 2002 and 2006.

Predictive values are suggested to be the most intuitive summary measure of test accuracy (see Chapter 2). The use of PVs in meta-analysis is markedly less than sensitivity and specificity, less than the use of LRs and between 2002 and 2006 less than the DOR. The variation of PVs with prevalence, and the impact this has on heterogeneity, may be deterring review authors, particularly if combined with a lack of appreciation that prevalence to some extent is a proxy for spectrum and therefore affects all test accuracy measures (see chapter 2). Derivation of PVs from average sensitivity and specificity was used by two reviews in this review. The validity of direct derivation of pooled PVs has only recently been explored [86] and has the potential to impact on the use of this metric in the future.

It is interesting to note that the use of test errors in narrative reviews is comparable to PVs and LRs although the extent to which this is driven by primary study reporting is unclear.

More than half of reviews reported more than three outcomes. Although there may be a need to use and report complimentary outcome measures, (for example global measures may be used to compare tests and pre to post test probability to illustrate the potential diagnostic

impact of testing), this was rarely explicitly articulated by review authors and mostly the rationale for choice of outcome measures was unclear. There is currently no guidance that attempts to link the setting, test application and test role with synthesis approach and choice of outcome measures. Such guidance would encourage reporting of outcomes that highlight unique contextual considerations pertinent to individual test evaluations and would therefore ensure a better contextual fit of the test accuracy evidence base.

4.7.3.5 The downstream consequences of test results

The use of outcome measures should be linked to decision making. Although summary measures distinguishing between the two dimensions of accuracy were more frequently used than global measures, less than half of included reviews made an attempt to link test accuracy to decision making by differentiating between the two dimensions. In addition less than half of reviews explicitly acknowledged the downstream consequences of test results as a means to discuss the implications of false positives and false negatives.

In the existing research environment, which is characterised by a paucity of RCTs of test and treat combinations [90] it becomes all the more important to consider the downstream consequences of test results on patient outcomes when reporting test accuracy evaluations.

4.7.4 Implications for the conduct and reporting of test accuracy reviews

Systematic reviews offer the opportunity to formulate a focused question, identify primary studies of relevance to that question and synthesise a volume of evidence according to the intended application and role of the test under evaluation. This is particularly important where primary research is characterised by ill defined objectives in relation to the application and role of an index test or tests and without consideration of the potential variation in spectrum in the population to be tested. Indeed it is claimed that the potential contribution of systematic

reviews to the test accuracy evidence base is compromised by the quality and reporting of primary studies (see section 1.3).

A recent review of meta-analyses of diagnostic or predictive tests by the Agency for Health Care Research and Quality [120] suggested that over the period 1996 and 2009 substantial improvements in literature review methods, quality assessment and statistical analysis methods employed have taken place. The authors note that improvements in quality assessment are associated with the use of quality item checklists which concurs with the findings of this methodological review (4.7.1.2).

However this review suggests that currently, systematic reviews and meta-analyses of test accuracy are characterised by ill defined objectives which is reflected in review synthesis and reporting of review findings. Key pieces of information that may help determine the relevance of the review to readers are absent or not highlighted sufficiently in the review abstract or review objectives.

Inadequate question formulation underpinned by a lack of appreciation of spectrum effects on estimates of test accuracy appears the most probable explanation for this finding which will exacerbate any inadequacies in primary studies. Inadequacies in question formulation raise the issue as to the degree to which synthesis is data led rather than addressing questions of most clinical importance. Lack of clarity at the question formulation stage precludes judgement about the rationale for the review, the proposed role of the index test in a care pathway and therefore the utility of the information provided by the review findings.

It has been suggested that failure to investigate heterogeneity and wide variation in methods used [7] may be a reflection of the complexity and continuing development of methods for undertaking meta-analyses of diagnostic tests. However a lack of recognition of the degree to which test performance varies with clinical context and inadequacies in question formulation may be additional explanations. Question formulation and structuring the review process, including a priori statement of variables to be investigated as potential sources of heterogeneity should be complementary processes. Against a complex and developing

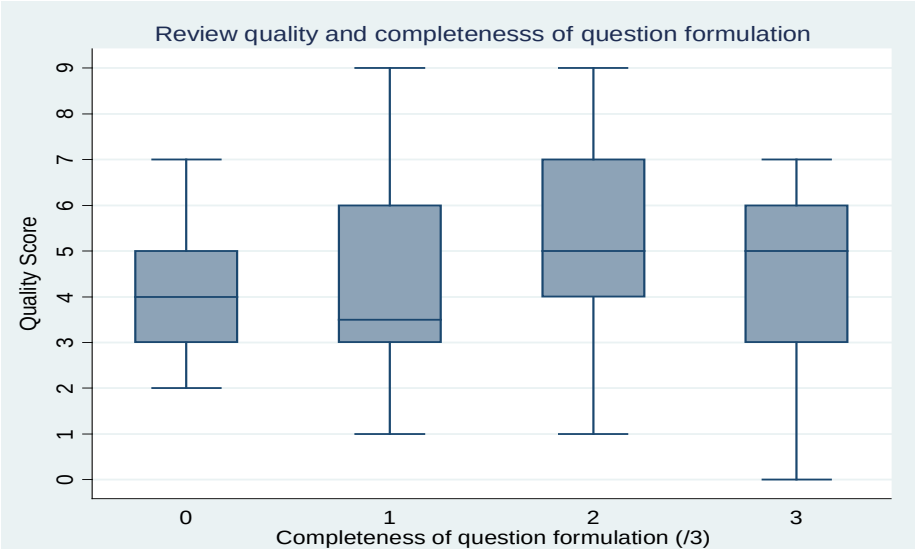
methodological framework, refining a review question as far as possible to ensure the provision of clinically relevant and focused information becomes increasingly important.

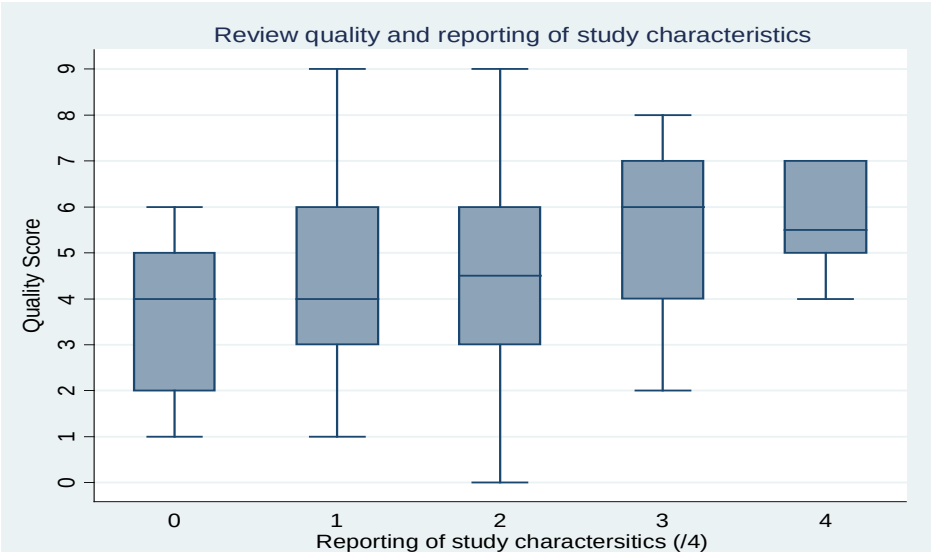
It is proposed that engaging stakeholders, including the end users of the findings from test accuracy evaluations, will help to ensure appropriate and focused test accuracy review objectives [120]. However between 94% and 99% of included reviews included a clinician as co-author. A possible explanation for this observation may be the use of clinical methodologists rather than clinical topic experts on review teams. However the personal experience of those undertaking training for Cochrane review groups and authors suggests that even clinical topic experts have difficulty formulating test accuracy review questions, particularly contextualising the role of index tests in testing pathways.

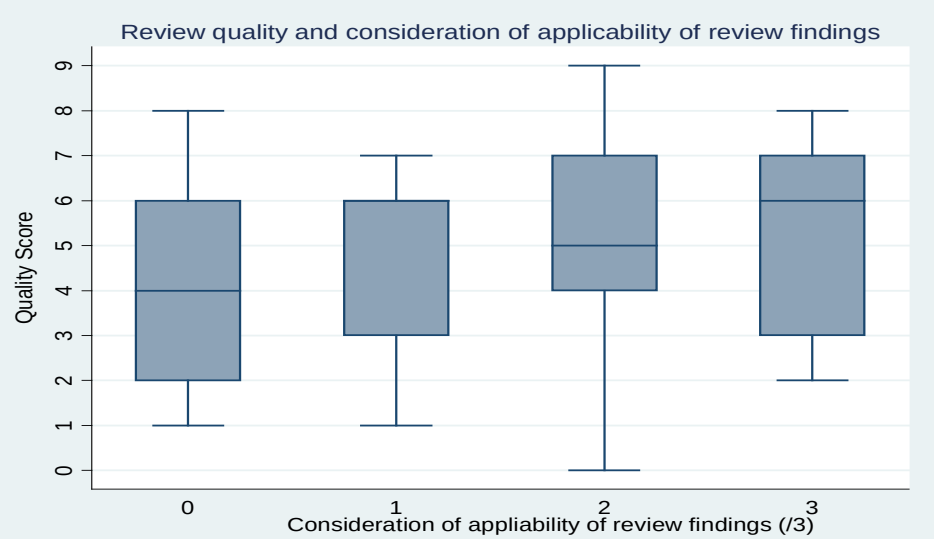
Figures 4.15 to 4.17 (see also appendix 4.7), illustrate the relationship between the quality of included reviews, as measured by nine quality items taken from the original QUORUM [111] and AMSTAR [110] checklists and each of completeness of review question formulation (as evidenced by explicit mention of one or more of the index test application, index test role and any tests performed prior to the index test), completeness of review reporting of study characteristics (as evidenced by reporting of one or more of tests received prior to the index test, patient presentation (symptomatic or asymptomatic), prevalence of the target disorder and quality of included studies) and consideration of applicability of review findings (as evidenced by discussion of one or more of spectrum, prevalence and index test characteristics of the included studies). These measures were chosen from the more comprehensive assessment used in this review as measures that might be considered key to test accuracy reviews, regardless of review topic (see 4.5.3; 4.5.4; 4.5.6). There appears to be no relationship between the quality of included reviews and review reporting, consideration of applicability of included studies, or completeness of review question formulation. Assessment of the internal validity of systematic reviews according to existing guidance does not appear to be a good reflection of the degree to which review authors have

considered the external validity of their findings. It is therefore likely that inadequacies in test accuracy question formulation at the current time are due to the developmental stage of test accuracy evaluation methodology, limited dissemination of methods to review authors and a lack of reporting guidelines specific to systematic reviews of test accuracy.

Figs 4.15-4.17


Review quality and completenesss of question formulation
Quality Score
0
1
2
3
4
5
6
7
8
9
0
1
2
3
Completeness of question formulation (/3)


Review quality and reporting of study characteristics
Quality Score
0
1
2
3
4
5
6
7
8
9
0
1
2
3
4
Reporting of study charactersitics (/4)


Review quality and consideration of applicability of review findings
Quality Score
0
1
2
3
4
5
6
7
8
9
0
1
2
3
Consideration of appliability of review findings (/3)

Chapter 5: Survey of understanding and application of test accuracy measures

5.1 Abstract

Background

Increase in test use over recent decades has occurred despite disappointing results from test accuracy evaluations. Difficulties with understanding and application of test accuracy information are purported to be important contributors to this observed evidence 'gap'. Empirical research to date is based on the premise that formal probability revision is a necessary pre-requisite for informed diagnostic decision making and is characterised by self selected samples with recent experience or expertise in test evaluation. The survey aimed to describe how clinicians apply existing test accuracy metrics for diagnostic decision making.

Methods

An incentivised, electronic survey was used. Informed application of test accuracy information was evaluated by asking respondents to indicate their management decision following presentation of nine different representations of the same test accuracy information to a common hypothetical scenario. Quantitative and qualitative synthesis was employed based on closed an open responses to management decisions.

Results

204 General Practitioners (response rate 95%) did not appear to be self selected on the basis of academic position, involvement in policy or experience in test evaluation. Sensitivity and specificity, the annotated 2x2 diagnostic table and predictive values were reported to be familiar metrics by the most respondents. Likelihood ratios the DOR and AUC were familiar to less than 1/3 of respondents. Application of test accuracy metrics resulted in marked variation in responses to both positive and negative test results although greater inconsistency and management uncertainty was observed following presentation of a negative test result in comparison to a positive test result. Formal probability revision was not

a feature of the diagnostic decision making process. Test errors (false negatives and false positives) were prominent as part of the translational pathway from quantitative summary estimates of test accuracy to management decisions. Summary measures that separate the two dimensions of test accuracy in the absence of prevalence information (for example sensitivity and specificity) appeared to result in a misplaced emphasis in one or other of false positive or false negative test errors. Presenting test accuracy data using the 2x2 diagnostic table or a pictograph attenuated this effect.

Conclusion

Choice of test accuracy metric appears to have a profound effect on diagnostic decision making. Understanding, contextual factors and motivational biases are likely to be contributing factors to the observed variability. It is unclear to what extent any advantage of test accuracy metric for informed decision making is based on familiarity as opposed to their intuitive nature. Simultaneous illustration of both dimensions of test accuracy in order to facilitate informed diagnostic decision making requires further exploration.

5.2 Survey rational and aims

The rationale for undertaking primary research was to begin to address gaps in the existing literature about how clinicians use test accuracy information and to refine emerging hypotheses about characteristics of test accuracy metrics that facilitate their application. The target audience, content and distribution methods of the final survey reflect the findings from the literature reviews (chapter 2), mapping the epidemiological characteristics of existing test accuracy reviews (chapter 3) and a pilot electronic survey, distributed using NHS e-mail accounts, to general practitioners in the Birmingham and Black Country (BBC) region of the UK in 2009. The pilot survey was assessed by the National Research Ethics Service as a service evaluation and therefore did not require NHS ethical review. The survey did receive Research & Development approval from each of the seven Primary Care Trusts (PCTs) initially sampled.

5.3 Survey Aims and Objectives

Aims

The aims of the survey were to describe how a representative sample of primary care clinicians use sources of test accuracy information and to evaluate whether and how existing test accuracy metrics are understood and applied.

Objectives

- To identify which sources of test accuracy information are used by primary care clinicians and facilitators and barriers to their use
- To evaluate the utility of existing test accuracy metrics as measured by self-reported familiarity with terminology, perceived ability to define or explain metrics and self-reported use of metrics in clinical practice

- To investigate how a range of existing test accuracy metrics are applied to a hypothetical testing scenario and in particular whether there is consistency in the application of different metrics.
- To evaluate the within-person relationship between perceived understanding of test accuracy metrics and their application in a hypothetical testing scenario.

5.4 Survey Methods

5.4.1 Sampling and questionnaire distribution

The lack of familiarity of clinicians with the subject area was considered a major disincentive relative to other aspects of questionnaire design that influence response rate and are amenable to change. Most studies identified during the review of empirical test accuracy literature (see chapter 2) were undertaken on motivated, educated participants, often during educational events. A major aim of this survey was therefore to capture a representative sample of practicing clinicians. In addition, due to under-representation of primary care in the existing literature (see chapters 2, 3 and 4) and because involvement in the earlier stages of the diagnostic work up is likely to result in a more diverse experience of testing, a general practitioner (GP) sample was chosen in preference to a secondary care sample.

As part of the initial BBC pilot survey, face to face methods of distribution were explored and abandoned on the basis that only educational events and GP tutor and trainers' forums were identified. However subsequent electronic distribution of this initial pilot survey to 1600 GPs via NHS e-mail accounts resulted in <3% response rate, of which 49% of respondents had received training in relation to testing in the preceding three years. Reliance on PCT communication teams for distribution and competing demands on NHS e-mail accounts were identified as major problems with this dissemination method. For the final survey, an incentivised, electronic survey hosted by doctors.net.org; a professional network of ~200,000 General Medical Council registered doctors with access to approximately 27 000 of 41 000

GPs across the UK was chosen as the distribution method most likely to achieve a large and as representative a sample as possible. Doctors.net.org offer 'electronic Surfing Rewards' as an incentive to participate in research which can be exchanged for products and high street vouchers.

5.4.2 Questionnaire content

5.4.2.1 Questionnaire structure and presentation

The risk communication literature suggested the lack of an observed association between preference and comprehension of risk measures was a result of heuristics applied to familiar presentation formats. In the final questionnaire assessment of *perceived* understanding (familiarity with test accuracy terminology and confidence in defining or explaining test accuracy metrics and graphics) was therefore examined separately to an assessment of understanding as measured by application of test accuracy information to a hypothetical scenario. In order to minimize invalid responses and to mitigate against a poor response rate, skip logic was employed such that if a respondent indicated that they were not familiar with a test accuracy metric or graphic they were not required to answer questions concerning self-reported confidence in defining or explaining the metric or graphic, or self-reported use of the metric or graphic in practice. However respondents' understanding as measured by application of test accuracy information provided to them in hypothetical scenarios was examined regardless of their stated familiarity with that metric or graphic. The rational for this approach was to allow investigation of whether the method of presentation of test accuracy information modifies management decisions following a test result, regardless of familiarity. Open comments were invited for all questions but in contrast to closed questions were not a requirement to proceed through the questionnaire.

In order to investigate the degree to which use of test accuracy information by respondents was informed the questionnaire presented a hypothetical, unnamed, new triage test for

referral of women for investigation of ovarian cancer. The design of the scenarios was informed by the BBC pilot. An unnamed test was used following the observation in the pilot that use of a named test (CA125 as a triage test for referral for specialist investigation for ovarian cancer) resulted in respondents drawing on published commentaries and evidence summaries about the CA125 test for their management decisions, rather than the test accuracy information provided to them.

In order to reduce context-specific 'noise' and potential framing effects other than those that might be associated with test accuracy presentation itself (for example framing effects may be introduced by presentation of test errors as opposed to correct test results) a neutrally framed scenario identical with the exception of the method of presentation of test accuracy, was used.

5.4.2.2 Self-reported use of test accuracy information

The extent to which clinicians seek and use quantitative estimates of test accuracy and pre-test probability does not appear to have been addressed by existing literature (see chapter 2). Open responses from the BBC pilot indicated that clinicians relied heavily on colleagues when making decisions about test use. The final questionnaire therefore included questions concerning frequency of use of a range of test accuracy information sources common to primary and secondary test accuracy research and potential barriers to their use.

5.4.2.3 Assessment of familiarity and perceived understanding of test accuracy metrics and graphics

The emphasis in the literature reviewed in chapter 2 was the ability of healthcare professionals to manipulate a limited range of test accuracy summary metrics for the purpose of probability revision. Test accuracy metrics more common to systematic reviews were neglected. This survey investigated the familiarity of respondents with test accuracy metrics common to primary test accuracy evaluations (sensitivity, specificity, predictive values (PVs),

likelihood ratios (LRs)) as well as those more common to systematic reviews (the Receiver Operator Characteristic (ROC) curve, the Area Under the Curve (AUC) and the Diagnostic Odds Ratio (DOR).

5.4.2.4 Informed application of test accuracy information

The majority of empirical research attempting to evaluate clinician understanding and application of test accuracy metrics (as distinct from familiarity with metrics) is characterised by a requirement for respondents to undertake formal probabilistic reasoning. This is unlikely to be representative of how test accuracy information is used in practice and may not be a necessary pre-requisite for the appropriate use of test accuracy information for decision making. Consistency of respondents' management decisions following provision of test accuracy information and a test result (positive or negative), in combination with respondents' open comments, were therefore used as a proxy for informed application of test accuracy metrics in this survey (see 5.5.5 below). Open responses from the initial pilot survey indicated that respondents were distinguishing between the ability of a test in two dimensions: detecting disease and ruling out disease. The final questionnaire was therefore designed to allow responses separately for each dimension of accuracy: *'If the test came back positive would you refer for further investigation?'; 'If the test came back negative would you be confident not to investigate further at this time?'*

Separation of test accuracy dimensions also allowed an assessment of any distinction made between false positive and false negative test errors and relative ease of application of negative and positive test results; the latter being an issue raised by the literature reviews (2.5.3.3; 2.9.3.3).

Self reported variation in tolerance of test errors was evident in the BBC pilot, (tolerance of false negatives was reported to be markedly less in screening compared to diagnostic applications of tests and tolerance of test errors was less for tests used to diagnose more serious disease compared to less serious disease). In order to investigate the extent of

variation in test-treat thresholds across the sample elicitation of respondents' tolerance of test errors for the testing context used in the scenarios was elicited after they had submitted responses to the hypothetical scenarios (see section 5.4.6 below).

Test accuracy presentation formats evaluated in hypothetical scenarios

The literature reviews presented in chapter 2 include few examples of direct comparisons of informed application of commonly used summary test accuracy metrics. The few existing direct comparisons have advantaged LRs over other metrics by the use of plain language explanations and there has been minimal evaluation of understanding of PVs despite them being promoted as more intuitive test accuracy metrics. The priority for this survey was therefore a comparison of the informed application of summary test accuracy metrics likely to be more familiar to practising healthcare professionals: sensitivity, specificity, LRs and PVs. In addition, an annotated 2x2 diagnostic contingency table was included as a natural frequency presentation format and sensitivity and specificity and PVs were presented using a normalized frequency presentation format in addition to their more conventional percentage probabilistic representation. The inclusion of frequentist representations of probabilistic information was included in recognition of the considerable body of evidence suggesting that these may be more accessible for probability revision (2.5.3.2). An additional consideration is that frequentist representations communicate test errors more explicitly than conventional test accuracy summary measures and use of test errors appeared a prominent aspect of diagnostic decision making in the BBC pilot survey. Finally, in response to the suggestion from the review of the risk communication literature that use of multiple presentation formats are preferred over the use of any single format alone, an annotated pictographic (numeric, verbal and graphic) representation of test accuracy was included.

The target sample for this survey comprised practising healthcare professionals and the hypothetical scenarios required the application of test accuracy information for diagnostic decision making at the bedside. The expectation was that test accuracy metrics not

distinguishing between the two dimensions of test accuracy would not be helpful for decision making in this context. However the DOR was included in order to gain insight into respondents' understanding of the properties of global test accuracy measures distinct from those metrics that allow estimation of disease probability following a test result (positive or negative).

A paper based version of the complete survey can be found in appendix 5.1

5.4.3 Synthesis

Survey results were collated in Microsoft Office Excel 2007. Synthesis was mostly descriptive. Chi squared tests for paired data were undertaken in STATA IC11.

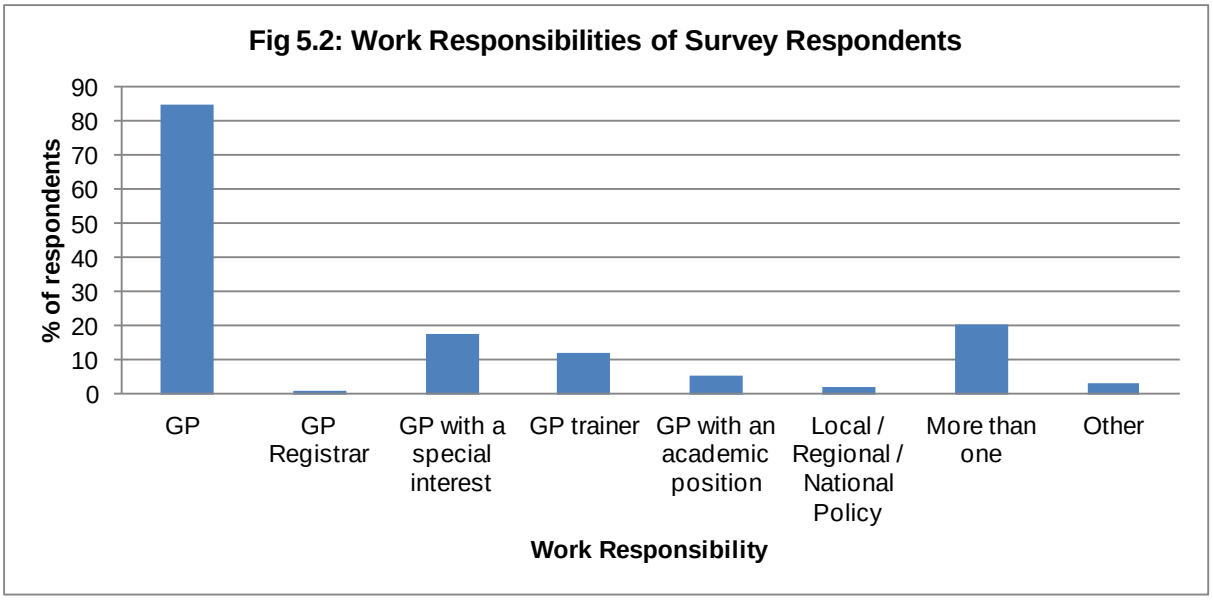
5.5 Results

5.5.1 Description of survey participants

A total of 222 UK GPs accessed the incentivised survey link via Doctors net. The survey distribution was designed to ensure responses were geographically representative. Seven respondents accessing the link met eligibility criteria but were excluded as they came from regions with sufficient representation; 11 respondents met eligibility criteria but of these three declined to participate once the survey topic area (test accuracy measures and their use in practice was revealed (see appendix 5.1) and eight respondents agreed to participate but did not complete the survey. Two hundred and four of 215 eligible participants (95%) completed the survey in full and the analysis is based on these 204 complete responses.

Sixty four percent of respondents were male; 75% of respondents were in full time employment and the number of years since qualification in the specialty ranged from 0-41 (median 14 years) see figure 5.1.

The distribution of work responsibilities across respondents is detailed in figure 5.2:



Although respondents who are GP trainers, who hold academic positions or who are involved in policy development may have a greater familiarity with evaluation of test accuracy and test accuracy metrics, few respondents (11%) were in any of these positions. In addition only 13% of respondents had undertaken training that included test accuracy interpretation in the last three years.

Figure 5.3 illustrates the geographical distribution of respondents. Stratified sampling was used by the survey host in order to reflect the geographical distribution of GPs across the UK.

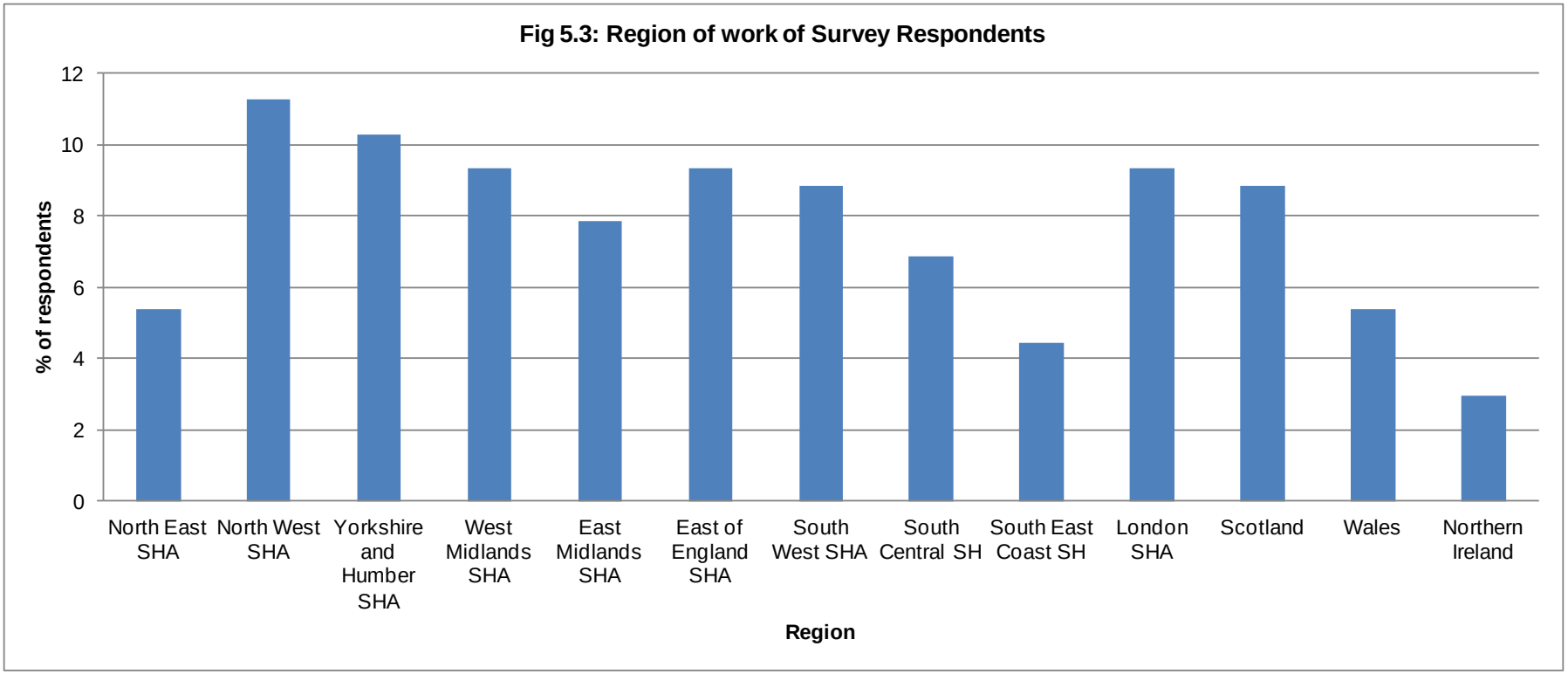
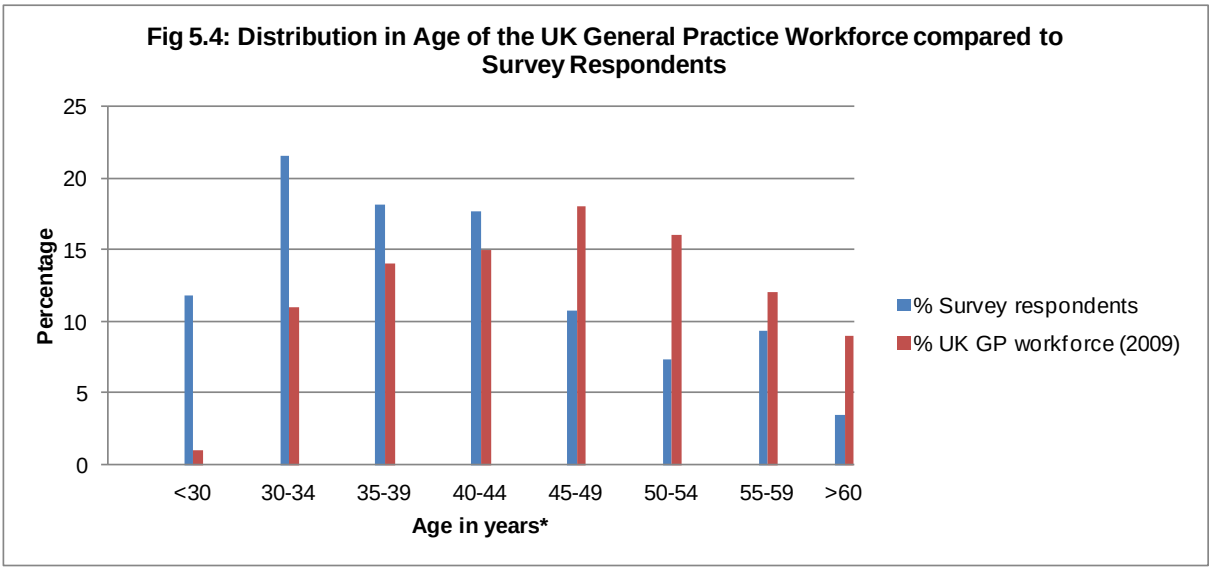


Fig 5.3: Region of work of Survey Respondents
% of respondents
12
10
8
6
4
2
0
North East SHA
North West SHA
Yorkshire and Humber SHA
West Midlands SHA
East Midlands SHA
East of England SHA
South West SHA
South Central SH
South East Coast SH
London SHA
Scotland
Wales
Northern Ireland
Region

Some demographics of respondents to this survey can be compared with workforce surveys of UK GPs [121,122]. Respondents to this survey were more likely to be male compared to the GP workforce (64% compared to 57% respectively). However there are fewer full time female GPs [121] and this may at least in part account for the lower percentage of female practitioners responding to the survey. Respondents to this survey appear to be younger compared to the UK GP workforce. Twenty two percent of GPs in the UK are older than 55 years in age compared to an estimated 12% of this survey sample [121] (see figure 5.4). However this comparison is limited by the fact that years spent practising in general practice and not age in years was recorded for survey respondents (see foot notes to figure 5.4). Using years spent in medical practice is likely to underestimate age in years of survey respondents as career breaks or variable periods of time spent practicing prior to gaining specialist certification in general practice will not have been accounted for. The discrepancy in age between survey respondents and the UK GP workforce is therefore likely to be less than that suggested by figure 5.4.



