

Towards greater clarity for the analysis of imaging studies: Development & validation of an alternative to the area under the receiver-operator characteristic curve.

*Submitted for the postgraduate research degree of PhD,
March 2015*

Division of Medicine, UCL

Candidate:

Professor Steve Halligan MB BS, MD, FRCP, FRCR

Head, UCL Centre for Medical Imaging,

3rd Floor East
250 Euston Road
London NW1 2PG

Telephone: 0044 2034567890 (Ext: 9094)

Direct Line: 0044 20 3447 9094

email: s.halligan@ucl.ac.uk

UCL student number: HALLI02

Declaration of interests:

SH acted as a consultant for Medicsight plc, a company who developed CAD software for CT colonography, until 2010. He is currently consultant for iCAD, a company who also develop CAD software for CTC.

Declaration:

I, Professor Steve Halligan confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. Specifically, a full breakdown of responsibilities is presented in the "Contribution and acknowledgements" section of the thesis.

Thesis abstract

This thesis arose from a 2006 study performed by the author and his collaborators that attempted to gain regulatory approval for computer-assisted detection (CAD) software. The USA Food & Drug Administration (FDA) obliged us to use the change in the area under the receiver-operator characteristic curve (ROC AUC) as our primary outcome. Despite its wide dissemination in radiology research, we found implementation of ROC AUC very problematic. This thesis explores the hurdles we encountered and argues for an alternative approach.

Chapter 1 describes the rationale for and against ROC AUC as a measure of diagnostic performance. An alternative analysis based on net benefit is proposed on the basis that it is more transparent and simpler to interpret.

Chapter 2 uses the net benefit method to analyse a multi-reader multi-case (MRMC) study of CAD for CT colonography. The analysis requires an estimate of relative misclassification costs for false-negative versus false-positive diagnoses; “ W ”. This study used a conservative value for W , arrived at via consensus.

In **Chapter 3** an evidence-based value for W in the context of screening for colorectal cancer and polyps by CT colonography is arrived at via a discrete choice experiment (DCE) of patients and healthcare workers.

Chapter 4 uses the value for W obtained in Chapter 3 in a net benefit analysis to compare observer performance in two MRMC studies of CAD for CT colonography.

Chapter 5 obtains W by DCE for a different clinical context – detection of extracolonic pathology by CT colonography.

Chapter 6 describes a systematic review that aims to determine whether reporting of MRMC ROC AUC methods in the radiological literature is comprehensive.

Chapter 7 then provides guidelines for the comprehensive reporting of MRMC ROC AUC studies.

The thesis finishes with a summary of the work performed and suggestions for further research.

Dedication

To Julia and Sarah.

I am so very proud of you both: May your medical careers be as rewarding as mine.

Table of Contents

Candidate:.....	1
Thesis abstract	2
Dedication.....	3
List of Tables.....	7
List of Figures	9
List of abbreviations/glossary.....	11
Chapter 1: Disadvantages of using the area under the receiver operator characteristic curve (ROC AUC) to assess imaging tests: A discussion & proposal for an alternative approach.....	12
Abstract	12
Introduction.....	13
The ROC plot	13
ROC AUC.....	16
MRMC ROC studies	21
What are the advantages of ROC AUC?.....	22
What are the disadvantages of ROC AUC?.....	23
<i>Clinical comprehension and relevance</i>	<i>23</i>
<i>Are sensitivity and specificity equally important?.....</i>	<i>24</i>
<i>Confidence scores may be inconsistent and unreliable.....</i>	<i>25</i>
<i>Curve extrapolation</i>	<i>29</i>
<i>Prevalence of abnormality.....</i>	<i>32</i>
Alternatives to ROC AUC.....	33
Advantages of net benefit methods.....	37
Summary	37
Chapter 2: Incremental benefit of computer-aided detection when used as a second- & concurrent-reader for CT colonography: Multi-observer study.	39
Abstract	39
Introduction.....	40
Methods.....	40
<i>Definition of ground truth</i>	<i>42</i>
<i>Readers and case interpretation.....</i>	<i>42</i>
<i>Statistical analysis.....</i>	<i>43</i>
Results	44
<i>Per-patient analysis.....</i>	<i>44</i>
<i>Per-polyp analysis.....</i>	<i>47</i>
<i>Interpretation time.....</i>	<i>48</i>
Discussion.....	48
Chapter 3: Patients' & healthcare professionals' values regarding true- & false-positive diagnosis when colorectal cancer screening by CT colonography: Discrete choice experiment.	52
Abstract	52
Introduction.....	53
Methods.....	54
<i>Ethical statement</i>	<i>54</i>
<i>Design</i>	<i>54</i>
<i>Pilot.....</i>	<i>56</i>
<i>Recruitment.....</i>	<i>56</i>
<i>Statistical analysis.....</i>	<i>57</i>

Results	60
<i>Non-traders</i>	60
<i>Cancer</i>	61
<i>Polyps</i>	62
<i>Willingness-to-pay</i>	62
Discussion.....	63
Chapter 4: Assessment of the incremental benefit of computer-aided detection (CAD) for interpretation of CT colonography by experienced and inexperienced readers.	67
Abstract	67
Introduction.....	68
Methods.....	68
<i>Data sources and readers</i>	68
<i>Data characteristics</i>	69
<i>Reading environment and CAD paradigm</i>	69
<i>Statistical analysis</i>	70
Results	72
<i>Per-patient analysis</i>	72
<i>Per-polyp analysis</i>	74
<i>Second-read CAD</i>	75
<i>Other analyses</i>	77
Discussion.....	77
Chapter 5: Detection of extracolonic pathology by CT colonography: A discrete choice experiment of perceived benefits versus harms.....	81
Abstract	81
Introduction.....	82
Methods.....	83
<i>Recruitment</i>	83
<i>Attributes</i>	83
<i>Experiment format</i>	86
<i>Pilot testing</i>	87
<i>Statistical analysis</i>	87
Results	88
<i>Non-traders</i>	89
<i>Radiological testing discrete choice experiment</i>	90
<i>Invasive testing discrete choice experiment</i>	92
Discussion.....	93
Chapter 6: Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: Systematic review with a focus on quality of data reporting.....	97
Abstract	97
Introduction.....	98
Methods.....	98
<i>Search strategy, inclusion and exclusion criteria</i>	98
<i>Data extraction</i>	99
<i>Analysis</i>	100
Results	100
<i>Study characteristics</i>	102
<i>Study design</i>	103
<i>Methods of reporting study outcomes</i>	104
<i>Model assumptions</i>	105
<i>Model fitting</i>	106
<i>Presentation of results</i>	107
Discussion.....	109

Chapter 7: Guidelines for the reporting of multi-reader multi-case imaging studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy.....	112
Abstract	112
Introduction.....	113
Methods.....	113
Results: The guidelines.	114
<i>Item #1: The imaging test</i>	114
<i>Item #2: Patients (cases) used for the study</i>	114
<i>Item #3 Study readers</i>	114
<i>Item #4 Readers' diagnostic task(s)</i>	115
<i>Item #5 Confidence rating scales</i>	115
<i>Item #6 Study design</i>	115
<i>Item #7 Unit of analysis</i>	116
<i>Item #8 Distribution of rating scores and data fitting</i>	116
<i>Item #9 Data analysis</i>	116
<i>Item #10 Data presentation</i>	117
<i>Item #11 Study interpretation</i>	117
Discussion.....	118
Chapter 8: Thesis summary and recommendations for future research .	121
Contribution and acknowledgements	128
Appendix 1: Indexed, peer-reviewed journal articles arising from this thesis.	131
References.....	132

List of Tables

Table 1: Data from a hypothetical study of diagnosis of colorectal cancer by CT colonography showing 50 patients with and 50 patients without cancer by reference standard, and the diagnostic rating score attributed by a radiologist observer who uses the following scale to rate their belief that cancer is present or absent: 1 – Definitely normal; 2 – Probably normal; 3 – Equivocal; 4 – Probably has cancer; 5 – Definitely has cancer.	15
Table 2: Example data for the situation where CT more sensitive than the test in Table 1 but which retains identical specificity.	17
Table 3: Example data for the situation where CT is more specific than the test in Table 1 but which retains identical specificity.	18
Table 4: Example data for the situation where CT is as sensitive as the test in table1 but where these gains are exactly offset by diminished specificity.	19
Table 5: Demographic and data acquisition details for the 112 patient cases interpreted.	41
Table 6: Per-patient sensitivity, specificity and CAD net effect for detection by CT colonography of patients with polyps of all sizes when readers were unassisted compared to second-read CAD and concurrent CAD. Data are shown for all 16 readers.	45
Table 7: CAD net benefit, per-patient sensitivity and specificity for detection by CT colonography of patients with polyps of all sizes and patients with polyps ≥ 6 mm when readers were unassisted compared to second-read CAD and concurrent CAD. Data are averaged across 16 readers.	47
Table 8: Per-polyp sensitivity for detection by CT colonography of polyps of all sizes, polyps ≥ 6 mm, and polyps ≤ 5 mm when readers were unassisted compared to second-read CAD and concurrent CAD. Data are averaged across 16 readers. Confidence intervals are 98.3% to account for multiple testing.	48
Table 9: Overview of attributes and levels presented in cancer (1A) and polyp (1B) discrete choice experiments.	55
Table 10: Demographic characteristics and household annual income of patient and professional participants, including non-traders (% have been rounded).	60
Table 11: Tipping points and relative weighting for cancer and polyp detection scenarios calculated for patients, professionals, and all participants combined (FP = false-positive diagnosis, TP = true-positive diagnosis).	61
Table 12: Patient and professionals' willingness to pay (WTP) for a 0.10 (10%) increase in test sensitivity without any reduction in specificity, for detection of cancer or clinically significant polyps.	63
Table 13: Per-patient results for net benefit of CAD assistance when used in concurrent mode for interpretation of CT colonography by inexperienced and experienced readers. For all comparisons differences are calculated as performance with CAD assistance minus performance when unassisted. All data are percentages. Using CAD changed the proportion of readers correctly identifying cases with polyps in 83% of cases for both inexperienced and	

experienced readers; detection of patients with polyps increased in 70% and 57% of cases over 10 and 16 readers respectively.73

Table 14: Per-polyp sensitivity for CAD assistance when used in concurrent mode for interpretation of CT colonography by inexperienced and experienced readers. For all comparisons differences are calculated as performance with CAD assistance minus performance when unassisted. All data are percentages.....	75
Table 15: Effect of CAD assistance when used in second-read mode for interpretation of CT colonography by experienced readers. For all comparisons differences are calculated as performance with CAD assistance minus performance when unassisted. All data are percentages.	76
Table 16: Attributes and levels presented in the “radiological testing” (1A) and “invasive testing” (1B) experiments.....	85
Table 17: Demographic characteristics of patient and professional participants.	89
Table 18: Tipping points and number of FP deemed acceptable in each scenario, for patients, professionals, and the two groups combined.....	90
Table 19: Citations for the 49 papers (contributing 51 studies) included in the systematic review. Details are also provided for the 15 articles excluded from the systematic review after reading the full-text, along with primary reasons for their exclusion (multiple reasons for exclusion were possible).	101

List of Figures

- Figure 1: ROC plot of the data shown in Table 1. The diagnostic thresholds from which the curve is built are labelled on the curve; for example, “Threshold 4” indicates the sensitivity and corresponding false positive rate at the diagnostic threshold of “Probably has cancer” (Threshold explanations are described in the legend for Table 1). The empiric ROC AUC is 0.828 (all ROC curves drawn using Eng J. ROC analysis: web-based calculator for ROC curves. Baltimore: Johns Hopkins University 2006. Available from: <http://www.jrocfite.org>).16
- Figure 2: ROC plot of data from Table 2. The empiric AUC is 0.891.18
- Figure 3: The Roc curve for these data is shown below. The empiric ROC AUC is 0.891. The curve has moved to the left compared to Figure 1.19
- Figure 4: ROC plot for the data presented in table 4. The empiric ROC AUC is 0.50.20
- Figure 5: Plot of data from Figure 1. The x-axis represents confidence scores and the y-axis their frequency. Patients with cancer are represented by the dashed line and patients without cancer by the solid line. The two distributions cross at a confidence score of 3, indicating that this is the diagnostic threshold that best separates patients with and without cancer. Moving this boundary to the right or left of the graph is akin to raising or lowering the diagnostic threshold respectively. If CT were perfect at discriminating between patient groups, the two distributions would not cross. The less good a test at discriminating between diseased and non-diseased patients, the more the two distributions will overlap.21
- Figure 6: Histogram of confidence ratings ascribed by 10 radiologists in a prior study of CT colonography(17). The dark brown bars represent ratings for 107 patients (of whom 60 had colon polyps) when using computer-assisted detection (CAD) whereas the light brown bars represent ratings when unassisted. The distribution is bimodal: The highest peak occurs for patients who received zero scores both with and without CAD. There is a second broader, more continuous distribution for patients, with most scores being 50 or more and a peak at approximately 70.28
- Figure 7: Data extrapolation: ROC plots each for an individual reader using CT colonography without CAD. Green dots indicate real data points underlying curve fitting. ROC curve are shown extrapolated from these data using LabMRMC (red dotted line) and Proproc software (blue solid line). Plots for five individual readers are shown, labeled 1 to 5.30
- Figure 8A: Data extrapolation: ROC plots of individual readers using CT colonography with- and without CAD. These data are reproduced from an analysis of the study described in Chapter 2, sponsored by Medicsight plc, which was used to obtain FDA approval successfully. The primary outcome measure is diagnosis of polyps of any diameter on a per-segment basis. It can be seen that the data points cluster in the bottom left-hand area of the ROC plot space, and that the AUC is dominated by extrapolated data curves. Two types of extrapolation are shown. Figure 8A shows the data for all 17 readers. Figure 8B shows the data for a single reader (R14) markedly emphasising both the clustering and the fact that the method used for extrapolation markedly influences the ultimate AUC.31
- Figure 9: Example question from the cancer detection scenario. Each tally mark represents one of 5000 potential outcomes for a patient undergoing screening: True positive (blue), false negative (yellow), true negative (white), or false positive (red). Participants were informed that if they were to undertake the test in question, their odds of receiving any of the outcomes were represented by the chance of

picking any of these tally-marks at random “like roulette”. Data are also represented numerically using both relative and absolute percentages. This scenario corresponds to the ‘tipping point’ for patients and professional respondents: On average, participants favoured the enhanced test (test B) in view of its additional sensitivity up to, but not beyond, this level of additional false positives.....58

Figure 10: Cumulative graph of participants’ tipping points for trading absolute numbers of true-positive versus false-positive diagnoses.59

Figure 11: Example question from the “invasive testing” experiment. A hypothetical screening population of 600 individuals were presented and participants invited to choose between a test generating a variable rate of false-positives (pink) but a 1 in 600 chance of finding an early stage extracolonic malignancy (green) or a test that generated no false-positives but no chance of finding an extracolonic malignancy (yellow). Participants were informed that their chance of receiving any particular result could not be predicted in advance and was essentially random. Relative and absolute percentages were presented. This image corresponds to the median “tipping point” for patients and professionals combined: On average, unrestricted CTC (Test A on the Figure) was preferred to this level of false-positive invasive tests, but not beyond.87

Figure 12: Cumulative plot of tipping-points expressed as absolute numbers of additional unnecessary tests for patients (A) and professionals (B) in the “radiological testing” experiment. Each grey dot shows an individual’s tipping-point. Large red square shows the median value, corresponding to “an average participant”. Blue squares show 25 and 75 percentage points.....91

Figure 13: Cumulative plot of tipping-points expressed as numbers of additional unnecessary tests for (A) patients and (B) professionals in the “invasive testing” experiment: Color-codes are the same as in figure 12.....92

Figure 14: Bar chart showing data extracted by the systematic review relating to study readers, design, and the confidence scales used to build ROC curves.....103

Figure 15: Bar chart showing data extracted by the systematic review relating to the presentation of individual study results.108

List of abbreviations/glossary

2D/3D	Two dimensional/three dimensional
AC	Ascending colon
AUC	Area under the curve
BCSP	Bowel cancer screening programme
BI-RADS	Breast imaging and reporting data system
CI	Confidence interval
CRC	Colorectal cancer
CT	Computed tomography
CTC	Computed tomographic colonography
DBM	Dorfman Berbaum Metz
DC	Descending colon
DCE	Discrete choice experiment
DOR	Diagnostic odds ratio
FDA	Food and Drug Administration
FN	False negative
FP	False positive
GBP	Great British pound
IQR	Inter-quartile range
IRB	Institutional review board
NPV	Negative predictive value
pAUC	Partial area under the curve
PPV	Positive predictive value
REC	Research ethics committee
ROC	Receiver operator characteristic
RS	Reference standard
SC	Sigmoid colon
SD	Standard deviation
TC	Transverse colon
TN	True negative
TP	True positive
WC	Weighted comparison
WTP	Willingness to pay

Chapter 1: Disadvantages of using the area under the receiver operator characteristic curve (ROC AUC) to assess imaging tests: A discussion & proposal for an alternative approach.

This Chapter has been published as:

Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology* 2015;25:932-9.

Winner, Gold Medal for best GI paper published in European Radiology.

Abstract

Aim To describe the disadvantages of the area under the receiver operating characteristic curve (ROC AUC) to measure diagnostic test performance. To propose an alternative based on net benefit.

Methods Narrative review supplemented by data from a study of computer-assisted detection for CT colonography.

Results We identified problems with ROC AUC: Confidence scoring by readers was highly non-normal and score distribution bimodal. Consequently, ROC curves were highly extrapolated with AUC mostly dependent on areas without patient data. AUC depended on the method used for curve-fitting. ROC AUC does not account for prevalence or different misclassification costs arising from false-negative and false-positive diagnoses. Change in ROC AUC has little direct clinical meaning for clinicians. An alternative analysis based on net benefit is proposed, based on the change in sensitivity and specificity at clinically relevant thresholds. Net benefit incorporates estimates of prevalence and misclassification costs, and is clinically interpretable since it reflects changes in correct and incorrect diagnoses when a new diagnostic test is introduced.

Conclusions ROC AUC is most useful in the early stages of test assessment whereas methods based on net benefit are more useful to assess radiological tests where the clinical context is known. Net benefit is more useful for assessing clinical impact.

Introduction

Most of our working week as radiologists is concerned with diagnostic tests: we interpret medical images with the aim of detecting disease. The choice between one test or another (e.g. CT or MRI?) will depend on a variety of factors including availability and cost. However, for the most part, the choice is influenced by how effectively the test and its interpretation by a radiologist detects or excludes the disease being considered by the referring clinician. In this sense, the radiologist is acting as a “classifier”(1), whose task is to sort patients into disease-positive and disease-negative cases. Sensitivity, the ability of a test to identify patients with disease, is a measure of diagnostic test accuracy that is very familiar to radiologists. At the same time we must also consider specificity, the ability of a test to identify patients who do not have disease. Sensitivity and specificity are inextricably linked and usually move in different directions. Most obviously, if we designated every image we interpreted as positive for disease, then we would have 100% sensitivity but 0% specificity, which means all normal patients would be subjected to unnecessary further investigation and possibly treatment, which would be inconvenient, illogical, precipitate anxiety, and be extremely costly! Conversely, if we called every image negative, then specificity would be perfect but sensitivity would be 0% and we would never diagnose any pathology. Although sensitivity and specificity are “two sides of the same coin” and should virtually always be considered together, it is sometimes difficult to do so, especially when comparing between different tests. For example, if one test has high sensitivity and another high specificity, which is best? Combining both sensitivity and specificity into a single measure of diagnostic accuracy facilitates comparisons between different tests. For radiologists, the most familiar combined measure is the area under the receiver-operator-characteristic curve, usually abbreviated to “ROC AUC” (2).