Notes to Fig: 5.4: *Age of survey respondents was estimated from reported years specialising in General Practice: (years in general practice <5=<30 yrs; years in general practice 5-9 =30-34 years; years in general practice 10-14=35-39 yrs; years in general practice 15-19=40-44 yrs; years in general practice 20-24=45-49 yrs; years in general practice 25-29 =50-54 yrs; years in general practice 30-34=55-59 yrs; years in general practice >35 = > 60 years)

5.5.2 Test accuracy information sources used by respondents

"Please estimate how often you use the following test accuracy information sources as part of your clinical work"

Figure 5.5 illustrates how respondents stated they used different sources for information about test accuracy. Clinical experience was used as a source of test accuracy information by most respondents, (99 (49%)). Ninety three respondents (46%) considered that the laboratory normal range conveyed information about the accuracy of a test; this was elaborated on in free text responses:

"Textbooks/research papers/guidelines are usually unhelpful as the ranges used in them are not always relevant to the tests performed at the local hospital lab. and (they) may even use different units esp(ecially) if (the)research or textbook (is) from (a) different country."

"Usually not appropriate to look through textbooks or published articles for results - we have good relationship with our labs and departments and use their data for test accuracy information."

Clinical guidelines and colleagues were also frequently used as sources of test accuracy information whilst research papers were used least. One hundred and eighty eight of 204 respondents, (92%) stated that they used other information sources, the majority of these, (76/188), used web-based resources. Specific sites mentioned by respondents were www.gpnotebook.co.uk , www.patient.co.uk, Egton Medical Information Systems Ltd (EMIS) Mentor (www.emis-online.com/mentor-www.doctors.net.uk , the British Medical Journal, the British National Formulary (BNF), the Royal College of General Practitioners, BMJ clinical evidence, www.clinicalevidencebmj.com and NHS clinical knowledge summaries, www.cks.nhs.uk, (see appendix 5.2). On the basis of brief internet browsing of these online resources the diagnostic information provided is in the form of guidelines for assessing presenting complaints and managing healthcare conditions or in the form of evidence based summaries with an emphasis on the effectiveness of interventions in the absence of numerical test accuracy information. Some of these online resources do provide evidence

based medicine training related to test evaluation, for example BMJ clinical evidence: http://clinicalevidence.bmj.com/ceweb/resources/EBMtraining.jsp

Other information sources included direct contact to discuss results with the local laboratory (n=2), patient generated information (n=1), direct contact with specialists in secondary care (n=2), local protocols (n=1), the ‘general press’ (n=1), ‘GP magazines’ (n=1) and continuing professional development (n=3).

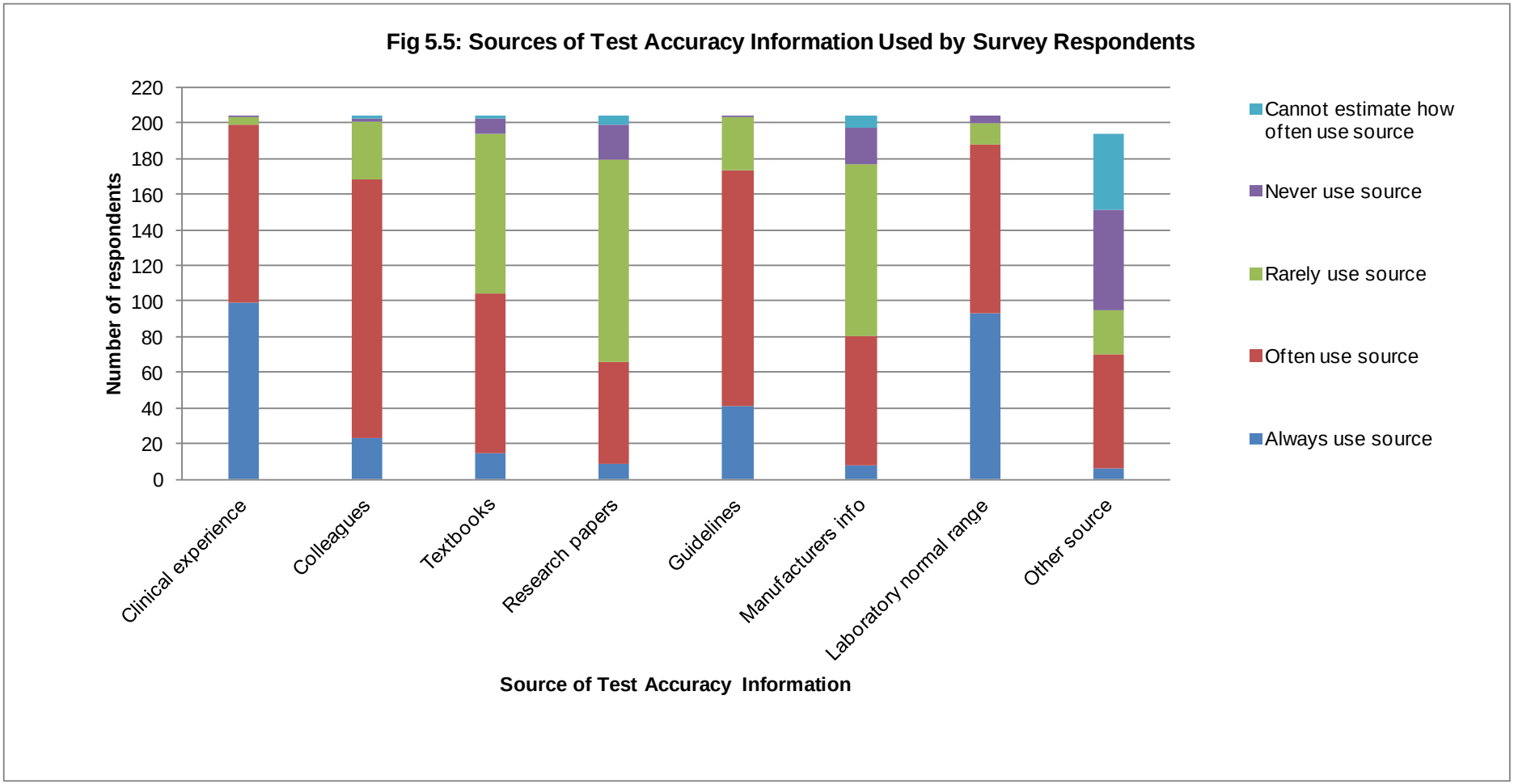


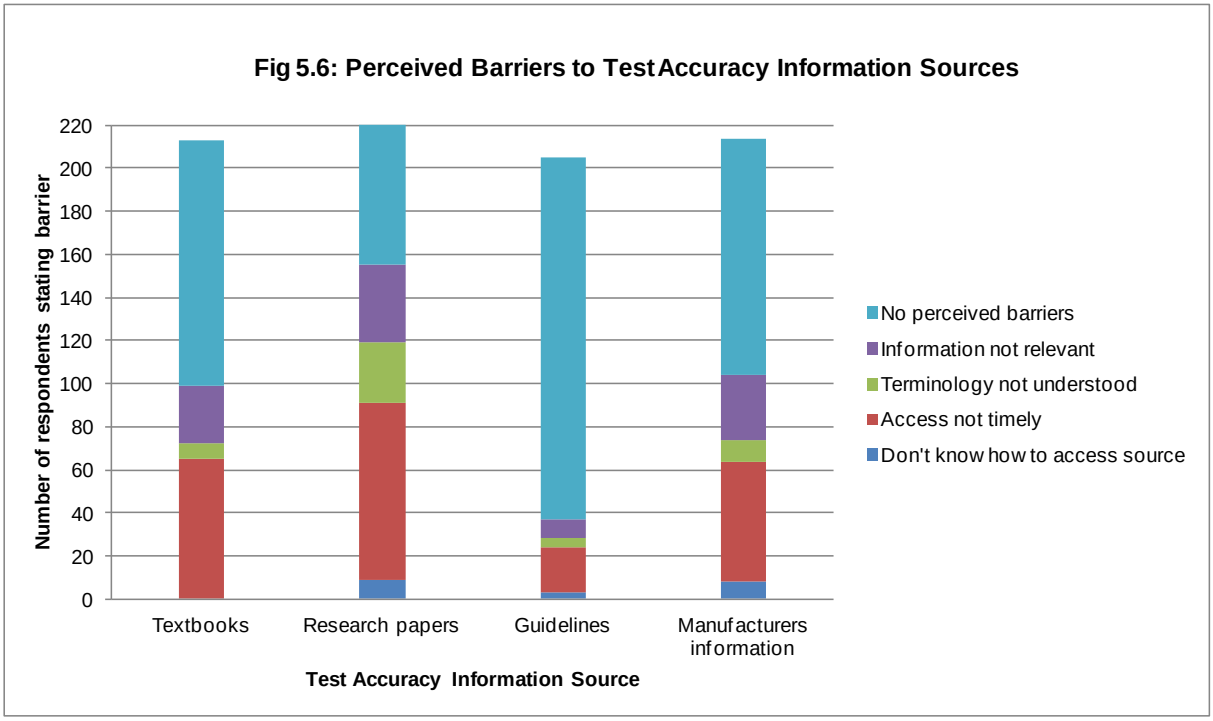
Fig 5.5: Sources of Test Accuracy Information Used by Survey Respondents
Number of respondents
220
200
180
160
140
120
100
80
60
40
20
0
Clinical experience
Colleagues
Textbooks
Research papers
Guidelines
Manufacturers info
Laboratory normal range
Other source
Source of Test Accuracy Information
Cannot estimate how often use source
Never use source
Rarely use source
Often use source
Always use source

5.5.3 Barriers to use of information sources

"Please indicate which statements apply to the following test accuracy information sources: Text books; Research papers; Guidelines; Manufacturers' information:

- *I don't know how to access the source*
- *I can't access the source at the time I need information*
- *The source uses terminology I don't understand*
- *The source does not contain information relevant to my practice*
- *None of the statements apply"*

Figure 5.6 illustrates the perceived barriers to use of test accuracy information sources.



Guidelines were perceived to present the least barriers to use with 164 (80%) of respondents suggesting there were no barriers to their use. Research papers were perceived to present the most barriers to their use. The most commonly suggested barrier to use of all resources was timeliness of access (53% to 66% of all barriers per source) followed by relevance of information (23% to 29% of all barriers per source). Ability to understand terminology used was suggested to account for 7% to 18% of all barriers per source with research papers prominent in this respect.

Free text responses revealed a mistrust of industry generated information:

“Manufacturers information is sometimes hard to trust.”

“Of manufacturers information: (it is) often skewed to make the drug look good.”

“Accuracy is often dependant on the source and funding of the source.”

The internet and guidelines were viewed positively with respect to timeliness and currency of information and textbooks negatively:

“Time is often the greatest factor.”

“No problem following guidelines/reading textbooks. Main problem is finding relevant guidelines when needed.”

“Difficult to access original research in a timely manner.”

“(I) use the internet the most as (it is) quick.”

“Guidelines most easy to access.”

“Text books - OK if I have a problem that I can look up in library/at home, not so good generally if I'm in surgery.”

“Sometimes difficult to do a relevant search of research papers and would tend to look for 'summary' of evidence.”

“With guidelines I accept that they have clarified the test accuracy information source and rely on them - the same with text books.”

“It is usually easier to access information via internet than rely on printed material which may not be up to date.”

“Text books are becoming less useful as out of date quickly and can gain information via internet.”

“Textbooks - often out of date.”

Explanations as to why terminology was not understood included difficulty with the statistical aspects of test accuracy information:

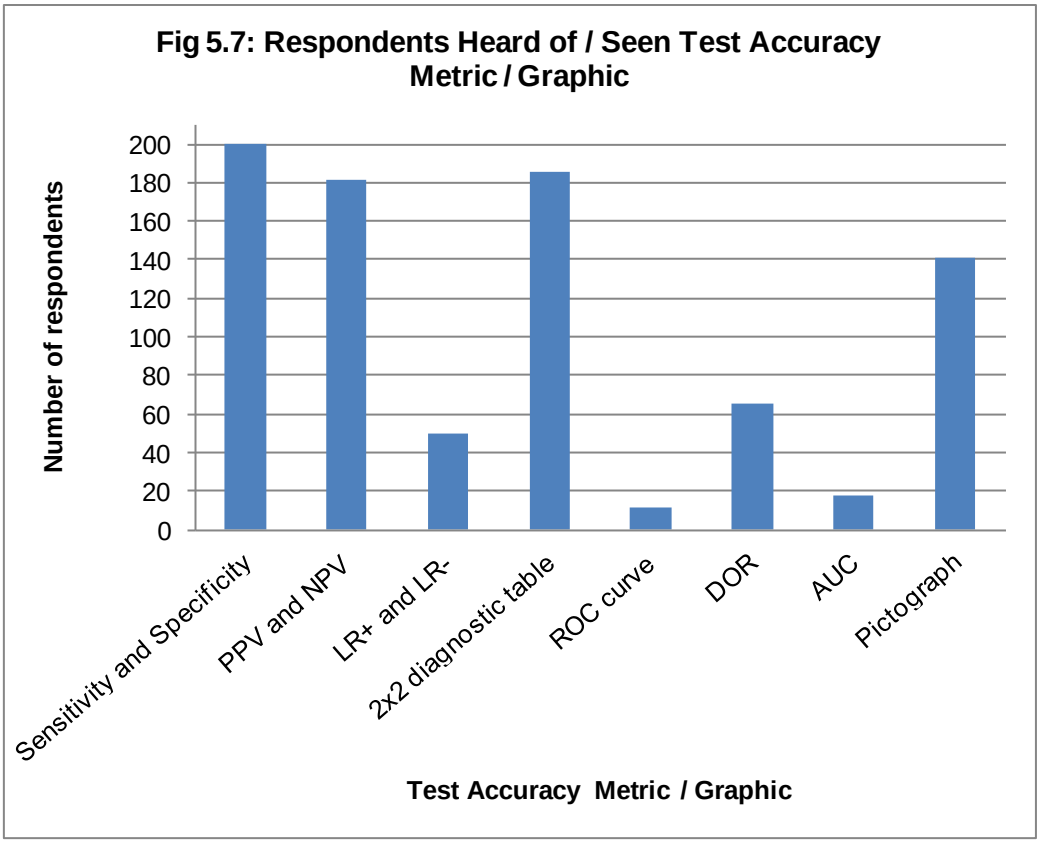
“(Research papers) - *I don't know if my knowledge of stats is good enough to use this source appropriately.”*

5.5.4 Utility of test accuracy metrics for clinical decision making

5.5.4.1 Familiarity with test accuracy metrics and graphics

"Have you heard of the measures sensitivity and specificity?"

Figure 5.7 demonstrates the familiarity of survey respondents with test accuracy metrics and graphics.



Notes to Fig 5.7: illustrative diagrams of the annotated 2x2 diagnostic contingency table and an example of a pictograph were provided for respondents

Sensitivity and specificity, the annotated 2x2 diagnostic table and PVs were reported to be familiar metrics by the most respondents. Sixty nine percent of respondents had seen an annotated pictograph which does not reflect the use of this graphic to communicate test accuracy information in the research literature. It is likely that familiarity with the pictograph is associated with its use in the risk communication literature (5.5.4.4) and this is supported by open responses (see below). Only 26% of respondents indicated that they had heard of LRs.

Test accuracy metrics more commonly used in systematic reviews of test accuracy (DOR, AUC and ROC curves) were familiar to 32% of respondents or less; the relatively greater reported familiarity of the DOR is likely to represent confusion with the odds ratio (OR).

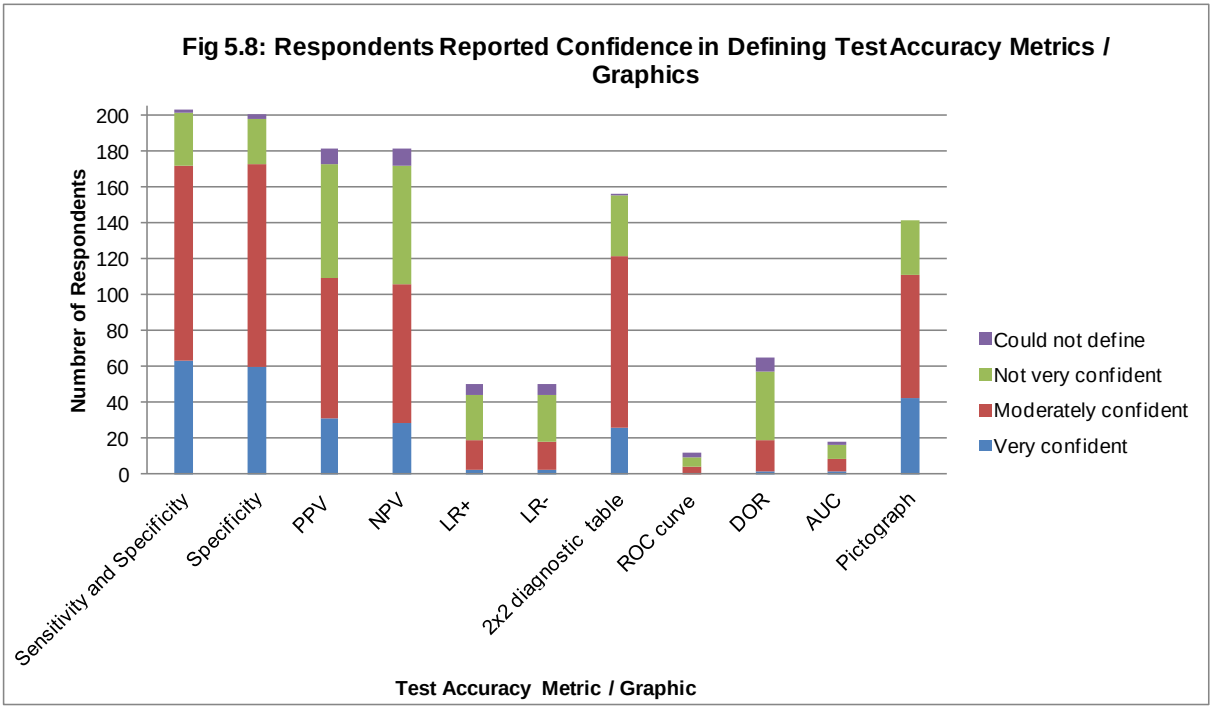
5.5.4.2 Perceived ability to define or explain test accuracy metrics

"How confident would you be in defining sensitivity and specificity to a colleague?

- *Very confident*
- *Moderately confident*
- *Not very confident*
- *Could not define"*

Figure 5.8 demonstrated the self-reported confidence of respondents in defining or explaining test accuracy metrics and graphics.

Fig 5.8: Respondents Reported Confidence in Defining Test Accuracy Metrics / Graphics



Respondents' reported confidence in defining metrics followed a similar pattern to reported familiarity. Of those respondents that had heard of a metric or graphic, sensitivity, specificity, the pictograph and the annotated 2x2 diagnostic table were those that respondents reported having most confidence in defining, followed by PVs (see table 5.9). There was little

discrepancy between respondents' reported confidence in defining those metrics representing one or other of the two dimensions of test accuracy with slightly more (2%) of respondents who had heard of sensitivity and specificity reporting being very or moderately confident in defining specificity compared to sensitivity, (1%) more of respondents who had heard of the positive predictive value (PPV) and the negative predictive value (NPV) reporting being very or moderately confident in defining PPV compared to NPV and (2%) more of respondents who had heard of the positive likelihood ratio (LR+) and the negative likelihood ratio (LR-) reporting being very or moderately confident in defining LR+ compared to LR-.

Table 5.9: Percentage of respondents who had heard of a test accuracy metric who reported being able to define that metric

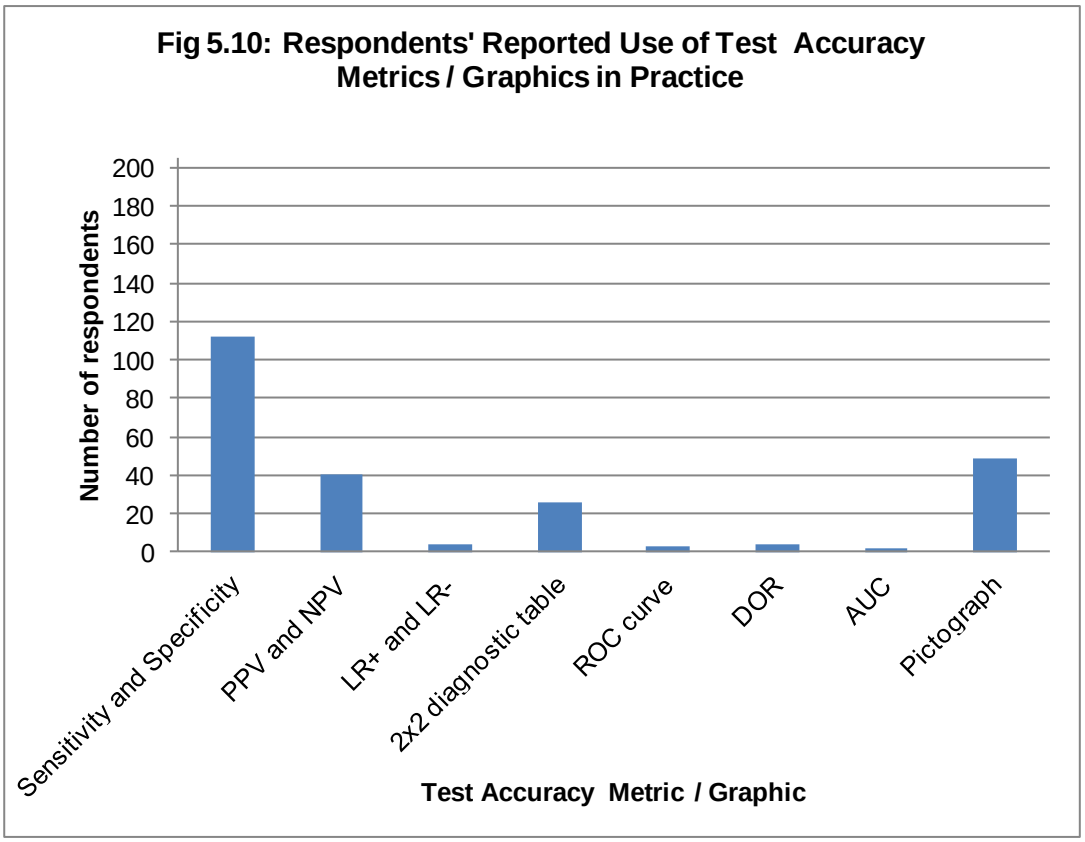
Metric / Graphic	Number of respondents heard of metric	% of respondents heard of metric reporting very or moderate confidence in defining
Sensitivity	203	85%
Specificity	200	87%
Pictograph	141	79%
2X2 diagnostic table	156	78%
Positive Predictive Value (PPV)	181	60%
Negative Predictive Value (NPV)	181	59%
Area Under the Curve (AUC)	18	44%
Positive Likelihood Ratio (LR+)	50	38%
Negative Likelihood Ratio (LR-)	50	36%
Receiver Operator Characteristic Curve (ROC)	12	33%
Diagnostic Odds Ratio (DOR)	65	29%

Less than 50% of respondents who had heard of the AUC, LRs, the ROC and the DOR responded that they would be very or moderately confident in defining them.

5.5.4.3 Use of test accuracy metrics in practice

- *"Do you use sensitivity and specificity in clinical practice? (Yes / No)"*
- *"Please comment on how you use the measure sensitivity in practice"*
- *"Please comment on how you use the measure specificity in practice"*

Figure 5.10 illustrates which metrics and graphics respondents reported using in practice.



As expected, reported use of metrics and graphics follows a similar pattern to respondents' confidence in defining them, with the exception that PVs were reported to be used more than the diagnostic 2x2 table, despite a relatively greater proportion of respondents stating that they could explain the 2x2 table. A pattern across all metrics and graphics is that their reported use in practice is much less (approximately 1/3) the number of respondents' stating they would be very or moderately confident in defining them. This observation may be a result of reporting patterns of metrics and graphics in the test accuracy literature or

alternatively a reflection of differences in the utility of metrics and graphics specifically for diagnostic decision making. Problems with timely access to test accuracy information may also explain the observed discrepancy between familiarity and use across all metrics.

5.5.4.4 Open responses concerning use of test accuracy metrics / graphics in practice

Open responses concerning use of test accuracy metrics and graphics were received from between one (<1%) and 102 (50%) of respondents.

Sensitivity and specificity (N=102 (50%))

Several respondents related their use of sensitivity and specificity to specific tests and this was mostly in the context of ruling out disease which is probably a reflection of the primary care sample [(TTA8)].

Sensitivity/rule out:

"*The d-dimer test, high sensitivity so if negative I am happy the patient does not have a DVT* (deep vein thrombosis)."

"D-dimer sensitive but not very specific."

"In discussion with patients how reassuring a negative result - eg PSA (prostate specific antigen) *will mean that the (patient) should not have a particular condition - eg prostate cancer."*

"In explaining the usefulness of some screening tests to patients."

"We use it to explain test results to pts e.g. when explaining rheumatoid factor results."

"Eg when we use inflammatory markers, they are sensitive for inflammation but not specific."

(Sensitivity)..."*I think about it when ordering a test and when assessing the result eg PSA is a very sensitive test but it still misses some patients with prostate cancer and one has to remember that when deciding what to do with a particular patient...*(Specificity) *It is especially important that a test for a serious illness is very specific as otherwise you may treat someone with a potentially toxic agent unnecessarily. PSA test eg is not specific to prostate cancer and is raised in BPH* (benign prostatic hypertrophy) *and uti* (urinary tract infection)."

(Sensitivity)..."*Urine dipstix etc"*

(Sensitivity)..."*D-Dimer."*

(Sensitivity and Specificity)...*"I don't measure* (them)*, but place more emphasis on more sensitive tests."*

"I prefer tests to be sensitive in order to reduce false negatives and pick up the disease that I am looking for - sensitive tests give me peace of mind. I also prefer specific tests ie few false positives but I can live with them and work round them."