The ROC plot

A ROC curve is a plot of the true-positive rate (y-axis) of a test against the corresponding false-positive rate (x-axis), i.e. sensitivity against 1-specificity. A fundamental building block of the ROC curve is the concept of test performance at different “diagnostic thresholds”. Some diagnostic tests give a simple “yes/no” answer. One such example might be urinalysis for glucose; glucose is either present or absent. However, analysis of plasma glucose is different – there is a normal range of values for fasting blood glucose, 3.9 to 5.8 mmol/L. Diagnosis of diabetes becomes increasingly more likely as fasting blood glucose rises above this but, for example, not everyone with a level of 6.3 mmol/L will have diabetes; i.e. there is a “grey” area where diagnosis

is uncertain. It follows that the proportion of patients who truly have diabetes and those who do not will change with the threshold used for a positive diagnosis.

Interpretation of radiological images is one diagnostic scenario where diagnostic thresholds vary, and where the thresholds are made and used by a human observer. Take diagnosis of colon cancer by CT colonography as an example. Imagine a study where a radiologist is confronted by scans from 100 patients, 50 of whom have colon cancer (i.e. a prevalence of abnormality of 50%). We would certainly expect a competent radiologist to get the diagnosis right more often than not but there will be occasions when this does not happen. Subtle cancers may not be seen or perhaps they are seen but misinterpreted as colonic spasm. Occasionally, even “obvious” tumours are missed and some cancers will be too small to be resolved adequately by the scan. As radiologists, we are all too familiar with the concept of uncertainty in our diagnosis. A measure of this uncertainty can be captured by asking the radiologist to attribute a confidence score to his/her diagnosis for each individual case. The following categories might be appropriate for colon cancer: “definitely normal”, “probably normal”, “equivocal”, “probably cancer”, “definitely cancer”. The BI-RADS score for mammography is a well-known example of this type of rating used in daily diagnostic practice; “negative”, “benign”, “probably benign”, “suspicious”, “highly suggestive of malignancy”(3). In some situations, a rating system from 0 (definitely no disease) to 100 (definitely disease) is used in an attempt to elicit finer rating detail(4). Such confidence scores are an amalgam of whether disease is or is not resolved by the test (technical adequacy) and whether the radiologist has or has not seen the abnormality subsequently and then interpreted it correctly (diagnostic accuracy).

Of course, in reality, each patient either has cancer or not, so the issue arises of how to extract a binary diagnosis from such rating scales. For example, to determine how effective CT colonography is for detection of cancer, we might apply both CT and an independent reference test to patients with and without the disease, thus providing a CT diagnosis and a “ground-truth” or “Gold-standard” diagnosis for each patient. We could then apply a diagnostic threshold (cut-point) to the scale of diagnostic certainty, i.e. a point on the rating scale at which (and above) the patient is believed to have cancer and below which they do not. This result is then compared with the reference standard diagnosis, thereby calculating the number of correct (true-positive) and incorrect (false-negative) diagnoses for patients with and without disease at each threshold. For example, if we apply a threshold at “definitely cancer” then the large majority of patients labeled as such by CT will prove to have cancer (if CT is accurate). However, because of the uncertainties noted above, there will likely be many patients

labeled at the “probably cancer” threshold and below who also have the disease but who will be missed with a diagnostic threshold set at “definitely cancer”. Dropping the diagnostic threshold to “probably cancer” will therefore increase the proportion of positive patients identified; i.e. sensitivity increases. However, at the same time more patients without disease will be erroneously labeled as positive; i.e. the false-positive fraction will increase (decreased specificity). Plotting the proportion of true-positive against false-positive patients at each diagnostic threshold will build the ROC curve.

Table 1 below shows a data table for our hypothetical study of CT colonography:

Table 1: Data from a hypothetical study of diagnosis of colorectal cancer by CT colonography showing 50 patients with and 50 patients without cancer by reference standard, and the diagnostic rating score attributed by a radiologist observer who uses the following scale to rate their belief that cancer is present or absent: 1 – Definitely normal; 2 – Probably normal; 3 – Equivocal; 4 – Probably has cancer; 5 – Definitely has cancer.

Reference diagnosis	Rating score				
	1	2	3	4	5
Cancer	2	5	10	18	15
No cancer	15	18	10	5	2

If we apply the diagnostic threshold for cancer at 5 (“definitely cancer”) then 15 patients with cancer are diagnosed correctly (true-positive), as are 48 without (true-negative). However 35 patients with cancer have been “missed” (false-negatives) and 2 patients without cancer diagnosed with disease (false-positives); i.e. sensitivity = 0.30 and specificity = 0.96.

Dropping the diagnostic threshold to include those patients labeled “probably cancer” results in 33 true-positives at the cost of 7 false-positives. At this threshold sensitivity = 0.66 and specificity = 0.86.

A diagnostic threshold dropped to include “equivocal” patients means 43 true-positives and 17 false-positives; sensitivity = 0.86, specificity = 0.66.

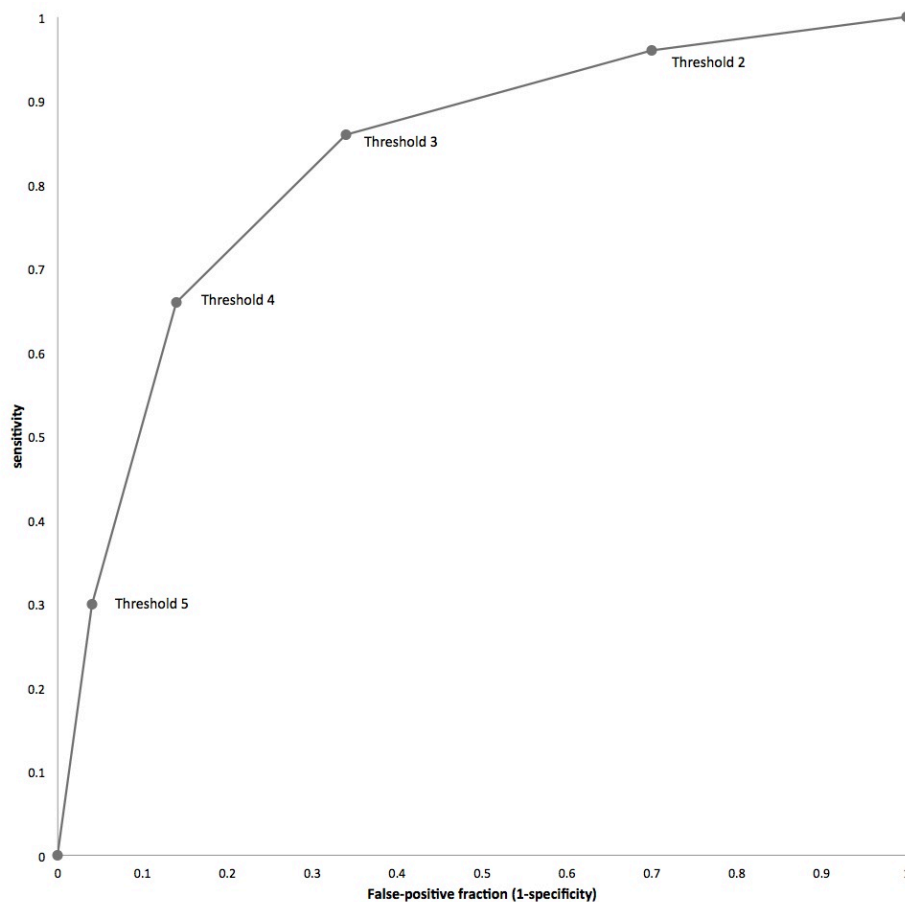
A diagnostic threshold dropped to include “probably not cancer” patients means 48 true-positives and 35 false-positives; sensitivity = 0.96, specificity = 0.30.

Of course, dropping the diagnostic threshold to include “definitely not cancer” (a small proportion of whom may actually have cancer) will mean that all possible thresholds

entail a positive diagnosis; i.e. there are 50 true-positives but also 50 false-positives; sensitivity = 1.0, specificity = 0.0.

The figure below is a graph that plots these sensitivity/specificity pairs for each diagnostic threshold, i.e. a ROC plot.

Figure 1: ROC plot of the data shown in Table 1. The diagnostic thresholds from which the curve is built are labelled on the curve; for example, “Threshold 4” indicates the sensitivity and corresponding false positive rate at the diagnostic threshold of “Probably has cancer” (Threshold explanations are described in the legend for Table 1). The empiric ROC AUC is 0.828 (all ROC curves drawn using Eng J. ROC analysis: web-based calculator for ROC curves. Baltimore: Johns Hopkins University 2006. Available from: <http://www.jrocfite.org>).



ROC AUC

As explained in the section above, the ROC plot therefore describes the diagnostic performance of a test over the whole range of possible thresholds. Performance is usually summarised across all these thresholds using ROC AUC, corresponding to the area under the curve. The maths are complex but, in simple terms, they calculate how

likely it is that the test will rank two patients, one with disease and one without, in the correct order of their likelihood of having disease, across all possible thresholds(1, 5). A more intuitive way to express AUC is the chance that a randomly selected patient with disease will be ranked above a randomly selected patient without disease(2). Thus, the greater the AUC, the better the test at achieving this separation. A perfect test would have 100% sensitivity with a false-positive fraction of 0. This point lies at the extreme top left hand corner of the ROC plot space and a test this accurate at all thresholds would have an AUC of 1.0. If the ROC curve is a straight line connecting the extreme bottom-left (sensitivity, FPR: 0,0) and top-right (1,1) corners of the ROC plot space (sometimes called the “chance diagonal”) this describes a test with no discriminatory ability; AUC = 0.5 (equivalent to picking a test result by tossing a coin). The AUC for the data in Figure 1 is 0.828, indicating a test that is better than chance at discriminating patients with disease.

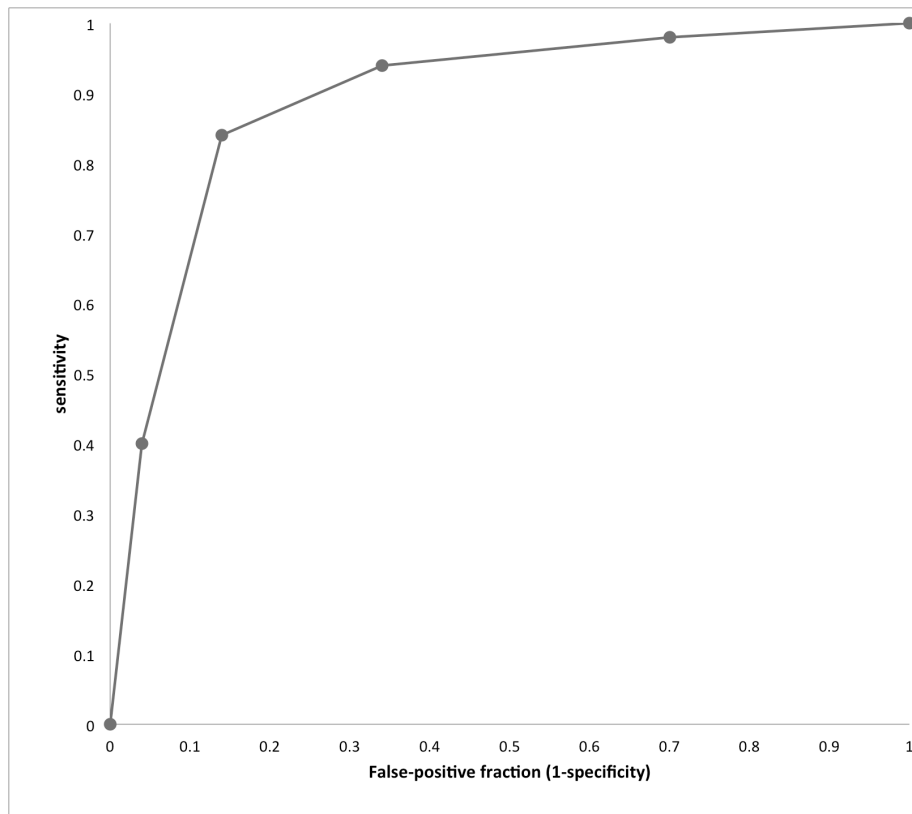
In table 2 below, test sensitivity has been inflated while specificity remains the same as in Table 1.

Table 2: Example data for the situation where CT more sensitive than the test in Table. 1 but which retains identical specificity.

	Rating score				
Reference diagnosis	1	2	3	4	5
Cancer	1	2	5	22	20
No cancer	15	18	10	5	2

The ROC plot for these data is shown in Figure 2 below; the curve has moved upwards compared to figure 1 and ROC AUC rises to 0.891.

Figure 2: ROC plot of data from Table 2. The empiric AUC is 0.891.

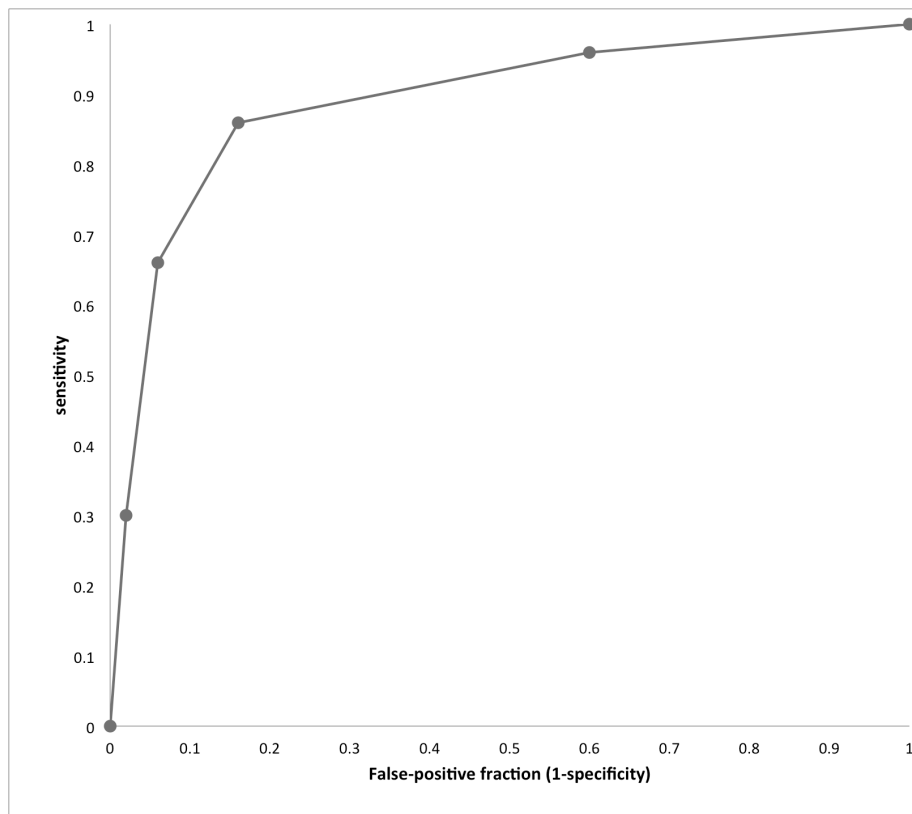


Conversely, in table 3 below, test specificity has been inflated by the same amount but sensitivity remains the same as in table 1; The AUC again rises to 0.891 again but the shape of the curve is different to Figure 2, having moved towards the y-axis, reflecting the fact that specificity rather than sensitivity has improved.

Table 3: Example data for the situation where CT is more specific than the test in Table. 1 but which retains identical specificity.

	Rating score				
Reference diagnosis	1	2	3	4	5
Cancer	2	5	10	18	15
No cancer	20	22	5	2	1

Figure 3: The Roc curve for these data is shown below. The empiric ROC AUC is 0.891. The curve has moved to the left compared to Figure 1.



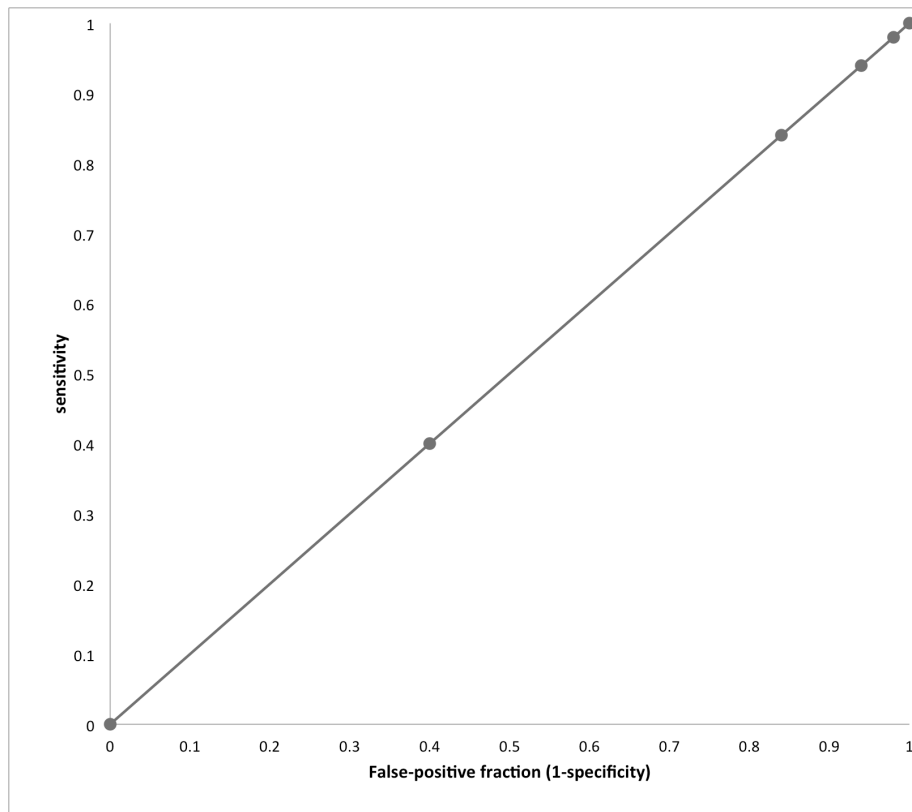
In table 4 below, sensitivity has been raised to the level seen in table 2 but these gains are exactly mirrored by loss of specificity.

Table 4: Example data for the situation where CT is as sensitive as the test in table1 but where these gains are exactly offset by diminished specificity.

	Rating score				
Reference diagnosis	1	2	3	4	5
Cancer	1	2	5	22	20
No cancer	1	2	5	22	20

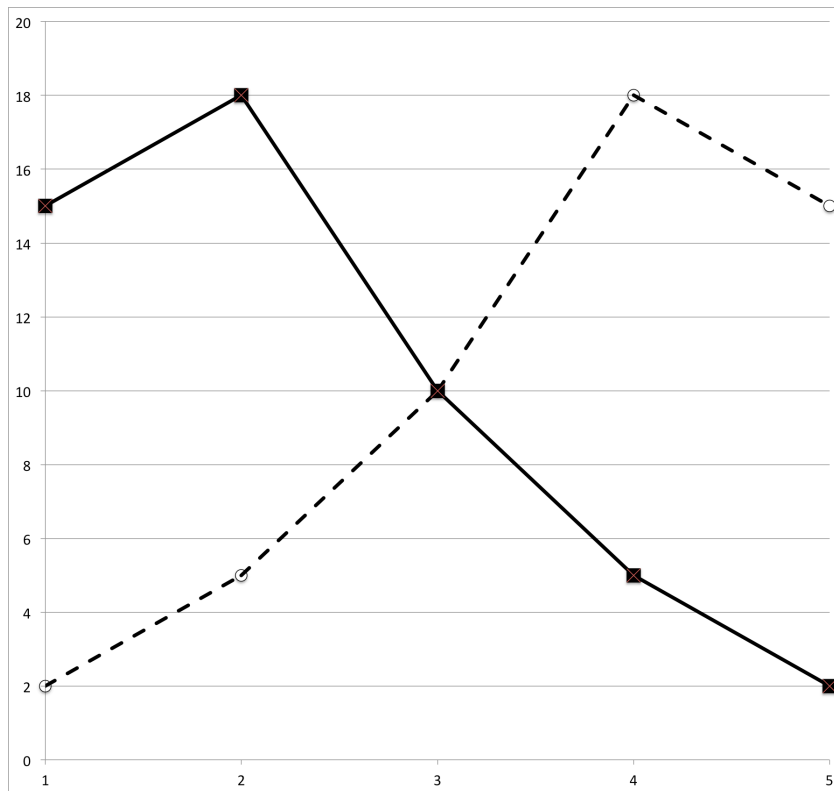
Figure 4 below shows the ROC plot arising from these data. This time the curve is actually a straight line and the empiric AUC is 0.5, suggesting the test has no discriminatory value.

Figure 4: ROC plot for the data presented in table 4. The empiric ROC AUC is 0.50.



When faced by an individual patient and their test result in clinical practice, we must settle on a threshold that denotes a positive test. The ROC plot is actually a composite of two distributions of diagnostic certainty, i.e. for patients both with and without disease. Figure 5 below shows the distribution of confidence scores for diseased and normal patients plotted separately (data taken from Figure 1). The two distributions cross at a confidence score of 3. The threshold that best separates patients with and without disease can be estimated by looking at the ROC plot (Figure 1) and identifying the threshold on the curve that lies nearest to the top left-hand corner of the plot space. This “best operating point” can be moved to maximise sensitivity at the expense of specificity, and vice-versa, but is conventionally placed where the test optimally separates diseased from non-diseased patients.

Figure 5: Plot of data from Figure 1. The x-axis represents confidence scores and the y-axis their frequency. Patients with cancer are represented by the dashed line and patients without cancer by the solid line. The two distributions cross at a confidence score of 3, indicating that this is the diagnostic threshold that best separates patients with and without cancer. Moving this boundary to the right or left of the graph is akin to raising or lowering the diagnostic threshold respectively. If CT were perfect at discriminating between patient groups, the two distributions would not cross. The less good a test at discriminating between diseased and non-diseased patients, the more the two distributions will overlap.



MRMC ROC studies

When performing studies of radiological tests, it is usually desirable to have as many readers as possible who each read multiple cases. Such studies are known as “multi-reader, multi-case” studies (MRMC)(6, 7). The MRMC design is popular because once a radiologist has viewed 20 cases there is less information to be gained by asking him to view a further 20 than by asking a different radiologist to view the same 20. This procedure enhances the generalisability of study results (i.e. the extent to which study results reflect the “real-world” situation) and having multiple readers interpret multiple cases enhances statistical power. Because multiple radiologists view the same cases, “clustering” occurs. For example, small lesions are generally seen less frequently than larger lesions, i.e. reader observations are clustered within cases. Similarly, more experienced readers are likely to perform better across a series of cases than less experienced readers, i.e. results are correlated within readers. For MRMC studies complex bootstrap resampling and multilevel modelling is needed to account for

clustering, linking results from the same observers and cases, so that 95% confidence intervals are not spuriously narrow (8).

If 100 different radiologists were asked to rate the CT scans from our hypothetical study, then we would likely get 100 different ROC curves. Several factors underpin this: Some radiologists will be more experienced than others, some will have more intrinsic competence, and all will have differing internal thresholds (also known as implicit thresholds) for calling disease – we are all aware of colleagues who have reputations as “over-callers” or “under-callers”. The net result is that because diagnosis for the *same* image may differ depending upon who is interpreting it, there is no single ROC curve for any imaging modality that incorporates human interpretation. The curve is therefore an amalgam of both the intrinsic technical ability of the test to resolve disease and the ability of the radiologist to detect it. On average, we would expect the more able and/or experienced radiologists to have a greater AUC than those who are less able and/or experienced.

What are the advantages of ROC AUC?

ROC AUC is a single metric and so it is easy to compare between different tests without having to conceptually “juggle” with sensitivity and specificity values that usually move in different directions. Proponents state that by measuring and averaging across all possible diagnostic thresholds the total “inherent” performance of a test is determined. This is achieved irrespective of an individual threshold used for diagnosis, and ROC is constant across the prevalence of abnormality in the dataset(4).

Some proponents argue that it is most sensible to have a method that simply compares diagnostic accuracy across all possible diagnostic thresholds, and one that is independent of prevalence. For example, Zweig and Campbell wish to look at the performance of a laboratory test in isolation across all scales of measurement since they argue this provides the best assessment of intrinsic test accuracy(9). They go on to argue that different misclassification costs and prevalence (both discussed in the sections below) are important but should not be incorporated into ROC AUC since doing so detracts from clarity around the intrinsic accuracy of the test itself(9). The fact that no a priori knowledge is needed regarding the diagnostic thresholds at which a test would likely be used in clinical practice could be regarded as an advantage of ROC AUC. This is especially the case where such clinical diagnostic thresholds are simply not established.

Also, the ROC curve accounts for different thresholds existing between readers, known as “response criteria”. The assumption is that each individual technology has its own intrinsic ROC curve and that different readers will lie at different points on this same curve. The visual nature of data means that it is easy to compare different curves, especially they are displayed on the same scale.

What are the disadvantages of ROC AUC?

Clinical comprehension and relevance

ROC AUC combines across sensitivity and specificity values yielding a single metric that facilitates comparison between diagnostic tests. However, sensitivity and specificity are familiar to clinicians and easy for them to understand. In contrast, ROC AUC has little meaning for clinicians (especially non-radiologists), patients, or health-care providers. While it can be appreciated that a test whose AUC is 0.9 is “better” than one whose AUC is 0.8, how does this translate into gains in terms of patients diagnosed with and without disease?

It is well-established that the capabilities of a diagnostic test are best understood by patients and clinicians when presented in terms of gains and losses to individuals(10), and are most meaningful when these gains/losses are presented at diagnostic thresholds relevant to clinical practice, along with the appropriate clinical context and representative disease prevalence. ROC AUC is not directly reconcilable to individual patients and so lacks clinical interpretability, and it is obscure what change in AUC is clinically important. The ROC curve itself suggests an operating point that best separates diseased from non-diseased patients, and this can be explained in terms of gains and losses to individual patients, but ROC AUC plays no role once an individual threshold has been identified. Moreover, clinicians are uninterested in diagnostic performance across all possible thresholds. Rather, they are only interested in those thresholds that are clinically relevant to their decision-making. How a test performs at near 100% sensitivity or specificity is clinically absurd in most cases. Interpreting the entire AUC gives clinically illogical or irrelevant thresholds the same weight as those that are reasonable and important. Moreover, two different tests may have the same AUC overall but have very different performance characteristics at the diagnostic threshold most appropriate to clinical practice. The use of “partial” AUC (PAUC), which considers the area under segments of the curve near to important clinical thresholds(11) is proposed as a means to focus performance onto the most relevant

thresholds. However, choosing a relevant range of thresholds may often prove more problematic than simply picking a single clinically relevant threshold.

Are sensitivity and specificity equally important?

On average, the calculation of ROC AUC treats sensitivity and specificity as equally important. Notably, the conventional operating point occurs where the test optimally separates diseased and non-diseased patients (Figure 5). But what if sensitivity and specificity are not equally important? On reflection, it is obvious that the clinical consequences of gains/losses in sensitivity are not equivalent to those for specificity. Taking our CT colonography example, ROC AUC on average regards a true-positive diagnosis of cancer as equally beneficial as a true-negative diagnosis, and a false-negative diagnosis as equally detrimental as a false-positive. It follows that, when using ROC AUC as a performance measure, a false-positive diagnosis of cancer on CT colonography is valued on average as equivalent to a missed diagnosis of cancer. The consequence for the first patient is an unnecessary colonoscopy that will ultimately prove to be normal whereas the consequence for the second is delayed treatment or even death. No-one would argue that these two situations are clinically equivalent yet they are treated as equivalent when ROC AUC is used as a measure of diagnostic performance. Supporting this, a mammographic study found that women were prepared to tolerate up to 500 false-positive diagnoses in order to achieve a single additional true-positive diagnosis of cancer(12). Indeed, Lusted's original statistical decision analysis for diagnosis of active tuberculosis rated the cost of one false negative as equivalent to 400 false positive diagnoses(13).

Take the example given in Figure 4: Here sensitivity at a diagnostic threshold of a confidence score of 3 or more was raised to 94% for patients with cancer but was offset by equally diminished specificity with the results that the AUC is 0.5. Would patients and doctors agree the test is of no value at all? While the author would not argue that sensitivity should be enhanced at the cost of such enormously diminished specificity, there are many clinical scenarios where the modest gains in sensitivity achieved by a new test would be exchanged willingly for a proportionally larger drop in specificity. Analyses that do not account for this risk finding the new test of no value when patients and their doctors might consider otherwise. ROC AUC does not allow for any clinical differentials in "misclassification cost" for patients with and without disease and, furthermore, such costs are not clinically equivalent across the whole range of possible diagnostic thresholds(1). While in practice it may be difficult to quantify precisely the exact weights of false-positive versus false-negative misclassifications

(because these may change over time and vary according to external cultural and economic conditions) it is likely that something will be known about this ratio(14).

While it is widely believed that ROC AUC weighs changes in sensitivity and specificity equally (i.e. any gain in test sensitivity would be exactly negated by an equivalent drop in specificity), this is only true for one point on the curve, where the gradient equals one. Other points on the curve assign different weights for sensitivity and specificity, which are determined by the shape of the curve, without reference to any available clinically meaningful information. It must be clarified that where the gradient equals 1 then while an increase in sensitivity and a similar decrease in specificity do carry equal weight, that does not mean that a false-negative and a false-positive are considered equally undesirable – this only applies when the prevalence of abnormality is 0.50 (see section on prevalence below). In most clinical situations, prevalence is <0.10 , with the result that there are more than 9 extra false-positive events for each false-negative avoided. It is necessary to correct for prevalence odds in order to translate ROC increments into consequences for individual patients.

Test specificity is low in situations where most of the AUC is derived from the right hand side of the plot. For example, a 5% improvement in sensitivity contributes less to the AUC at values of high specificity, than the same improvement at low specificity. Thus AUC would consider a test that increased sensitivity at low values of specificity superior to a test that increased sensitivity at high values of specificity, which is not clinically meaningful. For example, when screening, better tests increase sensitivity at high values of specificity so that the programme is not overwhelmed by false-positive detections(15).

It should be noted that it is possible to implement misclassification costs into ROC AUC analysis, as described by Zweig and Campbell(9), but the procedure is far from straightforward and, furthermore, does not appear to have filtered into study analyses (see systematic review, Chapter 6).

Confidence scores may be inconsistent and unreliable

Confidence scores are needed to build ROC plots when human observers interpret the test output, and radiology is the prime example. Confidence scores are not necessary where the test result is on a continuous scale such as occurs for laboratory tests, such as the blood glucose example already given – no human interpretation is needed and the value for blood sugar can simply be plotted against the reference diagnosis for the

patient (diabetic or not) across the whole range of values. In this context, ROC AUC can be very valuable to determine the intrinsic accuracy of a test, a fact noted by laboratorians(9). However, to build ROC plots from interpretation of medical image data by radiologists, it is necessary to assign a confidence score that reflects observers' belief that the image is abnormal or not. This fundamental principle underpins the shape of the ROC curve but there is no evidence that scores assigned in this way are consistent and reliable. Scores can vary across radiologists for reasons completely unrelated to diagnostic certainty. For example, a study asking what is meant by "high confidence", found that radiologists gave 10 different interpretations including, "the image quality is good", "the finding is obvious", and "the finding is familiar"(16). Confidence scales must be ordinal, i.e. levels should be ranked in a meaningful order with a constant difference between them. However, radiologists may assign scales nominally. For example, BI-RADS 2 (benign abnormality) does not imply a greater suspicion of cancer than BI-RADS 1 (no abnormality), leading to variability and error(3).

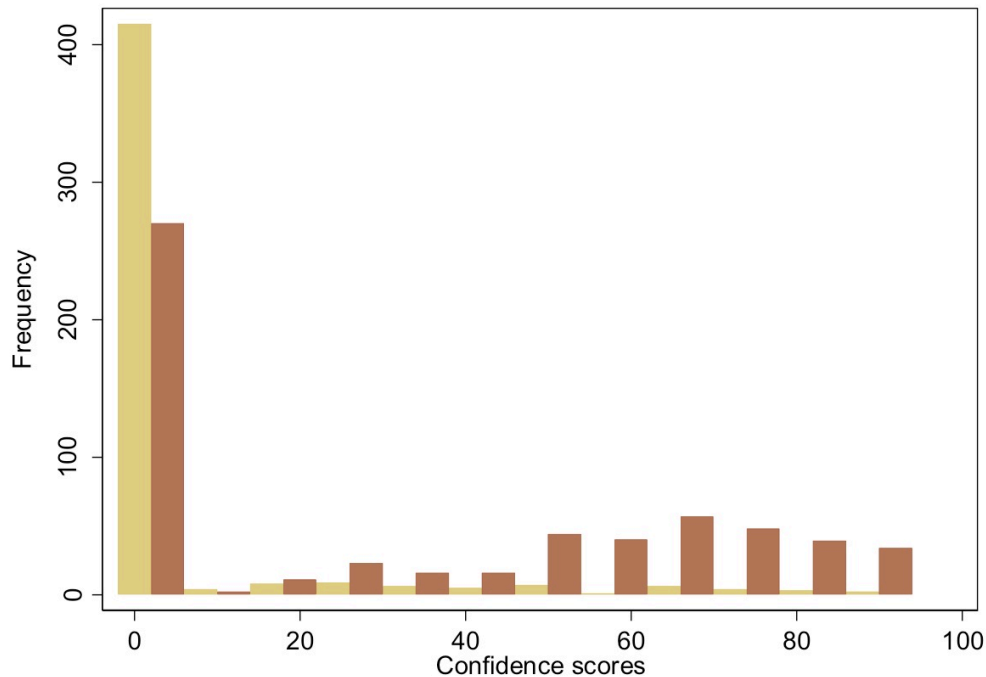
Consistent scoring is perturbed further by the multi-faceted nature of the radiological task. Take pulmonary nodules for example. A confidence score may be assigned to the probability that a nodule is present or absent, and also to whether a nodule (if present) is benign or malignant. A third score may be attached to where the abnormality is located. Thus there are three potential tasks – detection, localisation and characterisation. It is a prerequisite for ROC AUC analysis that confidence scores are distributed normally or can be transformed to a normal distribution. This is particularly difficult to achieve for identification tasks because, having once perceived an abnormality, readers are unlikely to then state they did so with low confidence. For example, the author, wishing to determine the utility of computer-assisted-detection (CAD) for diagnosis of colorectal polyps (17) was obliged by the USA Food and Drug Administration (FDA) to use ROC AUC as the primary outcome measure for the purpose of licensing the software. Adhering to guidance (6, 18), the author asked readers to score for the presence/absence of colorectal polyps at the patient level using a 100-point scale, with 100 reflecting complete confidence that a polyp was present and 0 the certain belief that no polyp was present. At the same time readers were asked to classify patients as normal or abnormal and, if believed abnormal, a confidence score was assigned to this belief. There were 60 patients with polyps and 47 normal patients (17). Having believed they had identified a polyp, confidence scores assigned by readers were influenced strongly by polyp size, with larger polyps attracting higher scores. In contrast, where no polyp was identified patients were assigned no score. Furthermore, by definition observers do not see false-negative

polyps and so were not able to score these. In true-negative patients, confidence scores assigned to a polyp by readers can only apply to false-positive detections. In studies with 50% prevalence of abnormality, half of the patients may attract no per-polyp confidence score. It is possible to impose a score of zero when coding the data but this introduces a second scoring method that is inconsistent with reader scores and imposes two different scoring methods on two different groups of patients. Confidence scores were therefore bi-modal and highly non-normal because, in effect, there were two different distributions, one continuous and one binary (Figure 6). It is often suggested that extensive scales (e.g. 100-points) and encouragement to use the whole range available will broaden distributions (4) but this contradicts clinical practice where binary decisions are appropriate. Gur et al (19) have stated that, "even when observers provide a distribution of confidence ratings, it may be more representative of the subtleness of the depicted abnormality rather than the confidence that the observer actually 'saw' or did not 'see' it."

The binormal distribution is most often used to construct the ROC curve; i.e. it is assumed that confidence scores for both disease positive and negative patients are normally distributed, or can be transformed by a monotonic distribution (20). When publically available MRMC software is used for ROC AUC modelling, this often requires assumptions of normality for confidence scores or their transformations when ROC curve fitting methods are used, especially for the older and most prevalent DBM (Dorfman Berbaum Metz) software; many studies, especially older research, use this software. However, where confidence scores are not normally distributed these software methods are not recommended (21-24). Although Hanley shows that ROC curves can be reasonable under some distributions of non normal data (25), concerns have been raised, particularly in imaging detection studies that measure clinically useful tests with good performance to distinguish well-defined abnormalities. In tests with good performance two factors make estimation of ROC AUC unreliable. Firstly, a "good" test means that true-positive and true-negative diagnoses tend to be confident with the result that readers' scores are often at each end of the confidence scale. Accordingly, the confidence score distributions for normal and abnormal cases have very little overlap (19, 21-23, 26). Secondly, tests with good performance also have few false positives making ROC AUC estimation highly dependent on confidence scores assigned to possibly fewer than 5% or 10% of cases in the study (21). Thus, a wide range of confidence scores would be unexpected in a test that performed well, for example one that was ready for clinical implementation. However, a wide range of scores is necessary to build the ROC curve. When scores are not normally distributed, even if non parametric approaches are used to estimate ROC AUC, this lack of

normality may indicate additional problems with obtaining reliable estimates of ROC AUC (19, 21-23, 26).