Specificity/rule in:

"Explaining the pitfalls of a screening test eg PSA (prostate specific antigen) *which may not be specific to prostate cancer."*

"When want to diagnose specific medical condition like sle (systemic lupus erythematosus) *etc."*

(Specificity)... *"Screening eg faecal occult blood."*

(Specificity)... "*Troponin."*

Several respondents used the open response option to define sensitivity and specificity

without any indication of how the metrics were used in practice:

"(Sensitivity)...*The percentage of sick people who are correctly identified as having the condition...*(Specificity)...*The percentage of healthy people who are correctly identified as not having the condition."*

"(Sensitivity)...*The proportion of actual positives which are correctly identified...*(Specificity)... *The proportion of negatives which are correctly identified."*

"(Sensitivity)...*Reflects a tests being positive in the presence of disease true positive divided by true positive plus false negative...*(Specificity)... *Reflects the test being negative in absence of disease false negative divided by false negative plus true positive."*

(Sensitivity)... *"Tp/Tp+FN"...*(Specificity) *"TN/TN+FP"*

Some definitions of sensitivity and specificity offered by respondents were more closely

linked to how the respondents applied them in practice:

(Specificity)... "*Deciding whether test will point towards one likely disease, or many possible diseases.*"

(Sensitivity)... *"How accurate a test comes out positive when the patient is positive"*... (Specificity)..."*How accurate a test comes out negative when the patient is negative."*

(Sensitivity)... "*The likelihood of detecting the disease"...*(Specificity)... "*The likelihood of detecting other conditions."*

Several responses indicted that sensitivity and specificity were used when communicating with patients about tests:

"Sometimes use the term when discussing value of test with patient."

"Explaining to patients the limitations of a test. Deciding if and when to use a test."

"I explain the concepts to patients in order for them to understand that a test is rarely definitive."

Positive and negative predictive values PVs (N=36/204 (18%))

As with sensitivity and specificity, several respondents related their use of PVs to specific tests although the emphasis on ruling out disease was not evident:

(PPV)... "*Talk about things like rheumatoid factor"...*(NPV)... "*When explaining e.g. d-dimer."*

(PPV)... "*eg consolidation on X ray when pneumonia is suspected...*(NPV)... *"eg normal chest X ray when TB is suspected."*

(PPV)... "*When thinking of elements of a history that would lead to a diagnosis"...* (NPV)... "*When thinking of tests to exclude certain conditions eg celiac."*

(PPV)... "C*ardiac enzymes"...*(NPV)... "*d-dimer"*

As for sensitivity and specificity several respondents illustrated their use of PVs by defining the metrics:

(PPV)... "*Number of people who have a screening test that test(s) positive that have the disease"....* (NPV)... "*Number of people who have a screening test* (that tests negative) *correctly identified as being free from disease."*

(PPV)... "*Proportion of subjects with positive test results who are correctly diagnosed"...* (NPV)... "*Proportion of subjects with a negative test result who are correctly diagnosed."*

(PPV)... "*People with positive result who have disease...*(NPV)...*People with negative test who do not have disease."*

(PPV)... "*Looking at a test, I say 'If the test says YES, how likely is it that that is a TRUE YES?"...* (NPV)... *"Looking at a test, I say 'If the test says NO, how likely is it that that is a TRUE NO?'"*

(PPV)... "*TP/TP+FP"...*(NPV)... "*TN/TN+FN."*

Positive and negative likelihood ratios (LR+ and LR-) (N=4/204(2%))

The low number of open responses for use of LRs reflects the low number of respondents across the whole sample indicating that they use these metrics in practice (see figure 5.10). Two of the four responses concerning LRs were definitions of these metrics with no indication of how they were used:

(LR+)... "*the probability of a person who has the disease testing positive divided by the probability of a person who does not have the disease testing positive"...*(LR-)... "*the probability of a person who has the disease testing negative divided by the probability of a person who does not have the disease testing negative."*

"LR+ = Sensitivity / False positive rate ...LR- = *False negative rate / Specificity.*"

One respondent related the use of LRs to the value of the clinical history:

"*When taking a history to ascertain likelihood of diagnosis.*"

This may be a reflection of the promotion of LRs in general practice settings to evaluate the utility of multiple components of the clinical history and examination (2.4.1.6).

One open response relating to use of LRs in practice referred to guidelines for interpretation but it was unclear whether the reference was to a generic guideline to assist with interpretation (for example Fagan's nomogram [63] or a rule for interpreting the magnitude of likelihood ratios [62], or whether the reference was specific to a disease or laboratory:

"We refer to the guideline for positive likelihood ratio / negative likelihood ratio or phone the lab for advice."

2x2 Diagnostic table (N=24/204 (12%))

The predominant response was use of the 2x2 table as a method for explaining test accuracy metrics and test results to colleagues, patients and students:

"When discussing screening e.g. prostate screening with colleagues or patients."

"Teaching trainees."

"Teaching medical students."

“eg PSA testing, to explain possible outcomes to patient asking for the test.”

“Explaining the importance of a screening test to a patient.”

“Patient education when (usually) trying to reassure them after a bad result.”

“Sometimes draw table in explaining to patients.”

Whilst several respondents indicated that they used the 2x2 table as the basis for deriving and interpreting test accuracy metrics:

“That is in my head.”

“Calculate sensitivity.”

Receiver Operator Characteristic (ROC) Curve (N=1/204(0.4%)), Area Under the Curve (AUC) (N=2/204 (0.9%)), Diagnostic Odds Ratio DOR) (N=3/204(1%))

All responses concerning the ROC curve, AUC and DOR were general, non specific statements or responses demonstrating limited understanding of these metrics:

(DOR) “*To explain mammograms.”*

(ROC) “*Interpreting clinical / experimental studies.”*

(AUC) *“Bells curve.”*

As for LRs, the low number of open responses for the ROC curve, AUC and DOR reflects the low number of respondents across the whole sample reporting that they use these metrics in practice, (see figure 5.10).

Annotated pictogram (N=45/204 (22%))

Annotated pictograms were the second most commonly used metric / graphic across the whole sample after sensitivity and specificity (see figure 5.10) and invited the second highest volume of open responses regarding use in practice. The majority of open responses referred to the use of pictograms in communicating risk more generally as opposed to the uncertainty conveyed by test accuracy, reflecting the findings of the literature reviews (2.5):

"Predicting the cardiovascular risk and risk reduction which is achieved."

"To explain Framingham scores."

"10 year heart disease prediction."

"…on our computer system (eg cardiovascular risk) - discussing potential treatment with patients."

*"Explain extra risk of breast cancer in HRT (*hormone replacement therapy*)."*

"Warfarin counselling in AF (atrial fibrillation)*."*

"Smoker at risk of cancer."

"Pill and breast cancer."

The use of pictograms for explaining difficult concepts to patients, such as the number needed to treat (NNT) was also prominent:

"Explaining difficult nnt etc concepts to patients."

"Statins NNT."

A minority of open responses concerning the pictogram referred to its application to testing in practice:

"PSA (prostate specific antigen) *testing and breast screening."*

"In PSA (prostate specific antigen) *and Ca 125."*

"Smiley face chart for PSA (prostate specific antigen) *screening."*

"Very useful to explain value of test results to patients, or value of screening, etc."

In summary, the frequency of open responses to the use of test accuracy metrics in practice mirrors the relative use of metrics across the whole sample which suggests open responses were not unduly influenced by respondent fatigue, even though questions were not randomised. In addition, reference to specific tests when illustrating how the most frequently used metrics were used (sensitivity, specificity, PPV, NPV, and the pictogram) is evidence of the construct validity of open responses. Although the low response rate for open comments

(median of 7% across eight metrics) would suggest that they represent those more knowledgeable in the area of test accuracy, open responses including respondents' own definitions of test accuracy metrics are useful examples of language that might be accessible to clinicians more generally.

Many respondents used the open responses to illustrate their ability to define test accuracy metrics and illustrated this with practical examples including clear differentiation between the two dimensions of test accuracy. There was limited evidence of confusion about definitions of metrics and graphics although the voluntary nature of the open responses is likely to have deterred those with less confidence. The exceptions were metrics more commonly associated with systematic reviews of test accuracy (DOR, AUC and the ROC curve) and this could be seen as evidence that few if any respondents understand these metrics.

The 2x2 diagnostic contingency table and annotated pictogram were prominent as graphics used in practice to explain difficult concepts to patients, students and colleagues. Although the majority of practical examples for the pictogram concerned communication of risk more generally, there were examples where practitioners had used this graphic to communicate test accuracy.

5.5.5 Comparison of application of nine different test accuracy presentation formats to a common testing scenario

Table 5.11 and figure 5.12 illustrate the responses of the 204 GPs to a series of nine scenarios concerning a hypothetical new biological marker for ovarian cancer available to primary care physicians:

> *"The questions that follow will ask you to apply test accuracy information to clinical scenarios. Each scenario reflects a primary care setting where the prevalence of ovarian cancer in **asymptomatic**, post-menopausal women is ~3%.*
>
> *A new biological marker for ovarian cancer has been identified and is available as a blood test for use in primary care. A 57 year old asymptomatic woman presents to you concerned about her risk of ovarian cancer and you perform the blood test at her request."*
> *TEST ACCURACY INFORMATION PRESENTED IN ONE OF NINE DIFFERENT FORMATS*
> *"If the test came back positive would you refer the woman for further investigation?*
>
> *If the test came back negative would you be confident not to investigate further at this point in time?"*

Each scenario was identical with the exception of the test accuracy metric used to convey a constant estimate of test accuracy.

The test accuracy information provided across all scenarios was that a positive test result would increase the probability of having ovarian cancer (as indicated by sensitivity, LR+, PPV and the number of false positive test results) to a greater extent than a negative test result would decrease the probability of having ovarian cancer (as indicated by specificity, LR- , NPV and the number of false negative test results). It was anticipated that the management decision indicated by respondents for each scenario would be modified by variation in the test-treat (or for this hypothetical scenario test-refer) threshold across the sample. Therefore test-refer thresholds were elicited from respondents at the end of the survey in order not to influence responses to the hypothetical scenarios.

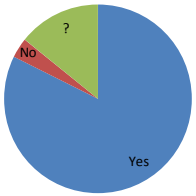
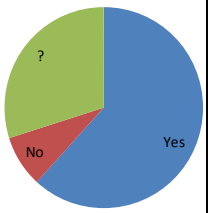
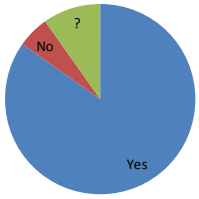
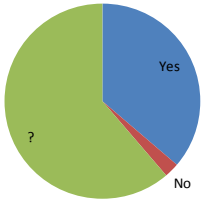
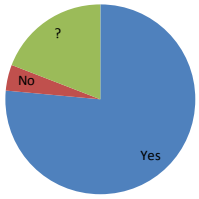
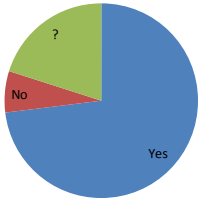
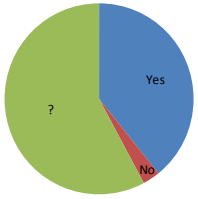
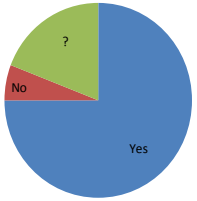
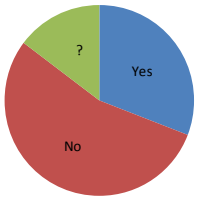
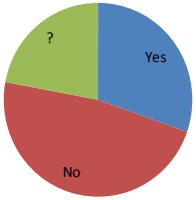
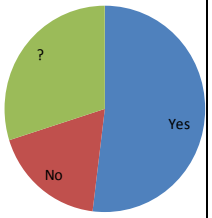
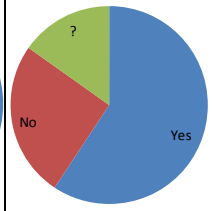
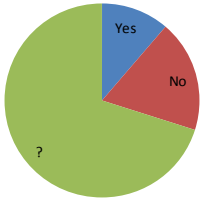
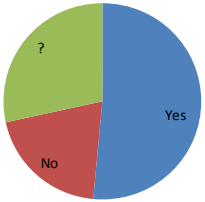
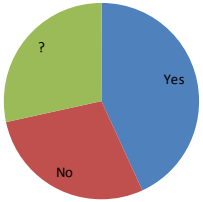
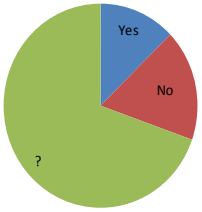
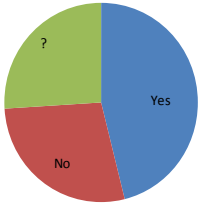
5.5.5.1 Across sample application of nine different test accuracy presentation formats to a common testing scenario

There was marked inconsistency in responses to both positive and negative test results across the sample. Respondents had particular difficulty applying LRs and the DOR with 61% and 58% of respondents respectively indicating that they did not know what management decision they would take on the basis of a positive test result and 70% of respondents indicating that they did not know what management decision they would take on the basis of a negative test result for both metrics.

Table 5.11: Responses to hypothetical scenarios using nine different test accuracy presentation formats

	Sensitivity & Specificity (%)	**Sensitivity & Specificity: normalised frequencies**	**Predictive values (%)**	**Predictive values: normalised frequencies**	**Likelihood Ratios**	**Pre-post test probability**	**Annotated 2x2 Diagnostic Table**	**DOR with a guide to interpretation**	**Annotated Pictograph**
	Sens 76% Spec 98%	76/100 diseased test+ve 98/100 disease-free test -ve	PPV 54% NPV 99%	54/100 test +ves have disease 99/100 test –ves disease- free	LR+ 38 LR-0.2	Pre-post +ve test 3% to 54% Pre-post –ve test 3% to 0.6%	Annotated 2x2 Diagnostic Table	DOR 190 with a guide to interpretation	Annotated Pictograph
*"If the test came back **positive** would you refer the woman for further investigation?"*									
Yes	93%	82%	62%	85%	36%	76%	73%	39%	75%
No	4%	3%	8%	5%	2%	4%	7%	3%	6%
Don't know	3%	14%	30%	10%	61%	19%	20%	58%	19%
*"If the test came back **negative** would you be confident not to investigate further at this point in time?"*									
Yes	31%	30%	52%	59%	11%	51%	43%	13%	46%
No	54%	48%	18%	25%	19%	20%	28%	18%	28%
Don't know	15%	22%	30%	15%	70%	28%	28%	70%	26%
	Sens 76% Spec 98%	76/100 diseased test+ve 98/100 disease-free test -ve	PPV 54% NPV 99%	54/100 test +ves have disease 99/100 test –ves disease- free	LR+ 38 LR-0.2	Pre-post +ve test 3% to 54% Pre-post –ve test 3% to 0.6%	Annotated 2x2 Diagnostic Table	DOR 190 with a guide to interpretation	Annotated Pictograph
	Sensitivity & Specificity (%)	**Sensitivity & Specificity: normalised frequencies**	**Predictive values (%)**	**Predictive values: normalised frequencies**	**Likelihood Ratios**	**Pre-post test probability**	**Annotated 2x2 Diagnostic Table**	**DOR with a guide to interpretation**	**Annotated Pictograph**

Fig 5.12: Responses to hypothetical scenarios using nine different test accuracy presentation formats

	Sensitivity & Specificity (%)	Sensitivity & Specificity: normalised frequencies	Predictive values (%)	Predictive values: normalised frequencies	Likelihood Ratios	Pre-post test probability	Annotated 2x2 Diagnostic Table	DOR with a guide to interpretation	Annotated Pictograph
	*"If the test came back **positive** would you refer the woman for further investigation?"*								
Biological marker +ve									
Biological marker –ve									
	*"If the test came back **negative** would you be confident not to investigate further at this point in time?"*								
	Sensitivity & Specificity (%)	**Sensitivity & Specificity: normalised frequencies**	**Predictive values (%)**	**Predictive values: normalised frequencies**	**Likelihood Ratios**	**Pre-post test probability**	**Annotated 2x2 Diagnostic Table**	**DOR**	**Pictograph**

Notes to table X: Yes; No; Don't know

Management decisions following a positive test result

The response to a positive test result when presented as sensitivity and specificity (%), sensitivity and specificity expressed as normalised frequencies and PVs expressed as normalised frequencies, resulted in the most consistency across the sample (> 85% respondents indicating they would refer on the basis of a positive test result). Presentation of pre to post-test probabilities derived from LRs, the annotated pictograph and the annotated 2x2 diagnostic table resulted in 76%, 75% and 73% of respondents respectively indicating they would refer on the basis of a positive test result. PVs presented in percentage terms resulted in less consistency across the sample with 30% of respondents indicating they did not know whether they would refer on the basis of a positive test result. Across all test accuracy presentation formats a minority of respondents indicated that they would not refer on the basis of a positive test result.

Management decisions following a negative test result

There is greater inconsistency in respondents' management decisions following a negative compared to a positive test result.

Sensitivity and specificity presented both in percentage terms and using normalised frequencies resulted in the majority of respondents indicating that they would pursue further testing despite a negative test result (54% and 48% respectively). The management decision indicated by the majority of respondents for the remainder of the test accuracy presentation formats (PVs (%), PVs (normalised frequency presentation), pre to post-test probability, the annotated 2x2 diagnostic table and the annotated pictograph), was a decision not to pursue further investigations following a negative test result (52%, 59%, 51%, 43% and 46% respectively).

The literature reviews presented in chapter 2 suggested that health professionals may have more problems with the application of negative test results (2.5.3.3). However evaluation of

the application of negative test results was sparse compared to evaluation of the application of positive test results. Table 5.13 presents a comparison of 'Don't know' management responses following positive and negative test results for each of the nine hypothetical scenarios. The table illustrates that with the exception of one scenario (PVs %), a greater proportion of respondents indicated they were not able to make a decision about referral, (responding 'don't know'), following a negative test result in comparison to a positive test result. This difference was significant at the 5% level for 5/9 scenarios. The differential between ability to apply positive and negative test results is least marked for a normalised frequency presentation of PVs and the annotated pictogram.

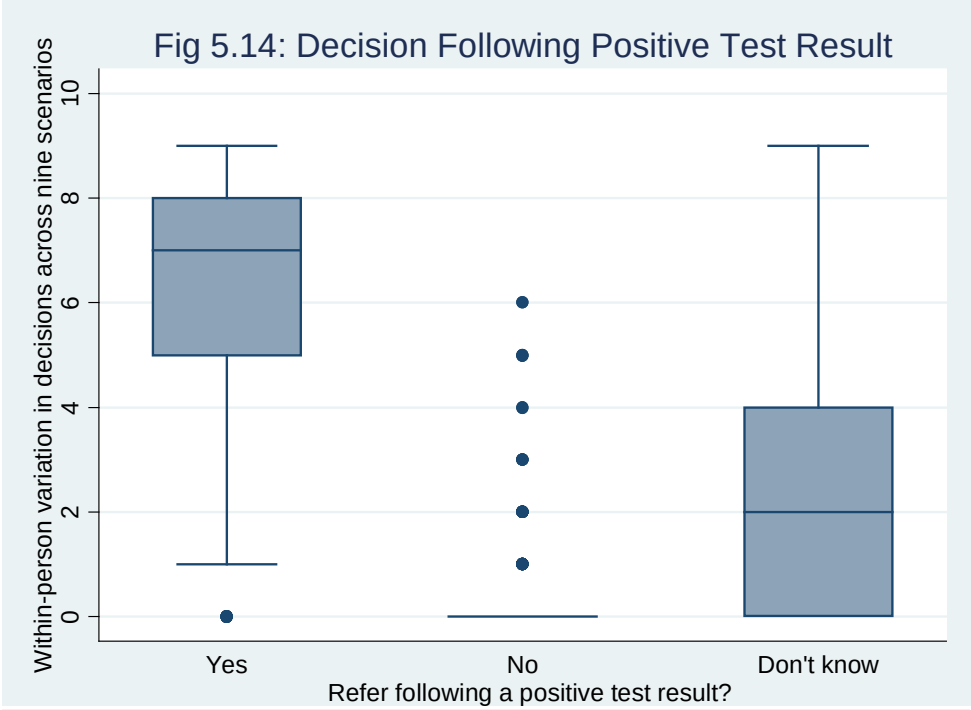
Table 5.13: Comparison of 'don't know' management responses following positive and negative test results for each of nine scenarios

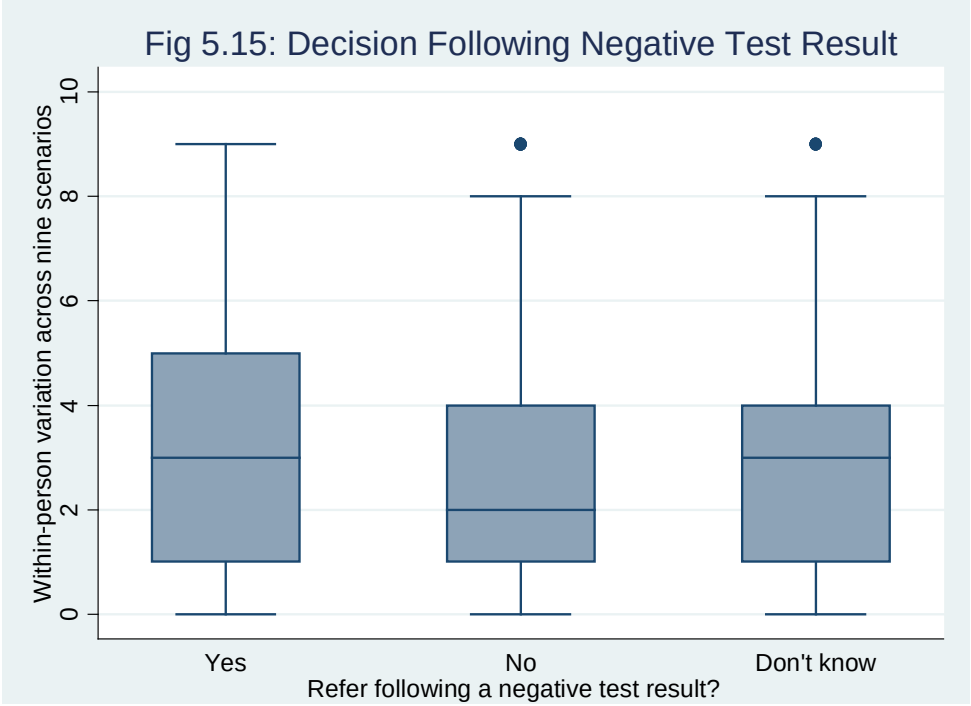
% 'Don't know' responses	Sensitivity & Specificity (%)	Sensitivity & Specificity (normalised freq.)	Predictive Values (%)	Predictive Values (normalised freq.)	Likelihood Ratios	Pre-post test probability	Annotated 2x2 Diagnostic Table	Annotated Pictograph	Diagnostic Odds Ratio
+ve test result	3%	14%	30%	10%	61%	19%	20%	19%	58%
-ve test result	15%	22%	30%	15%	70%	28%	28%	26%	70%
P value (Chi-square test)	0.00	0.04	1.00	0.10	0.06	0.03	0.05	0.10	0.01

5.5.5.2 Within-person application of nine different test accuracy presentation formats to a common testing scenario

Investigating the degree of decision making consistency for individual respondents controls for between-person variation in test-refer thresholds and tolerance of test errors.

Fig 5.14 and 5.15: Within Person Variation in Management Decisions Across Nine Scenarios Depicting a Constant Estimate of Test Accuracy





Figures 5.14 and 5.15 are concordant with the across sample distribution of management decisions (see table 5.11 and figure 5.12 above).

There is greater consistency in management following a positive test result with only six respondents (outliers) indicating they would not refer (for between one and six of nine scenarios) (figure 5.14). The median number of scenarios for which respondents indicated they would refer following a positive test result was seven, (inter-quartile range 5-8 of nine scenarios) and the median number unsure of their management decision following a positive test result was two, (inter-quartile range 0-4 of nine scenarios). On the basis of this examination of within-person responses it appears that the relatively greater consistency in management decisions following a positive test result observed across the whole sample appears to reflect within-person consistency rather than an artefact caused by random within-person variation across scenarios.

Following a negative test result the median number of scenarios for which respondents indicated they would refer in any case was two, (inter-quartile range 1-4 of nine scenarios) (figure 5.15). The median number of scenarios for which respondents indicated they would not refer following a negative test result was three, (inter-quartile range 1-5 of nine scenarios). The median number of scenarios for which respondents indicated they would be unsure of their management decision was three, (inter-quartile range 1-4 of nine scenarios). On the basis of this examination of within-person responses it appears that the inconsistent pattern of responses following a negative test result observed across the whole sample is real and reflects within-person *in*consistency rather than an artefact caused by random within-person variation across scenarios.

5.5.5.3 Open responses to application of nine different test accuracy presentation formats to a common testing scenario

Open responses concerning scenario management decisions were received from between 6% (for the last scenario) and 28% (for the first scenario) indicating respondent fatigue. Figure 5.16 illustrates the relationship between a respondent's median confidence in their ability to define or explain the test accuracy metrics/ graphics presented across the nine scenarios and the total number of open responses they provided across the nine scenarios. For the purposes of this analysis respondents were assigned to the 'could not define' category if they reported not having heard of a metric (see details of the skip logic application as part of the questionnaire design (5.4.2.1).

Fig 5.16: Median self-reported confidence (ability to define metrics) and number of open responses

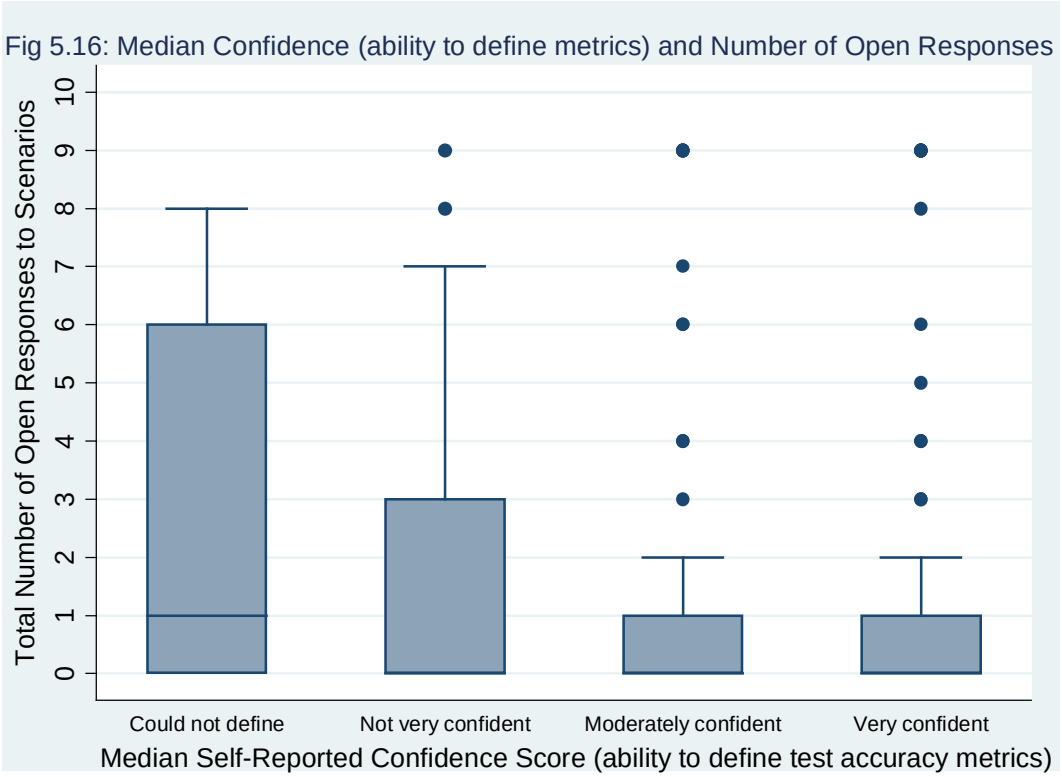


Figure 5.16 demonstrates that the inter-quartile range of open responses provided appears to decrease with an increase in the median self reported confidence of respondents to define

test accuracy metrics. It does not appear that respondents more confident in their ability to define metrics were also those more likely to offer open responses in support of their scenario management decisions.

With the exception of the LR and DOR scenarios, where the majority of open responses were concerned with lack of familiarity with these metrics, two themes emerged from free text responses to scenarios: the importance of shared decision making with patients and a requirement for additional diagnostic information including a feeling of obligation to pursue further testing even in the presence of negative test results.

Shared decision making

A common comment, particularly concerning negative test results, was the importance of involving patients in the decision making process:

"Would discuss further each case with p(atien)t"...(Not to investigate following a negative test result)... "*only after in depth discussion."*

"The NPV of 99% indicates a low risk of false negatives and I would discuss risks with patient and feel that further screening is not appropriate."

"If she were asymptomatic I would explain to her the short comings of the test and symptoms that she should report, etc."

"Again I would discuss with patient and inform them there is a 1% false negative rate and combined with a low prevalence it would be reasonable not to investigate further."

"Probably confident w(ith) lowering to 0.6%, but p(atient)t would need to be happy w(ith) that level of risk."

"Assuming patient aware of possibility of false negative test."

*"I would explain that 46% with a positive result are unaffected and if there is no FH (*family history*) and she is asymptomatic, these are favourable factors but would refer for specialist advice and further testing/imaging."*

"Patient choice as well - but if she wanted further referral I would do this."

The annotated pictogram attracted positive comments as a tool to facilitate shared decision making:

"Useful for showing patient."

"I would use the diagram to show the patient and explain result."

"This I understand and explain to patient is easy."

Requirement for additional diagnostic information

Several respondents commented that they would like additional diagnostic information to inform their management decision with the clinical history and examination prominent in this respect:

"Only 76% sensitive (therefore) *need to assess the whole hx* (history) *and risk."*

"Other clinical information will guide."

"Need clinical context and examination."

"It is sensitive, but not specific and no one test predicts."

"We cannot base diagnosis on a blood test alone. Ultrasound and examination would be useful in assessment probably prior to referral."

"As sensitivity is lower then (I) *would consider referral for USS* (ultrasound scan) *especially if FH* (a family history of ovarian cancer)*."*

"A lot depends on the clinical findings at examination."

Obligation to pursue further testing

Several respondents indicated that they would feel obliged to investigate further even in the event of a negative test result. This may reflect the specific clinical context depicted in the scenarios and the relatively greater importance placed on false negative test errors:

"I would refer -ve result here...would be difficult to defend if subsequently turned out to have ovarian carcinoma."

"Small number of false positives, so +ve results indicate disease highly likely. Although much larger proportion of false -ves, a wrong decision would be difficult to defend, especially with 3% prevalence of disease in this population."

"But would probably investigate (on the basis of a negative test result) *but* (I am) *aware all further tests may be negative."*

Some respondents identified that pursuing further investigations in the presence of a negative test result suggests that decision making should be at the point of consideration of test use as distinct from decision making following test results. This was contextualised with respect to the test and population depicted in the scenarios: the consequences of the higher false negative rate (lower sensitivity) for detection of a serious disease and high absolute numbers of false positives in a low prevalence population:

"She is asymptomatic. If you are going to Ix (investigate) *asymptomatic p(a)t(ient)s with neg(ative) tests, there is no point using the test."*

"I would try to explain the use of the test in an asymptomatic patient who I would otherwise not have been investigating for ovarian cancer."

"Would have to investigate further with positive test...would have concerns over doing test in view of high false positive rate."

"You would have to refer a positive result wouldn't you! I think counselling before the test (aka PSA (prostate specific antigen)*) would be the way forward as so many false positives."*

"If someone tests +ve even with a poor predictive test then I feel you cannot ignore it and not investigate further."

"Would need referral in view of high specificity. I am not sure I would have done the test in the first place in view of the low sensitivity."

"If negative, then her risk is lower than the normal population risk of 3%, and as I don't screen for the normal population, I wouldn't screen for her."

"Woman is asymptomatic, and after a negative test her likelihood of ovarian cancer is even less then general population of same age."

"The prevalence of ovarian cancer in the symptomatic population is only 3% and the risk of false negative result is 10:1360 and therefore I feel it is appropriate not to investigate further."

Understanding of Test Accuracy Metrics

Several respondents justified their management decisions by providing a definition of the metric being used. For sensitivity and specificity the majority of respondents emphasised the false negative rate:

"High specificity, so +ve indicates disease present Low-ish sensitivity - cannot exclude disease on basis of -ve result."

"Will miss about 1/4 +ve cases so would need f/u (follow up).*"*

"Only 76% sensitivity therefore will be negative in 24% of cases (of ovarian cancer).*"*

"The 98% specificity means that if positive it is almost certain the woman has ovarian cancer. The 76% sensitivity means it'll miss a quarter of cancers - to rule out the cancer, you need a better test. As a screening test, it is (to use the technical term) crap."

"I would not be confident not to investigate due to the sensitivity of 76%."

"24% women with ovarian cancer would test negative."

"24% false negative - too high."

"I would worry about the potential 24% who may be missed."

"I think the test sensitivity is poor - ie there are a lot of falsely negative results - ie less useful in terms of reassuring asymptomatic women."

A minority of open responses to the sensitivity and specificity scenarios suggested some confusion over application of the SpPIN and SnNOUT rules [(TTA27)]:

"Test with high specificity means fewer false negatives, ie a positive result is likely to be true."

"High specificity so if test -ve unlikely to have the disease, quite high sensitivity so +ve test needs Ix (investigation).*"*

"Specificity 98% not adequate for "rule out" test."

When presented with predictive values a greater emphasis was placed on false positive test errors, in contrast to the emphasis on false negative test errors observed for sensitivity and specificity:

"Sufficiently high risk to check +ve tests and sufficiently low risk 1% to monitor –ve (test results)."

If positive there is a 54% chance of cancer so need to investigate. If asymptomatic and negative I would not refer."

"Some concern over low positive predictive value,"

"A lot of healthy women would be investigated due to a positive result."

"Negative test reassuring. Would have concerns over doing test in view of high false positive rate."

"If positive she needs a better test to say whether is or isn't cancer: the odds are about 50:50; if -ve then cancer is highly unlikely, so 'see again if symptoms persist'."

This change in respondent emphasis on test errors may have arisen due to the considerable difference in magnitude between the PPV (54% or 54/100) and NPV (99% or 99/100) when presented in percentage or normalised frequency format due to the low disease prevalence in the scenario population. By contrast the difference between sensitivity (76% or 76/100) and specificity (98% or 98/100) when presented as either percentage or normalised frequency format was less marked.

Several respondents stated they had difficulty interpreting PVs when presented as a percentage:

"Not familiar with terminology here, presume PPV and NPV correspond with sensitivity and specificity but I would need to check."

"Not used to working with these figures."

"Can't understand this."

"Don't understand the terminology."

In addition there was evidence of reference class confusion between sensitivity and specificity and PVs when these metrics were presented using percentage presentation formats:

(PPV and NPV %)..."This test could lead to a lot of anxiety and distress in women who test positive although its specificity is quite low."

(PPV and NPV %)..."Test not sensitive enough-only 1% false negative."

In common with respondents' illustrations of their understanding of sensitivity, specificity and PVs, test errors were used by respondents when explaining the information contained in the 2x2 diagnostic table.

"Small number of false positives, so +ve results indicate disease highly likely. Although much larger proportion of false -ves, a wrong decision would be difficult to defend, especially with 3% prevalence of disease in this population."

"High false negative rate."

"I am concerned that so many cases are missed."

"1 in 4 women are incorrectly tested negative...a media pr nightmare for gps."

"The prevalence of ovarian cancer in the symptomatic population is only 3% and the risk of false negative result is 10:1360 and therefore I feel it is appropriate not to investigate further".

"Risk of true +ve being +ve is 26/57 = high enough to refer."

"Some concern over relative high false positives compared with true positives."

For some respondents the 2x2 diagnostic table was effective in conveying the implications of disease prevalence on the absolute numbers of test errors which in turn led to an attenuation of the emphasis placed on one or other of false positives or false negatives as observed with predictive values or sensitivity and specificity respectively:

"1303 women had neg(ative) tests. You cannot send all of these for further Ix(investigation) - *you would swamp the system. The 10 false -ves will just have t(o) c(ome) in if* (they develop) *symptoms."*

"Too many false positives-they nearly equal the true positives. It is much better at helping you predict who does not have ovarian cancer but still too many false negatives."

"Test still misses 30% of cancers yet only 10 in 1293 women."

Several respondents used the 2x2 table to derive summary test accuracy metrics to inform their management decisions:

"Strong negative predictive value."

"*I tend to convert most info*(rmation) *to sensitivity and specificity.*"

The scenario depicting pre to post test probability received few responses. Some respondents responded that this presentation of test accuracy information was accessible and might be helpful in dialogue with patients:

"Easier to understand the terms."

“Again difficult to rule out on basis of -ve test - here I would be happier to advise patient that probability of having disease was 0.6% if test -ve and would leave if she understood and was happy with explanation.”

Some respondents did not understand the information provided by pre-post test probabilities:

“I wouldn't be able to make a decision or discuss results with a patient using this data”

“Not sure what these values actually mean.”

The annotated pictogram also received few open responses, although all comments about this graphic as a diagnostic decision making tool were positive:

“Very clear diagram.”

“Useful for showing patient.”

“Looks much better with this presentation!!”

The metrics that appeared least well understood by respondents were the DOR and LRs; 91% (21/23) and 77% (20/26) of comments respectively were concerned with respondents’ lack of familiarity with these metrics:

“Don't understand this value - where does 190 lie on scale of 1 to infinity?”

“DOR 190 - is this good or bad?”

“Would need guidelines to follow here because I have no experience of the DOR.”

“Can't understand this presentation of test accuracy at all.”

“Would have to look further into DOR to know what this meant.”

“I do not understand the LR terminology.”

“Not confident with measures.”

“Not sure what LR - or + means.”

“Not used to these measurements.”

“I have no experience of using likelihood ratios so would have to research before deciding on next course of action.”

“I do not know what these ratios mean.”

One respondent, unfamiliar with the DOR, suggested that having information about the DOR of familiar tests would help with the interpretation of the magnitude of the DOR presented for the unknown test in the scenario:

"Would like to know DOR values for other tests which have been in widespread use for some time."

A further respondent identified the limitations of the DOR in distinguishing between the 2 dimensions of test accuracy:

"I think other statistical ways would help to clarify whether the test is useful for a negative or a positive test."

Only one open respondent illustrated some understanding of the interpretation of likelihood ratios:

"The LR+ is quite high at 38 so high possibility she has the disease. The LR- is not low enough for me to feel confident not to investigate."

Probabilistic versus normalised frequency representation

- **Plain language, normalised frequency representation of sensitivity and specificity:**

"Of every 100 women with ovarian cancer, 76 would test positive (be detected by the test) but 24 would test negative (be missed.

Of every 100 women without ovarian cancer, 98 would test negative (receive a correct diagnosis) but 2 would test positive (be falsely labelled as having cancer)."

- **Plain language, normalised frequency representation of positive and negative predictive values:**

"Of every 100 women who test positive, 54 will have ovarian cancer but 46 will not.

Of every 100 women who test negative with the marker 99 will not have ovarian cancer but 1 will have ovarian cancer and be missed."

Open responses suggested that plain language, normalised frequency representations of sensitivity and specificity and PVs were successful in mitigating against reference class confusion observed with percentage representations of these metrics. Characteristics of plain language normalised frequency representations that facilitate informed application are likely to include the incorporation of information about the reference class and explicit quantification of test errors.

Respondent preference for guidelines

A small number of respondents across all scenarios stated a preference for guidelines rather than quantitative test accuracy information to inform their decision making:

"This is very confusing and would need definite guidelines to tell us what to do."

"I would need referral guidelines."

"Would depend on local /national guidelines."

"Looks like the false negative rate is low - I would await guidelines."

One respondent expressed the opinion that probabilistic information incited a feeling of gambling with patients' lives. This suggests either ignorance of the inevitability of test errors or an intolerance of any level of false negatives in this testing situation:

"It's clinical medicine...not based on any form of probability. That's Gambling with lives."

Probability revision

One respondent demonstrating an appreciation of the importance of pre-test probability for the clinical application of test accuracy information and attempted to undertake probability revision using information presented in the PVs scenarios, both percentage and normalised frequency representations:

"If +ve then 54% chance of having condition, so in total population 100-54 = 46% are false +ve. True +ve are 3%, 99% identified by test. So if +ve 3/(46+3) chance of being true +ve = 6%, so possibly worth referring."

Inter-relationship between different test accuracy metrics

There was limited evidence from open responses that respondents had an appreciation of the relationship between different test accuracy metrics:

(Open responses to normalised frequency representation of PVs)...*"Poor positive predictive value/specificity but impossible not to recommend investigating a positive result. Good negative predictive value/ sensitivity at 1%."*

Several respondents recognised the similarity of test accuracy estimates across scenarios:

"Similar to previous worked example showing higher risk = referral; lower than population risk = wouldn't refer the negative."

"Same scenario again (very clever)."

Whilst several respondents also recognised that different presentation formats were having an influence on their management decisions:

"Is this the same data being presented with different indices? Scary how presenting the same data differently induces different behaviour!"

"It's slightly scary how the way this is presented can change the way you feel about the results."

"Interesting - when asked this question earlier I would have referred -ve result patient, but realising now I can confidently say she has only a 0.6% chance of having the disease I would explain this to her."

(Open response to sensitivity and specificity normalised frequencies)...*"Too many false neg(ative)s for me to feel comfortable when presented in this way."*

(Open response to annotated pictogram)...*"Interesting change in my own responses -i would still refer for positive test, but more worried now if it is the right test in the first instance."*

Summary of open responses to hypothetical scenarios

Open responses to the hypothetical scenarios provide an insight into respondents' understanding and application of different test accuracy metrics. Responses do not appear to be biased with respect to respondents' perceived ability to define metrics and included examples of misunderstanding and inability to interpret and apply metrics. Understanding and ability to apply test accuracy information appeared to vary substantially amongst respondents.

A common theme that emerged was a feeling of obligation to pursue testing in the face of uncertainty and a desire to share uncertainty with patients. This is coherent with commentaries suggesting that fear of litigation and a quest to reduce uncertainty are features of medical culture that contribute to excessive testing behaviour (1.2).

Respondents appeared most familiar with sensitivity and specificity and could use these metrics to assist with decision making. This is internally consistent with responses to closed questions earlier in the questionnaire (5.5.4.1). The use of test errors was a prominent feature of the application of sensitivity and specificity for decision making.

In contrast PVs did not appear to be as well understood by respondents as earlier closed responses had indicated. An annotated pictogram and an annotated 2x2 diagnostic table received positive comments by the minority of respondents providing open responses, particularly with respect to their usefulness for shared decision making. There was also

evidence that respondents were able to accurately use graphics to assist with decision making and in order to derive summary metrics. Metrics resulting in the most decision making uncertainty were the DOR and LRs. The DOR is not a metric promoted for use for decision making in individual patients but open responses did not suggest that this sample of clinicians were aware of its limitations in this respect. By contrast LRs are specifically promoted as an aid to probability revision and decision making at the bedside. Open responses suggested that respondents were not familiar with guidelines for interpretation of LRs [62] and a plain language explanation of the change in disease probability pre and post-test did not help respondents with their decision making. The familiarity of sensitivity and specificity and their associated heuristics (SpPIN and SnNOUT) [(TTA27)] appear dominant as test accuracy metrics for decision making, despite recognised limitations with their application (2.4.1.3; 2.4.1.4). Although the 2x2 diagnostic table prompted some respondents to question their management decisions made on the basis of sensitivity and specificity, there was no evidence that this was underpinned by recognition of the limitations of sensitivity and specificity as a guide to the clinical utility of a test at extremes of prevalence. Empirical test accuracy literature reviewed in chapter 2 (2.4.3.2) suggested that normalised frequencies offered little or no advantage over percentage representations for probability revision as neither of these representations incorporates base rate information. Open responses in this survey suggested that plain language normalised frequency representations of test accuracy are effective in removing the reference class confusion observed with summary test accuracy measures such as sensitivity, specificity, PPV and NPV. Normalised frequency representations may have the potential to facilitate the use of test accuracy evidence, particularly if probability revision is neither prominent or a necessary pre-requisite for informed diagnostic decision making.

Open responses suggested that sensitivity and specificity representations of test accuracy resulted in respondents placing a greater emphasis on false negative test errors, whilst PVs resulted in respondents placing a greater emphasis on false positive test errors. Although the

impact of prevalence on post-test probability will have in part have contributed to this observation in the low prevalence scenarios used in this questionnaire, this is an important observation that requires further investigation.

The 2x2 diagnostic table was the closest approximation to a natural frequency representation of test accuracy presented to respondents in this survey. The differential emphasis placed on false negative and false positive test errors with sensitivity and specificity and PV representations of test accuracy appeared attenuated by the 2x2 table and this presentation also prompted comments concerning the impact of prevalence on the absolute number of test errors.

5.5.6 Tolerance of test errors (false positives and false negatives)

"When tests are applied in a screening / triaging context, asymptomatic individuals who test positive undergo further definitive testing (for example following referral to secondary care). Individuals who test negative usually do not receive any further testing unless they re-present with new symptoms. The clinical significance of false positive test errors (individuals without disease who test positive) depends on the risks associated with further investigation. The clinical significance of false negative test errors (individuals with disease who test negative) depends on the risks associated with missed, untreated disease. If a test was being used to screen asymptomatic individuals for a potentially serious condition, such as cancer, please indicate on the scale below an acceptable level of missed disease (% false negative test results) and an acceptable level of healthy individuals wrongly labeled as having disease (% false positive test results) that you would tolerate from the test before you would consider it accurate enough to be used for this purpose."

Figure 5.17 illustrates variation in the percentage of test errors indicated as acceptable in the context of screening or triaging for a serious disease across the 204 respondents. With the exception of 13 (6%) respondents, the acceptable percentage of false negative test errors was less than false positive test errors which might be expected for this scenario. A minority of individuals (4 for false negative test errors and 15 for false positive test errors) indicated acceptable error rates up to 40%, suggesting this question may have been misunderstood by these respondents.

Fig 5.17: Acceptable % of test errors indicated by respondents when triage testing / screening for a serious disease

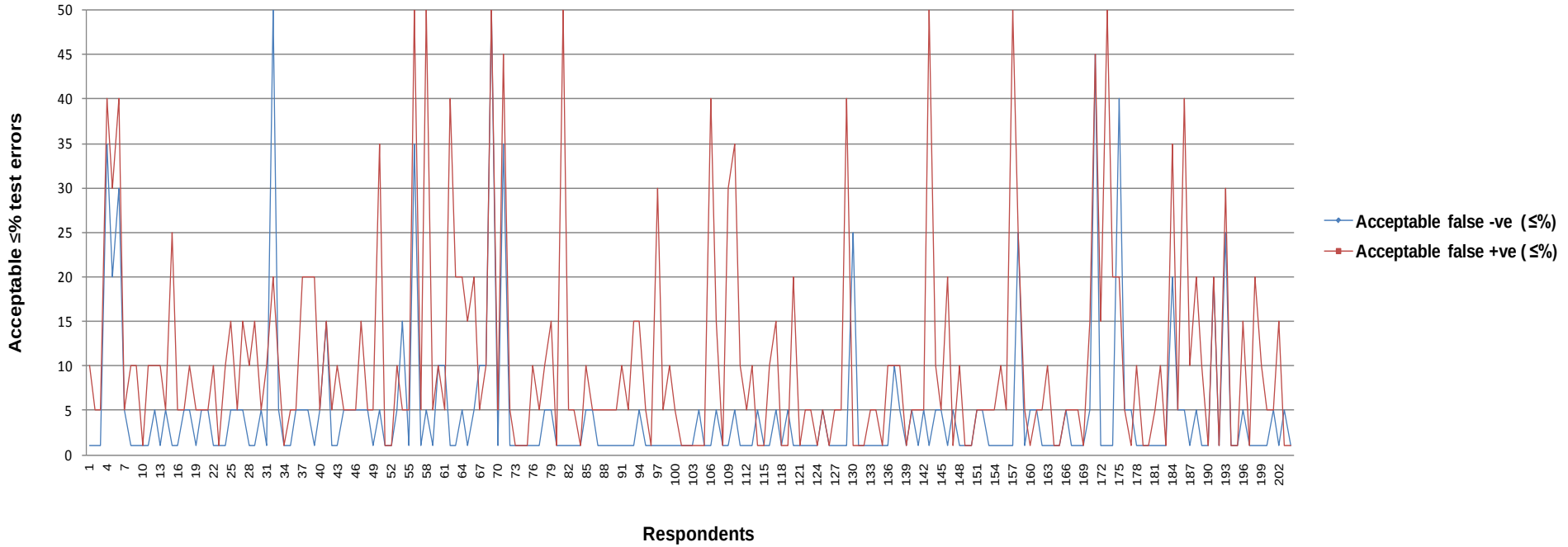
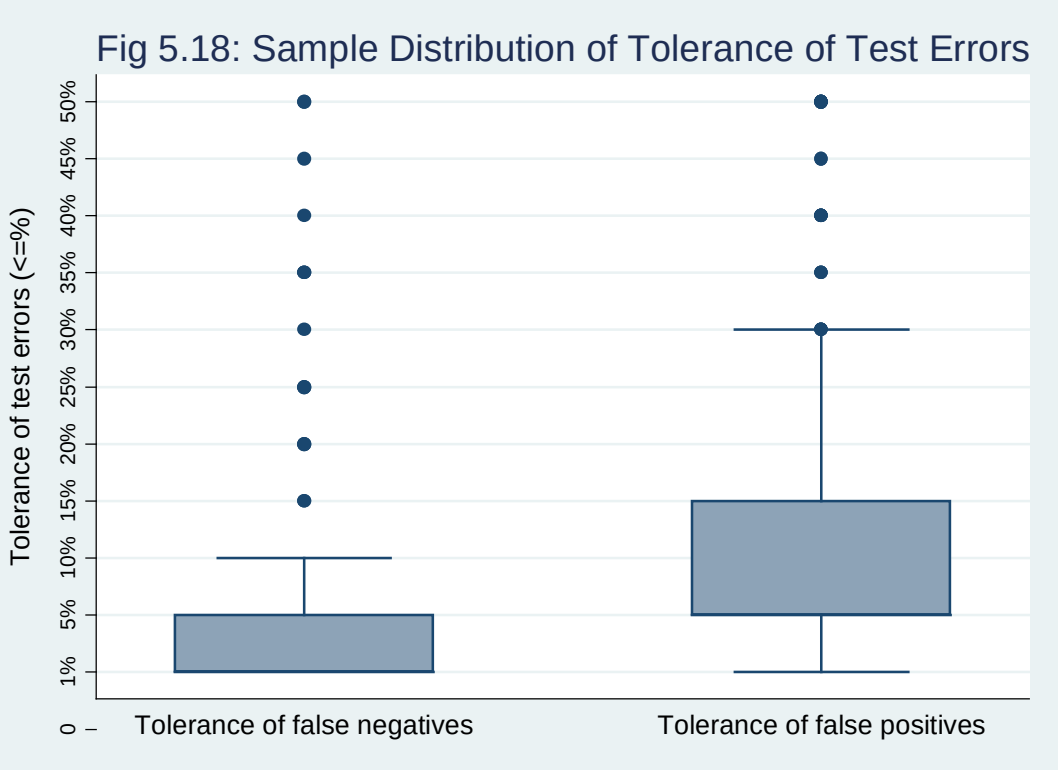


Figure 5.18 illustrates the sample distribution of tolerance of test errors for triage/ screening for 'serious disease'. The median tolerance of false negatives was ≤ 1%, with a relatively narrow inter-quartile range (4%) compared to a median tolerance of false positives of 5%, inter-quartile range 10%.

Fig 5.18: Sample distribution of test errors



Open responses to tolerance of test errors when screening / triage testing for a serious disease such as cancer

Open responses concerning tolerance of test errors were provided by 23 of 204 respondents (11%). Respondents who offered a definitive opinion of the relative importance of false positive and false negative test errors mostly supported the greater importance of false negative test errors in the testing context presented to respondents (screening / triage for a serious disease):

"false negatives more significant than false positives in this context. (I) would have checked a <0.1% box if there was one!"

"Missing diseased individuals potentially much more serious issue than over-investigation of healthy individuals."

"Do not want to be too falsely reassured by negative tests in asymptomatic population but false positive rates can be higher as presumably further investigations can be carried out if clinically indicated or appropriate."

"In an ideal world - the lower the numbers the better but false +ve safer than false -ve."

The importance of not missing disease was also suggested to be relatively more important to patients. Respondents indicated that it was easier to deal with the possibility of false positives compared to false negatives when communicating test results to patients:

"It is not good to miss cases - patients wish (for) a very sensitive test."

"A false reassurance can be dangerous, if you KNOW that many (even most) +ves are false positives, then that is safer, and not difficult to explain to patients."

There was evidence that respondents recognised the inevitability of test errors and that evaluation of the performance of a new test needed to be undertaken mindful of the accuracy of existing practice:

"Admittedly tests are not 100% accurate."

(Respondent indicating ≤1% as an acceptable number of false negative test errors)...*"Or < 5 % false neg - the test may still be better than no test for some people."*

"Depends on seriousness and usual presentation of the disease in question: for ovarian cancer which has few symptoms, accuracy in testing is paramount."

"...but in real life I accept for example the FOB (faecal occult blood) *testing for bowel ca(ncer) and welcome its input."*

"This also depends what other clinical tests / findings may be used as often it is a mixture of tests / findings that is used."

"Use in conjunction with clinical hx and exam makes clinical situation more specific/sens."

However, despite the fact that the majority of respondents indicated the relative greater importance of false negatives in the testing context presented, there was an appreciation of the potential consequences of false positive test results that should be considered:

"The acceptance of false positives depends on the available follow up, eg the more invasive the follow up the lower the acceptable false positive rate."

“We want a test that minimises anxiety (and extra work) as much as possible.”

“Acceptable level of false positives would depend on morbidity associated with further testing.”

In summary the open responses supported the lower tolerance for false negatives as indicated by a lower acceptable proportion on a continuous scale from 0-50% and a smaller degree of variation in the acceptable number of false negatives across the sample. There is evidence that respondents recognise the inevitability of test errors and that their consequences are context dependent.

5.5.7 Relationship between tolerance of test errors and management decisions

Figures 5.19 and 5.20 illustrate the results of an investigation of the relationship between respondents reported tolerance of test errors (5.5.6) and the management decision indicated by respondents (to refer or not to refer) for the sensitivity and specificity scenarios (percentage and normalised frequency presentation), the annotated pictograph scenario and the annotated 2x2 diagnostic table scenario. These four of the nine test accuracy metrics used across scenarios were selected for this investigation on the basis that they contain information on test accuracy with disease as reference class. Test accuracy metrics that do not have presence or absence of disease as reference class, (LRs, PVs, pre to post-test probability and the DOR) were not included in this investigation as they are not directly comparable with test error rates. Respondents choosing a 'don't know' management response to sensitivity and specificity, 2x2 table and pictograph scenarios were not considered in this investigation.

Tolerance of false negatives

For the purposes of this analysis it is assumed that those respondents who understood probabilistic and normalised frequency representations of sensitivity and specificity, the 2x2 table and an annotated pictograph and indicating a tolerance of false negative test errors ≤25% (5.5.6) would be expected to refer in the presence of a negative test result for the test in the scenarios with a sensitivity of 76% (false negative rate of 24%). Similarly, it is assumed that those respondents indicating a tolerance of false negative test errors of >25% would be expected not to refer in the presence of a negative test result for the test in the scenario with a sensitivity of 76% (false negative rate of 24%).

False positive test errors

For false positive test errors it is assumed that respondents indicating a tolerance of false positive test errors ≤5% (5.5.6) would be expected not to refer in the presence of a positive test result for the test in the scenarios with a specificity of 98% (false positive rate of 2%) whilst respondents indicating a tolerance of false positive test errors of >5% would be expected to refer in the presence of a positive test result for the test in the scenarios with a specificity of 98% (false positive rate of 2%). This assumption ignores motivational biases that may result in a feeling of pressure to refer in the presence of a positive test result as indicated in open responses (5.5.5.3).

Within person investigation of relationship between tolerance of test errors and management decisions for individual test accuracy metrics

Figures 5.19 and 5.20 illustrate that overall, for between 36% and 62% of respondents there was no concordance between stated tolerance of test errors and management decisions for individual test accuracy metrics. Figure 5.19 illustrates that following a negative test result, sensitivity and specificity appear to result in a slightly increased concordance between respondents' indicated tolerance of test errors and management decisions compared to the 2x2 table and annotated pictograph.

Figure 5.20 illustrates that following a positive test result the distinction between the summary measures sensitivity and specificity and the 2x2 table and annotated pictograph are no longer evident.

The slight increase in concordance for sensitivity and specificity for negative test results is likely to be a result of a combination of factors including familiarity with and use of these metrics in practice (5.5.4) and the fact that false positive and false negative test errors and sensitivity and specificity both encourage artificial partition of the two dimensions of test accuracy from each other and from the effects of disease prevalence in the tested population (2.4.1.3).