Figure 6: Histogram of confidence ratings ascribed by 10 radiologists in a prior study of CT colonography(17). The dark brown bars represent ratings for 107 patients (of whom 60 had colon polyps) when using computer-assisted detection (CAD) whereas the light brown bars represent ratings when unassisted. The distribution is bimodal: The highest peak occurs for patients who received zero scores both with and without CAD. There is a second broader, more continuous distribution for patients, with most scores being 50 or more and a peak at approximately 70.



The issue of true-negative results and how these can confound confidence scores can be examined further. For example, while true-negative can apply to a patient who has no abnormality, it may also apply to a patient who has an abnormality but which is, for example, considered benign rather than malignant. For example, the classic paper by Lewin et al (27) describes 4,945 mammographic screening examinations. Readers used the BI-RADS scale, with zero scores given not only to cases where no abnormality was identified but also to cases with an abnormality that required further analysis at 6-months follow-up (i.e. likely benign). While a decision whether a perceived abnormality is benign or malignant may use a confidence scale such as BI-RADS, when using a scale to indicate whether a lesion is present or not, there are an infinite number of locations where a lesion may be deemed not-present. A better test would be expected to improve confidence scores but there is no opportunity to improve on a score of zero for cases with no abnormality. In contrast, cases with benign lesions do present an opportunity for improved confidence scores, for example where better resolution switches an equivocal finding to benign. Also, changes in the proportion of true-negative classifications due to negative patients vs negative lesions will affect the

distribution of zero scores in the dataset. Because zero scores do not contribute to the shape of the ROC curve, AUC only summarises a subset of study data and excludes a large proportion of patients. In our colonography study (17) only 15% to 47% (varying by reader) of the 107 patients actually contributed to the shape of the ROC curve, and hence the AUC. Harrington states, "Under ROC reporting rules, the radiologist reports confidence levels only for a finding actually seen, or for a finding of normality. But seeing nothing with a given confidence level is not the same, for image quality purposes, as seeing something with that confidence level. As a result, ROC analysis is largely silent (or misleading) on one of the most important aspects of an imaging system's performance - the ability to avoid misses"(16).

It is clear from the data I present that issues with confidence scores apply particularly when attempting to score the presence of an abnormality as opposed to situations where its character should be scored (e.g. benign vs. malignant). ROC paradigms have been developed to tackle this specifically: The "free-response" ROC paradigm (FROC) attempts to tackle the issue of lesion presence/absence at a specific location by dividing the medical image into pre-defined regions, with the radiologist then asked to indicate whether there is or is not a lesion in each individual region. The FROC curve is the plot of the lesion localisation fraction against the non-lesion localisation fraction. Jackknife alternative free response ROC curve (JAFROC) is used for analysis of MRMC data arising from such localisation studies. The issue for FROC and its derivatives is that the larger the number of locations specified, the more complex the reading and subsequent analysis. For example, a chest x-ray or mammogram may be divided into quadrants, but this procedure does not mimic the real-world clinical scenario where it is vitally important to localise an abnormality with greater precision, e.g. within a lung lobe. For our CT colonography example, in reality there are innumerable points on the endoluminal colonic surface where a potential polyp might be found.

Curve extrapolation

Issues around assigning confidence scores consistently have direct consequences on the shape of the ROC curve. In the author's 2006 study of CT colonography (17), because very few false-positive polyps were reported, data points were clustered towards the lower left hand portion (0,0) of the ROC plot space (Figure 7). In order to complete a curve across all possible thresholds, it must be extrapolated beyond the last available data point. In situations such as our prior study, the majority of the AUC is then dominated by a region where there is no data. Furthermore, changing the

statistical method used to extrapolate the curve can have a profound effect on the calculated AUC (8)(Figures 7 and 8). Gur and colleagues have pointed out that, “selection of a specific analysis approach could affect the study conclusion” (28), noting that the problems associated with extrapolation occur, “when observers tend to be more decisive”. Indeed, many algorithms will not fit curves to “degenerate” data at all (i.e. data where there are no false-positive detections). Also, because false-positive diagnoses are infrequent, their scores exert disproportionate influence on curve shape as opposed to the more numerous true-positive scores. Thus the AUC is dominated by a small portion of the observed data. For example, in our prior study the median number of patients with false positive scores in unassisted reads was just 2 of 107 patients (17).

Figure 7: Data extrapolation: ROC plots each for an individual reader using CT colonography without CAD. Green dots indicate real data points underlying curve fitting. ROC curve are shown extrapolated from these data using LabMRMC (red dotted line) and Proproc software (blue solid line). Plots for five individual readers are shown, labeled 1 to 5.

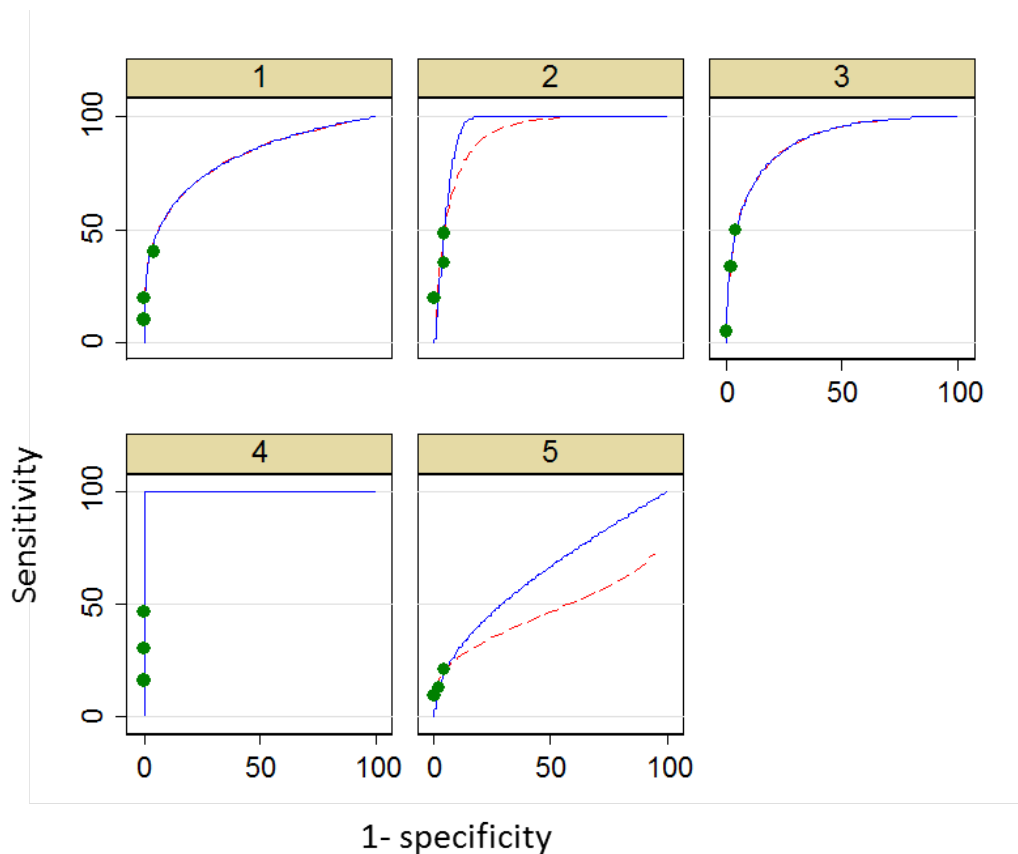


Figure 8

Figure 8A: Data extrapolation: ROC plots of individual readers using CT colonography with- and without CAD. These data are reproduced from an analysis of the study described in Chapter 2, sponsored by Medicsight plc, which was used to obtain FDA approval successfully. The primary outcome measure is diagnosis of polyps of any diameter on a per-segment basis. It can be seen that the data points cluster in the bottom left-hand area of the ROC plot space, and that the AUC is dominated by extrapolated data curves. Two types of extrapolation are shown. Figure 8A shows the data for all 17 readers. Figure 8B shows the data for a single reader (R14) markedly emphasising both the clustering and the fact that the method used for extrapolation markedly influences the ultimate AUC.

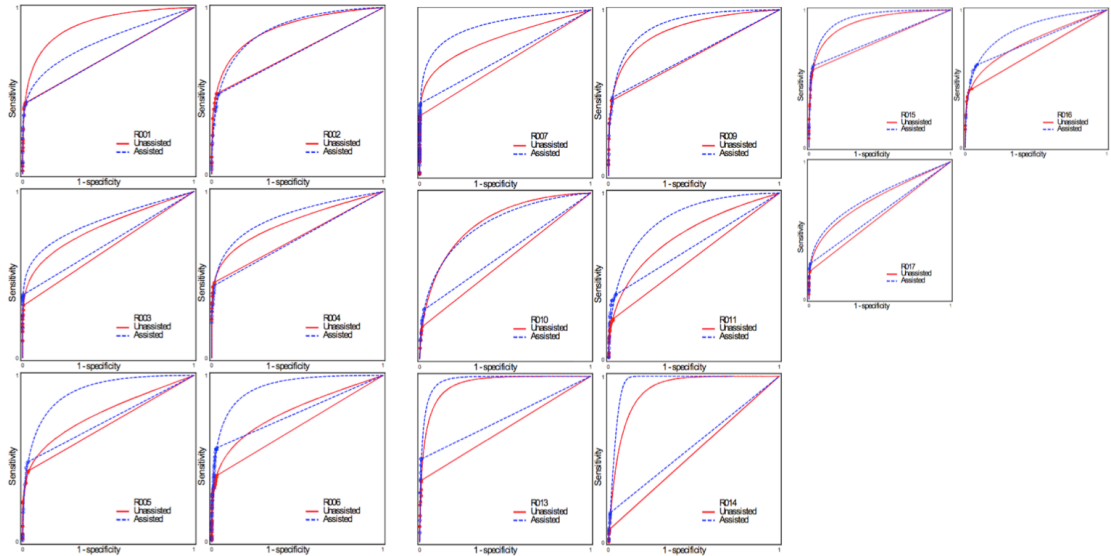
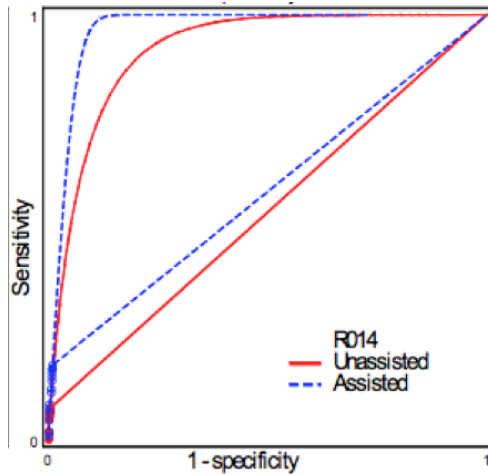


Figure 8B:



Prevalence of abnormality

As noted above, a stated advantage of ROC AUC is that it is independent of prevalence of abnormality. However, test performance usually changes with prevalence so ROC AUC is uninformative regarding diagnostic performance at differing prevalence. AUC is calculated by ranking pairs of patients, one with disease and one without, which implicitly suggests a prevalence of 50%. In reality, the prevalence of disease in the study dataset is ignored and AUC is the same regardless. While sensitivity and specificity are independent of prevalence, the absolute numbers of patients with and without disease will change with prevalence. For example, in high prevalence situations a test has more “chance” of encountering a patient with disease and vice-versa, i.e. in high-prevalence situations the number of diseased patients will increase for a given increase in sensitivity, and the converse applies to low-prevalence situations (e.g. screening). Therefore, if a new test changes sensitivity and/or specificity with respect to the standard test, then the raw numbers of patients diagnosed with and without disease will change, contingent on prevalence. To be useful as a performance measure overall, ROC AUC needs to incorporate realistic disease prevalence so that absolute changes in patient numbers are clinically interpretable. While sensitivity and specificity are prevalence independent, these measures keep positive and negative patients separate, whereas ROC AUC does not.

It is worth noting that while it is often stated that ROC AUC is unaffected by change in prevalence, this property arises from a model assumption. While it is true that the AUC is unaffected when one applies a different disease prevalence to an otherwise identical group, this does not hold when the group is different, for example patients referred by General Practitioners versus specialist clinics; i.e. diagnostic accuracy measured via ROC AUC in different populations of different prevalence can give different curves (although sometimes the AUC does not differ by much).

Alternatives to ROC AUC based on net benefit

I have described several problematic issues regarding use of ROC AUC as a measure of diagnostic performance of imaging studies in certain circumstances. These encompass conceptual issues (e.g. confidence scores may not be meaningful), non-trivial statistical issues (non-normal distributions and problems with data extrapolation), practical issues (many patients do not contribute to ROC AUC), and ethical issues (patients' and doctors values cannot be incorporated) with ROC AUC for analysis. An alternative should be easy to comprehend and express, incorporate explicit weightings for the value of gains in sensitivity vs loss of specificity, and account for disease prevalence. In particular, it should be possible to ascribe "costs" to the misclassification of true-positive versus true-negative patients in order to account for the different clinical consequences of false-positive and false-negative diagnoses.

The need for such an alternative to ROC AUC is well-recognised and several novel methods have been proposed, although these have not enjoyed substantial penetrance into the radiological literature, probably because ROC AUC is considered so pre-eminent (29, 30). Some authors have suggested simply moving the ROC curve operating point from one that optimises separation of events and non-events towards one event or the other, depending on the relative costs of misclassification (31). As noted already, clinicians are comfortable with using a test at a pre-specified diagnostic threshold (vs. across a range of thresholds), since this is more comprehensible and clinically relevant. Furthermore, since tests are often evaluated when close to clinical implementation, information is usually available regarding the most clinically appropriate threshold at which the test is likely to be used. It is therefore possible to set a diagnostic threshold for test positivity, determine sensitivity and specificity for disease at that threshold, and then compare this with the results obtained for an alternative test (or the same test used under different conditions, for example with and without CAD assistance) used at the same threshold. Such a comparison would present the change in sensitivity and corresponding change in specificity for the two tests/conditions when they are compared. Because comprehension is improved when sensitivity and specificity are combined into a single metric, it is then necessary to arrive at a method with which to achieve this. Simple summation of the change in sensitivity and change in specificity to give a "net benefit" could be performed but this would not take into account any differing misclassification costs. Therefore, in order to arrive at a comparison that is clinically useful, it is necessary to incorporate contextual information regarding the relative importance of false negative and false positive diagnoses, and prevalence of disease.

A search of the available literature reveals that “Reclassification Improvement”, “Weighted Net Reclassification Improvement”, and “Relative Utility” have all been advocated as measures that account the differing consequences of correct and incorrect diagnosis. A review of several different net benefit, threshold-based measures and the relationship between them has been provided by van Calster and colleagues (31), published subsequent to work starting on this thesis. Many of these measures, such as weighted-comparison (32) and Net Reclassification Index with two categories (33), are based directly on the difference in sensitivity and specificity between the two tests being assessed. Van Calster concluded that many of these measures, “are transformations of each other and hence always lead to consistent conclusions” (31).

Moons and co-workers investigate the situation where a single diagnostic test is applied at a threshold where the decision is simply to treat or not treat. They consider the consequences (i.e. clinical costs, loss of utility) of treating and not treating a patient with disease. They also consider the costs of treating a patient without disease, and quantify and incorporate this. For example, they state, “a ratio of net risks of 10 means that it is 10 times worse to withhold treatment from a diseased patient than to treat a non-diseased patient”, i.e. a benefits/costs ratio (32). Given a particular treatment threshold and treatment strategy, they show that the benefits and risks of subsequent decisions can be calculated and expressed as one parameter that they describe as “expected risks, ER” (32). They show that ER for any diagnostic test is equivalent to:

$$(p \times ERD+) + ([1 - p] \times ERD-)$$

where p = disease prevalence

ERD+ = expected risks to a patient with disease

ERD- = expected risks to a patient without disease

The sensitivity and specificity of the test can be calculated by estimating the probability of disease from each test result using a logistic model and dichotomising the range of estimated probabilities at the treatment threshold (P_t). Estimated probabilities greater than P_t is defined as a positive test result, and less than P_t as a negative test result.

To determine the better of two diagnostic tests (or the same test under different conditions), the difference in their expected risks at a specific threshold can be compared; $d(ER)$. They show that:

$$d(ER) = [(sensitivity1 - sensitivity 2) + (1-p/p)] \times [P_t/(1-P_t)] \times [specificity 1 - specificity 2]$$

So, if Δ sensitivity is the change in sensitivity and Δ specificity is the change in specificity, we get:

$$D(ER) = [\Delta\text{sensitivity} + (1-p/p)] \times [Pt/(1-Pt)] \times \Delta\text{specificity}$$

This equation provides an index for the comparison of the diagnostic performance of two tests (or the same test under different conditions, for example with and without CAD).

Vickers (34) states that:

$$\text{Net benefit} = TP - FP \times (pt/1-pt)$$

Where pt = the threshold at which the test is used.

$$TP = \text{sensitivity} \times p$$

$$FP = (1-\text{specificity}) \times (1-p)$$

$$\text{Thus net benefit} = \text{sensitivity} \times p - ((1-\text{specificity}) \times (1-p) \times (pt/1-pt))$$

The change in sensitivity (Δ sensitivity) and corresponding change in specificity (Δ specificity) would apply if two tests were compared.

We wished to investigate net benefit in situations where it is known at which threshold a test was to be used, specifically in binary situations (e.g. polyp present/polyp absent) as opposed to evaluation over a range of different thresholds. Taking the example of using CAD to interpret CT colonography, it is therefore possible to reframe the net benefit equation as follows:

$$\text{Net benefit} = \Delta\text{sensitivity} + [\Delta\text{specificity} \times (1/W) \times (1-p/p)]$$

Where Δ sensitivity is the change in sensitivity and Δ specificity is the change in specificity when using CAD assistance. Net benefit will be positive overall if CAD is beneficial; a zero value would indicate no benefit and a negative value would mean a net loss. We would expect CAD to increase sensitivity but decrease specificity. However, as I have explained, an increase in sensitivity may be regarded as

particularly desirable and therefore worth more than the negative consequences of a corresponding fall in specificity. In order to account for this a weighting factor “W” is used to diminish the effect of reduced specificity, achieved by multiplying Δ specificity by $1/W$ (i.e. the larger the value of W the less effect exerted by a given fall in specificity). This is analogous to the expression $pt/1-pt$, where W is equivalent to using the test in a binary, single-threshold fashion.

p is the estimated prevalence of abnormality in the target population (i.e. the population in whom the test is to be used ultimately in clinical practice). It is necessary to incorporate a correction for prevalence because sensitivity and specificity are used to derive TP and FP patients. When disease prevalence is low, true-negative diagnosis is easier to achieve since most subjects are disease-free. $1-p$ gives the proportion of disease-free subjects and dividing this by p gives the odds of having disease-free patients diagnosed over and above diseased patients. When performing a clinical trial it is often assumed that the prevalence of abnormality in the trial dataset mirrors that in the target population but this is often not the case because of a need to increase power for positive subjects (screening is the most obvious example, a situation where positive subjects are encountered relatively infrequently). My current thinking is that when using the net benefit equation we need to use a prevalence weighting that reflects the target population because the raw numbers of patients who would be TP and FN due to any changes in sensitivity and specificity induced by the new test are determined by the prevalence of abnormality. There is an essentially philosophical issue when analysing the results of a research study – do you report how the new test performed in the trial (i.e. when used on high prevalence, enriched data) or report its likely performance in clinical practice (where prevalence is likely to be lower)? Both are technically correct but have different implications and will be examined in Chapters 2 and 4.