Fig 5.19: Within-person agreement between tolerance of false negatives and management decision following a negative test result (individual metrics)

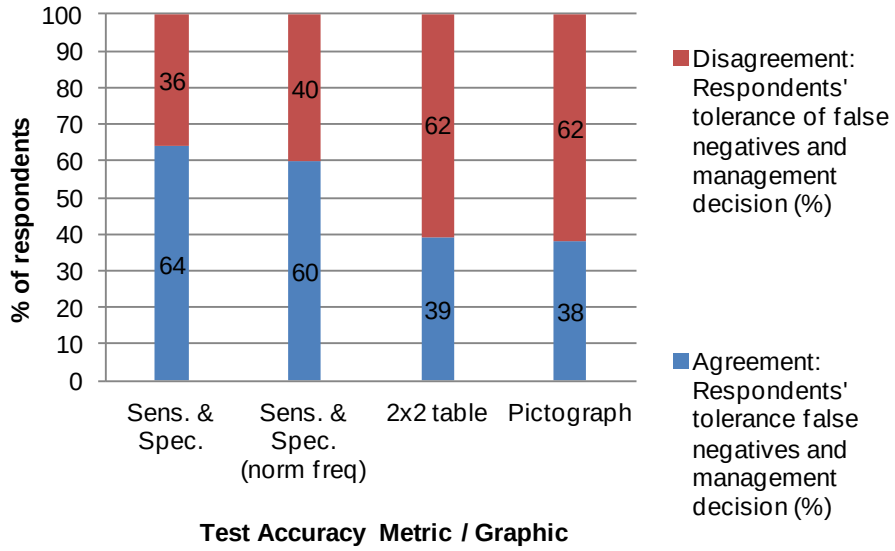
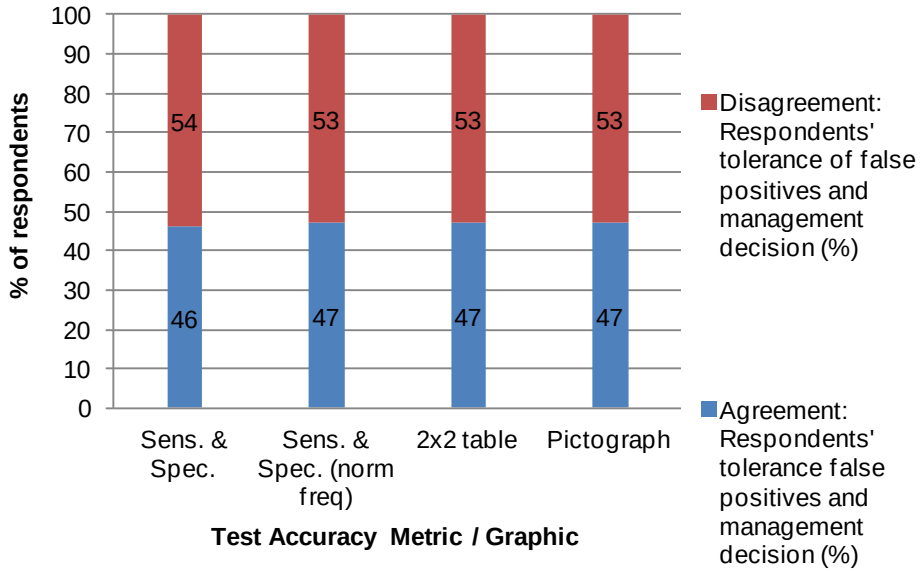


Fig 5.20: Within-person agreement between tolerance of false positives and management decision following a positive test result (individual metrics)



5.5.8 Relationship between reported understanding and application of test accuracy metrics in scenarios

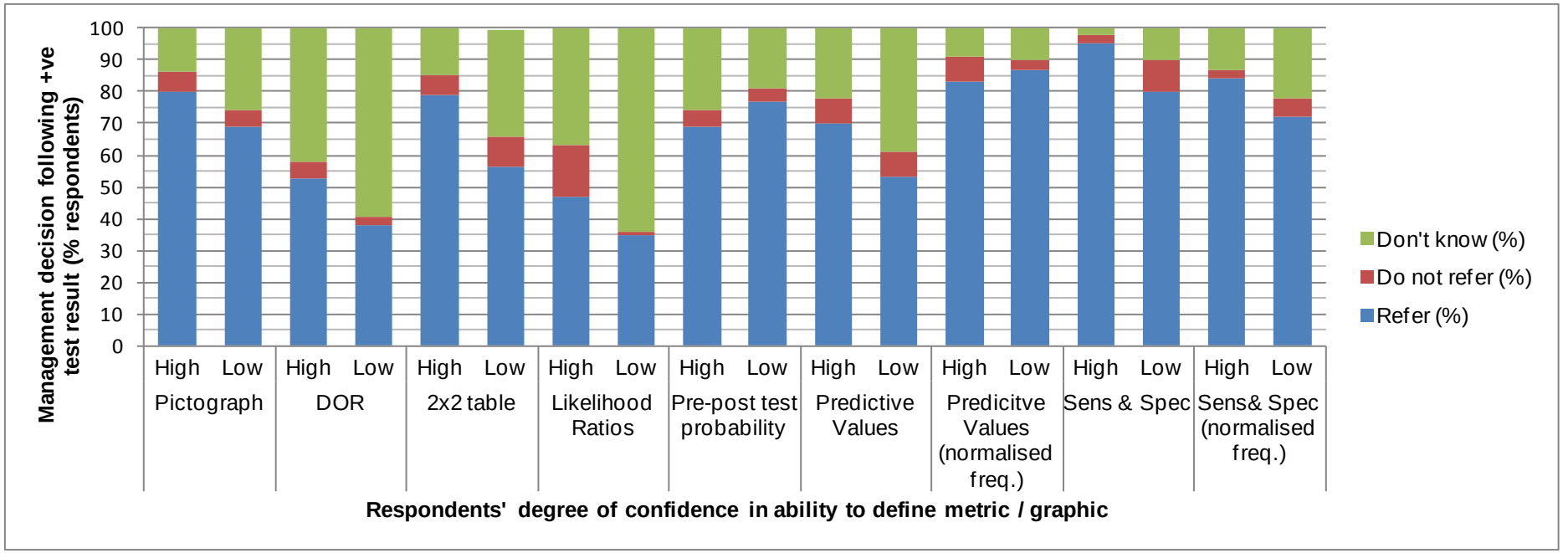
Figure 5.21 illustrates the relationship between respondents familiarity with test accuracy metrics: high confidence, (very confident could define or moderately confident could define) and low confidence: (not very confident could define or could not define or not heard of) and the indicated management decision following a *positive* test result (to refer, not to refer or don't know whether to refer). Figure 5.22 illustrates the relationship between respondents stated ability to define metrics and the indicated management decision following a *negative* test result (to refer, not to refer or don't know whether to refer). For the small number of respondents where there was a discrepancy in familiarity between different dimensions of test accuracy (sensitivity versus specificity; PPV versus NPV; LR+ versus LR-) (see 5.5.4.2) the less familiar response was used for that metric.

5.5.8.1 Management decision following a positive test result

Figure 5.21 illustrates that for the majority of test accuracy metrics and graphics, the percentage of respondents indicating management uncertainty following a positive test result is higher for respondents with a low degree of confidence in their ability to define or explain that metric or graphic compared to respondents with a high degree of confidence. The increase in management uncertainty in low confidence respondents was largely at the expense of decisions to refer following a positive test result: fewer low confidence respondents indicated they would refer following a positive test result compared to high confidence respondents. The percentage of respondents deciding not to refer following a positive test result was relatively small and affected little by metric or degree of confidence. Exceptions to this pattern were observed for a normalised frequency representation of PVs where the impact of respondent confidence on management decisions following a positive test result is less marked than for other metrics. Presentation of pre to post-test probabilities resulted in an increase rather than a decrease in referrals in low confidence respondents.

However this may be explained by the fact that assessment of respondent confidence was based on stated confidence defining LR+ and LR- rather than pre to post-test probability information.

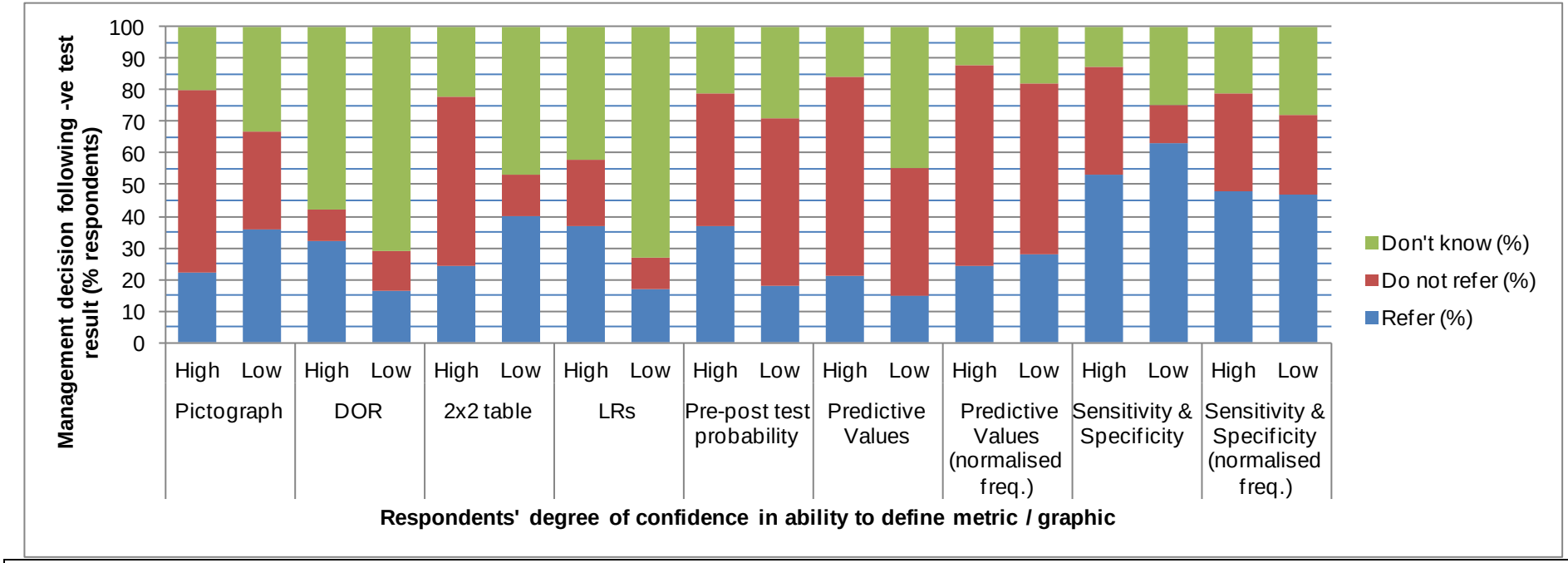
Fig. 5.21 Respondents' confidence in ability to define test accuracy metric and management decision following a positive test result



MANAGEMENT DECISION FOLLOWING POSITIVE TEST RESULT																		
Management decision following a positive test result	Pictograph		DOR		2x2 table		Likelihood Ratios		Pre-post test probability		Predictive Values (%)		Predictive Values (NF)		Sensitivity & Specificity (%)		Sensitivity & Specificity (NF)	
	Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)	
	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
Don't know	14	26	42	59	15	34	37	64	26	19	22	39	9	10	2	10	13	22
Do not Refer	6	5	5	3	6	9	16	1	5	4	8	8	8	3	3	10	3	6
Refer	80	69	53	38	79	57	47	35	69	77	70	53	83	87	95	80	84	72
TOTAL (N)	111	93	19	185	151	53	19*	185*	19*	185*	106†	98†	106†	98†	172‡	32‡	172‡	32‡

Notes to fig 5.21: *Degree of confidence taken from respondents' ability to define 'positive likelihood ratio' and 'negative likelihood ratio'; † Degree of confidence taken from respondents' ability to define 'positive predictive value' and 'negative predictive value'; ‡Degree of confidence taken from respondents' ability to define 'sensitivity' and 'specificity'; NR: normalised frequencies

Fig. 5.22: Respondents' confidence in ability to define test accuracy metric and management decision following a negative test result



MANAGEMENT DECISION FOLLOWING NEGATIVE TEST RESULT																		
Management decision following a negative test result	Pictograph		DOR		2x2 table		Likelihood Ratios		Pre-post test probability		Predictive Values (%)		Predictive Values (NR)		Sensitivity & Specificity (%)		Sensitivity & Specificity (NR)	
	Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)		Confidence (%)	
	High	Low	High	High	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
Don't know	20	33	58	71	22	47	42	73	21	29	16	45	12	18	13	25	21	28
Do not Refer	58	31	10	13	54	13	21	10	42	53	63	40	64	54	34	12	31	25
Refer	22	36	32	16	24	40	37	17	37	18	21	15	24	28	53	63	48	47
TOTAL (N)	111	93	19	185	151	53	19*	185*	19*	185*	106†	98†	106†	98†	172‡	32‡	172‡	32‡

Notes to figure 5.22: *Degree of confidence taken from respondents' ability to define 'positive likelihood ratio' and 'negative likelihood ratio'; † Degree of confidence taken from respondents' ability to define 'positive predictive value' and 'negative predictive value'; ‡Degree of confidence taken from respondents' ability to define 'sensitivity' and 'specificity'; NR: normalised frequencies

5.5.8.2 Management decision following a negative test result

Figure 5.22 illustrates that management uncertainty following a negative test result is also higher for low confidence compared to high confidence respondents. However there was no consistent relationship observed between lower confidence and management decisions following a negative test result in contrast to the relatively consistent trade off of a decision to refer with management uncertainty observed following a positive test result (figure 5.21 above),

In summary, decreasing confidence increased management uncertainty following both positive and negative test results. Following a positive test result the increase in management uncertainty was largely at the expense of decision to refer, whilst for negative test results an increase in management uncertainty did not result in any consistent trade off. This observation may reflect the specific testing context depicted, where the consequences of false negative test results were viewed as relatively more important than those associated with false positive test results (see 5.5.6 above). Alternatively this observation may reflect a generic difficulty applying negative test results as suggested by the literature reviews (2.5.3.3): an insensitivity to negative test results, reflecting an emphasis on the presence rather than the absence of symptoms and disease in clinical practice.

5.6 Strengths and Limitations: Survey of understanding and application of test accuracy measures in primary care

A major strength of this survey, particularly in comparison with previous empirical research in this area, is the size and representativeness of the sample. It is unlikely that selection bias is entirely absent and the survey would need replicating to check the external consistency of findings. However respondents do not appear to be self-selected on the basis of expertise or experience in the topic area. In addition, although opportunities for comparison are limited, the survey findings do not contradict earlier empirical research.

In the interests of achieving a reasonable and representative response rate the questionnaire was limited in terms of what it could cover. The findings can therefore not be generalised to different healthcare settings where the importance attached to test errors, the magnitude of test errors, pre-test probabilities and experience with any particular test will vary. The purposive sampling of GPs in the UK may also preclude generalisation to other health care settings where testing culture may differ. Findings based on hypothetical scenarios may not be generalisable to the actual clinical encounter.

The survey findings demonstrate internal consistency and where possible within-person response patterns have been compared to assist with identification of real from artefactual associations caused by variation within the sample. However these analyses were not comprehensively conducted across all outcomes and were based on aggregated data (median confidence across metrics /sub-categories of metrics). It is therefore possible that additional sources of variation were missed.

Respondent fatigue was evident in the open responses to scenarios and in the absence of randomisation of question order it is possible that the compulsory closed responses were also affected, resulting in less considered and repetitive answers.

5.7 Discussion: Survey of understanding and application of test accuracy measures in primary care

The survey of understanding and application of test accuracy measures was primarily exploratory in nature with the aims of addressing gaps in the existing empirical literature and refining emerging hypotheses about features of test accuracy measures that facilitate their understanding and application. In particular the survey aimed to represent a sample of practicing clinicians from a generalist setting; to describe whether and how test accuracy information is used by clinicians in practice; to assess the perceived utility of existing test accuracy metrics and to assess whether there is consistency in the application of different test accuracy metrics across a common hypothetical scenario.

5.7.1 Representation of practising clinicians in a generalist setting

Survey respondents appear more likely to be female although the effect of confounding by part-time practice could not be assessed. In addition older practitioners may be underrepresented by the survey. The extent of any age bias may be postulated to overestimate the ability of UK GPs to understand and apply test accuracy measures as the evidence based medicine movement is only two decades old.

Importantly only 11 of 222 GPs accessing the survey questionnaire electronically failed to participate once the research topic area was revealed or withdrew part way through the questionnaire. The survey sample appears to be representative of GPs working in the UK and not self selected on the basis of previous training or interest in the topic area. Similarly the small proportion of respondents offering open responses in the survey does not appear to be self selected on the basis of confidence in the topic area.

5.7.2 Sources of test accuracy information used by clinicians

Personal clinical experience, colleagues, guidelines, web based resources and local laboratories were prominent sources of test accuracy information and the timeliness and currency of these information sources were cited as facilitating characteristics.

The popularity of guidelines may stem from the fact that they translate quantitative test information into recommendations and as such their use may not require an understanding of probabilistic information or the ability to undertake probability revision. The view that guidelines represent an accurate and trusted source of test accuracy information was a theme in open responses for those respondents encountering difficulty applying quantitative test accuracy information. Indeed it may be unrealistic and inappropriate to expect clinicians to formally work with probabilities in the clinical encounter when deciding to use a test or when applying test results. In addition a theme in open responses to scenarios is that a test is rarely applied in isolation. Guidelines may therefore represent an important tool to encourage evidence based testing and greater consistency of test use. However informed assessment of guideline credibility and applicability is underpinned by an understanding of how recommendations have been reached. Guidelines therefore cannot be considered a replacement for an understanding of test properties.

The web based resources cited in open responses were mostly condition rather than symptom based guidelines or were evidence based summaries concerned with clinical effectiveness without quantitative test accuracy information. The prominence of local laboratories as a source of test accuracy information in this sample appeared, at least in part, underpinned by a misunderstanding that the normal range communicates the accuracy of a test. However the timeliness and ease of access of these information resources make them a potentially powerful education and dissemination medium for test accuracy information.

Although personal clinical experience and colleagues were the sources of test accuracy information reported to be most frequently used by respondents, their value needs to be

judged in the light of an assessment of the ability of clinicians to interpret and apply test accuracy information (see 5.7.4 below).

5.7.3 Perceived utility of existing test accuracy metrics

The utility of existing test accuracy metrics (sensitivity and specificity, PVs, LRs, the 2x2 diagnostic table, the ROC curve, DOR and AUC) and an annotated pictogram more commonly used in the communication of risk were assessed on the basis of familiarity, perceived ability to define and self reported use in practice. Clinicians may have a preferred presentation format based on familiarity rather than understanding (2.6.1; 5.4.2.1) and triangulation allowed investigation of any association between these three utility measures. In addition the relationship between respondents' perceived understanding of metrics and their actual application was examined (see 5.7.4.1).

Sensitivity and specificity, the annotated 2x2 diagnostic table and PVs were reported to be familiar metrics by the most respondents. It is likely that this pattern of familiarity reflects exposure to these metrics, including their use in sources of test accuracy sources used by respondents. Certainly the familiarity of respondents with the pictograph is associated with its use in the risk communication literature. Importantly LRs and test accuracy metrics more commonly used in systematic reviews of test accuracy (DOR, AUC and the ROC Curve) were familiar to 32% of respondents or less.

Perceived ability to define metrics and use of metrics in practice followed a similar pattern to familiarity. The exception was that despite a greater reported confidence in defining the 2x2 diagnostic table, PVs were reported to be used more in practice. This finding is likely to be influenced by greater exposure to PVs in the literature. No research documenting the use of test accuracy metrics in primary evaluations of test accuracy has been identified in support of this hypothesis although sensitivity and specificity were the most common metrics used in

test accuracy reviews as documented by three recent methodological reviews [6,7] see (4.7.3.4).

5.7.4 Application of nine different test accuracy presentation formats to a common testing scenario

Application of test accuracy metrics across 9 hypothetical testing scenarios resulted in marked variation in responses to both positive and negative test results. Greater inconsistency and management uncertainty was observed following presentation of a negative test result in comparison to a positive test result. The greater inconsistency for negative test results was corroborated by a within-person analysis of responses, arguing against the observation being an artefact of random variation at sample level.
Sensitivity and specificity (percentage and normalised frequency representations) resulted in the most consistency in management decisions across the sample for both positive and negative test results. Respondents had particular difficulty applying LRs and the DOR.

5.7.4.1 The impact of respondents' understanding of test accuracy metrics and graphics on management decisions

An assessment of the degree to which the observed pattern of responses to scenarios can be explained by respondents' understanding and informed application of test accuracy metrics as distinct from other contextual factors and motivational biases was informed by investigating the within-person relationship between self reported confidence in ability to define metrics and management decisions (5.5.8) and the within-person relationship between self reported tolerance of test errors and management decisions (5.5.7).

Relationship between reported ability to define metrics and management decisions

Lower median confidence in reported ability to define metrics was associated with an increase in uncertainty as measured by the proportion of 'Don't know' management decisions

across scenarios for both positive and negative test results. This suggests informed application of test accuracy metrics was impacting on management decisions.

Relationship between reported tolerance of test errors and management decisions

Within-person association between tolerance of test errors and management decisions suggested no consistent relationship. However greater consistency in agreement between tolerance of false positive test errors and management decisions following positive test results was observed with combination graphics (2x2 table and pictograph) and greater consistency in agreement between tolerance of false negative test errors and management decisions following negative test results was observed with representations of sensitivity and specificity. This is congruent with effects observed in the open responses to scenarios (5.5.5.3) suggesting that different test accuracy presentations may have the effect of emphasising one or other of false positives and false negatives.

5.7.4.2 Open responses as an insight to understanding

Open responses to the hypothetical testing scenarios also offer an insight into the degree to which management decisions are informed by understanding and ability to apply test accuracy metrics. Open responses suggested that sensitivity and specificity appeared to be understood by most respondents and this is congruent with closed responses to scenarios where use of these metrics resulted in the most consistency in management decisions. PVs resulted in relatively more examples of misunderstanding by respondents and in particular normalised frequency representations of these metrics resulted in some respondents confusing PVs with sensitivity and specificity. One explanation for this observation may be reference class confusion (2.4.3.2). Alternatively this observation may be a consequence of the relatively greater familiarity of respondents with sensitivity and specificity and the use of these as default metrics where there is uncertainty.

An annotated 2x2 diagnostic table and an annotated pictogram appeared to be accessible metrics for decision making and these graphics also attracted positive comments concerning their potential as aids to shared decision making with patients. As an approximation of a natural frequency presentation of test accuracy, the 2x2 diagnostic table also appeared to effectively communicate the relationship between pre test probability and test errors. Indeed test errors were prominent as part of the translational pathway from quantitative estimates of test accuracy conveyed by summary metrics and graphics to management decisions.

There was no indication from open responses that normalised frequency representations of test accuracy offered any advantage over percentage representations, which is congruent with closed responses (see above). This concurs with the published literature (2.4.3.2) which suggests that the advantage of frequentist over percentage representations of uncertainty for probability revision depends crucially on the fact that natural frequencies do not require decision makers to consider base rates (prevalence or pre-test probability), in contrast to normalised frequency representations.

The DOR, and LRs consistently attracted comments that conveyed lack of familiarity with these metrics. Whilst the DOR is not a metric aligned to decision making in the clinical encounter, the LR is a metric that has been promoted for use in this context since the mid 1990s (2.4.1).

5.7.4.3 The effects of context and motivational biases on the application of test accuracy metrics

Context

The context in which respondents were required to apply test accuracy information was fixed across scenarios in order to eliminate other factors that might modify decision making aside from the test accuracy metric presented. The specificity of the test presented in the scenarios was high and false positives regarded as the least important test error in the testing context presented (5.5.6) therefore a decision to refer in the presence of a positive test might be

considered a straightforward one and easy to communicate with patients. In contrast the sensitivity of the test presented in the scenarios was relatively low and false negatives were regarded as the more important test error in the testing context presented (5.5.6). The relatively greater importance attached to false negatives is likely to increase both the complexity of any dialogue with patients and decisions about referral which in turn may result in greater variation in referral decisions following a negative as opposed to a positive test result.

The importance of prevalence and both dimensions of test accuracy when judging the clinical utility of a test

Consideration of one dimension of test accuracy alone as a measure of a test's clinical utility may mislead and the limitations of reliance on a test's sensitivity or specificity alone in this respect are recognised (2.4.1.4). The relatively greater consistency in decision making observed with sensitivity and specificity in this survey (to refer even in the event of a negative test result) is likely in part to have been explained by consideration of sensitivity alone as a measure of a test's ability to rule out disease (SnNOUT) [(TTA27)] in this particular testing context where false negative test errors were perceived as relatively more important. However an additional observation of this survey is the fact that the 2x2 diagnostic table and an annotated pictograph attenuated this pattern, with the majority of respondents deciding not to pursue further investigations following a negative test result. The simultaneous presentation of both dimensions of test accuracy, conveying the trade off between the two dimensions of test accuracy and incorporating information about prevalence are one of the main tenets underlying the utility of LRs. This survey demonstrates the importance of the complementary nature of this information for diagnostic decision making but suggests that presentation of these relationships using the 2x2 diagnostic table or an annotated graphic may be more effective at communicating these relationships than summary metrics.

Motivational biases

The effect of context on the relative weight given to clinician utility relative to patient utility will shape how motivational biases affect decision making. Motivational biases may occur if respondents are able to apply the information presented in the hypothetical scenarios presented in this survey:

- Management decisions that are discrepant with test results (a decision not to refer in the presence of a positive test result or a decision to refer in the presence of a negative test result) may occur if respondents are able to incorporate information about test errors in their decision making.
- The relative importance of false negative test errors is likely to result in greater weight given to management decisions associated with negative test results if respondents are able to incorporate information about test errors in their decision making.
- Omission bias (1.5.2.2) is likely to operate preferentially for positive test results if respondents are able to incorporate information about the relative magnitude of test errors in their decision making: acting (an act of commission) on false positives will result in further investigation and anxiety for those tested, whereas acting on false negatives will result in increased detection of those who might benefit from treatment.

The nature of the relationship between an understanding of test accuracy metrics, contextual factors and individual motivational biases on decision making is inevitably complex and the summation of effects of these contextual influences on the balance between patient and clinician utility are difficult to quantify.

5.7.5 Discussion summary

In summary there is evidence from this survey of GPs that presentation of test accuracy using different metrics has a profound effect on diagnostic decision making. The ability to understand and apply test accuracy metrics, contextual factors and variation in individual motivational biases are all likely to be contributing factors to the observed variability.

Sensitivity and specificity are understood by a significant proportion of respondents and result in application of quantitative estimates of test accuracy. However it is unclear to what extent any advantage of sensitivity and specificity over other test accuracy metrics is due to familiarity and reporting practice as opposed to their intuitive nature. The 2x2 diagnostic table and annotated pictograms appear to have promise as alternative representations of test accuracy and importantly may encourage consideration of the relationship between the two dimensions of accuracy and the effect of prevalence on these. PVs do not appear to be well understood and in this survey respondents did not appear aware of the potential of PVs to communicate the impact of pre test probability on test errors. Further research is needed to elucidate the degree to which the variation in decision making with different test accuracy presentation formats observed in this questionnaire are context-specific or are a function of characteristics of presentation format.

LRs and metrics more common to systematic review of test accuracy were not familiar to most respondents in this survey. Metrics such as the DOR, the ROC curve and the AUC are not intended for use for diagnostic decision making at the bedside and therefore it might be expected that the majority of respondents were not able to apply them in the hypothetical testing scenarios presented in this survey. However the increase in the number of systematic reviews of diagnostic test accuracy (see chapter 3) will require practitioners to be familiar with the properties of these global metrics as a resource for evidence based diagnosis.

The lack of familiarity of respondents with LRs calls into question their usefulness as aids to diagnostic decision making and the effectiveness of educational interventions implemented as part of their promotion over the preceding decade.

Chapter 6: Discussion: Systematic reviews and meta-analyses of test accuracy: developing methods that meet practitioners' needs

6.1 Main findings: summary and implications for practice

This thesis has contributed substantially towards redressing the imbalance in research concerned with systematic reviews of test accuracy. Test accuracy research to date has been dominated by methodological and statistical considerations at the expense of consideration of contextualisation of review findings and investigation of the extent to which the evidence base is accessible to decision makers. It has been observed that important barriers to the use of evidence include accessibility: timely access and possession of skills necessary for interpretation and application of information, as well as the acceptability and applicability of research evidence: the provision of information relevant to the decision context [30]. Contextual information is important both because decision makers need to recognise the relevance of evidence to their needs and because the utility attached to outcomes consequent on decisions are contextually dependent.

In recognition of recent attempts to shape the content of test evaluations both in terms of developing more rigorous testing policy [95] and encouraging a comparative approach to evaluation [96,123] this thesis is timely in reviewing the characteristics of the existing test accuracy systematic review evidence base and evaluating the features of test accuracy metrics that facilitate understanding, in order to inform such initiatives.

6.1.1 Evaluation of the familiarity, use, understanding and application of test accuracy information for decision making (chapters 2 and 5)

Whilst there is widespread belief that clinicians have difficulty applying test accuracy information, this has not been based on a systematic interrogation of the evidence base. As a result it has not been possible to date to quantify or characterise the extent of the problem in order to improve understanding and application. The literature reviews presented in

chapter 2 of this thesis have captured a breadth of perspectives pertinent to the understanding and application of test accuracy evidence and the findings provide a framework for further research.

The survey of general practitioners presented in chapter 5 was primarily exploratory in nature with the aims of addressing gaps in the existing empirical literature and refining emerging hypotheses about features of test accuracy measures that facilitate their understanding and application. The primary care setting was chosen as a more diverse testing environment in comparison to secondary care and in order to address the under-representation of generalist settings in the test accuracy evidence base. The survey is an important addition to the existing evidence base in terms of its representativeness, particularly, in contrast to existing empirical literature the survey sample does not appear self selected on the basis of interest, expertise or experience in the topic area.

6.1.1.1 Familiarity with test accuracy metrics

The observation that familiarity and preference for a metric or presentation format is not a reliable measure of comprehension is suggested to be a function of decision makers having preferred presentation formats that are based on familiarity, including example exposure to metrics in the published literature, rather than understanding. In addition familiarity is likely to encourage the use of heuristics which in turn introduce the potential for error, including cognitive biases. It might be expected that perceptual differences observed between ratio and absolute risk metrics observed in the literature concerned with communication of intervention risks (a relative magnification of perception of effect with ratio measures) will present themselves in the test accuracy literature as recent initiatives (see 6.2.1 below), methodological advances and a rise in the number of systematic reviews of test accuracy will present more opportunities for a comparative approach to test accuracy evaluation and use of comparative test accuracy metrics.

In common with the empirical literature reviewed, sensitivity and specificity, and PVs were reported to be familiar metrics by the most survey respondents (Chp 5) which mirrors their use in the published literature (4.7.3.4). Interestingly LRs were not familiar to survey respondents which may be a reflection of the fact that these metrics have not been widely adopted by the test accuracy research community or that they are difficult to comprehend and apply for diagnostic decision making. Metrics more commonly used in systematic reviews of test accuracy (DOR, AUC and the ROC Curve) were familiar to 32% of respondents or less which has implications for the accessibility of test accuracy reviews at the current time.

6.1.1.2 Self reported use of test accuracy evidence for decision making

Although the validity of self reported use of metrics is reliant on respondents' understanding of metrics, one of the questions raised by the literature reviews in chapter 2 was the extent to which published estimates of test accuracy and pre-test probability are used for diagnostic decision making in practice. Although PVs were reported to be used by 80% of clinicians in one study, sensitivity and specificity were reported to be used by <4% clinicians and ROC curves and LRs by <1%. In addition there was evidence that decision makers were drawing on personal clinical experience rather than published estimates of test accuracy and pre-test probability.

Similarly, in the general practice survey, personal clinical experience, colleagues, guidelines, web based resources and local laboratories were prominent sources of test accuracy information in the survey sample. The popularity of guidelines may originate from the fact that they do not require application of quantitative test accuracy information. However whilst the practical application of diagnostic research findings may be best served by clinical guidelines, this is not a replacement for making the explicit quantification of diagnostic probabilities a more frequent occurrence in clinical practice. The credibility of guidelines is underpinned in part by an understanding of how recommendations have been reached.

Reliance on guidelines to promote evidence based testing should be underpinned by an appreciation of generic attributes of tests: the inevitability of test errors; the fact that tests merely adjust the probability of disease rather than eliminating uncertainty (the concepts of pre and post-test probability) and the fact that accuracy is not a fixed property of tests. Reliance on guidelines to promote evidence based testing is also dependent on the ability of guideline developers themselves to understand and apply test accuracy information; the aptitude of the policy making community in this area is unknown at the present time as the policy maker perspective was absent from the empirical and non empirical test accuracy literature and was outside the remit of the survey. In addition, if guidelines are to underpin the promotion of evidence based testing, challenges will include capturing the complexity of the diagnostic work up which is not restricted to determining the probability of any single disease and is symptom rather than disease based investigation.

6.1.1.3 Understanding and application of test accuracy evidence

Investigation of the characteristics of test accuracy information that facilitate diagnostic decision making has been shaped by an implicit assumption that formal probability revision is a necessary pre-requisite. This is despite evidence that published estimates of pre-test probability and test accuracy are unlikely to be used for formal probability revision in practice. Self reported use of Bayes' theorem for diagnostic decision making was less than 3% in one study [(ETA29)]; estimation of pre-test probability and accuracy of named tests is observed to be inaccurate and highly variable and provision of quantitative information about test accuracy does not appear to improve probability revision.

The discipline of psychology is dominant in this area of research which consistently demonstrates the greater accessibility of frequency formats over probabilistic representations for probability revision.

A consequence of the emphasis on probability revision and presentation format (frequency versus probability) is that investigation of the utility of summary metrics more common to medical research for diagnostic decision making has been limited. For example the utility of PVs (which are a direct expression of post-test probability) and the 2x2 diagnostic table (a natural frequency expression of test accuracy), has not been investigation. Only two studies investigating the utility of summary metrics in medical settings identified in the literature reviews and provide no evidence to distinguish sensitivity and specificity and likelihood ratios for improving the accuracy of probability revision.

The pertinence of the existing evidence base therefore depends on the extent to which formal probability revision is a *necessary* pre-requisite for informed diagnostic decision making; an assumption which is challenged by the findings of the general practice survey presented in chapter 5.

Open responses to the general practice survey suggest that summary test accuracy metrics common to the medical literature can result in informed diagnostic decision making without recourse to formal probability revision. It is however unclear whether this is a consequence of familiarity rather than inherent characteristics of individual metrics that facilitate understanding. Importantly test accuracy metric had a profound impact on diagnostic decision making which has implications if the use of quantitative test accuracy estimates is to be encouraged in practice. Open responses suggested several factors were likely to be contributing to the observed variation:

-differences in understanding and ability to apply metrics

-methods for combining the 2 dimensions of test accuracy (encouraging comparison of the magnitude of test errors (false positives and false negatives); the degree to which test errors are made explicit and the impact of prevalence on the absolute numbers of test errors

-contextual factors (the downstream consequences of test results)

This observation is consistent with those made in the non-empirical literature presented in chapter 2, highlighting the need to distinguish between the two dimensions of test accuracy

due to the contextually dependent importance of test errors whilst also recognising that considering any one dimension in isolation has the potential to mislead with respect to the utility of a test. These findings suggest that explicit quantification of test errors (not currently a feature of existing summary test accuracy metrics) may represent a possible mechanism for aligning test accuracy more closely to clinical decision making.

6.1.1.4 The interplay between contextual factors and test accuracy metric on comprehension

The nature of the relationship between comprehension of test accuracy metrics, the impact of presentation format and contextual factors impacting on patient and professional utility for decision making is inevitably complex and the summation of effects of these influences are difficult to quantify and predict. Further research is needed to unpick the contribution of motivational and contextual influences distinct from comprehension. Although requiring decision makers to manipulate probabilities as a measure of comprehension of test accuracy information has construct validity, it is unlikely to be representative of how test accuracy information is used in practice or a necessary pre-requisite for informed use of test accuracy information for decision making. In addition an awareness of the potential contextual and perceptual modifiers of test accuracy information is essential in order to ascertain the degree to which educational interventions are likely to impact on decision making, to ensure a productive dialogue with clinicians who are the recipients of guidelines concerning test use and to ensure that choice of test accuracy metrics adequately capture both dimensions of test accuracy.

6.1.2 Contextualisation of the test accuracy evidence base (chapters 3 and 4)

6.1.2.1 Epidemiological characteristics of existing test evaluation reviews

Systematic reviews are an important resource for summarising existing knowledge and underpin guideline development and needs assessment for research activity. Systematic

reviews and meta-analyses of test evaluations are increasing in number and prominence as a resource for diagnostic decision making.

Chapter 3 demonstrates that reviews of test accuracy dominate the test evaluation landscape. Within this body of research, tests more commonly applied in secondary care and certain disease topic areas predominate and there is a lack of comparative test accuracy evaluation. Based on the epidemiological characteristics of test evaluations it is unlikely that the existing evidence base reflects the clinical need for evidence.

6.1.2.2 The contextual fit of test accuracy evaluations

Question formulation

It is particularly important that test evaluation is grounded on well defined clinical questions as many tests are applied far removed from a final diagnosis and by a variety of professionals. Lack of clarity at the question formulation stage precludes judgement about the utility of the information provided by review findings. Detailed scrutiny of test accuracy reviews presented in chapter 4 reveals ill-defined objectives which are reflected in question formulation, review synthesis and reporting of findings. The place of index tests within a testing pathway is mostly not articulated by consideration of test role, (add, replace, triage), healthcare setting, patient presentation, prior tests or current testing practice. The importance of clear question formulation for planning a synthesis framework, including investigation of heterogeneity and the contextualisation of review findings needs promoting. The investigation presented in chapter 4 suggests that at the current time inadequacies in question formulation and the subsequent impact on contextualisation of test accuracy review findings may be undermining the potential for statistical and methodological advances in meta-analysis of test accuracy to positively impact on diagnostic decision making.

Existing guidance pertinent to the contextualisation of test accuracy review questions is limited to recent initiatives of the Cochrane Collaboration and personal experience suggests that review authors, including clinicians, find question formulation in this area difficult. This is

supported by the findings presented in chapter 4. The relative lack of familiarity of the research community with the architecture of test accuracy evaluation and a lack of appreciation of spectrum effects on test accuracy are likely to be contributing to these observed difficulties. Difficulty with question formulation is likely also to reflect testing culture, which has been characterised by indiscriminate introduction of new testing technology without sufficient consideration of the intended role of a test in a testing pathway (see chapter 1). Dissemination of a framework within which to consider evaluations of test accuracy will therefore need to be accompanied by a more critical approach to the adoption of new tests.

Contextualisation of review findings

In the absence of information about clinical context, the results of test accuracy evaluations will mislead and encourage a lack of consideration of the importance of spectrum effects. In addition consideration of the downstream harms and benefits of positive and negative test results is necessary to link diagnostic accuracy with clinical decision making.

Reporting of study characteristics, consideration of the applicability of review findings and consideration of the downstream consequences of test results was observed to be poor in the reviews scrutinised as part of this research, despite the fact that ninety four percent included a clinician as a co-author.

Lack of contextual fit and poor reporting of included primary studies did not appear to be responsible for lack of contextualisation of review findings, on the basis that a minority of reviews reported limitations in the primary evidence base. In addition, lack of contextualisation of review findings is unlikely to be due to limitations in review reporting as lack of clarification at question formulation stage was not rectified at subsequent review stages. Lack of contextualisation of review findings is therefore likely to represent deficiencies in methodological approach rather than being entirely a function of limitations in the primary evidence base and restrictions on reporting imposed by publication.

A structured assessment of external validity according to a clearly focused review question should be promoted as complementary to, and as important as, an assessment of internal validity. In addition any guidance concerned with contextualisation will be more effective if it is coherent with statistical and epidemiological developments in this area.

Use of outcome measures in reviews

The majority of reviews reported between 2 and 3 outcomes although a substantial minority (26/99) reported four or five outcomes. Lack of consideration of context, including the downstream consequences of test results in the reviews scrutinised suggests that use of multiple metrics is not underpinned by an appreciation of the particular dimension or dimensions of test accuracy that each metric conveys. Without an appreciation of the relationship between different test accuracy metrics, review authors are likely to use metrics indiscriminately without conveying those aspects of a test's performance that are most important in any given clinical context. Evidence reviewed in chapter 2 suggests that multiple numerical metrics are perceived as unhelpful. There is therefore a need for development of guidance to assist with the optimum use of test accuracy outcome metrics to support decision making.

Test evaluation in generalist settings

The traditional test accuracy evaluation framework does not fit the early stages of the testing pathway well and the research community needs to consider whether this is acting to discourage test evaluation in generalist settings, including evaluation of the clinical history and examination. Greater emphasis on the purpose of testing may help to identify disorders that are priority 'rule outs' in generalist settings and for which evaluation of a single dimension of test accuracy across these multiple priority target disorders is justifiable. A similar approach could be taken for priority 'rule- ins'.

6.1.3 Summary: Implications for practice

The evidence presented here suggests that quantitative estimates of test accuracy are not a prominent resource for diagnostic decision making in practise. Formal probability revision is unlikely to be commonplace in practice or a necessary pre-requisite for informed decision making. However the extent to which use of test accuracy metric impacted on decision making in the hypothetical testing scenarios presented in chapter 5 has important implications for practice.

The thesis findings raise questions about the extent to which existing use of metrics is based on familiarity rather than ease of application. Counter to what might be expected from the review of empirical and non empirical literature presented in chapter 2, sensitivity and specificity appeared to result in the most informed decision making in the survey presented in chapter 5. Sensitivity and specificity are prominent metrics in test accuracy research and medical education; a situation that is likely to be historical and supported by a widespread misperception that they represent fixed properties of tests. Further, empirical investigation of the accessibility of existing test accuracy metrics has been based on the premise that formal probability revision should be aspired to and is common place in practice. The effect of these perspectives is likely to have stifled meaningful empirical investigation of the utility of existing metrics for diagnostic decision making and the development of novel methods for communicating test accuracy. The findings from this thesis suggest that presentation formats that encompass prevalence and quantify, in frequentist terms, the number of test errors relative to correct test results have the potential to facilitate informed diagnostic decision making.

The concept of evidence based testing is challenged by the findings of this thesis, although it is unclear the extent to which this is a result of problems with comprehension and lack of

familiarity with test accuracy evaluations and outcome measures on behalf of decision makers or deficiencies in the existing test accuracy evidence base.

6.2 Application of research findings

It is extremely likely that the large observed degree of variation in testing behaviour and any inappropriate rise in testing is in part a reflection of the lack of contextualisation of the evidence base and the accessibility of test accuracy metrics for decision makers.

6.2.1 Existing initiatives

Researchers have begun to draw attention to the fact that in the absence of information about clinical context, presentation of test accuracy evaluations may be misleading [28] and it is suggested that examination of the downstream harms and benefits of positive and negative test results is necessary to link diagnostic accuracy with clinical decision making [44]. Researchers are being directed to consider of the role of tests in testing pathways as a means of encouraging a comparative approach to test accuracy evaluation [96,123]. Explicit links between important aspects of question formulation have recently received attention [124] and recent guidance for undertaking medical test reviews from the Agency of Healthcare Research and Quality [120] stress the need for a clear distinction between internal and external validity.

In addition, advances in statistical techniques that allow a more sophisticated and robust investigation of heterogeneity [125,126] are now more widely disseminated and have the potential to support reviews in addressing questions of maximal clinical relevance.

The GRADE (Grading of Recommendations Assessment, Development and Evaluation) working group (http://www.gradeworkinggroup.org/) have highlighted the importance of explicit consideration of the role of tests being evaluated in order to place the use of individual tests within the diagnostic pathway for a condition. Test accuracy is considered as

an intermediate outcome and the GRADE group also recommend that the impact of all possible tests results (true negatives, true positives, false negatives, false positives) on patient outcomes (tests and treatments received and ultimately morbidity and mortality) should be considered [127,128]. However the GRADE initiative is concerned with the development of frameworks within which to consider evidence for guideline developers, rather than targeting those undertaking test accuracy reviews or using test accuracy information. Thus the impact of the work of GRADE to date is likely to have a limited impact on encouraging contextualisation of test accuracy evaluations.

The QUADAS instrument for the quality assessment of primary studies to be included in reviews has recently been updated (QUADAS 2) [129] to include a clear differentiation between internal validity (risk of bias) and external validity (applicability). Although these developments should encourage review authors to consider the contextual fit of included studies, subjective judgments about external validity based on signalling questions are not a replacement for the provision of information to allow an independent assessment of applicability by readers. In the absence of reporting guidelines for systematic reviews of test accuracy, the potential for review authors to consider the use of QUADAS 2 a replacement for a more detailed consideration of clinical context should be considered. The STARD initiative [38] concerned with reporting of primary test accuracy research, has been used by some review authors alongside the QUADAS instrument to guide an assessment of external validity of included studies (personal communication).

Initiatives concerned with the evaluation of diagnostic technologies in the NHS have followed recommendations from Lord Darzi's next stage review [95]. The Centre for Evidence based Purchasing (CEP) superseded the Device Evaluation Service (DES) in 2005 and provides information to assist with purchasing decisions in the UK National Health Service for all medical devices with a CE mark, including diagnostic devices. This information can now for

the first time include reviews of effectiveness and cost-effectiveness. In addition, the NICE (National Institute for Health and Clinical Excellence) Diagnostic Appraisals Programme is nearing the end of its first year. The programme is not restricted to evaluations of test accuracy but includes assessments of the impact of diagnostic technologies across care pathways and should encourage a more critical adoption of new tests and scrutiny of existing testing pathways.

The Cochrane Handbook for Diagnostic Test Accuracy Reviews [96] is an important resource for those undertaking test accuracy reviews. However the handbook is incomplete as at 2012 and the personal experience of those undertaking training for Cochrane review groups and authors suggests that some issues pertinent to the contextualisation of reviews may not be clearly articulated. In addition, software developed for production of Cochrane reviews of diagnostic test accuracy [130] has important limitations in this respect; for example it is currently not possible to produce a summary table of study characteristics to allow readers to assess the applicability of included studies or to present the results of quality assessment or study characteristics for studies sub-grouped according to important potential causes of heterogeneity.

6.2.2 Contribution of the research findings

Dissemination of these research findings are required in order to challenge beliefs prominent in the existing literature which rely heavily on the views of clinical academics. Examples include the accessibility of LRs for decision making, the intuitive nature of PVs and the intractable nature of misunderstanding in this area.

Chapter eleven of the Cochrane handbook will offer guidance on the interpretation of results and inform the development of a summary of results table for Cochrane Diagnostic Test Accuracy Reviews. A steering group is to oversee the development of the chapter which will

draw on relevant research concerned with the comprehension, application and contextualisation of test accuracy evaluations. It is anticipated that the work of the steering group will result in guidance on applicability, presentation of test accuracy information, consideration of the downstream consequences of test results and the critical and complimentary use of test accuracy metrics as well as highlighting areas in need of further research. This thesis represents an invaluable resource for the development of this chapter and the author is a member of the steering group.

Improving the clinical relevance of questions about test accuracy should have a positive impact on all aspects of the conduct of systematic reviews of test accuracy and will assist with the interpretation of results and assessment of the impact of findings. Explicit links between important aspects of question formulation, such as the role in which index tests are being evaluated, and supporting statistical techniques are stressed as part of the guidance for undertaking Cochrane DTA reviews and are a key aspect of training of review authors. The author is member of the Cochrane Diagnostic Test Accuracy Working Group and recently appointed member of the Diagnostic Test Accuracy Editorial Team and will therefore be in a position to disseminate findings of this research in order to support an improvement in the contextual fit of test accuracy reviews. In addition the findings of this research have been presented at the most recent annual Cochrane colloquium [131,132]. Dissemination to the wider research community (out with the Cochrane Collaboration) is likely to be slower but will be facilitated by publication of the findings of this research. Raising awareness about the importance of question formulation to ensure contextual fit and therefore clinical relevance of systematic reviews of test accuracy is important in order to ensure that reviews fulfil their potential to improve the test accuracy evidence base in parallel with the recognised improvements in primary test accuracy studies needed.

The thesis' findings themselves may be sufficient to inform the development of a number of focused educational initiatives. Guidance to those conducting test accuracy reviews might include:

- providing a brief definition of test accuracy metrics as part of the reporting of review findings
- explanation of the different and complementary nature of test accuracy metrics: this might be supported by an annotated 2x2 diagnostic table to illustrate inter-relationships
-illustration of the relationship between the two dimensions of test accuracy and the effect of prevalence on the magnitude of test errors

Such guidance should also result in a more critical and considered use of test accuracy metrics by researchers.

No reporting guidance is available for systematic reviews and meta-analyses of test accuracy studies. The deficiencies highlighted by the review of test accuracy reviews (chapter 4) represents an important addition to existing knowledge concerning the impact of primary test accuracy study quality and reporting practice [38,41] as well as the internal validity of the systematic review process itself (PRISMA) [133].

Assessment of the quality of the systematic reviews included in the methodological review using tools developed for the conduct of systematic reviews and meta-analysis of interventions (AMSTAR) [110] and (QUOROM) [111] was not associated with quality of question formulation or contextualisation of review findings. On the basis of this finding it appears that existing guidance for assessing the quality of conduct and reporting of systematic reviews does not capture important design and reporting features pertinent to systematic reviews of test accuracy. The potential for reporting guidelines to influence the conduct of systematic reviews is particularly great as reporting and conduct are closely entwined.

Although there is no generic guidance on how to develop a reporting guideline, Moher and colleagues [134] propose a strategy underpinned by an executive guideline group, a systematic review of the literature and the need for and a face to face meeting consensus meeting with

key stakeholders. The importance of a dissemination strategy utilising multiple and simultaneous publication as well as extensive post-publication activities to encourage guideline endorsement is emphasised. Widespread dissemination of reporting guidelines for RCTs (the CONSORT statement) accompanied by endorsement by journal editors appears to have had an impact on research conduct [135,136].

6.3 Strengths and Limitations

Strengths

A major strength of this thesis in comparison to existing work is the representative nature of the systematic review (chapters 3 and 4) and survey (chapter 5) samples.

Investigations of test accuracy review methodology to date have been conducted on review samples selected on the basis of topic, (for example Mallet 2006) [106] or convenience samples drawn from single, quality assured review databases [6,7].The review findings presented in Chapters 3 and 4 represent a comprehensive review of five diverse systematic review databases in order to gauge the volume and type of test evaluations and their epidemiological characteristics. This comprehensive repository of reviews allowed the selection of a smaller sample of reviews for detailed scrutiny representing a range of tests, a range of healthcare settings and conducted to variable quality standards.

A major strength of the survey presented in chapter 5 compared with previous empirical research in this area, is the size and representativeness of the sample. Respondents do not appear to be self-selected on the basis of expertise or experience in the topic area.

In addition this thesis includes the first systematic enquiry of the published literature concerning the understanding and application of test accuracy information undertaken to date. The breadth of the bibliographic search strategy used to inform chapter 2, as well as a check of face and content validity at an international test evaluation symposium is likely to

have captured key issues that have been considered by empirical and non empirical researchers regarding the accessibility of existing test accuracy metrics for decision making to date. Adopting a qualitative approach to the investigation of understanding and application of test accuracy metrics presented in the published literature has offered a unique insight into presentation formats that may act as facilitators or barriers to their use.

The internal validity of the thesis' findings is supported by internal consistency observed across its individual elements. Opportunities for comparison concerning understanding and application of test accuracy information between the published literature (chapter 2) and the survey in primary care (chapter 5) are limited by disparities in sample composition and investigative approach. However themes observed in both investigations include a feeling of obligation to pursue further testing to reduce uncertainty (5.5.5.3; 2.4.2), problems with probability revision (2.5.3; 5.5.5.3), confusion of reference class (2.5.3.2; 5.5.5.3), preference for graphical and mixed numerical and verbal presentations of probability (2.6.1.2; 5.5.5.3), and unfamiliarity with and use of LRs and summary metrics more common to systematic reviews (2.5.1.1; 5.5.4.2; 5.5.4.3).

Limitations

The review of non empirical and empirical test accuracy literature represents the perspectives of experts rather than practising clinicians and it is probable that comprehension, accuracy of perception and ability to manipulate risks has been overestimated with less clear impact on preference for metric and presentation format. In addition, the lack of representation of the perspective of generalist settings in the published literature (chapter 2) is mirrored by a paucity of test accuracy evaluations in primary care (chapters 3 and 4). The importance of multidisciplinary perspectives and difficulties associated with searching for literature concerned with test accuracy present challenges to any attempt to comprehensively capture literature relevant to consideration of the accessibility of existing test accuracy metrics and graphics for decision making. However the

exclusive use of published literature will have resulted in under-representation of practising clinicians for the non-empirical evidence base reviewed. Indeed 25 of 30 non empirical papers were authored by clinicians, of which 16/25 were affiliated with an academic institution (2.4).

The results of an assessment of the contextualisation of test accuracy reviews presented in chapter 4 is likely to have presented an overoptimistic picture as a result of the subjective nature of data extraction, the clinical and methodological expertise of the author and the broad assessment framework used which may not be optimal for the topic of any single review. It is therefore likely that the assessment of the extent to which test accuracy reviews addressed testing context is overoptimistic and recommended remedial action more pertinent.

The depth afforded to the analysis in the survey was at the expense of a relative lack of representation of the policy maker perspective in the published literature as well as comprehensive consideration of different testing contexts. The findings of the survey are therefore not generalisable to policy making environments. The extent to which they are reflective of the secondary care setting is less clear.

6.4 Research recommendations

6.4.1 Evaluation of the understanding and application of test accuracy information for decision making

The results of this thesis leave considerable uncertainty regarding the most accessible method for integrating test accuracy evidence into decision making. This uncertainty is present at two levels; the first the degree to which decision makers seek and use research based estimates of test accuracy and the second, the presentation format or combination of presentation formats that is optimal for the promotion of informed diagnostic decision making.

The contribution of this thesis has been to characterise the existing evidence base, highlight deficiencies and to challenge the prevailing view that ability to undertake probability revision is a necessary pre-requisite for informed diagnostic decision making.

Although the survey conducted in primary care represents an important addition to the evidence base it was primarily exploratory in nature and its major strength lies in its ability to inform future research. The scope of the primary research possible in the time available was limited and test accuracy metrics, testing context and the decision maker sample were selected to best address limitations of the existing published literature whilst supporting the evolution of the evidence base. In particular the content of the survey has been driven by an assessment of the utility of existing metrics, presentation formats and sources of test accuracy information and not by an evaluation of the needs of decision makers. Based on the contribution of the open responses to the interpretation of the findings of this questionnaire qualitative research should play a key role in further research aiming to elucidate sources of variation in understanding and application of test accuracy metrics. Specific research questions arising from the survey include:

- The extent to which sources of test accuracy evidence and estimates of pre-test probability are sought in practice, their perceived usefulness and facilitators and barriers to their use.
- The consistency of the findings of this thesis concerning the use, understanding and application of test accuracy information in representative clinical samples across different clinical disciplines in order to distinguish the effect of contextual factors and motivational biases on decision making distinct from methods of presenting test accuracy information. For example contextual variables may be specific to professional groups, reflecting their position in the referral pathway or particular challenges associated with a patient group.
- The effect of contextual modifiers on test and test-treat thresholds in order to inform assessment of the potential impact of a test's accuracy in specific clinical testing

contexts. Quantification of test and test-treat/test-refer thresholds in specific clinical contexts may also assist with prioritisation of test accuracy evaluations.

- Understanding and application of test accuracy presentation formats that include:

  - explicit provision of information about test errors as a possible mechanism for aligning estimates of test accuracy more closely to clinical decision making.

  - the 2x2 diagnostic table as a natural frequency expression of test accuracy and in particular investigation of whether natural frequency expressions offer any advantage over PVs which provide a direct estimate of setting-specific post-test probability

  - graphical displays of test accuracy information

  - mixed (combinations of verbal, graphical and numerical) presentation formats

- The importance of the ability of clinicians to undertake formal probability revision as a facilitator of effective diagnostic decision making at the bedside. Research efforts should be directed at developing and assessing the utility of test accuracy information that incorporates context specific information about prevalence in order to negate the need for probability revision in clinical practice.
- Clarification of the complex relationship between spectrum, prevalence and test accuracy in order to determine whether disease reference class measures (for example sensitivity, specificity) and metrics derived from these measures (LRs) offer any advantage over PVs in this respect. Indeed recent research suggests that directly deriving PVs from meta-analysis produces similar estimates to PVs derived indirectly from summary estimates of sensitivity and specificity [86]. Clarification of the nature of the relationship between spectrum, prevalence and test accuracy metric would also

inform spectrum and prevalence standardised test accuracy estimates as one method for promoting the contextual fit of the evidence base.

- The degree to which the policy maker community comprehend and can apply test accuracy metrics. Indeed such an investigation would be timely with the introduction of new initiatives such as the NICE Diagnostic Appraisal Programme and the development of the Cochrane Database of Diagnostic Test Accuracy Reviews. Preliminary research in this area suggests that lack of familiarity with the architecture of test accuracy research may contribute to problems with the comprehension of test accuracy metrics in test accuracy systematic reviews [137].
- The extent to which variation in tolerance of test errors contributes to variation in diagnostic decision making. Exploration should include testing scenarios that capture variation in the risk associated with the testing process itself and variation in the consequences of test errors as a result of the seriousness of the target condition and the risks associated with further testing or treatment.

6.4.2 Contextual fit of test accuracy evidence

The methodological review presented in chapter 4 aimed to capture how testing context was being incorporated into systematic review methods and represented in the reporting of review findings. The research strongly suggests that deficiencies in the methodological approach to question formulation underpin deficiencies in the contextual fit of existing test accuracy reviews, magnifying the complexity associated with synthesis of primary test accuracy studies. The process of identifying a representative sample of test accuracy reviews allowed investigation of the epidemiological characteristics of the existing test accuracy evidence base. The pattern of research activity is unlikely to be addressing clinical

need with under-representation of tests applicable to generalist settings and no clear explanation for activity concentrated in certain disease topic areas. Initiatives that are supported by the findings from this thesis include:

- The development of links between identified priorities for test accuracy evaluation and research activity. The recently introduced NICE diagnostic appraisals programme should improve the clinical relevance of test accuracy evaluation. In addition prioritisation of test accuracy research could be informed by expansion of initiatives such as the James Lind Alliance Priority Setting Partnership (http://www.lindalliance.org/Introduction.asp) that aims to prioritise research on the basis of uncertainties shared by clinicians, patients and carers and The Database of Uncertainties about the Effects of Treatments (DUETS) which has been established to identify and publish patients' and clinicians' questions about effectiveness that are not answered by existing systematic reviews, (www.library.nhs.uk/DUET).
- Dissemination of the epidemiological characteristics of existing review databases as part of resources for test accuracy review authors. These databases represent a resource for primary studies of test accuracy [138]. The field is a dynamic one and revisiting the databases originally considered for chapters 3 and 4 of this thesis revealed considerable changes over a 4 year period.
- The development of test accuracy review reporting guidelines (considered above 6.2.2). Guideline development should reflect the findings of any further research concerned with improving the accessibility of test accuracy metrics. In addition educational initiatives designed to promote an understanding of the inter-relationship between existing test accuracy metrics and a more critical and considered use of test accuracy metrics (6.2.2) have the potential to encourage more imaginative presentation formats and may provide data on which to refine reporting guidelines, particularly reporting of meta-analyses of test accuracy. The success of reporting guidelines for test accuracy reviews should be evaluated by a re examination of the

conduct and reporting of systematic reviews of test accuracy, although a minimum delay of 5 years between dissemination and evaluation is likely to be required [97].

6.5 End piece

This thesis represents an important contribution to knowledge in the area of test accuracy evaluation. Existing empirical research concerning the understanding and application of test accuracy information is not driven by the needs of decision makers and is based on an unproven assumption that probability revision is a necessary pre-requisite for informed diagnostic decision making. Original primary research has demonstrated that quantitative estimates of test accuracy are unlikely to be used by clinicians at the current time. Initiatives to encourage evidence based testing need to be developed mindful of the fact that choice of test accuracy metric appears to have a profound influence on decision making.

Comprehension of test accuracy metrics, the use of interpretation heuristics promoted in support of evidence based testing and characteristics of metrics themselves - particularly those concerned with conveying the relationship between the two dimensions of test accuracy and pre-test probability, are likely to be important modifiers of decision making behaviour.

The existing test accuracy evidence base does not appear to reflect the clinical need for information. A historical pattern of research activity, a lack of comparative test accuracy evaluations and deficiencies in question formulation are undermining the potential contribution of systematic reviews of test accuracy to promote evidence based testing.