A related point is that net benefit applies to each positive case detected rather than the net benefit per subject tested. It may be argued that the latter is more appealing because all individuals referred for testing are considered on an equal footing. In order to calculate the average consequences for all subjects it is necessary simply to multiply the net benefit formula by its divisor, i.e. Wp .

It is implicit from the net benefit equation that it is necessary to specify a diagnostic threshold in order to arrive at a given change in sensitivity and a corresponding change in specificity with which to populate the equation. Use of the net benefit equation therefore requires some *a priori* knowledge regarding the clinically relevant threshold above which the diagnostic test would be deemed positive, and below which it would be deemed negative.

Net benefit for a MRMC study would be calculated using a multilevel approach much as is used in meta-analysis, treating each reader as if they were an “individual “study”. Bootstrap methods can be used to obtain 95% confidence intervals empirically, to take into account the correct clustering of results within readers and cases in these MRMC studies. A significant benefit for a new test is defined as a positive net effect whose 95% confidence interval does not include zero. To enhance clinical interpretability, net benefit can be converted into an equivalent increase in true-positive and false-positive patients per 100 examined. Thus, in a series of 200 patients, 100 of whom have polyps, using CAD would result in the detection of approximately 7 additional true-positive patients at a cost of an additional 2 to 3 false-positives.

Perhaps, the most challenging aspect of this approach is the need to make assumptions regarding the relative misclassification costs of false-negative and false-positive diagnoses. While the precise value of W will often be unknown, it is likely that some insight will be available, even if just from expert opinion.

Advantages of net benefit methods

Like ROC AUC, net benefit combines sensitivity and specificity in a single metric, facilitating comparisons between tests but misclassification costs are transparent and incorporated in the analysis explicitly. Further, where the value of W is unknown or not known with precision, a range of weightings can be assigned via sensitivity analysis to examine the effect of using different values for W . Using net benefit based on decisions made in everyday practice can potentially avoid problems due to variations related to interpretation of diagnostic confidence scores. Prevalence is accounted for and there is no need to fit the available study data or extrapolate beyond it. Ultimately, the measure has clinical relevance and is interpreted easily; certainly easier than ROC AUC in the author’s opinion. In particular, study data are expressed in terms of false negative and false positive patient diagnoses (specified as difference in sensitivity and difference in specificity).

Summary

Diagnostic test accuracy studies should provide evidence in a comprehensible and intuitive format that facilitates choice of test for clinicians, their patients, and healthcare providers. Results should be reported in the context of clinical management decisions made at clinically sensible and important thresholds, preferably in terms of patients. For comparisons of tests, differences in true-positive and false-positive diagnoses

should be reported, and it is important that any statistical analysis should be able to incorporate misclassification costs that account for the fact that false-negative and false-positive diagnoses are rarely clinically equivalent. We have argued that in certain circumstances ROC AUC does not achieve these aims because the measure lacks clinical interpretability, incorporates confidence scales and thresholds that are clinically nonsensical, cannot account for the differing clinical implications of false-negative and false-positive diagnoses, and does not account for disease prevalence. The author and his collaborators propose an alternative measure based on net benefit that satisfies these requirements.

Arguing for ROC AUC, Zweig and Campbell(9) state that, "The ROC plot provides a more global comprehensive view of the test, independent of prevalence", going on to point out that, "sensitivity and specificity are properties inherent to the test; predictive value and efficiency (percentage of correct results) are properties of the application once the context (decision threshold and prevalence) is established". I agree, and believe ROC AUC is most useful in the early stages of diagnostic test assessment, especially for tests not requiring subjective interpretation (e.g. laboratory tests where no human reader is required). However, most radiological research investigates tests or applications that are ready for clinical use (or indeed already in use), so the context *is* already established. For example, the diagnostic threshold that constitutes a positive result is known frequently. Because of this, meaningful evaluation must incorporate how the test influences results for individual patients, at prevalence applicable to daily practice, incorporating an explicit assessment of the differing misclassification costs of false-negative and false-positive diagnoses. Also, the data should be comprehensible and intuitive to facilitate choices for clinicians, their patients, and healthcare providers. ROC AUC cannot achieve these aims easily, and is beset by non-trivial statistical problems induced by the confidence scales used to build the ROC curve. By contrast, alternative net benefit methods provide meaningful and clinically interpretable results.

Chapter 2: Incremental benefit of computer-aided detection when used as a second- & concurrent-reader for CT colonography: Multi-observer study.

This Chapter has been published as:

Halligan S, Mallett S, Altman DG, McQuillan J, Proud M, Beddoe G, Honeyfield L, Taylor SA. Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: Multiobserver study. *Radiology* 2011;258:469-76

Abstract

Aim: To quantify the changes in reader performance levels, if any, during interpretation of computed tomographic (CT) colonographic data when a computer-aided detection (CAD) system is used as a second or concurrent reader.

Methods: After institutional review board approval was obtained, 16 experienced radiologists searched for polyps in 112 patients, 56 of whom had 132 polyps. Each case was interpreted on three separate occasions by using an unassisted (without CAD), second-read CAD, or concurrent CAD reading paradigm. Reading paradigm and case order were randomised, with a minimal interval of 1 month between consecutive interpretations. Readers' findings were compared with the reference-truth interpretation. Mean per-patient sensitivity and mean per-patient specificity with CAD were compared with those achieved unassisted. The primary outcome measure was the net benefit for CAD assistance compared to unassisted interpretation for identification of patients with polyps of any size, where an increase in per-patient sensitivity was considered to be clinically more important than an equivalent decrease in specificity.

Results: The mean per-patient sensitivity for identification of patients with polyps of any size increased significantly with second-read CAD (mean increase, 7.0%; 95%CI: 4.0%, 9.8%) and concurrent CAD (mean increase, 4.5%; 95%CI: 0.8%, 8.2%). The mean per-patient specificity did not decrease significantly with second-read CAD (mean decrease, 2.25%; 95%CI: 2.52%, 0.1%) or concurrent CAD (mean decrease, 2.2%; 95%CI: 2.46%, 0.2%). These data gave a net benefit for second-read CAD of 6.2% (95%CI 3.1% to 9.3%), indicating a significant increase. The net benefit for concurrent CAD was 3.8% (95%CI -0.03% to 7.6%), indicating a non significant increase.

Conclusion: Use of second-read CAD significantly improves readers' per-patient and per-polyp detection. Concurrent CAD is less effective.

Introduction

Computer-aided-detection (CAD) to aid interpretation of CT colonography has recently become commercially available from both CT scanner manufacturers and independent software developers. CAD aims to improve diagnostic performance by alerting radiologists, via visual prompts, to pathology that might otherwise be missed. Initial assessments have focussed on the stand-alone sensitivity of CAD for colorectal polyps, estimating the incremental benefit that might be achieved in clinical practice via indirect comparison with the sensitivity of unaided radiologists and colonoscopists (35, 36). However, it is increasingly recognised that radiologists may misinterpret CAD prompts; true-positive prompts may be ignored and false-positive prompts may be interpreted as true polyps. The true benefit of CAD in clinical practice is therefore likely to be overestimated by indirect comparisons and direct comparisons that incorporate radiologists' interpretations are needed.

It has been hypothesised that CAD may diminish the need for prior reader experience (17) but results from such readers have been disappointing, despite undoubted benefit from CAD assistance (17). An estimate of the effect of CAD is therefore needed in readers who have some prior experience of CT colonography interpretation.

Furthermore, the reading paradigm used for CAD may also influence lesion detection: A study of 10 radiologists found marginal evidence for enhanced detection of small polyps when a second-read CAD was used rather than concurrent CAD (37), but more powerful studies are needed to answer this question with precision. Also, artificial environments, where readers interpret large numbers of consecutive studies in a short period, may introduce a "laboratory effect" and findings may not be generalisable. By performing a multi-reader study undertaken in a representative environment and using experienced radiologists, we aimed to quantify the incremental benefit of CAD for interpretation using both second-read and concurrent paradigms. Because of the problematic issues relating to ROC AUC analysis outlined in Chapter 1 of this thesis, the present study used the net benefit equation for analysis of our primary outcome measure.

Methods

Following Institutional Review Board (IRB) or Research Ethics Committee (REC) approval, CT colonography data acquired between January 2002 and June 2005 were accrued from 3 USA and 2 European centres. Both symptomatic and asymptomatic screening patients were included. Prone and supine CT colonography had been

performed following full bowel purgation in accordance with published standards(38). Each patient subsequently underwent optical colonoscopy. Both positive and negative (normal) cases were accrued from centres contributing multiple patients, in temporally contiguous blocks to enhance generalisability of findings(39). Patients with known polyposis syndromes and/or known cancer were excluded, since the CAD system was not developed specifically to detect cancer. Non-diagnostic studies, defined as non-visualisation of any colonic segment after the prone and supine studies had been assessed together, were also excluded (i.e. technical problems that would likely result in recall for further examination in normal clinical practice). Only non-diagnostic studies were excluded – there was no attempt to select only technically excellent studies since this would result in over-optimistic CAD performance. 116 patients were accrued and 112 used for analysis after exclusions due to data incompatibility (Table 5).

Table 5: Demographic and data acquisition details for the 112 patient cases interpreted.

	USA center 1	USA center 2	USA center 3	European center 1	European center 2
Number of patients	55	23	1	25	8
Mean age (years)	60.2	61.1	54	59.6	59.2
Male	31	22	0	12	4
Female	24	1	1	13	4
Symptomatic	0	11	0	0	0
Asymptomatic	55	12	1	25	8
Positive patients	30	17	1	8	0
Negative patients	25	6	0	17	8
Total number of polyps	64	50	1	17	0
Polyp size (mm)	2 to 20	2 to 17	25	3 to 15	0
Total no polyps ≥6mm	36	11	1	8	0
Tagging	Fluid and faecal	None	None	None	Fluid
Detector-rows	4	8	4	4	4
Slice/recon interval (mm)	2.5, 1.5	2.5, 1.25	1.25, 0	3.2, 1.6	1.25, 0.25

Definition of ground truth

Three radiologists experienced in CT colonography interpretation (minimum 200 endoscopically validated cases), interrogated each case to establish a “ground-truth” against which CAD and reader performance could be judged subsequently. Cases were interpreted without CAD. The original colonoscopy and histology reports were available and each case was read independently in its entirety twice in order to identify colonoscopic false-negatives. Polyps of all sizes were searched for and CT diameter and co-ordinates/segment noted. Polyps identified by colonoscopy but not visualised on CT were excluded because their co-ordinates could not be specified. Disagreement was resolved by consensus. There were 132 polyps in the 56 positive patients, 56 of which were ≥ 6 mm (Table 5).

Readers and case interpretation

Sixteen radiologists from 10 different hospitals interpreted the 112 cases. Each reported CT colonography in their clinical practice (mean prior experience 264 cases, range 50 to >1000). Eleven (69%) had completed a training course previously. Two commercially available visualisation platforms were used (either Viatronix V3D, Viatronix Inc, Stony Brook, NY, USA or Vitrea2 4.0 Colon, Vital Images Inc, Minnetonka, MN, USA) into which a CAD system was integrated (ColonCAD API 3.1, Medicsight PLC, Hammersmith, London, UK), and readers trained in its use. The CAD system had not been developed using any of the individual patient data used in the validation dataset but 116 other patient datasets from two of the contributing centers were used to develop the algorithm (USA center 1; 88 cases. USA center 2; 28 cases). Readers were unaware of the prevalence of abnormality and had no knowledge of standalone CAD performance characteristics. They were instructed that CAD might indicate polyps but might also make false-positive marks, and that the ultimate diagnosis should be based on their interpretation. Readers interpreted each of the 112 cases on three separate occasions, temporally separated by at least one month. Readers used three reading paradigms: unassisted, second-read CAD, concurrent CAD(40). CAD was not used when unassisted. For the second-read paradigm, readers applied CAD only following a full, unassisted interpretation(40). For the concurrent paradigm, CAD assistance was available to readers from the outset of their interpretation(17). Case and paradigm ordering were randomised between readers by the study statistician (SM). Each batch of 112 cases was read in a hospital environment over at least one month, to simulate normal clinical workload and avoid a laboratory effect. Readers noted all polyps, irrespective of size, recording maximal

transverse diameter, anatomical segment, CT co-ordinates, and interpretation time on a study report sheet that included a JPEG image-grab of any polyp detected.

Statistical analysis

Readers' case-classifications for each reading paradigm were compared with the ground-truth for each case. Two data-monitors performed this procedure, aided by ground-truth records of diameter, segment, CT coordinates, and JPEG image-grabs. In this way, each patient classification by readers was graded as either true-positive, false-positive, true-negative, or false-negative. Individual polyps noted by readers were also classified as true-positive or false-positive.

Our primary aim was to determine if the number of correctly classified patients increased significantly with CAD assistance without unacceptably diminished specificity. Our primary outcome was a comparison of accuracy for second-read CAD compared to unassisted interpretation. We anticipated CAD would increase detection of patients with polyps but also increase the number of normal patients categorized as false-positive (i.e. decrease specificity). As outlined in Chapter 1, because the author and his collaborators believe ROC AUC analysis unsuitable for the present analysis, our primary outcome was based on a net benefit measure, again as outlined in Chapter 1. We defined the "net benefit of CAD" as the difference in sensitivity using CAD plus the difference in specificity, the latter weighted by two factors, (i) W , defined as the relative increase in value of an additional correctly identified true-positive patient compared to the reduction in value of an additional false-positive patient and, (ii) an adjustment for prevalence of abnormality (proportion of true-positives) in the dataset. i.e:

$$\text{Net effect of CAD} = \Delta\text{sensitivity} + \Delta\text{specificity} * (1/W) * ([1-\text{prevalence}]/\text{prevalence})$$

Because no value for W was available in the existing literature at the time the analysis was first performed (a deficiency addressed in the following Chapter of this thesis), the author settled on a value of 3 following face-to-face discussion between the research team, and considering expert opinion. Given equivalent values derived from the mammographic literature(12), we believed 3 to be a very conservative estimate indeed. The prevalence of abnormality in the dataset was 50% and at the time of this research, that was the value we used for the analysis. We defined a significant improvement with CAD as a net benefit measure that was positive and whose 95% confidence interval did not include zero.

Average estimates were calculated from 1999 bootstrap samples generated by random sampling patients for each reader. Meta-analysis with equal weighting per reader was used to obtain an average across all readers. Confidence intervals were calculated by

taking the 2.5% and 97.5% percentiles of the cumulative distribution of the 1999 estimates. 1999 bootstraps samples were used because that number allowed us to achieve one of the original values as the 2.5% and 97.5% (i.e. 5% of values summed between the low and high tails of the distribution) of values.

Our primary analysis set a diagnostic threshold for the net benefit equation that investigated the change in sensitivity and specificity with CAD compared to baseline for patients with any size of polyp, since this is known to be an important clinical decision threshold (see Discussion). However, since larger polyps are known to be more at risk from malignant transformation than smaller polyps, we also calculated the CAD net benefit measure restricted to patients with polyps $\geq 6\text{mm}$ (i.e. the diameter that conventionally separates “small” from “medium” sized polyps). We also calculated per-polyp sensitivity for all polyps, those $\geq 6\text{mm}$, and those $\leq 5\text{mm}$. 98.3% confidence intervals were used to adjust for multiple secondary outcomes. We additionally investigated the effect on interpretation time in 15 readers (one reader was excluded due to missing data) since there is evidence that concurrent CAD may be time-efficient(17). Data were analysed by Dr Susan Mallett using STATA (StataCorp, College Station, TX).

Results

The stand-alone detection characteristics of CAD in the 112 patients were as follows: CAD correctly identified polyps in 40 of the 56 (71.4%) patients with polyps: Sixty nine of 132 polyps (52.3%) of all sizes were detected, 40 of 56 polyps $\geq 6\text{mm}$ (71.4%) and 14 of 15 polyps $\geq 10\text{mm}$ (93.3%). There were an average of 13.6 false-positive prompts per positive patient (i.e. prone and supine studies combined) and 14.1 per negative patient.

Per-patient analysis

Per-patient sensitivity, specificity and CAD net effect for individual readers for detection by CT colonography of patients with polyps of all sizes when readers were unassisted compared to second-read and concurrent CAD are shown in Table 6.

Table 6: Per-patient sensitivity, specificity and CAD net effect for detection by CT colonography of patients with polyps of all sizes when readers were unassisted compared to second-read CAD and concurrent CAD. Data are shown for all 16 readers.

Reader	Sensitivity			Specificity			CAD effect	
	Unaided	Second read	Concurrent read	Unaided	Second read	Concurrent read	Second Read*	Concurrent read*
1	68	59	57	93	89	86	-10.1	-13.1
2	71	75	70	86	82	86	2.4	-1.8
3	45	50	43	95	98	96	6.5	-1.2
4	63	57	55	93	91	93	-5.9	-7.1
5	75	71	66	84	86	79	-3.0	-10.1
6	63	80	80	86	84	91	17.3	19.6
7	48	63	50	98	98	100	14.3	2.4
8	39	46	48	93	96	95	8.3	9.5
9	75	68	64	93	89	86	-8.3	-13.1
10	38	52	61	91	89	80	13.7	19.1
11	55	68	77	80	71	68	9.6	17.3
12	59	75	61	95	91	96	14.9	2.4
13	54	64	66	96	96	96	10.7	12.5
14	18	36	29	96	91	91	16.1	8.9
15	80	84	80	91	88	86	2.4	-1.8
16	71	84	86	70	59	75	9.0	16.1

*The CAD effect is the net benefit achieved with CAD when used as both a second- and concurrent-reader, compared to baseline values without CAD. A positive value indicates a positive net benefit with CAD and a negative value a negative net benefit.

The net benefit for second-read CAD averaged across readers was 6.2% (95%CI 3.1% to 9.3%), indicating a significant increase. Average reader sensitivity increased significantly by 7% (95%CI 4.0% to 9.8%) whereas the average decrease in specificity was non significant; -2.5% (95%CI -5.2% to 0.1%); i.e. in a series of 200 patients (100 with polyps) readers using CAD would correctly identify 7 additional patients with polyps at the cost of approximately 2 to 3 false-positives.

The net benefit for concurrent CAD was 3.8% (95%CI -0.03% to 7.6%), indicating a non significant increase. Average reader sensitivity increased significantly by 4.5% (95%CI 0.8 to 8.2) with a non significant decrease in specificity of -2.2% (95%CI -4.6 to 0.2); i.e. in a series of 200 patients (100 with polyps) readers using CAD would correctly identify 4 or 5 additional patients with polyps at the cost of approximately 2 false-positives.

A sensitivity analysis was performed to investigate how the primary outcome for second-read CAD varied with different weightings for the relative value of true-positive and false-positive patients (W), and polyp prevalence. The CAD net benefit measure remained significant with the primary outcome re-analysed with equal weightings for the relative value of true-positive and false-positive patients (i.e. $W = 1$), and polyp prevalence of 50%: 4.5% (95% CI, 0.45% to 8.6%). A significant net benefit of 4.5% (95%CI, 0.45% to 8.6%) also persisted at a prevalence of abnormality of 25% and our conservative relative weighting of 3 for the relative value of a true-positive patient versus a false-positive. A significant benefit for second-read CAD remained with the analysis restricted to patients with polyps ≥ 6 mm; CAD net benefit measure 6.7% (95% CI 2.7% to 10.9%) (Table 3). There was no clear evidence that concurrent CAD increased correct categorisation for these patients (Table 7).

Table 7: CAD net benefit, per-patient sensitivity and specificity for detection by CT colonography of patients with polyps of all sizes and patients with polyps ≥ 6 mm when readers were unassisted compared to second-read CAD and concurrent CAD. Data are averaged across 16 readers.

Analysis		Unassisted by CAD % (95% CI)	Second read CAD % (95% CI)	Concurrent CAD % (95% CI)
Patients with all polyps	Average sensitivity	57.6% (49.9 to 65.1)	64.5% (56.5 to 72.2)	62.2% (53.9 to 69.8)
	Increase in average sensitivity versus unassisted	n/a	7.0% (4.0 to 9.8)	4.5% (0.8 to 8.2)
	Average specificity	90.1% (86.2 to 93.3)	87.6% (83.5 to 91.3)	87.8% (84.2 to 91.2)
	Increase in average specificity versus unassisted	n/a	-2.5% (-5.2 to 0.1)	-2.2% (-4.6 to 0.2)
	CAD net benefit	n/a	6.2% (3.1 to 9.3)	3.8% (-0.03 to 7.6)
Patients with polyps ≥ 6mm	Average sensitivity	65.8% (56.9 to 73.8)	72.8% (63.7 to 80.9)	70.1% (61.1 to 78.4)
	Increase in average sensitivity versus unassisted	n/a	7.1 (3.0 to 11.1)	4.2 (-0.5 to 8.9)
	Average specificity	93.0% (90.7 to 95.2)	92.2% (89.5 to 94.7)	92.3% (89.4 to 94.9)
	Increase in average specificity versus unassisted	n/a	3.9 (-2.3 to 9.0)	-0.6 (-2.3 to 0.9)
	CAD net benefit	n/a	6.7% (2.7 to 10.9)	3.9% (-0.8 to 8.8)

Per-polyp analysis

Detection of polyps of all sizes increased significantly when using both paradigms: The average increase in sensitivity was 7.3% (98.3%CI 3.9% to 10.4%) for second-read CAD and 4.0% (98.3%CI 1.1% to 7.7%) for concurrent CAD. The benefit for second-read CAD persisted with analysis restricted to polyps ≥ 6 mm, increasing by an average sensitivity of 9.0% (98.3%CI 4.9% to 12.8%), but did not for concurrent CAD (average sensitivity increase 2.9% (98.3%CI -1.1% to 8.0%). Both paradigms conveyed a significant benefit with the analysis restricted to polyps ≤ 5 mm, with mean increase in

sensitivity of 5.9% (98.3%CI 3.2% to 9.1%) and 4.8% (98.3%CI 2.2% to 7.9%) respectively (Table 8).

Table 8: Per-polyp sensitivity for detection by CT colonography of polyps of all sizes, polyps ≥ 6 mm, and polyps ≤ 5 mm when readers were unassisted compared to second-read CAD and concurrent CAD. Data are averaged across 16 readers. Confidence intervals are 98.3% to account for multiple testing.

Analysis	Unassisted by CAD % (98.3% CI)	Second read CAD % (98.3% CI)	Concurrent CAD % (98.3% CI)
Average sensitivity for all polyps	30.5% (24.2 to 37.1)	37.6% (29.3 to 45.1)	34.5% (27.7 to 42.1)
Increase in average sensitivity versus unassisted for all polyps	n/a	7.3% (3.9 to 10.4)	4.0% (1.1% to 7.7%)
Average sensitivity for polyps ≥ 6 mm	50.8% (40.5 to 60.4)	60.0% (48.3 to 69.7)	53.9% (43.2 to 64.2)
Increase in average sensitivity versus unassisted for polyps ≥ 6 mm	n/a	9.0% (4.9 to 12.8)	2.9% (-1.1 to 8.0)
Average sensitivity for polyps ≤ 5 mm	15.4% (10.9 to 20.6)	21.4% (14.8 to 29.0)	20.4% (14.2 to 27.8)
Difference in average sensitivity for polyps ≤ 5 mm versus unassisted	n/a	5.9% (3.2 to 9.1)	4.8% (2.2 to 7.9)

Interpretation time

Mean unassisted reading time was 7.8 minutes (95%CI 7.5 to 8.1). When using concurrent CAD and second-read CAD, readers on average took 0.93 minutes longer (95%CI 0.66 to 1.19) and 2.2 minutes longer (95%CI 1.87 to 2.66) respectively than when unassisted.

Discussion

While CAD for mammography has been available for many years, providing the opportunity for large multi-observer studies, reader studies of CAD for CT

colonography are relatively recent. Most studies have used relatively few readers and cases. Petrik and colleagues used four readers and 60 patient studies, finding that CAD increased sensitivity for polyps $\geq 6\text{mm}$ but with equally diminished specificity (24). Baker used seven less-experienced readers, but only 30 patient studies, also finding that CAD significantly improved sensitivity (22). So, larger studies are needed, as are those that address which CAD reading paradigm is preferred, in particular “second-read” or “concurrent”? Second-read is the classic CAD paradigm, whereby the reader performs a full, unassisted interpretation before activating CAD (40). The CAD marks are then reviewed with the aim of identifying lesions missed initially. Readers are not supposed to alter their decision regarding lesions detected by them when unassisted; specifically, if CAD “fails” to annotate a lesion previously believed present by the reader, the lesion should be retained for the clinical report. Second-read CAD impairs workflow since CAD assistance is additional to usual interpretation. In normal practice it is also possible that readers may not pay due vigilance when unassisted, choosing to activate CAD prematurely (40). The “concurrent” paradigm is an attempt to address these issues by integrating the assisted and unassisted interpretations: CAD is activated from the outset and annotated and unannotated regions scrutinised simultaneously. In a previous study the author and co-workers compared concurrent and unassisted paradigms in 10 readers each interpreting 107 colonography studies, finding that CAD increased sensitivity without a significant drop in specificity (17). However, very little research has compared concurrent and second-read paradigms directly; Taylor and colleagues asked 10 radiologists to read 25 datasets using both paradigms, finding significantly improved odds of polyp detection for both, but highest for second-read CAD (37). However, benefit was defined by improvement relative to the initial unassisted component of the second-read paradigm, which does not truly reflect the unassisted situation.

At the time of writing, the present study is the largest to date and has quantified the incremental benefit of CAD assistance for both second-read and concurrent paradigms via comparison to a temporally separated unassisted interpretation; The 16 readers were representative of the target audience for CAD and performed 5,376 individual interpretations in total. Furthermore, the study was performed in environments and over timescales generalisable to daily clinical practice. Via the net benefit analysis, we found that second-read CAD significantly improved reader detection of patients with polyps without a clinically unacceptable decrease in specificity. Furthermore, this effect persisted with sensitivity and specificity weighted equally, as is essentially the case for ROC AUC, but was even more convincing when we accounted for the fact that true-positive detection is more clinically important than false-negatives by using a weighting factor of 3. In contrast, the benefit with concurrent CAD was not significant.

Although we found that concurrent CAD increased per-polyp detection significantly, this was due to enhanced detection of those measuring $\leq 5\text{mm}$ whereas second-read CAD also improved detection of polyps $\geq 6\text{mm}$. The reasons as to why concurrent CAD is less effective is unclear. Since the CAD marks made during each paradigm read are identical, the effect cannot be a consequence of fewer true-positive marks during the concurrent read. Decreased vigilance paid to the unannotated endoluminal surface (including unannotated polyps) during concurrent reading may be an explanation. Alternatively, correctly annotated polyps may have been dismissed inappropriately by the reader although it is unclear why this should apply more to the concurrent paradigm. Further analysis of whether polyps missed during concurrent interpretation (vs. second read interpretation) were annotated correctly by CAD may help clarify.

We chose a per-patient analysis for our primary outcome because the decision to subsequently refer for colonoscopy is usually based on the maximal diameter of any polyps detected rather than their absolute number. The location and number of polyps is therefore secondary to correct classification of a patient as an individual with or without polyps. Also, individual patients with multiple polyps will exert undue overall influence on a per-polyp analysis. The disadvantage is that a larger study is required to adequately power a per-patient analysis. We chose to include all polyp diameters in our primary analysis because while all opinion leaders that polyps $\geq 10\text{mm}$ should be removed, removal of smaller polyps is controversial. Current guidelines suggest that polyps 6-9mm are important and should be removed, and authors have highlighted the issue of measurement error for both colonography and colonoscopy (41). An analysis of all polyps will capture those polyps 6mm or larger that might fall below this threshold due to measurement error. We did however perform secondary analyses restricted to patients with polyps measuring $\geq 6\text{mm}$, finding that second-read CAD remained significantly beneficial.

This is the author's first study to use the net benefit measure outlined previously in Chapter 1; net benefit was used as opposed to ROC AUC because the author and his collaborators found the latter unworkable in a prior multi-reader, multi-case study of CT colonography (17). Net benefit analysis is predicated by the difference in sensitivity using CAD plus the difference in specificity weighted by the perceived relative value of a true- and false-positive detection, adjusted for prevalence of abnormality. Our analysis used a multi-reader, multi-case design and combined two co-primary outcomes from two measures (i.e. change in sensitivity and change in specificity) that move in opposite directions. No extrapolation beyond the data was required and a range of relative weightings could be assigned and were explicit (unlike ROC AUC analysis). Weightings can be adjusted to reflect different patient preferences and also

to investigate uncertainties in relative weightings. The author believes the weighting used was very conservative; neither patients nor health-care providers are likely to regard a patient with a missed diagnosis as equivalent to a false-positive. For example, in breast screening a gain in sensitivity has been rated 150 to 500 times more highly than a corresponding loss in specificity (42). Nevertheless, the author and his collaborators arrived at a value for W of 3 after discussion between the research team; there was no precise value for W available in the current literature at the time the study was performed. Because of this, the following Chapter attempts to arrive at an evidence-based value for W for detection of colorectal cancer and polyps, established via discrete choice experiments. Later Chapters will extend this work beyond the colon to detection of extracolonic disease by CT colonography.

It is also worth noting that we used a prevalence of 50% for the net benefit equation since that was the prevalence of abnormality in the dataset. For the reasons outlined in Chapter 1, our current thinking is that it may be more appropriate to use a prevalence of abnormality that reflects the target population rather than the study population (the latter is often enriched to increase power for positive subjects, as was the case in this study). Chapter 4 re-analyses the data using a prevalence more representative of the target population.

In summary, second read CAD significantly improves per-patient and per-polyp detection by readers when interpreting CT colonography, without a clinically unacceptable drop in specificity. Concurrent CAD is less effective. In contrast to ROC AUC, we found it was possible to implement the net benefit analysis described in the previous chapters for analysis of a multi-reader multi-case study of CT colonography.

Chapter 3: Patients' & healthcare professionals' values regarding true- & false-positive diagnosis when colorectal cancer screening by CT colonography: Discrete choice experiment.

This Chapter has been published as:

Boone D, Mallett S, Zhu S, Yao GL, Bell N, Ghanouni A, von Wagner C, Taylor SA, Altman DG, Lilford RJ, Halligan S. Patients' and healthcare professionals values regarding true- and false-positive diagnosis when colorectal cancer screening by CT colonography: Discrete choice experiment. PLoS One 2013;8:e80767.

Abstract

Purpose: To establish the relative weighting given by patients and healthcare professionals to gains in diagnostic sensitivity versus loss of specificity when using CT colonography (CTC) for colorectal cancer screening.

Materials and Methods: Following ethical approval and informed consent, 75 patients and 50 healthcare professionals undertook a discrete choice experiment in which they chose between "standard" CTC and "enhanced" CTC that raised diagnostic sensitivity 10% for either cancer or polyps in exchange for varying levels of specificity. We established the relative increase in false-positive diagnoses participants traded for an increase in true-positive diagnoses.

Results: Data from 122 participants were analysed. There were 30 (25%) non-traders for the cancer scenario and 20 (16%) for the polyp scenario. For cancer, the 10% gain in sensitivity was traded up to a median 45% (IQR 25 to >85) drop in specificity, equating to 2250 (IQR 1250 to >4250) additional false-positives per additional true-positive cancer, at 0.2% prevalence. For polyps, the figure was 15% (IQR 7.5 to 55), equating to 6 (IQR 3 to 22) additional false-positives per additional true-positive polyp, at 25% prevalence. Tipping points were significantly higher for patients than professionals for both cancer (85 vs 25, $p<0.001$) and polyps (55 vs 15, $p<0.001$). Patients were willing to pay significantly more for increased sensitivity for cancer ($p=0.021$).

Conclusion: When screening for colorectal cancer, patients and professionals believe gains in true-positive diagnoses are worth much more than the negative consequences of a corresponding rise in false-positives. Evaluation of screening tests should account for this.

Introduction

Understanding diagnostic test performance is essential for evidence-based practice (43, 44), particularly for screening where risks and benefits are balanced finely. No screening test is 100% sensitive and the consequence is readily understood; false-negative tests will delay or prevent cure. Specificity is important when screening because most people are disease-free. A false-positive test means healthy individuals may undergo invasive procedures causing anxiety, morbidity, and even mortality (45). Tests that increase the proportion of people with disease who test true-positive (increase sensitivity) usually simultaneously increase the proportion of people without disease who test false-positive (diminish specificity). For example, computer-aided-detection (CAD)(46), digital imaging (47), and shorter screening intervals (48) all increase mammographic sensitivity for breast cancer but decrease specificity.

When comparing two diagnostic tests, interpretation is sometimes difficult if one has high sensitivity and the other high specificity: which is best? A combined measure of sensitivity and specificity facilitates interpretation: As outlined in Chapter 1 of this thesis, examples include net benefit or the area under the receiver-operator characteristic curve (ROC AUC) (49-53). The author has suggested that an advantage of net benefit measures is that they can incorporate relative values for gains in true-positive diagnoses versus costs of false-negative diagnoses, whereas ROC AUC cannot. However, few studies have quantified these costs, and none have done so for CT colonography, with the result that the experiment in Chapter 2 of this thesis used a value of 3 for W based solely on a conservative estimate from expert opinion.

Studies that have investigated similar misclassification costs suggest that they are valued very differently by patients; one study found women would accept 500 false-positive mammograms to avoid a single missed cancer (12). While qualitative research suggests that attendees value sensitivity over specificity when screening for colorectal cancer (54, 55) this has not been quantified with precision. Ignoring these preferences may underestimate test benefit. For example, the Medicaid/Medicare decision to not reimburse CT colonography (CTC) did not consider that screenees may still value gains in sensitivity over and above diminished specificity (56). To clarify this issue we established the relative weighting given by patients and healthcare professionals to additional true-positive diagnoses compared to additional false-positive diagnoses (i.e. gains in sensitivity versus loss of specificity) when using CTC for colorectal cancer screening; i.e. this Chapter aims to determine the value of W for the net benefit equation described in Chapter 1 and used for analysis of the primary outcome in Chapter 2.

Methods

Ethical statement

Ethics committee approval was granted by the local institutional ethical review board of University College Hospitals London; all participants gave written informed consent.

Design

We designed and conducted a discrete choice experiment (DCE)(57-59), according to recent guidelines(59). In particular, patients may value sensitivity so highly that even small changes can mask the influence of other attributes (59). Also, specificity is conceptually challenging, and patients are often unaware that false-positive diagnoses can even occur (54). Therefore, to simplify decision-making we used a “probability equivalence” design to establish attitudes to sensitivity and specificity alone, without the influence of other attributes. We presented sensitivity and specificity in terms of differing numbers of true-positive and false-positive diagnoses by imaging. A hypothetical “enhanced” CTC screening test was presented against “standard” CTC and participants expressed their preference between the two. Sensitivity and specificity for cancer for “standard” CTC was 85% and 95% respectively and 80% and 85% for polyps ≥ 6 mm. “Enhanced” CTC raised sensitivity for cancer to 95%, equivalent to detecting one additional cancer per 5000 screenees (cancer prevalence 0.2%)(60, 61). “Enhanced” CTC raised sensitivity for polyps to 90%, equivalent to detecting 125 additional people with polyps per 5000 (polyp prevalence 25%)(62). We aimed to raise sensitivity by 10% while avoiding a perfect test (i.e. sensitivity 100%), which is unrealistic. Specificity of “enhanced” CTC was dropped in increments across the scenarios, to 10% for cancer and 20% for polyps (Table 9). Such extremely low specificity is unlikely in everyday practice but necessary to calculate “tipping points”, i.e. the level at which an individual is willing to “trade” one attribute for the other. In the present case, the tipping point was the maximum reduction in specificity that participants were prepared to trade for a 10% absolute (vs. relative) increase in sensitivity.

Table 9: Overview of attributes and levels presented in cancer (1A) and polyp (1B) discrete choice experiments.

Table 9A; Cancer scenario.

Cancer question number	"standard" CTC		"Enhanced" CTC		Additional true positive detections per 5000 screening examinations	Additional false positive detections per 5000 screening examinations	FP tipping point (specificity: Standard CTC minus enhanced CTC) (%)
	Sensitivity for cancer (%)	Specificity for cancer (%)	Sensitivity for cancer (%)	Specificity for cancer (%)			
1c*	85	95	95	95	1	0	0
2c*	85	95	95	95	1	0	0
3c	85	95	95	90	1	250	5
4c	85	95	95	80	1	750	15
5c	85	95	95	70	1	1250	25
6c	85	95	95	50	1	2250	45
7c	85	95	95	40	1	2750	55
8c	85	95	95	30	1	3250	65
9c	85	95	95	20	1	3750	75
10c***	85	95	95	10	1	4250	85

Table 9B; Polyp scenario.

Polyp question number	"standard" CTC		"Enhanced" CTC		Additional true positive detections per 5000 screening examinations	Additional false positive detections per 5000 screening examinations	FP tipping point (specificity: standard CTC minus enhanced CTC) (%)
	Sensitivity for polyps (%)	Specificity for polyps (%)	Sensitivity for polyps (%)	Specificity for polyps (%)			
1p*	80	85	90	90	125	-250	-5
2p*	80	85	90	85	125	0	0
3p**	80	85	90	80	125	250	5
4p**	80	85	90	80	125	250	5
5p	80	85	90	70	125	750	15
6p	80	85	90	60	125	1250	25
7p	80	85	90	50	125	1750	35
8p	80	85	90	40	125	2250	45
9p	80	85	90	30	125	2750	55
10p***	80	85	90	20	125	3250	65

*Questions both favour enhanced CTC for both sensitivity and specificity. Respondents choosing standard CTC were considered to have misunderstood the task.

**Questions are identical to test for internal consistency.

***Participants choosing enhanced CTC in response to question 10 were considered potential non-traders; i.e.. they considered detection of a single additional cancer worth 4250 additional colonoscopies.