Background Reference List

1 Gray JAM. Evidence Based Healthcare. New York: Churchill Livingstone; 1997.
2 Hernandez-Aguado I. The winding road towards evidence based diagnoses. *Journal of Epidemiology & Community Health* 2002; **56**:323-325.
3 Reid CM, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *N* 1995; **274**:645-651.
4 Deville WL, Buntinx F, (a). Guidelines for conducting systematic reviews of studies evaluating the accuracy of diagnostic tests. In: Knottnerus JA, editor. The Evidence Base of Clinical Diagnosis. London: BMJ books; 2002. p. 145-165.
5 Deeks J. Systematic reviews of evaluations of diagnostic and sreening tests. In: Egger M, Davey Smith G, Altman DG, editors. Systematic Reviews in Health Care: Meta-analysis in context. London: BMJ Publishing Group; 2001. p. 248-282.
6 Honest,H, Khan,KS. Reporting of measures of diagnostic accuracy in systematic reviews of diagnostic literature. Access date: 23 Jan. 2003, URL: http://www.biomedcentral.com/1472-6963/2/4
7 Dinnes J, Deek, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. [9(12)]. 2005. Health Technology Assessment.
Ref Type: Serial (Book,Monograph)
8 Asch DA, Patton JP, Hershey JC. Knowing for the sake of knowing: the value of prognostic information. *Med Decis Making* 1990; **10**:47-57.
9 Zaat JOM, van Eijk TM. General practitioners' uncertainty, risk preference and use of laboratory tests. *Med Care* 1992; **30**:846-854.
10 Dekay ML, Patina-Echeverini D, Fischbeck PS. Distortion of probability and outcome information in risky decisions. *Organisational Behaviour and Human Decision Processes* 2009; **109**:79-92.
11 Pauker SG, Pauker SP. Expected-utility perspectives on Defensive testing: Torst, Tradeoffs, and Thresholds - Is Defensive Medicine Defensible? *Med Decis Making* 1998; **18**(29):29-31.
12 Schiff GD, Hasan O, Seijeoung K, Abrams R, Cosby K, Lambert B, *et al*. Diagnostic Error in Medicine: analysis of 583 physician-reported errors. *Archives of Internal Medicine* 2009; **169**(20):1881-1887.
13 Fowkes FGR. Containing the use of diagnostic tests. *British Medical Journal* 1985; **290**:488.
14 van Walraven C, Naylor D. Do we know what inappropriate laboraory utilisation is? *Journal of the American Medical Association* 1998; **280**(6):550-558.
15 Rink E, Hilton S, Szczepura A, Fletcher J, Sibbald B, Davies C, *et al*. Impact of introducing near patient-testing for standard investigations in general practice. *British Medical Journal* 1993; **307**:775-778.
16 Verrilli D, Welch HG. The impact of diagnostic testing on therapeutic interventions. *Journal of the American Medical Association* 1996; **275**:1189-1191.
17 Winter A, Ray N. Paying accurately for imaging services in Medicare. *Health Affairs* 2008; **27**(6):1479-1490.
18 Heidenreich P. Assessing the value of a diagnostic test. *Arch Intern Med* 2009; **14**:1262-1264.
19 Kassirer JP. Our stubborn quest for diagnostic certainty: a cause of excessive testing. *New England Journal of Medicine* 1989; **320**:1489-1491.
20 Leurquin P, van Casteran V, de Maeseneer J, for the Eurosentinel Study Group. Use of blood test in general practice: a collaborative study in eight European countries. *British Journal of General Practice* 1995; **45**:21-25.

21 Verstappen WH, ter Riet G, Dubois WI, Winkens R, Grol RP, Van der Weijden T. Variation in test ordering behaviour of GPs: professional or context-related factors? *Family Practice* 2004; **21**(4):387-395.
22 Smellie WSA, Galloway MJ, Chinn D, Gedling P. Is clinical practice variability the major reason for diferences in pathology requesting patterns in general practice? *Journal of Clinical Pathology* 2002; **55**(3):312-314.
23 Kristiansen IS, Hjortdahl P. The General Practitioner and Laboratory Utilisation: Why Does It Vary? *Family Practice* 1992; **9**:22-27.
24 Solomon DH, Hashimoto H, Daltroy L, Liang MH. Techniques to improve physicians' use of diagnostic tests: a new conceptual framework. *Journal of the American Medical Association* 1998; **280**(23):2020-2027.
25 Avorn J. Regulation of Devices: Lessons canbe learnt from drug regulation. *British Medical Journal* 2010; **341**:947-948.
26 Linden L, Vondeling H, Packer C, Cook A. Does the National Institute for Health and Clinical Excellence only appraise new pharmaceuticals? *International Journal of Technology Assessment in Health Care* 2007; **23**(3):349-353.
27 Wennberg DE, Kellet MA, Jickens JD, et al. The association between local diagnostic testing intensity and invasive cardiac procedures. *Journal of the American Medical Association* 1996; **275**:1161-1164.
28 Whiting P, Toerien M, de Salis I, Sterne JAC, Dieppe P, Egger M, *et al*. A review identifies and classifies reasons for ordering diagnostic tests. *Journal of Clinical Epidemiology* 2007; **60**(10):981-989.
29 Sood R, Sood A, Gosh AK. Non-evidence based variables affecting physician's test-ordering tendencies: a systemtic review. *The Netherlands Journal of Medicine* 2007; **65**(5):167-177.
30 Bryan S, Williams I, McIver S. Seeing the NICE side of cost-effectiveness analysis: a qualitative investigation of the use of CEA in NICE technology appraisals. *Health Economics* 2007; **16**:179-193.
31 Baumann AO, Deber RB, Thompson GG. Overconfidence among physicians and nurses: the micro-certainty, macro-uncertainty phenomenon. *Soc Sci Med* 1991; **32**(2):167-174.
32 Cochrane,AL. Random reflections on health services. The Nuffield Provincial Hospitals Trust. London: The Royal Society of Medicine Press Ltd, 1999.; 1972. Report No.:
33 Knottnerus JA. The Evidence Base of Clinical Diagnosis. 1st ed. London: BMJ Publishing Group; 2002.
34 Leeflang MMG, Deeks J, Gatsonis CA, Bossuyt PM, (a). Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008; **149**(12):889-897.
35 Rutjes AWS, Reitsma JB, Di Niso M, Smidt N, van Riijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006; **174**(4):469-476.
36 Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, *et al*. Challenges in systematic reviews of diagnostic technologies. *Ann Intern Med* 2005; **142**:1048-1055.
37 Smidt N, Rutjes AW, van der Windt DAWM, Ostelo RWJG, Reitsma JB, et al. Quality of reporting of diagnostic accuracy studies. *Radiology* 2005; **235**:347-353.
38 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou P, Irwig L, *et al*. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *British Medical Journal* 2003; **326**:41-44.
39 Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and Reporting Characteristics of Systematic Reviews. *PLoS Medicine* 2007; **4**(3):447-455.
40 Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *Journal of Clinical Epidemiology* 1995; **48**(1):119-130.
41 Lijmer JG, Mol BW, Heisterkamp SH, Bonsel GJ, Prins MH, van de Meulen JHP, *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association* 1999; **282**(11):1061-1066.

42 Oosterhuis WP, Niessen RW, Bossuyt PM. The science of systematic reviewing studies of diagnostic tests. [Review] [40 refs]. *Clinical Chemistry & Laboratory Medicine* 2000; **38**(7):577-588.
43 Khan KS, Khan SF, Nwosu CR, Chien PFW. Misleading authors' inferences in obstetric diagnostic test literature. *American Journal of Obstetrics and Gynaecology* 1999; **181**:112-115.
44 Cornell J, Mulrow CD, Localio AR. Diagnostic Test Accuracy and Clinical Decision Making. *Ann Intern Med* 2008; **149**(12):904-906.
45 Hamm RM. Clinical Intuition and Clinical Analysis: Expertise and the Cognitive Continuum. In: Dowie J, Elstein A, editors. Professional Judgement: A reader in clinical decision making. Cambridge: Cambridge University Press; 1988. p. 78-105.
46 Lusted LB. Uncertainty and indecision. *Med Decis Making* 1984; **4**:397-399.
47 Kassirer JP, Kuipers BJ, Gorry GA. Toward a theory of clincial expertise. In: Dowie J, Elstein A, editors. Professional Judgement: a reader in clinical decision making. Cambridge: Cambridge University Press; 1988. p. 212-225.
48 Heneghan C, Glasziou P, Thompon A, Rose P, Balla J, Lasserson D, *et al*. Diagnostic strategies used in primary care. *British Medical Journal* 2009; **338**:1003-1006.
49 Eddy DM, Clanton CH. The art of diagnosis: solving and clinicopathological exercise. In: Dowie J, Elstein A, editors. Professional Judgement: A reader in clinical decision making. Cambridge: Cambridge University Press; 1988. p. 200-211.
50 Dowie J. The research practice gap and the role of decision anaysis in closing it. *Health Care Analysis* 1996; **4**:1-14.
51 Egglin KP, Feinstein AR. Context Bias. A Problem in Diagnostic Radiology. *Journal of the American Medical Association* 1996; **276**(21):1752-1755.
52 Hayward RA, Wilson MC, Tunis SR, Bass EB, Guyatt G, for the Evidence Based Medicine Working Group. User's guides to the medical literature VIII:how to use clinical practice guidelines, A: are the recommendations valid? *Journal of the American Medical Association* 1995; **274**:70-74.
53 Wilson MC, Hayward R, Tunis SR, Bass EB, Guyatt G, for the Evidence Based Medicine Working Group. User's guides to the medical literature VIII:how to use clinical practice guidelines, B: what are the recommendations and will they help me in caring for my patient? *Journal of the American Medical Association* 1995; **274**:1630-1632.
54 Einhorn HJ, Hogarth RM. Behavioural decision theory: processes of judgement and choice. *Annual Reviews in Psychology* 1981; **32**(53):88.
55 Matchar DB, Orlando LA. The relationship between test and outcome. In: Price CPaCRH, editor. Evidence-Based Laboratory Medicine: principles, practice and outcomes. Washington DC.: AACCPress; 2007. p. 53-66.
56 Kassirer JP. Diagnostic reasoning. *Ann Intern Med* 1989; **110**:893-900.
57 Simon HA. Rational decision making in business organisations. *American Economic Review* 1979; **69**(493):513.
58 Bordage G. Why did I miss the diagnosis? Some cognitive explanations and educational implications. *Academic Medicine* 1999; **74**:S138-S142.
59 Elstein A, Schwartz A. Clinical Problem solving and diagnostic decision making: a selsctive review of the cognitive research lierature. In: Knottnerus JA, editor. The Evidence Base of CLinical Diagnosis. London: BMJ books; 2002. p. 179-195.
60 Calman K. Cancer:science and society and the communication of risk. *British Medical Journal* 1996; **313**:799-802.
61 Calman K, Royston G. Risk language and dialects. *British Medical Journal* 1997; **315**:939-942.
62 Jaeschke R, Guyatt G, Sackett DL, (b). User's Guides to the Medical Literature IV. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring For My Patients. *Journal of the American Medical Association* 2006; **271**(9):703-707.

63 Fagan TJ. Nomogram for Bayes's Theorem. *New England Journal of Medicine* 1975; **293**:257.
64 Wyatt JC, Altman D. Prognostic models: clinically useful or quickly forgotten? *British Medical Journal* 1995; **311**:1539.
65 Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman D. Prognosis and prognostic research: what, why and how? *British Medical Journal* 2009; **338**:1317-1320.
66 Klein JG. Five pitfalls in decisions about diagnosis and prescribing. *British Medical Journal* 2005; **330**(7494):781-783.
67 Kahneman D, Tversky A. On the psychology of prediction. *Psychological Review* 1993; **80**:237-251.
68 Tversky A, Koehler DJ. Support theory: A nonextentional Representation of Subjective Probability. *Psychological Review* 1994; **101**:547-567.
69 Tversky A, Khaneman D. Judgement under Uncertainty: Heuristics and Biases. *Science* 1974; **185**:1124-1131.
70 Rottenstreich Y, Tversky A. Unpacking, repacking and anchoring: Advances in support theory. *Psychological Review* 1997; **104**:406-415.
71 Sox HC. Probability theory in the use of diagnostic tests. An introduction to critical study of the literature. *Ann Intern Med* 1986; **104**:60-66.
72 Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. *Econometrica* 1979; **47**(2):263-292.
73 Hozo I, Djulbegovic B. When is diagnostic testing inappropriate or irrational? Acceptable Regret Approach. *Med Decis Making* 2008; **28**:540-553.
74 Spranca M, Minsk E, Baron J. Omission and commission in judgement and choice. *Journal of Experimental Social Psychology* 1991; **27**(1):76-105.
75 Hayward RA, Kent DM, Vijan S, Hofer TP, Hayward RA, Kent DM, *et al*. Reporting clinical trial results to inform providers, payers, and consumers. *Health Affairs* 2005; **24**(6):1571-1581.
76 Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *New England Journal of Medicine* 1980; **302**:1109-1117.
77 Barnett-Page E, Thomas J. Methods for the synthesis of qualitative research: a critical review. *BMC Medical Research Methodology* 2009; **9**(59).
78 Britten N, Campbell R, Pope C, Donovan JMM, Pill R. Using meta-ethnography to synthesise qualitative research: a worked example. *Journal of Health Services Research and Policy* 2002; **7**:209-215.
79 Dixon-Woods M, Cavers D, Agarwal S, Annandale E, Arthur A, Harvey J, *et al*. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology* 2006; **6**(35).
80 The Evidence Based Medicine Working Group. Evidence Based Medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association* 1992; **268**:2420-2425.
81 Sackett DL, Richardson WL, Rosenberg W, Haynes B. Evidence Based Medicine: How to practice and teach evidence based medicine. First edition ed. Churchill Livingstone; 1997.
82 Whiting P, Westwood M, Burke M, Sterne J, Glanville J, (a). Systematic reviews should search a range of databases to identify primary studies. *Journal of Clinical Epidemiology* 2008; **61**:357-364.
83 Whiting P, Westwood M, Burke M, Sterne J, Harbourd R, Glanville J, *et al.* Can diagnostic filters offer similar sensitivity and specificity and a reduced number needed to read compared to searches based on index test and target condition?08 Jul; Department of Public Health, Epidemiology and Biostatistics, University of Birmingham. Birmingham, UK.: 2008

84 Ritchie G, Glanville J, Lefebvre C. Do published search filters to identify diagnostic test accuracy studies perform adequately? *Health Information Library Journal* 2007; **24**(3):188-192.
85 Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and Patients Make Sense of Health Statistics. *Association for Psychological Science* 2008; **8**(2):53-96.
86 Leeflang MMG, Hooft L, Reitsma JB, Deeks J, Bossuyt P. Bivariate meta-analysis of predictive vaules. Methods for Evaluating Medical Tests and Biomarkers. Symposium. Jul 1-2; Public Health Epidemiology and Biostatistics, University of Birmingham, Birmingham, UK: 2010
87 Leeflang MMG, Bossuyt P, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence based diagnosis. *Journal of Clinical Epidemiology* 2009; **62**:5-12.
88 Moons KG, van Es GA, Deckers JW, Habbema J, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratios and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* 1996; **8**(1):12-17.
89 Whiting PF, Sterne JAC, Westwood ME, Bachmann LM, Harbord R, Egger M, *et al*. Graphical presentation of diagnostic information. *BMC Medical Research Methodology* 2008; **8**(20).
90 Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks JJ. A capture-recapture analysis demonstrated that randomised trials evaluating the impact of diagnostic tests on patient outcomes are rare. *Journal of Clinical Epidemiology* 2011; **in press**.
91 Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomised controlled trials? *Ann Intern Med* 2006; **144**:850-855.
92 Mitchell R, Rinaldi F, Craig J. Performance of published search strategies for studies of diagnostic test accuracy (SDTAs) in MEDLINE and EMBASE.05 Oct 22-05 Oct 26; XIII Cochrane Colloquim. Melbourne Australia. 22nd-26th October 2005; 2005
93 Leeflang MMG, Scholten RJPM, Rutjes AW, Reitsma JB, Bossuyt PMM. Use of methodological search filters to identify diagnostic accuracy stuies can lead to the ommission of relevant studies. *Journal of Clinical Epidemiology* 2006; **56**:234-240.
94 Storz P, Kolpatzik K, Perleth MKS, Haussler B. Future relevance of genetic testing: a systematic horizon scanning analysis. *International Journal of Health Technology Assessment in Health Care* 2006; **23**(4):495-504.
95 Darzi,LoDK. High Quality Care For All: NHS Next Stage Review final report. Crown Copyright; 2008. Report No.: 7432.
96 Cochrane Test Accuracy Working Group. Cochrane Handbook for Diagnostic Test Accuracy Reviews. Access date: 17 May 2011, URL: http://srdta.cochrane.org/handbook-dta-reviews
97 Willis B, Quigley M. Uptake of newer methodological developments in teh deployment of meta-analysis in diagnostic test research: a systematic review. *BMC Medical Research Methodology* 2011; **11**(27).
98 Leeflang,MMG, Debets-Ossenkopp,YJ, Visser,CE, Scholten,RJPM, Hooft,L, Bijlmer,HA, *et al.* Galactomannan detection for invasive aspergillosis in immunocompromised patients. Access date: 4 Oct. 2011,
99 Mitchell, R., Eisinga, A., Glanville, J., Leeflang, M. M. G., and on behalf of the Cochrane Collaboration EMBASE introduces diagnostic test accuracy as an indexing term. 9-4-2011 http://www.cochrane.org/news/blog/embase-introduces-diagnostic-test-accuracy-study-indexing-term
100 Falk,G, Fahey,T. Diagnosis in General Practice: Clinical Prediction Rules. *British Medical Journal 2009;**338**: b2899.*

101 Irwig L, Bossuyt P, Glasziou P, Gatsonis CA, Lijmer JG. Designing studies to ensure that estimates of test accuracy will travel. In: Knottnerus JA, editor. The Evidence Base of Clinical Diagnosis. London: BMJ books; 2002. p. 95-116.
102 Knottnerus JA. Medical Decision making by general practitioners and specialists. *Family Practice* 1991; **8**:305-307.
103 Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, *et al*. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994; **120**(667):676.
104 Deville WL, Buntinx F, Bouter LM, Montori VM, deVet HCW, van de Windt AWM, *et al*. Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Medical Research Methodology* 2002; **2**(9).
105 Love T, Fahey T. Defining diagnosis: screening and decision making in primary care. *British Journal of General Practice* 2003; **53**(497):914-915.
106 Mallet S, Deeks J, Halligan S, Hopewell S, Cornelius V, Altman D. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *British Medical Journal* 2006; **333**:413-419.
107 CRD. DARE (Database of Abstracts of Reviews of Effectiveness). Access date: 7 July 11 A.D., URL: http://www.crd.york.ac.uk/CMS2Web/AboutDare.asp
108 The Cochrane Library. **Issue 3**[Chichester]. 2006. Wiley.
109 West Midlands Commissioning Support Unit. ARIF database. Access date: 8 July 11 A.D., http://www.arif.bham.ac.uk/databases.shtml . Accessed July 2011.
110 Shea B, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, *et al*. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology* 2007; **7**(10).
111 Moher D, Cook DJ, Eastwood S, Olkin E, Rennie D, Stroup DFft, *et al*. Improving the quality of reports of meta-analysis of randomised controlled trials: the QUOROM statement. *The Lancet* 1999; **354**:1896-1900.
112 Bayliss S, Davenport C. Locating systematic reviews of test accuracy studies: how five specialist review databases measure up. *International Journal of Technology Assessment in Health Care* 2008; **24**(4):403-411.
113 van Weel C. Evidence based interventions and comprehensive treatment. *The Lancet* 1999; **353**(13):916-918.
114 McAlister FA, Straus SE, Sackett DL. Why we need large, simple studies of the clinical examination: the problem and a proposed solution. *The Lancet* 1999; **354**(13):1721-1724.
115 UK National Screening Committee. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme. Access date: 17 Feb. 2012. URL: http://www.screening.nhs.uk/criteria
116 Altman DG. Practical Statistics for Medical Research. 1st ed. London: Chapman and Hall; 1991.
117 Hasselblad V, Hedges LV. Meta-analysis of Screening and Diagnostic Tests. *Quantitative Methods in Psychology* 1995; **117**(1):167-178.
118 Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the qulaity assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 2003; **3**(25).
119 Wilczynski NL. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication-before and after study. *Radiology* 2008; **248**(3):817-823.
120 Agency for Health Care Research and Quality. A comprehensive overview of the methods and reporting of meta-analyses of test accuracy (draft report). Rockville, MD: Agency for Health Care Research and Quality. Rockville, MD; 2011. Report No.:
121 Welsh Assembly Government Statistical Directorate. Statistical Bulletin: Workforce statistics for General Practitioners in Wales 1998-2009. 2009. Report No.: SB 5O/2009.
122 BMA Health Policy and Economics Research Unit. Briefing Note: 2009 UK Medical Workforce. 2010. Report No.:

123 Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *British Medical Journal* 2006; **332**:1089-1092.
124 Hayen A, Macaskill P, Irwig L, Bossuyt P. Appropriate statistical methods are required to assess diagnostic test for replacement, add-on and triage. *Journal of Clinical Epidemiology* 2010; **63**:883-891.
125 Reitsma JB, Glas AS, Rutjes AW, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* 2005; **58**:982-990.
126 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001; **20**:2685-2884.
127 Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, *et al*. GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *British Medical Journal* 2008; **336**:1106-1110.
128 Hsu J, Brozek JL, Terracciano L, Kries J, Compalati E, Stein AT, *et al*. Application of GRADE: Making evidence-based recommendations about diagnostic tests in clincial guidelines. *Implementation Science* 2011; **6**:62-71.
129 Whiting PF, Rutjes AW, Westwood ME, Mallet S, Deery CH, Reitsma JB, *et al*. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; **155**:529-536.
130 Cochrane IMS. RevMan. Cochrane Collaboration, 2011 http://ims.cochrane.org/revman/
131 Davenport,C, Hyde,C, (a). To what extent is the clinical context considered in diagnostic test accuracy (DTA) reviews ?: A methodological review. (Abstract) *Cochrane Database of Systematic Reviews.* Abstracts of the 19th Cochrane Colloquium. 19th-22nd October 2011, Madrid, Spain. Supplement 2011, 9
132 Davenport,C, Hyde,C, (b). What do we know about interpretation and application of test accuray measures? (Abstract) *Cochrane Database of Systematic Reviews* Abstracts of the 19th Cochrane Colloquium. 19th-22nd October 201, Madrid, Spain. Supplement 2011, 11
133 Moher D, Liberati A, Tetzlaff J, Altman D, and the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *Ann Intern Med* 2009; **151**(4):264-269.
134 Moher,D, Schulz,PE, Simera,I, Altman,DG. Guidance for developers of health research reporting guidelines. *PloS Medicine* 2010;**7**(2): e1000217
135 Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reporting of RCTs in 2000 and 2006: a comparative study of articles indexed by PubMed. *British Medical Journal* 2010; **340**(c723).
136 Plint AC, Moher D, Morrison A, Schulz KF, Altman DG, Hill C, *et al*. Does the CONSORT checklist improve the quality of reporting of randomised controlled trials? A systematic review. *Medical Journal of Australia* 2006; **185**:263-267.
137 Zhelev,Z, Garside,R, Hyde,C. Investigating and improving the understanding of Cochrane diagnostic test accuracy reviews (DTARs). (Abstract) *Cochrane Database of Systematic Reviews* Abstracts of the 19th Cochrane Colloquium. 19th-22nd October. Madrid, Spain. Supplement 2011, 8
138 de Vet HC, Eisinga A, Riphagen II, Aertgeerts B, Pewsner D. Searching for Studies. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. The Cochrane Collaboration: 2008. p. 1-60.

Reference List: Theoretical Test Accuracy Literature (TTA)

TTA1 Akobeng AK. Understanding diagnostic test 1: sensitivity, specificity and predictive values. Acta Paediatrica 2007(a); 96(3):338-341.

TTA2 Akobeng AK. Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Paediatrica 2007(b); 96(4):487-491.

TTA3 Benish WA. Mutual information as an index of diagnostic test performance. Methods Inf Med 2003; 42(3):260-264.

TTA4 Bianchi MT, Alexander B. Evidence based diagnosis: does the language reflect the theory? BMJ 2006; 333:442-445.

TTA5 Daniel BL, Daniel TM. Graphic representation of numerically calculated predictive values: an easily comprehended method of evaluating diagnostic tests. Med Decis Making 1993; 13(4):355-358.

TTA6 Doust J. Using probabilistic reasoning. BMJ 2010; 339:1080-1083.

TTA7 Dujardin B, Van den Ende J, Van Gompel A, et al. Likelihood ratios: a real improvement for clinical decision making? European Journal of Epidemiology 1994; 10:29-36.

TTA8 Falk G, Fahey T. Diagnosis in General Practice: Clinical Prediction Rules. BMJ 2009; 339:[b2899]

TTA9 Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: frequency formats. Psychological Review 1995;102:684-704.

TTA10 Gigerenzer G. The psychology of good judgment: frequency formats and simple algorithms. Medical Decision Making 1996; 17(2):273-280.

TTA11 Gigerenzer G, Edwards A. Simple tools for understanding risks: From innumeracy to insight. BMJ 2003;327(741):744.

TTA12 Gill CG, Sabin L, Schmid CH. Why clinicians are natural bayesians. BMJ 2005; 330:1080-1083.

TTA13 Gorry GA, Pauker SG, Schwartz WB. The diagnostic importance of the normal finding.The New England Journal of Medicine 1978; 298:486-489.

TTA14 Grimes DA, Schulz KF. Refining diagnosis with likelihood ratios. Lancet 2005; 365:1500-1505.

TTA15 Halkin A, Reichman J, Schwaber M, Paltiel O, Brezis M. Likelihood ratios: getting diagnostic testing into perspective. Quaterly Journal of Medicine 1997;91:247-258.

TTA16 Henderson AR. Test accuracy is example of redundant information. BMJ 1998;316(7127):312.

TTA17 Hoffrage U, Gigerenzer G, Krauss S, Martignon L. Representation facilities reasoning: what natural frequencies are and what they are not. Cognition 2002; 84:343-352.

TTA18 Kassirer JP. Our stubborn quest for diagnostic certainty: a cause of excessive testing. New England Journal of Medicine 1989;320:1489-1491.

TTA19 Klein JG. Five pitfalls in decisions about diagnosis and prescribing. BMJ 2005; 330(7494):781-783.

TTA20 Knottnerus JA. Interpretation of Diagnostic Data - An Unexplored Field in General Practice. Journal of the Royal College of General Practitioners 1985; 35(275):270-274.

TTA21 Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ 2003;327:716-719.

TTA22 McCowan C, Fahey T. Diagnosis and diagnostic testing in primary care. British Journal of General Practice 2006;56(526):323-324.

TTA23 Miettinen OS, Henschke CI, Yankelevitz DF. Evaluation of Diagnostic Imaging Tests: Diagnostic Probability Estimation. Journal of Clinical Epidemiology 1998; 51(12):1293-1298.

TTA24 Moons KG, Harrell F. Sensitivity and Specificity Should be De-emphasized in Diagnostic Accuracy Studies. Academic Radiology 2003; 10:670-672.

TTA25 Pewsner D, Battaglia J, Minder CE, Harx A, Bucher HC, Egger M. Ruling a diagnosis in or out with "SpIN" and "SnOUT": a note of caution. BMJ 2004; 329:209-213.

TTA26 Richardson WS, Polashenski WA, Robbins BW. Could our pre-test probability estimates become evidence based? A prospective survey of hospital practice. Journal of General Internal Medicine 2003; 18:203-208.

TTA27 Sackett DL, Straus S. On some clinically useful measures of the accuracy of diagnostic tests. Evidence Based Medicine 1998;(May/June):68-70.

TTA28 Sonis J. How to use and interpret interval likelihood ratios. Family Medicine 1999; 31(6):432-437.

TTA 29 Sox HC. Probability theory in the use of diagnostic tests. An introduction to critical study of the literature. Ann Intern Med 1986; 104:60-66.

TTA30 Sox HC. Better care for patients with suspected pulmonary embolism. Ann Intern Med 2006(b); 144(3):210-212.

TTA31 Stengel D, Bauwens K, Sehouli J, Ekkernkamp A, Porzsolt F. A likelihood ratio approach to meta-analysis of diagnostic studies. Journal of Medical Screening 2003; 10:47-51.

TTA32 Summerton N. The medical history as a dignostic technology. British Journal of General Practice 2008; 58:273-276.

TTA33 Van den Ende J, Moreira J, Basinga P, Bisoffi Z. The trouble with likelihood ratios. The Lancet 2005; 366:548.

TTA34 Zaat JOM, van Eijk TM. General practitioners' uncertainty, risk preference and use of laboratory tests. Med Care 1992;30:846-854.

Reference List: Empirical Test Accuracy (ETA)

ETA1 Adab P, Marshall T, Rouse A, Randhawa B, Sangha H, Bhangoo N. Randomised controlled trial of the effect of evidence based information on women's willingness to participate in cervical cancer screening. Journal of Epidemiology and Community Health 2003; 57(8):589-593.

ETA2 Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. American Journal of Medicine 1981; 71:991-998.

ETA3 Borak J, Veilleux S. Errors of intuitive logic among physicians. Soc Sci Med 1982;16:1939-1947.

ETA4 Bramwell R, West H, Salmon P. Health professionals' and service users' interpretation of screening test results: experimental study. British Medical Journal 2006; 333(7562):284-286.

ETA5 Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? Quaterly Journal of Medicine 2003; 96(10):763-769.

ETA6 Casscells W, Schoenberger A, Grayboys T. Interpretation by physicians of clinical laboratory results. New England Journal of Medicine 1978; 299:999-1001.

ETA7 Christensen-Szalanski JJJ, Bushyhead JB. Physicians' Misunderstanding of Normal Findings. Med Decis Making 1983; 3:169-175.

ETA8 Cosmides L, Tooby J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition 1996; 58:1-73?

ETA9 Curley SP. Seeking and applying diagnostic information in a health care setting. Acta Psychologica 1990; 73(3):211-223.

ETA10 Davey HM, Limm J, Butow P, Barratt AL, Houssami N, Higginson R. Consumer information materials for diagnostic breast tests: women's views on information and their understanding of test results. Health Expectations 2003; 6(4):298-311.

ETA11 Dolan JG, Bordley DR, Mushlin AI. An evaluation of clinicians' subjective prior probability estimates. Med Decis Making 1986; 6(4):216-223.

ETA12 Egglin KP, Feinstein AR. Context Bias. A Problem in Diagnostic Radiology. N 1996; 276(21):1752-1755.