Because DCEs are difficult to comprehend, especially when administered via postal questionnaires (63), we used an interviewer-led design for patients, which clarifies understanding and permits qualitative exploration afterwards, especially with non-traders (64) - a “non-trader” is a participant who will not trade their preferred attribute at any cost. With respect to the present study, this would usually represent an individual who would accept any value of diminished specificity in order to achieve 10% increased sensitivity. A multimedia laptop presentation of colorectal cancer screening by colonoscopy and CTC was given, including information on survival benefit, that early detection was not always curative (65), and that false-positive CTC caused unnecessary colonoscopy. For clarity, only the most serious colonoscopic complication was presented, i.e. perforation in 1:500 (66, 67). Because inconsistent framing may introduce bias (68), both absolute and relative risks were displayed textually and graphically. Participants were asked to assume they were average risk for cancer/polyps and that polypectomy would reduce lifetime disease-specific mortality by 25% (69).

A random scenario was repeated to test consistency. A scenario with one option unquestionably superior for both sensitivity and specificity identified “irrational” responders. Finally, we incorporated “willingness-to-pay” (WTP) assessment: Standard CTC was pitched against CTC with sensitivity raised by 10% but with identical specificity. Participants were asked how much, if anything, they would pay for this.

Pilot

10 volunteers were piloted to confirm comprehensibility and inform sample size (70). While understanding attributes and levels, some did not trade (i.e. they judged the lowest specificity presented as acceptable). We therefore reprogrammed additional “stress-slides” (automatically triggered by responses accepting the lowest specificity for enhanced CTC), reinforcing potential harms, to assess whether heuristic bias anchored their decision; Seemingly irrational responses declined on repeat piloting of the same volunteers. Also, we found that participants were confused by considering cancer and polyp scenarios simultaneously, so the final survey presented separate cancer and polyp DCEs sequentially, each consisting of 10 scenarios.

Recruitment

We recruited consecutive consenting patients of screening age (>55 years), scheduled for non-cancer outpatient ultrasound/plain-radiographic investigations at a teaching hospital, identified via booking systems. Information/consent forms were mailed and responders interviewed on their appointment day. To avoid bias we excluded

respondents with a personal history of, or being investigated for, bowel cancer (71). All participants were offered a £10 gift voucher.

To investigate any attitudinal difference between patients and healthcare professionals, we recruited radiologists, gastroenterologists, surgeons, nurse-specialists, and radiographers who requested, performed, or interpreted colorectal imaging. To facilitate recruitment, healthcare professionals could complete the DCE online since we considered they were familiar with the concepts presented. Otherwise, a radiologist or clinical psychologist conducted DCEs, with scenarios presented in random order within the two DCEs. All participants were asked their age, ethnicity, education, and household income bracket.

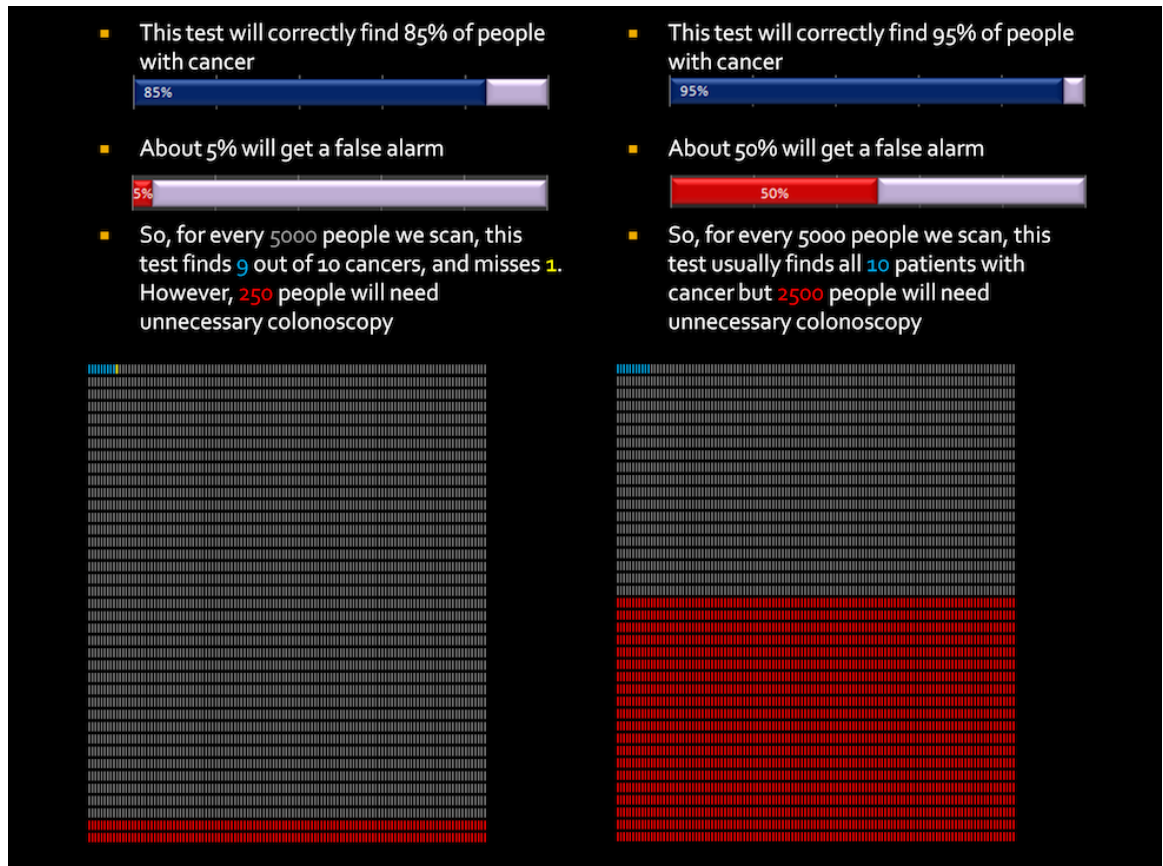
Statistical analysis

Our primary outcome was the reduction in specificity participants were willing to “trade” for 10% absolute (vs relative) increase in sensitivity. We defined the “tipping-point” as the highest increase in false-positive rate (FPR; 1-specificity) above baseline at which participants perceived the benefits of increased sensitivity were outweighed by potential harms. In the pilot this was 45% for cancer (i.e. participants allowed specificity to fall from 95% to 50% on average). To determine the median tipping point $\pm 5\%$ at two-sided alpha 0.05 and 90% power required 96 participants:

$$N = 4\sigma^2 z_{crit}^2 / D^2 \text{ where } D=0.10, p=0.45, z_{crit}=1.960, \sigma = 0.25 \text{ (72).}$$

We pre-specified a secondary outcome comparing patients and professionals, for which 88 participants were required for 90% power to detect 15% difference. Because our pilot suggested data would be non-normal, we recruited a further 15% (72). The stress-slides were triggered by participants preferring “enhanced” CTC despite increasing FPR by 85% for cancer and 65% for polyps. The highest tipping-point was taken if they traded subsequently; others were deemed persistent non-traders. Because participants were presented simultaneously with sensitivity, specificity, pictorial descriptions of changes and numerical information on the absolute increase in false-positives compared to increase in true-positives (Figure 9), we framed our results in terms of false-positive vs true-positive diagnoses, as this is most easily understood. Some data rounding was used in the scenarios when events were rare.

Figure 9: Example question from the cancer detection scenario. Each tally mark represents one of 5000 potential outcomes for a patient undergoing screening: True positive (blue), false negative (yellow), true negative (white), or false positive (red). Participants were informed that if they were to undertake the test in question, their odds of receiving any of the outcomes were represented by the chance of picking any of these tally-marks at random “like roulette”. Data are also represented numerically using both relative and absolute percentages. This scenario corresponds to the ‘tipping point’ for patients and professional respondents: On average, participants favoured the enhanced test (test B) in view of its additional sensitivity up to, but not beyond, this level of additional false positives.



The median tipping-point was calculated for cancer and polyps, for patients/healthcare professionals combined and separately. Because patient and professional numbers differed we used values from 1999 bootstrap estimates of median and IQRs, where samples included equal numbers ($n=50$) of each group, therefore weighting patients and professionals equally. At the tipping point, the change in specificity equivalent to a 10% change in sensitivity was converted into a change in the absolute relative numbers of false-positive and true-positive diagnoses using the equation for net benefit described in Chapters 1 and 2 of this thesis (32, 53):

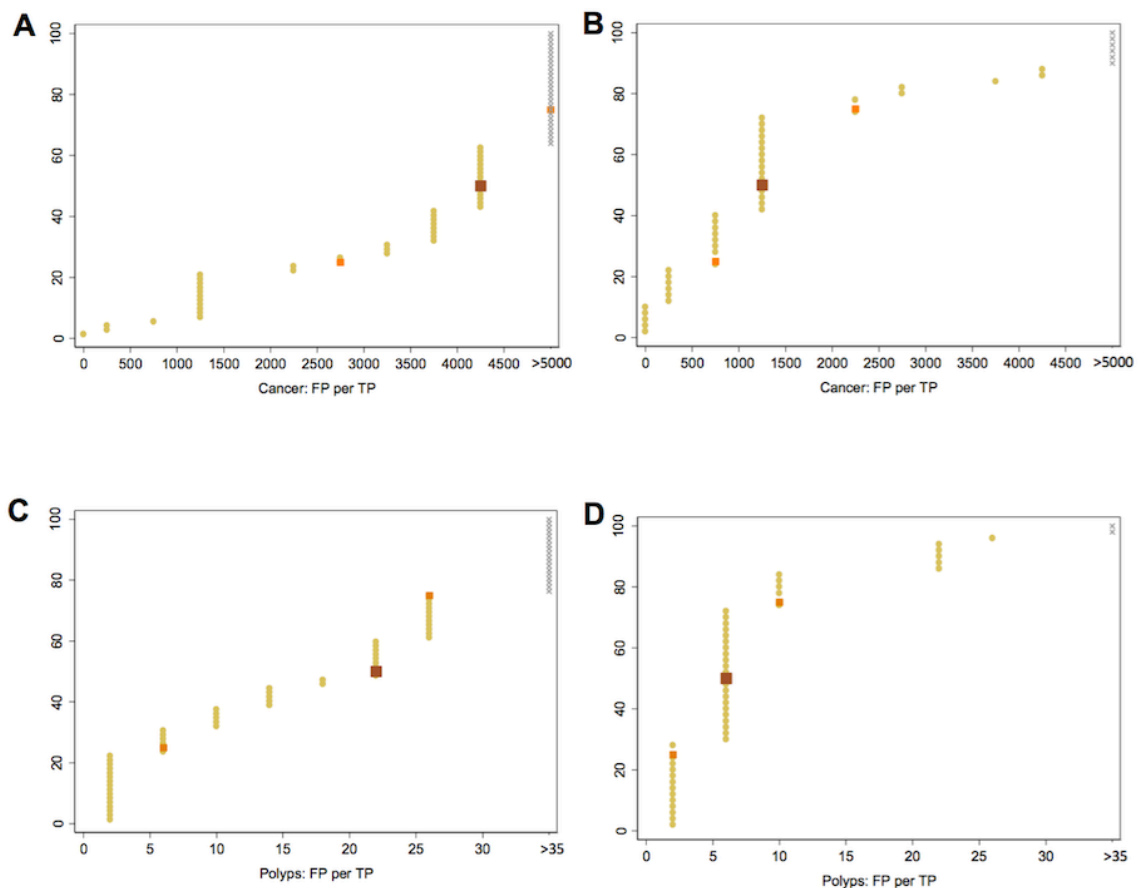
$$\text{net benefit} = \Delta \text{ sens} + [(1-\text{prevalence})/\text{prevalence}] \times (1/W) \times \Delta \text{ spec}$$

Where $\Delta \text{ sens}=10\%$, and $\Delta \text{ spec}=\text{median tipping point}$, and W is the relative weighting (the maximum number of additional false-positives traded per additional cancer or polyp detected). Prevalence was assigned 0.2% for cancer, 25% for polyps (60-62).

Tipping points were compared between patients interviewed by each researcher. Tipping points were highly non-normal so were summarised by medians and interquartile ranges (IQR 25% to 50%); the median can be interpreted as corresponding to an average participant. For tipping points and relative weighting of false-positive to true-positive, all non-traders were treated as requiring higher FP values than offered (Figure 10: grey values). The Mann-Whitney U test statistic and Wilcoxon signed rank sum test were used for unpaired and paired comparison, respectively. Statistical analysis was performed by Dr. Susan Mallett (using Stata V11.0, Stata Corporation, College Station, Texas).

Figure 10: Cumulative graph of participants' tipping points for trading absolute numbers of true-positive versus false-positive diagnoses. The x-axis shows the absolute number of false-positive diagnoses exchanged for a true-positive diagnosis. The y-axis is the % of participants. Each yellow dot shows an individual participant's trading point. Grey symbols indicate values assigned for participants who refused to trade. Brown dot shows median value representing "an average participant". Orange dots show 25 and 75 percentage points. Graphs are shown separately as follows:

- A; Patients, cancer scenario (n=72).
- B; Professionals, cancer scenario (n=50).
- C; Patients, polyp scenario (n=72).
- D; Professionals, polyp scenario (n=50).



Results

112 consecutive patients and 62 professionals were invited. 75 patients and 50 healthcare professionals participated, a response of 67% and 81% respectively (Table 2). Three patients could not complete the DCE leaving 122 for analysis (two medical professionals gave partial responses). Dr. Darren Boone interviewed 53, Ms. Nichola Bell interviewed 48; 21 responses were online. Compared to professionals, patients were older, discontinued education earlier, and had lower household income (Table 10).

Table 10: Demographic characteristics and household annual income of patient and professional participants, including non-traders (% have been rounded).

Characteristic	Patients (n=72)*	Professionals (n=50)**	Total (n=122)
	N (%)	N (%)	N (%)
Gender			
Female	49 (68)	24 (48)	73 (60)
Male	23 (32)	26 (52)	49 (40)
Age (year)			
25-34	0 (0)	26 (52)	26 (21)
35-54	0 (00)	23 (46)	26 (21)
55-59	18 (25)	1 (2)	16 (13)
60-69	40 (56)	0 (0)	40 (33)
70-79	14 (19)	0 (0)	14 (12)
Ethnicity			
White	49 (68)	33 (66)	82 (67)
Other	23 (32)	17 (34)	40 (33)
Household income/GBP/year			
< 10000	3 (4)	0 (0)	3 (3)
10001-20000	14 (19)	0 (0)	14 (11)
20001-30000	19 (26)	3 (6)	22 (18)
30001-40000	10 (14)	9 (18)	19 (16)
>40000	4 (6)	32 (64)	36 (30)
Declined to answer	22 (31)	6 (12)	28 (23)

Non-traders

For cancer detection 30 (25%; 27 patients, 3 professionals) participants were non-traders, 20 (16%; 18 patients, 2 professionals) of who were also non-traders for polyps. Non-traders were significantly more likely to be patients (27[38%] vs 3[6%]); $p < 0.001$, were significantly older (median age 64.5 vs 44.5; $p < 0.001$), and were less educated

than traders (15% vs 2% with no formal qualifications; $p<0.001$). There was no significant difference in gender (59% vs 61% female; $p=0.56$) or ethnicity (30% vs 33% non-white; $p=0.57$). Considering patients alone, non-traders ($n=27$) were older (median age 66.8 vs 60.1; $p=0.001$), less affluent (median household income GBP 10,001 to 20,000 vs. GBP 20,001 to £30,000 per annum; $p=0.03$. GBP = Great British Pound, = 1.2 Euros and 1.6 US Dollars at current exchange rates) and less qualified (median school-leaving age 16 vs. 18 yrs; $p=0.02$) than traders ($n=45$). For cancer and polyps respectively, 34% (16/47) and 35% (11/31) participants who were initially unwilling, subsequently traded following the stress-slides.

Cancer

Overall, the median tipping-point for cancer detection occurred at 45% drop in specificity (IQR 25 to >85%; Table 11). Thus, on average, a 45% drop in specificity was considered the maximal fall acceptable in exchange for 10% increased sensitivity. At population prevalence of 0.2%, this equates to 2250 (IQR 1250 to >4250) additional false-positive diagnoses per additional true-positive cancer. The average number of additional false-positives per additional true-positive detection was significantly higher for patients (median 4250 (IQR 2750 to >4250) than professionals (median 1250, IQR 750 to 2250, $p<0.001$), i.e. the average patient perceived a greater number of false-positives acceptable to gain an additional true-positive.

Table 11: Tipping points and relative weighting for cancer and polyp detection scenarios calculated for patients, professionals, and all participants combined (FP = false-positive diagnosis, TP = true-positive diagnosis).

	Tipping point (FP tipping point: max change in specificity acceptable for a 10% gain in sensitivity)		Relative weighting FP to TP (Average number of additional FP per additional TP detection)	
	Median	IQR	Median	IQR
Patients				
Polyp	55	15 to 65	22	6 to 26
Cancer	85	55 to >85	4250	2750 to >4250
Professionals				
Polyp	15	5 to 25	6	2 to 10
Cancer	25	15 to 45	1250	750 to 2250
Combined				
Polyp	15	7.5 to 55	6	3 to 22
Cancer	45	25 to >85	2250	1250 to >4250

Polyps

Overall, the median tipping-point for polyp detection was 15% (IQR 7.5 to 55; Table 11). Thus, on average, a 15% drop in specificity was considered the maximal fall acceptable in exchange for 10% increased sensitivity. At population prevalence of 25%, this equates to 6 (IQR 3 to 22) additional false-positive diagnoses per additional true-positive polyp. Again, the median number of additional false-positives per additional true-positive was significantly higher for patients (55, IQR 15 to 65) than professionals (15, IQR 5 to 25, $p<0.001$).

For patients and professionals combined, the average number of additional false-positives traded per additional true-positive detection was significantly higher for cancer than polyps (45 vs. 15; $p<0.001$), indicating that larger falls in specificity were perceived acceptable when testing for cancer.

There was no significant difference in overall median tipping point elicited by the radiologist or psychologist, ($p=0.57$) nor between medical professionals' data obtained face-to-face vs. online ($p=0.59$).

Willingness-to-pay

Three quarters of participants were willing to give a price range they would be willing to pay for a test with sensitivity raised by 10% but no loss of specificity. Median WTP was significantly higher for cancer than polyps: 201 to 500 GBP (IQR 101 to 200 GBP, to 501 to 1000 GBP) vs. 101-200 GBP (IQR 51 to 100, to 201 to 500 GBP), $p<0.001$, indicating participants felt cancer detection was worth more than polyp detection. There was no significant difference in WTP for polyp detection when patients and professionals were compared ($p=0.97$) but patients' WTP was significantly higher than professionals' for cancer detection: median 201 to 500 GBP (IQR 101 to 200 GBP, to 201 to 500GBP) vs. median 101 to 200 GBP (IQR 51 to 100 GBP, to 201 to 500 GBP, $p=0.036$). Moreover, median household income was significantly lower for patients than professionals (20,001 to 25,000 GBP vs. >40,000 GBP; $p=0.021$, Table 12), indicating that patient's values were particularly strongly held from a relative perspective. Most participants (27 of 32 participants) who declined to answer WTP, declined to answer for both polyps and cancers. Participants who declined gave, on average, higher values of false-positives per additional true-positive diagnosis.