ETA13 Evans J, Handley SJ, Perham N, Over DE, Thompson VA. Frequency versus probability formats in statistical word problems. Cognition 2000; 77:197-213.

ETA14 Gigerenzer G, Hoffrage U, Ebert A. AIDS counselling for low-risk clients. AIDS Care 1998; 40:197-211.

ETA15 Girotto V, Gonzalez M. Solving probabilistic and statistical problems: a matter of information structure and question form. Cognition 2001; 78(3):247-276.

ETA16 Grol R, Whitfield M, de Maeseneer J. Defensive attitudes in medical decision making of British, Dutch and Belgian practitioners. British Journal of General Practice 1990; 40:134.

ETA17 Hamm RM, Smith SL. The accuracy of patients' judgments of disease probability and test sensitivity and specificity. Journal of Family Practice 1998; 47(1):44-52.

ETA18 Heller RF, Sanders JE, Patterson L, McEldoff P. GPs' and Physicians' interpretation of risk benefit and diagnostic test results. Family Practice 2004; 21(2):155-159.

ETA19 Hinsz VB. Test-accuracy and base-rate information in the prediction of medical disease occurrence. Current Psychology 2005; 24(2):80-90.

ETA20 Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Academic Medicine 1998; 73(5):538-540.

ETA21 Houben PHH, Winkens RAG, Van der Weijden T, Vossen RCRM, Naus AJM, Grol RPTM. Reasons for ordering laboratory tests and relationship with frequency of abnormal results. Scandinavian Journal of Primary Health Care 2010; 28(1):18-23.

ETA22 Lyman GH, Balducchi L. Overestimation of tests effects in clinical judgement. Journal of Cancer Education 1994 (a); 8:297-307.

ETA23 Lyman GH, Balducchi L. The effect of changing disease risk on clinical reasoning. Journal of General Internal Medicine 1994 (b); 9:488-495.

ETA24 Nightingale SL. Risk preference and laboratory use. Med Decis Making 1987; 7:168.

ETA25 Noguchi Y, Kunihiko M, Imura H, Kiyota M, Fukui T. Quantitative Evaluation of the Diagnostic Thinking Process in Medical Students. Journal of General Internal Medicine 2002; 17(11):848-853.

ETA26 Phelps MA, Levitt MA. Pre-test probability estimates: a pitfall to the utility of evidence-based medicine? Academic Emergency Medicine 2004; 11:692-694.

ETA27 Poses RM, Cebul RD, Wigton RS. You can lead a horse to water - improving physicians' knowledge of probabilities may not affect their decisions. Medical Decision Making 1995; 15(1):65-75.

ETA28 Puhan MA, Steurer J, Bachmann LM, ter Riet G. A randomized trial of ways to describe test accuracy: the effect on physicians' post-test probability estimates. Ann Intern Med 2005; 143(3):184-189.

ETA29 Reid MC, Lane DA, Feinstein AR. Academic calculations versus clinical judgments: Practicing physicians' use of quantitative measures of test accuracy. The American Journal of Medicine 1998; 104(4):374-380.

ETA30 Sassi F, McKee M. Do clinicians always maximise patientoutcomes? A conjoint analysis of preferences for carotid artery testing? Journal of Health Services Research and Policy 2008; 13(2):61-66.

ETA31 Schwartz A, Hupert J. Medical students' application of published evidence: randomised trial. BMJ 2003; 326:536-538.

ETA32 Sox CM, Koepsell TD, Doctor JN, Christakis DA. Pediatricians' clinical decision making: results of 2 randomized controlled trials of test performance characteristics. Archives of Pediatrics & Adolescent Medicine 2006(a); 160(5): 487-492.

ETA33 Steurer J, Fischer E, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. British Medical Journal 2002; 324:824-826.

ETA34 Van den Ende J, Bisoffi Z, Van Puyumbroek H, Van der Stuyft P, Van Gompel A, Derese A et al. Bridging the gap between clinical practice and diagnostic clinical epidemiology. Pilot experiences with a didactic model based on a logarithmic scale. Journal of Evaluation in Clinical Practice 2006; 13:374-380.

ETA35 Villejoubert G, Mandel G. The inverse fallacy: an account of deviations from Bayes' theorem and the additivity principle. Memory & Cognition 2002; 30:171-178.

ETA36 Zaat JOM, van Eijk TM. General practitioners' uncertainty, risk preference and use of laboratory tests. Med Care 1992; 30:846-854.

Reference List: Empirical Risk: (ER)

ER1 Albada A, Ausems MG, Bensing JM, van DS. Tailored information about cancer risk and screening: a systematic review. Patient Education & Counselling 2009; 77(2):155-171.

ER2 Ancker JS, Senathirajah Y, Kukafka R, Starren JB. Design features of graphs in health risk communication: a systematic review. Journal of the American Medical Informatics Association 2006; 13(6):608-618.

ER3 Carling CL, Kristoffersen DT, Montori VM, Herrin J, Schunemann HJ, Treweek S et al. The effect of alternative summary statistics for communicating risk reduction on decisions about taking statins: a randomized trial. PLoS Medicine 2009; 6:e1000134.

ER4 Cuite CL, Weinstein ND, Emmons K, Colditz G. A test of numeric formats for communicating risk probabilities. Med Decis Making 2008; 28:377-384.

ER5 Edwards A, Hood K, Matthews E, Russell D, Russell I, Barker J et al. The effectiveness of one-to-one risk communication interventions in health care: a systematic review. Med Decis Making 2000; 20(3):290-297.

ER6 Edwards A. Presenting risk information - A review of the effects of "framing" and other manipulations on patient outcomes. Journal of Health Communication 2001; 6:61-82.

ER7 Edwards A, Unigwe S, Elwyn G, Hood K, Edwards A, Unigwe S et al. Personalised risk communication for informed decision making about entering screening programs. Cochrane Database Syst Rev. 2006 (a);(4):CD001865; PMID: 17054144.

ER8 Edwards A, Thomas R, Williams R, Ellner AL, Brown P, Elwyn G et al. Presenting risk information to people with diabetes: evaluating effects and preferences for different formats by a web-based randomised controlled trial. Patient Education & Counselling 2006(b); 63:336-349.

ER9 Edwards A, Gray J, Clarke A, Dundon J, Elwyn G, Gaff C et al. Interventions to improve risk communication in clinical genetics: systematic review. Patient Education & Counseling 2008; 71(1):4-25.

ER10 Epstein AE, Alper BS, Quill TE. Communicating evidence for participatory decision making. Decision Making 2004; 291(19):2359-2366.

ER11 Hembroff LA, Holmes-Rovner M, Wills CE. Treatment decision-making and the form of risk communication: Results of a factorial survey. BMC Medical Informatics and Decision Making 2004; 4, 2004. Article Number: 20.

ER12 Julian-Reynier C, Welkenhuysen M, Hagoel L, Decruyenaere M, Hopwood P, CRISCOM Working Group. Risk communication strategies: state of the art and effectiveness in the context of cancer genetic services. European Journal of Human Genetics 2003; 11(10):725-736.

ER13 Kuhberger A. The influence of framing on risky decisions: a meta-analysis. Organizational Behavior & Human Performance 1998; 75(1):23-55.

ER14 Lipkus IM, Crawford Y, Fenn N, Birdavolu M, Binder RA. Testing different formats for communicating colorectal cancer risk. Journal of Health Communication 1999; 4:311-324.

ER15 Lobb EA, Butow P, Meiser B. Women's preferences and consultants' communication of risk in consultations about familial breast cancer: impact on patient outcomes. Journal of Medical Genetics 2003; 40(5):e56.

ER16 McGettigan P, Sly K, O'Connell D, Hill S, Henry D. The Effects of Information Framing on the Practices of Physicians. Journal of General Internal Medicine 1999; 14:633-642.

ER17 Schwartz LM, Can Patients Interpret Health Information? An Assessment of the Medical Data Interpretation Test. Med Decis Making 2005; .25(3).

ER18 Sheridan SL, Pignone MP, Lewis CL. A randomized comparison of patients' understanding of number needed to treat and other common risk reduction formats. Journal of General Internal Medicine 2003; 18:884-892.

ER19 Siegrist M, Orlow P, Keller C. The effect of graphical and numerical presentation of hypothetical prenatal diagnosis results on risk perception. Med Decis Making 2008; 28:567-574.

ER20 Young SD, Oppenheimer DM, Young SD, Oppenheimer DM. Different methods of presenting risk information and their influence on medication compliance intentions: results of three studies. Clinical Therapeutics 2006; 28(1):129-139.

ER21 Zikmund-Fisher BJ, Ubel PA, Smith DM, Derry HA, McClure JB, Stark A et al. Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. Patient Education and Counselling 2008; 73:209-214.

Reference List: Methodological Review of Test Accuracy Reviews: TAR

TAR1. Agency for Health Care Research and Quality. Results of systematic review of research on diagnosis and treatment of coronary heart disease in women. AHRQ Evidence Report/Technology Assessment 2003;(80):1-4.

TAR2. Anderson REB. Meta-analysis of the clinical and laboratory diagnosis of appendicitis. British Journal of Surgery 2004; 91(1):28-37.

TAR3. Appel LJ. Ambulatory blood pressure monitoring and blood pressure self measurement in the diagnosis and management of hypertension. Annals of Internal Medicine 1993;118 (11):867-882.

TAR4. Austin MP LJ. Antenatal screening for postnatal depression: a systematic review. Acta Psychiatrica Scandinavica 2003; 107(1):10-17.

TAR5. Barlow J, Stewart-Brown S, Fletcher J. Systematic review of the school entry medical examination. Archives of Disease in Childhood 1998; 78:301-311.

TAR6. Bastian LA, Nanda K, Hasselblad V, Simel DL. Diagnostic efficiency of home pregnancy test kits: a meta-analysis. Archives of Family Medicine 1998; 7:465-469.

TAR7. Battaglia M. Accuracy of B-type natriuretic peptide tests to exclude congestive heart failure: systematic review of test accuracy studies. Archives of Internal Medicine 2006;166 (10):1073-1080.

TAR8. Becker DM, Philbrick JT, Bachhuber TL, Humphries JE. D- Dimer testing and acute venous thromboembolism. Archives of Internal Medicine 1996;156 (9):939-946.

TAR9. Berger MY, Lijmer JG, de KH, Prins A, Bohnen AM. Abdominal symptoms: do they predict gallstones? A systematic review. Scandinavian Journal of Gastroenterology 2000; 35:70-76.

TAR10. Berry J. Microalbuminuria testing in diabetes: is a dipstick as effective as laboratory tests? British Journal of Community Nursing 2003; 8:267-273.

TAR11. Brietzke SE, Katz ES, Roberson DW. Can history and physical examination reliably diagnose pediatric obstructive sleep apnea/hypopnea syndrome? A systematic review of the literature. Otolaryngology and Head and Neck Surgery 2004; 131(6):827-832.

TAR12. Chen SC, Bravata DM, Weil E, Olkin I. A comparison of dermatologists' and primary care physicians' accuracy in diagnosing melanoma. Archives of Dermatology 2001; 137:1627-1634.

TAR13. Chunn AA, McGee SR. Bedside diagnosis of coronary artery disease: a systematic review. American Journal of Medicine 2004; 117:334-43:334-343.

TAR14. Conde-Agudelo A. World health organization systematic review of screening tests for preeclampsia. Obstetrics and Gynecology 2004; 104(6):1367-1391.

TAR15. Cook RL, Hutchison SL, Ostergaard L, Braithwaite RS, Ness RB. Systematic review: noninvasive testing for chlamydia trachomatis and neisseria gonorhoea. Annals of Internal Medicine 2005; 142(11):914-925.

TAR16. de Brunn G, Graviss EA. A systematic review of the diagnostic accuracy of physical examination for the detection of cirrhosis. BMC Medical Informatics and Decision Making 2001; 1:6.

TAR17. Deville WL, Yzermans JC, van Duijn NP, Bezemer PD, van der Windt DA, Bouter LM. The urine dipstick test useful to rule out infections: a meta-analysis of the accuracy. BMC Urology 2004; 4:4.

TAR18. Dinnes J. The effectiveness of diagnostic tests for the assessment of shoulder pain due to soft tissue disorders: a systematic review. Health Technology Assessment 2003; 7(29).

TAR19. Dodd SR LGCJ. In a systematic review, infrared ear thermometry for fever diagnosis in children finds poor sensitivity. Journal of Clinical Epidemiology 2006; 59(4):354-357.

TAR20. Doust JA, Glasziou PP, Pietrzak E, Dobson AJ. A systematic review of the diagnostic accuracy of natriuretic peptides for heart failure. Archives of Internal Medicine 2004; 64(18):1978-1984.

TAR21. Fancher TL, White RH, Kravitz RL. Combined use of rapid D-dimer testing and estimation of clinical probability in the diagnosis of deep vein thrombosis: systematic review. BMJ 2004; 329:821-829.

TAR22. Fiellin DA, Carrington RM, O'Connor PG. Screening for alcohol problems in primary care: a systematic review. Archives of Internal Medicine 2000; 160:1977-1989.

TAR23. Flemons WW, Littnew MR, Rowley JA. Home diagnosis of sleep apnea: a systematic review of the literature: an evidence review cosponsored by the American academy of sleep medicine, the American college of chest physicians, and the American thoracic society. Chest 2003; 124(4):1543-1579.

TAR24. Fowler-Brown A. Exercise tolerance testing to screen for coronary heart disease: a systematic review for the technical support for the US Preventive Services Task Force. Annals of Internal Medicine 2004; 140(7):W9-24.

TAR25. Fransen GA JMMJLRJJ. Meta-analysis: the diagnostic value of alarm symptoms for upper gastrointestinal malignancy. Alimentary Pharmacology and Therapeutics 2004; 20(10):1045-1052.

TAR26. Garber AM, Solomon NA. Cost-effectiveness of alternative test strategies for the diagnosis of coronary artery disease. Annals of Internal Medicine 1999;(130):7-728.

TAR27. Gianrossi R, Detrano R, Colombo A, Froelicher V. Cardiac fluoroscopy for the diagnosis of coronary artery disease:a meta-analytic review. American Heart Journal 1990;120.

TAR28. Gisbert JP, Pajares JM. Diagnosis of helicobacter pylori infection by stoll antigen determination: a systematic review. The American Journal of Gastroenterology 2001; 96(10):2829-2838.

TAR29. Goodacre S, Sutton AJ, Sampson FC. Meta-analysis: the value of clinical assessment in the diagnosis of deep venous thrombosis. Annals of Internal Medicine 2005 (a); 143(2):129-139.

TAR30. Goodacre S, Sampson FC, Sutton AJ, Mason S, Morris F. Variation in the diagnostic performance of D-dimer for suspected deep vein thrombosis (DARE provisional record). Quaterly Journal of Medicine 2005 (b); 98:513-527.

TAR31. Gorelick MH, Shaw KN. Screening tests for urinary tract infection in children: a meta-analysis. Pediatrics 1999; 104:E54.

TAR32. Harris R. Screening adults for type 2 diabetes: a review of the evidence for the US Preventive Task Force. Annals of Internal Medicine 2003; 138: 215-299.

TAR33. Heim SW, Schectman JM, Siadaty MS, Philbrick JT. D-dimer testing for deep venous thrombosis: a meta-analysis. Clinical Chemistry 2004; 50:1136-1147.

TAR34. Hobbs FDR, Fitzmaurice DA, Wilson S, Hyde CJ, Thorpe GH, Earl-Slater ASM, et al. A review of near patient testing in primary care. Health Technology Assessment 1997;1(5):1-231.

TAR35. Huicho L. Fecal screening tests in the approach to acute infectious diarrhoea: a scientific overview. Pediatric Infectious Disease Journal 1996; 15(6):486-494.

TAR36. Huicho L, Campos-Sanchez M, Alamo C. Meta-analysis of urine screening tests for determining the risk of urinary tract infection in children. Pediatric Infectious Disease Journal 2002; 21:1-11.

TAR37. Ioannidis JP, Lau J. Technical report. Evidence for the diagnosis and treatment of acute uncomplicated sinusitis in children: a systematic overview. Pediatrics 2001; 108(8):E57.

TAR38. Jarvick JG Deyo RA. Diagnostic evaluation of low back pain with emphasis on imaging. Annals of Internal Medicine 2002; 137:586-97:586-597.

TAR39. Jorm AF. Methods of screening for dementia: a meta-analysis of studies comparing an informant questionnaire with a brief cognitive test. Alzheimer Disease and Associated Disorders 1997; 11:158-162.

TAR40. Kearon C, Julian JA, Newman TE, Ginsberg JS. Non-invasive diagnosis of deep venous thrombosis. Annals of Internal Medicine 1998; 128:663-677.

TAR41. Kim C, Kwok YS, Heagerty P, Redberg R. Pharmacologic stress testing for coronary disease diagnosis: a meta-analysis. American Heart Journal 2001; 142:934-944.

TAR42. Kotler TS, Diamond GA. Exercise thallium 201 scintigraphy in the diagnosis and prognosis of coronary heart disease. Annals of Internal Medicine 1990;(113):684-702.

TAR43. Kwok Y, Kim C, Grady D, Segal M, Redberg R. Meta-analysis of exercise testing to detect coronary artery disease in women. American Journal of Cardiology 1999; 83:660-666.

TAR44. Law J, Boyle J, Harris F, Harkness A, Nye C. Screening for speech and language delay: a systematic review of the literature. Health Technology Assessment 1998;2(9).

TAR45. Lee R. A meta-analysis of the performance characteristics of the free prostate-specific antigen test. Urology 2006; 67(4):762-768.

TAR46. Lewis NR, Scott BB. Systematic review: the use of serology to exclude or diagnose coeliac disease (a comparison of the endomysial and tissue transflutaminase antibody tests). Alimentary Pharmacology and Therapeutics 2006; 24(1):47-54.

TAR47. Linzer M. Diagnosing syncope. Part 1: value of history, physical examination and electrocardiography. Annals of Internal Medicine 1997; 126(12): 989-996.

TAR48. Loy CT, Irwig LM, Katelaris PH, Talley NJ. Do commercial serological kits for Helicobacter pylori infection differ in accuracy: a meta-analysis. American Journal of Gastroenterology 1996; 91:1138-1144.

TAR49. Maguire S, Mann MK, Sibert J, Kemp A. Can you age bruises accurately in children? A systematic review. Archives of Disease in Childhood 2005; 90(2):187-189.

TAR50. Mant JR, McManus RJ, Oakes RAL, Delaney BC, Barton PM, Deeks JJ, et al. Systematic review and modelling of the investigation of acute and chronic chest pain presenting in primary care. Health Technology Assessment 2004; 8(2).

TAR51. Marshall D, Johnell O, Wedel H. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. BMJ 1996; 312:1254-1259.

TAR52. Marx A, Pewsner D, Egger M, Nuesch R, Bucher HC, Genton B et al. Meta-analysis: accuracy of rapid tests for malaria in travellers returning from endemic areas. Annals of Internal Medicine 2005; 142:836-846.

TAR53. McGowan JH, Cleland JG. Reliability of reporting left ventricular systolic function by echocardiography: a systematic review of 3 methods. American Heart Journal 2003; 146(3):388-397.

TAR54. Mohseni-Bandpei MA, Watson MJ, Richardson B. Application of surface electromyography in the assessment of low back pain: a literature review. Physical Therapy Reviews 2000; 5:93-105.

TAR55. Mourad O. A comprehensive evidence based approach to fever of unknown origin. Archives of Internal Medicine 2003; 163: 545-551.

TAR56. Nayak S O, Liu.H., Grabe M, Gould MK, Allen E, Owens DK et al. Meta-analysis: accuracy of quantitative ultrasound for identifying patients with osteoporosis. Annals of Internal Medicine 2006; 144(11):832-841.

TAR57. Nelson HD, Nygren P. Screening for speech and language delay in preschool children: systematic evidence review for the US preventive services task force. Pediatrics 2006; 117(2):e298-3.

TAR58. Numans ME, Lau J, de Wit NJ, Bonis PA. Short-term treatment with proton-pump inhibitors as a test for gastroesophageal reflux disease: a meta-analysis of diagnostic test characteristics. Annals of Internal Medicine 2004; 140:518-527.

TAR59. Oei EH, Nikken JJ, Verstijnen AC, Ginai AZ, Hunink MG. MR imaging of the menisci and cruciate ligaments: a systematic review. Radiology 2003; 226:837-848.

TAR60. Ogilvie GS, Patrick DM, Schulzer M, Sellors JW, Petric M, Chambers K, et al. Diagnostic accuracy of self collected vaginal specimens for human papillomavirus compared to clinician collected human papillomavirus specimens: a meta-analysis. Sexually Transmitted Infections 2005; 81(3):207-212.

TAR61. O'Meara S, Nelson EA, Golder JE, Dalton JE, Craig D, Iglesias C on behalf of the DASIDU Steering Group. Systematic review of methods to diagnose infection in foot ulcers in diabetes .Diabetic Medicine 2006;23(4):341-347.

TAR62. Oosterhuis WP, Niessen RW, Bossuyt PM, Sanders GT, Sturk A. Diagnostic value of the mean corpuscular volume in the detection of vitamin B12 deficiency. Scandinavian Journal of Clinical and Laboratory Investigation 2000; 60:9-18.

TAR63. Owens DK, Holodniy M, Garber AM, Scott J, Sonnad S, Moses L et al. Polymerase chain reaction for the diagnosis of HIV infection in adults: a meta-analysis with recommendations for clinical practice and study design. Annals of Internal Medicine 1996; 124:803-815.

TAR64. Pasternack I, I, Malmivaara A, Tervahartiala P, Forsberg H, Vehmas T. Magnetic resonance imaging findings in respect to carpal tunnel syndrome. Scandinavian Journal of Work, Environment and Health 2003; 29:189-196.

TAR65. Peters R, Barlow J. Systematic review of instruments designed to predict child maltreatment during the antenatal and postnatal periods. Child Abuse Review 2003; 12:416-439.

TAR66. Petersen RC, Stevens JC, Ganguli M, Tangalos EG, Cummings JL, DeKosky ST. Practice parameter: early detection of dementia. Mild cognitive impairment (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology 2001; 56:1133-1142.

TAR67. Pignone M.P., Gaynes BN, Rushton JL, Orleans CT, Whitener BL, Mills C et al. Screening for depression in adults: a summary of the evidence for the US Preventive Services Task Force. Annals of Internal Medicine 2002; 136: 765-766.

TAR68. Pirozzo S, Papinczak T, Glasziou P. Whispered voice test for screening for hearing impairment in adults and children: systematic review. BMJ 2003; 327:967-970.

TAR69. Price CP, Newall RG, Boyd JC. Use of protein:creatinine ratio measurements on random urine samples for prediction of significant proteinuria: a systematic review. Clinical Chemistry 2005; 51:1577-1586.

TAR70. Ramsay J, Richardson J, Carter YH, Davidson LL, Feder G. Should health professionals screen women for domestic violence: systematic review. BMJ 2002; 325:314-318.

TAR71. Rappeport ED, Mehta S, Wieslander SB. MR imaging before arthroscopy in knee joint disorders? Acta Radiologica 1996; 37(5):602-609.

TAR72. Reed WW, Byrd GS, Gates RH, Howard RS, Weaver MJ. Sputum gram's stain in community-acquired pneumococcal pneumonia: a meta-analysis. Western Journal of Medicine 1996; 165:197-204.

TAR73. Reuchlin-Vroklage LM, Bierma-Zeinstra S, Benninga MA, Berger MY. Diagnostic value of abdominal radiography in constipated children: a systematic review. Archives of Pediatric and Adolescent Medicine 2005; 159(7):671-678.

TAR74. Riedemann JP, Munoz S, Kavanaugh A. The use of second generation anti-CCP antibody (anti-CCP2) testing in rheumatoid arthritis - a systematic review. Clinical and Experimental Rheumatology 2005; 23(5 Suppl 39):S69-S76.

TAR75. Rietveld RP, van Weert CPN, ter Riet G, Bindels JE. Diagnostic impact of signs and symptoms in acute infectious conjunctivitis: systematic literature search. BMJ 2003; 327:789.

TAR76. Rodgers M, Nixon J, Hempel S, Aho T, Kelly J, Neal D. Diagnostic tests and algorithms used in the investigation of haematuria: systematic reviews and economic evaluation. Health Technology Assessment 2006; 10(18).

TAR77. Ross SD, Allen IE, Harrison KJ, Kvasz M, Connelly J, Seinhait IA. Systematic review of the literature regarding the diagnosis of sleep apnoea. Rockville (MD). Agency for Health Care Policy and Research;1999. AHCPR publication no 99-E002.

TAR78. Schmitt B, Golub RM, Green R. Screening primary care patients for hereditary hemochromatosis with transferrin saturation and serum ferritin level: systematic review for the American College of Physicians. Annals of Internal Medicine 2005; 143:522-536.

TAR79. Scholten RJPM, Deville WLJM, Opstelten W, Bijl D, van der Plas CG, Bouter LM. The accuracy of physical diagnostic tests for assessing meniscal lesions of the knee a meta-analysis. The Journal of Family Practice 2001; 50(11):938-944.

TAR80. Schuijf JD, Bax JJ, Shaw LJ, de Roos A, Lamb HJ, van der Wall EE, Wijns W. Meta-analysis of comparative diagnostic performance of magnetic resonance imaging and multi-slice computer tomography for non-invasive coronary angiography. American Heart Journal 2006; 151(2):404-411.

TAR81. Scott DA, Loveman E, McIntyre L, Waugh N. Screening for gestational diabetes: a systematic review and economic evaluation. Health Technology Assessment 2002; 6:1-172.

TAR82. Scouller K, Conigrave KM, Macaskill P, Irwig L, Whitfield JB. Should we use carbohydrate-deficient transferrin instead of gamma-glutamyltransferase for detecting problem drinkers: a systematic review and meta-analysis. Clinical Chemistry 2000; 46:1894-1902.

TAR83. Selley S, Donovan J, Faulkner A, Coast J, Gillatt D. Diagnosis, management and screening of early localised prostate cancer. Health Technology Assessment 1997; 1:1-96.

TAR84. Singer DE, Nathan DM, Fogel HA, Schachat AP. Screening for diabetic retinopathy. Annals of Internal Medicine 1992;(116):660-671.

TAR85. Siu AL. Screening for dementia and investigating its causes. Annals of Internal Medicine 1991;(115):122-132.

TAR86. Stein PD, Hull RD, Patel KC, Olson RE, Ghali WA, Brant R et al. D-Dimer for the exclusion of acute venous thrombosis and pulmonary embolism: a systematic review. Annals of Internal Medicine 2004; 140:589-602.

TAR87. Stein PD, Beemath A. Multidetector computer tomography for the diagnosis of coronary artery disease: a systematic review. American Journal of Medicine 2006; 119(3):203-2.

TAR88. Storgaard H, Neilsen SD, Gluud C. The validity of the Michigan alcoholism screening test (MAST). Alcohol and Alcoholism 1994;29(5):493-502.

TAR89. Takata GS, Chan LS, Morphew T, Mangione-Smith R, Morton SC, Shekelle P. Evidence assessment of the accuracy of methods of diagnosing middle ear effusion in children with otitis media with effusion. Pediatrics 2003; 112:1379-1387.

TAR90. Tamariz LJ, Eng J, Segal JB, Krishnan JA, Bolger DT, Streiff MB et al. Usefulness of clinical prediction rules for the diagnosis of venous thromboembolism: a systematic review. American Journal of Medicine 2004; 117:676-684.

TAR91. Tu FF, As-Sanie S, Steege JF. Musculoskeletal causes of chronic pelvic pain: a systematic review of diagnosis: part I. Obstetrical and Gynecological Survey 2005; 60(6):379-385.

TAR92. Tugwell P, Dennis DT, Weinstein A, Wells G, Shea B, Nichol G et al. Laboratory evaluation in the diagnosis of Lyme disease. Annals of Internal Medicine 1997; 127:1109-1123.

TAR93. van den Hoogen HM, Koes BW, van Eijk JT, Bouter LM. On the accuracy of history, physical examination and erythrocyte sedimentation rate in diagnosing low back pain in general practice: a criteria-based review of the literature. Spine 1995; 20:318-327.

TAR94. van der Meer V, Neven AK, van den Broek PJ, Assendelft WJ. Diagnostic value of C reactive protein in infections of the lower respiratory tract: systematic review. BMJ 2005; 331:26.

TAR95. Wang WH, Huang JQ, Zheng GF, Wong WM, Lam SK, Karlberg J et al. Is proton pump inhibitor testing an effective approach to diagnose gastroesophageal reflux disease in patients with non-cardiac chest pain: a meta-analysis. Archives of Internal Medicine 2005; 165:1222-1228.

TAR96. Waugh JJ, Clark TJ, Divakaran TG, Khan KS, Kilby MD. Accuracy of urinalysis dipstick techniques in predicting significant proteinuria in pregnancy. Obstetrics and Gynecology 2004; 103:769-777.

TAR97. Whiting PF, Westwood ME, Watt IS, Kleijnen J. Rapid tests and urine sampling techniques for the diagnosis of urinary tract infection (UTI) in children under five years: a systematic review. BMC Pediatrics 2005; 5:1:4:1.

TAR98. Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M et al. Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. BMJ 2006; 332:875.

TAR99. Wiese W, Patel SR, Patel SC, Ohl CA, Estrada CA. A meta-analysis of the Papanicolaou smear and wet mount for the diagnosis of vaginal trichomoniasis. American Journal of Medicine 2000; 108:301-308.

TAR100. Zintzaras E, Gemenis AE. Diagnostic performance of antibodies against tissue transglutaminase for the diagnosis of celiac disease: meta-analysis. Clinical and Vaccine Immunology 2006; 13(2):187-192.