Table 12: Patient and professionals' willingness to pay (WTP) for a 0.10 (10%) increase in test sensitivity without any reduction in specificity, for detection of cancer or clinically significant polyps.
 GBP = Great British Pounds.

WTP/GBP	POLYP DETECTION					
	Patients (72)		Professionals (50)		Total (122)	
	n	%	n	%	n	%
<50	9	12	8	16	17	14
51-100	10	14	8	16	18	15
101-200	15	21	14	28	29	24
201-500	4	6	10	20	14	11
501-1000	10	14	4	8	14	11
>1000	0	0	0	0	0	0
Declined to answer	24	33	6	12	30	25
	CANCER DETECTION					
	Patients (72)		Professionals (50)		Total (122)	
	n	%	n	%	n	%
<50	5	7	5	10	10	8
51-100	3	4	7	14	10	8
101-200	10	14	12	24	22	18
201-500	14	20	9	18	23	19
501-1000	11	15	6	12	17	14
>1000	8	11	3	6	11	9
Declined to answer	21	29	8	16	29	24

Discussion

When screening for colorectal cancer and polyps, we found that patients and healthcare professionals both valued gains in diagnostic sensitivity over and above corresponding loss of specificity. On average, 2250 additional false-positives were considered worth trading for a single additional true-positive diagnosis of cancer and 6 additional false-positives for an additional true-positive diagnosis of a polyp. These values are vastly in excess of the (admittedly conservative) value of 3 used for W in Chapter 2, and are more in line with those from a study of mammography that found women willing to trade 500 false-positive mammograms (and their consequences) in order to diagnose a single additional cancer that would otherwise have been missed (12).

While it is known that patients value sensitivity over specificity for colorectal cancer screening (73, 74), we could find no data that quantified this for a radiological test, hence the necessity to use a conservative value for the analysis detailed in Chapter 2 of this thesis. CAD for CTC will increase sensitivity but at the cost of reduced

specificity, sometimes significantly (17, 75-77). However, as detailed in Chapter 1, the potential clinical consequences of missed cancer (death) are not equivalent to those of false-positive diagnosis (unnecessary colonoscopy); our findings confirm that both patients and healthcare professionals believe this. It is therefore important that analysis of research studies of diagnostic tests take account of this asymmetry in misclassification costs. Such values can be incorporated into net benefit measures as described in Chapters 1 and 2 of this thesis, since they incorporate relative weightings for different clinical costs (53, 78). By contrast statistical measures such as ROC AUC cannot assign different utilities to gains in sensitivity versus falls in specificity and so could find a new test of no value when both patients and professionals might think otherwise.

We used a discrete choice experiment, a relatively novel methodology for establishing preferences (79). Traditionally, preferences are elicited via ranking (80), with test attributes considered in isolation. Results are therefore predictable: Patients and professionals favour tests that are sensitive, specific, inexpensive, and non-invasive. However, this does not reflect the trade-offs demanded by everyday, real-life practice. DCEs are advocated increasingly by researchers because respondents indicate preferences between different test characteristics, which more accurately reflects real-world choices (57-59, 80-82). Because DCEs are complex, we delivered most experiments face-to-face to facilitate understanding and participation, which can increase the generalisability of results. Accordingly, most participants gave complete, consistent, meaningful responses. While interviewer bias is possible, we found no significant difference between responses obtained from the psychologist or radiologist. Further, there was no significant difference in responses obtained face-to-face vs. online.

To simplify and focus the cognitive task, we compared just two attributes, increase in true-positive and false-positive diagnoses by imaging (also expressed by sensitivity and specificity). In order to create an “enhanced” test that inflated sensitivity for cancer to 95% (perfect sensitivity would be unrealistic) we used a baseline sensitivity of 85% for standard CTC, which is likely an underestimate. However, in this type of experiment, the relative weighting given to attributes across different scenarios is key, not the absolute difference between them. Using this design we elicited the relative importance that participants placed on gains in sensitivity versus loss of specificity.

Although both groups valued gains in sensitivity over and above corresponding loss of specificity, this finding was stronger for patients. Healthcare professionals, especially

those who are medically qualified, will have a deeper understanding of the pros and cons of diagnostic testing; as noted earlier, some patients do not understand that false-positive diagnosis is even possible (54). Patients were older, discontinued education earlier, and had approximately half the annual household income of health professionals, yet patients ascribed a monetary value to enhanced sensitivity that was approximately twice that of professionals, demonstrating they consider sensitivity exceptionally important. Again, professionals may have been less willing to pay because they had a deeper understanding of the issues, for example that an earlier cancer diagnosis does not necessarily equate with cure, and knowledge that the large majority of polyps are destined never to become malignant.

If statistical analyses must account for discrepant weightings for sensitivity and sensitivity, a particularly interesting question is whose weightings should be used? Some will argue that healthcare professionals are the best option, notably medically qualified clinicians because they request tests, have the deepest understanding of pros and cons, and thus have the broadest and most informed perspective overall. Others will argue that society ultimately undergoes and pays for diagnostic testing, and so patients' perspectives are most appropriate. This issue warrants further research.

Our study has limitations. As noted previously, DCEs are challenging for participants (83), requiring motivation, literacy, and numeracy, which may introduce selection bias (64). We attempted to reduce this effect by using an interviewer rather than a postal questionnaire. Although we had power for our primary endpoint, larger and/or different samples will better investigate differences between subcategories of patients and healthcare professionals. Because we believed that we should not ignore particularly strongly held beliefs, we included non-traders via calculating median values; our estimates are therefore an underestimate. WTP estimates are also likely underestimates because of reluctance to state income. We followed guidelines for best practice of DCE studies (59) but suggest that strategies for design and analysis need further investigation (84, 85). Common to all hypothetical scenarios, subjects' actions in real life may not mirror those expressed in a DCE. Finally, the weightings we derived are specific to colorectal cancer screening. However, we believe they are likely to be similar to other scenarios that involve diagnosis of cancers with similar clinical consequences (12).

In summary, via discrete choice experiment we found that both patients and healthcare professionals believe gains in diagnostic sensitivity are worth vastly more than the perceived negative consequences of a corresponding loss of specificity, when

considering colorectal cancer screening. Gains in sensitivity over loss of specificity were valued more highly for cancer detection (vs. polyps) and by patients rather than healthcare professionals. These findings should influence the evaluation of new screening tests.

Chapter 4: Assessment of the incremental benefit of computer-aided detection (CAD) for interpretation of CT colonography by experienced and inexperienced readers.

Boone D, Mallett S, McQuillan J, Taylor SA, Altman DG, Halligan S. Assessment of the incremental benefit of computer-aided detection (CAD) for interpretation of CT colonography by experienced and inexperienced readers. PLoS ONE 2015;10: e0136624.

Abstract

Objectives: To quantify the incremental benefit of computer-assisted-detection (CAD) for inexperienced readers versus experienced readers of CT colonography.

Methods: 10 inexperienced and 16 experienced radiologists interpreted 102 CT colonography studies using temporally separated unassisted and concurrent-CAD paradigms. Readers were asked to indicate any polyps detected. Interpretations were compared against a reference standard for each case: 46 studies were normal and 56 had at least one polyp (132 polyps total). The primary outcome was the difference in CAD net benefit (a combination of sensitivity and specificity weighted in favour of sensitivity) between groups for detection of patients with any polyp.

Results: For detection of patients with any polyp, inexperienced readers' sensitivity rose from 39.1% to 53.2% with CAD and specificity fell from 94.1% to 88.0%, both statistically significant. Experienced readers' sensitivity rose from 57.5% to 62.1% and specificity fell from 91.0% to 88.3%, both non-significant. These data gave a significantly beneficial net benefit with CAD of 11.2% (95%CI 3.1% to 18.9%) for inexperienced readers compared with a non-significant 3.2% (95%CI -1.9% to 8.3%) for experienced readers, a difference of 7.9% (95%CI -0.9% to 16.6%).

Conclusions: Concurrent CAD resulted in a significant net benefit when used by inexperienced readers to identify patients with any polyp by CT colonography. The net benefit was nearly four times the magnitude of that observed in experienced readers. Experienced readers did not benefit significantly from concurrent CAD.

Introduction

Chapter 1 of this thesis detailed the use of a net benefit analysis as an alternative to ROC AUC for assessment of diagnostic test accuracy. The net benefit analysis was implemented in Chapter 2 and a discrete choice experiment used to generate an evidence-based value for W in Chapter 3. In this Chapter I further refine the net benefit analysis by incorporating the value of W generated in Chapter 3 and by using a value for prevalence that is more representative of the target population in everyday clinical practice rather than the prevalence within the study dataset (which was the method used in Chapter 2). The net benefit equation is then used to compare the diagnostic performance of experienced and inexperienced radiologists when using CAD to interpret CTC. While it has been hypothesised frequently that CAD may diminish the need for prior experience (17), the two largest reader studies of CAD published at the time of writing have relied solely on experienced radiologists; 19 (Dachman) and 16 readers (Halligan; Chapter 2) (75, 78). Few studies have compared experienced and inexperienced readers directly, and those that have done so are limited by their small size and low statistical power(86). For example, Mang and colleagues asked two “expert” and two “nonexpert” observers to interpret 52 patient datasets using CAD in a second-read paradigm, finding that CAD was only beneficial for the less experienced readers (87). The present study aimed to quantify the incremental benefit of CAD for inexperienced versus experienced readers by comparing data across two large multi-reader, multi-case studies of CT colonography, via the net benefit analysis.

Methods

Data sources and readers

We obtained original reader data acquired from two multi-reader, multi-case studies of CAD for CT colonography: the first was the study detailed in Chapter 2 (i.e. experienced readers) and the second was our prior study of CAD for CT colonography that first brought issues with ROC AUC to our attention (17). Both studies had full ethical committee approval for data sharing. The second study investigated 10 radiologist readers with no prior experience of CT colonography who interpreted 107 patient datasets both unaided and when using CAD in a concurrent paradigm (17).

Data characteristics

118 discrete patient cases were used for the two studies with 102 patient cases common to both. We included reader data from these 102 cases to enable paired comparisons of experienced and inexperienced observers without the need for imputation to account for missing data. Thus, any differences could be attributed directly to differences in experience rather than due to confounding because of different case mix. We calculated the difference between novices and experienced readers on a per case basis so allowing data clustering to be included in the analysis, generating more appropriate 95% confidence intervals. Cases were a mix of symptomatic and asymptomatic subjects aggregated from three USA and two European centres. Prone and supine CT colonography had been performed in each case using multidetector-row machines and following full bowel purgation, adhering to contemporaneous published guidelines for data acquisition (38).

A reference truth against which the CAD and reader output could be judged was established by three experienced readers (minimum 200 endoscopically validated cases) for each case as detailed in Chapter 2. None were readers in the studies. A pair read each case with the benefit of the original radiological report supplemented with colonoscopic and histologic data, and achieved consensus regarding the case classification and size and location of any polyp. Ultimately, of the 102 cases used for this study, 46 were judged normal and 56 had at least one polyp. There were 132 polyps in total: 15 polyps $\geq 10\text{mm}$, 41 polyps 6mm to 9mm, 76 polyps $\leq 5\text{mm}$, with 12, 25 and 19 cases where these were the largest polyps respectively. In 37 cases the largest polyp was at least 6mm.

Reading environment and CAD paradigm

In the study of inexperienced readers, readers interpreted all cases in a quiet environment without CAD over the course of one week and then repeated the interpretation two months later, this time using CAD in a concurrent paradigm(88). In the study of experienced readers, cases were read in three batches, each over one month, with a temporal separation of at least one month between batches, as noted in Chapter 2. All cases were read once in each batch, using one of three paradigms (unassisted, concurrent-CAD, second-read CAD), with paradigm and study sequence randomised between batches. Thus unassisted interpretation and concurrent-CAD interpretation of each individual case were common to both studies, with a temporal separation between reads of at least one month. For the concurrent paradigm, readers interpret CAD-annotated CT colonography data simultaneously with unannotated data

(37). The same CAD system was used for both studies, so that correctly annotated polyps and false-positive detections were the same (Colon CAD API 3.1, Medicsight, Hammersmith, UK). A proprietary CT colonography package was used to view CT colonography data for the study of inexperienced readers. For the study of experienced readers, CAD was implemented into commercially available workstations as explained in Chapter 2 (either Viatronix V3D, Stony Brook NY, USA, or Vital Images, Minnetonka, Minn, USA). None of the cases had been used previously to develop the CAD algorithm.

Readers were asked to indicate for each case whether they believed a polyp was present or not, irrespective of size. If they believed the case was positive, they were asked to indicate the segmental location of all polyps detected and note the CT coordinates. They also estimated the maximum diameter of each polyp using software callipers. Responses were made on study datasheets collated subsequently by a study coordinator.

Statistical analysis

The collated datasheet responses were compared to the reference truth diagnosis for each case so that each reader's response for each case could be classified as true-positive, true-negative, false-positive, or false-negative. Each individual polyp detected by readers was also categorised as true-positive or false-positive.

Our pre-specified primary outcome measure for the present study was the net benefit measure described in Chapter 1 and used in Chapter 2: a combination of sensitivity and specificity weighted in favour of sensitivity for detection of patients with any polyp, defined as:

$$\text{CAD net benefit} = \Delta\text{sensitivity} + [\Delta\text{specificity} \times (1/W) \times (1-p)/p]$$

where $\Delta\text{sensitivity}$ was the change in sensitivity, $\Delta\text{specificity}$ the change in specificity from baseline (i.e. assessment without CAD) for detection of patients with polyps of any diameter, achieved when using CAD assistance. For the present study, in contrast to Chapter 2, we used a value for p that reflected the proportion of patients with polyps in the population where the software was likely to be used (25%, i.e. asymptomatic screenees) rather than the prevalence of abnormality in the study dataset (50%). We also used a value of W that was derived from the discrete choice experiment performed in Chapter 3. Specifically, we used a value of 6, which was the tipping-point at which participants overall deemed the benefit of enhanced sensitivity to be outweighed by the negative consequences of diminished specificity; i.e. the present net

benefit analysis regarded one extra true-positive diagnosis as equivalent to 6 additional false-positives.

As for Chapter 2, overall net benefit was calculated by meta-analysis, treating each reader as if they were an individual study. Average estimates were calculated from bootstrap samples generated by means of random sampling of patient cases for each reader, bootstrapping positive and negative patients separately. Confidence intervals (CI) were calculated by taking the 2.5% and 97.5% percentiles of the cumulative distribution of the bootstrap estimates. We defined a significant beneficial effect for CAD as a positive net benefit whose 95% confidence interval did not include zero.

The following secondary outcomes were pre-specified for experienced and inexperienced readers, and the difference between them:

- Per-patient sensitivity and specificity when unassisted, when using concurrent CAD, and the change when using CAD, both for patients with all polyps and also patients with polyps ≥ 6 mm
- Per-polyp sensitivity when unassisted, when using CAD, and the change when using CAD, for patients with all polyps, polyps ≥ 6 mm, and polyps ≤ 5 mm (6mm is the diameter threshold that conventionally separates “small” from “medium” polyps; see Chapter 2).
- The mean number of patients correctly classified as true-positive solely as a consequence of false-positive detections.
- Mean reading time with and without CAD, and the difference between the two.

We also wished to speculate on the potential gain for inexperienced readers using CAD in a second-read paradigm by quantifying the difference in accuracy between concurrent and second-read CAD paradigms for experienced readers via existing data.

Average estimates were calculated from 2000 bootstrap samples generated by random sampling of patients and readers, retaining data clustering. Positive and negative patients were bootstrapped separately and the same bootstrap samples of cases used for both studies. Readers were bootstrapped separately for each study. Differences between inexperienced and experienced readers were calculated within each case prior to averaging across all cases. Meta-analysis with equal weighting per reader was used to obtain an average across all readers. For per-polyp sensitivity bootstrap analysis accounted for the clustering of multiple polyps per patient. Confidence intervals were calculated by taking the 2.5% and 97.5% percentiles of the cumulative distribution of the 2000 estimates. Although underpowered for analysis at the 1cm threshold, we calculated the median number of patients detected. Interpretation times

for experienced readers were based on 15 readers (one had missing data). Sensitivity and specificity, and changes in these are expressed as percentages. Results are reported with 95% confidence intervals (CI). Differences with confidence intervals not including zero were considered to be statistically significant.

Results

Per-patient analysis

For both inexperienced and experienced readers, using CAD changed the proportion of readers correctly identifying cases with polyps in 83% of cases; detection of patients with polyps increased in 70% and 57% of cases over 10 and 16 readers respectively. Per-patient sensitivity and specificity (with 95%CI) for readers when unassisted and when using CAD are shown in Table 13. There was a statistically significant mean gain in sensitivity for all polyps of 14.1% for inexperienced readers when using CAD (rising from 39.1% to 53.2%). Sensitivity for patients with all polyps was higher for experienced readers but the mean gain of 4.6% with CAD was not significant (rising from 57.5% to 62.1%). Inexperienced readers benefitted by a mean gain in sensitivity approximately three times that for experienced readers, a significant difference of 9.6% (95%CI 1.2% to 17.7%). The mean change in specificity of -6.1% with CAD was significant for inexperienced readers (falling from 94.1% to 88.0%) whereas the mean change in specificity of -2.7% with CAD was nonsignificant for experienced readers (falling from 91.0% to 88.3%). Thus, in a series of 200 patients (100 with polyps) inexperienced readers using CAD would on average correctly identify 14 additional patients with polyps, at the cost of approximately 6 additional false-positives, whereas experienced readers would identify 4 or 5 additional patients with polyps at cost of 2 or 3 additional false-positives. For our primary outcome, these data gave a mean CAD net benefit of 11.2% (95%CI 3.1% to 18.9%) for inexperienced readers versus 3.2% (95%CI -1.9% to 8.3%) for experienced, with a mean difference of 7.9% (95%CI -0.9% to 16.6%) between the two groups (Table 13).

Table 13: Per-patient results for net benefit of CAD assistance when used in concurrent mode for interpretation of CT colonography by inexperienced and experienced readers. For all comparisons differences are calculated as performance with CAD assistance minus performance when unassisted. All data are percentages. Using CAD changed the proportion of readers correctly identifying cases with polyps in 83% of cases for both inexperienced and experienced readers; detection of patients with polyps increased in 70% and 57% of cases over 10 and 16 readers respectively.

	Inexperienced readers [mean % (95%CI)]	Experienced readers [mean % (95%CI)]	Difference: Inexperienced – Experienced [mean % (95%CI)]
CAD net benefit measure (all polyps)	11.2 (3.1 to 18.9)	3.2 (-1.9 to 8.3)	7.9 (-0.9 to 16.6)
Unassisted sensitivity (all polyps)	39.1 (30.9 to 47.0)	57.5 (49.6 to 65.2)	-18.5 (-25.3 to -11.9)
Unassisted specificity (all polyps)	94.1 (90.0 to 97.4)	91.0 (87.0 to 94.8)	3.1 (-1.7 to 7.9)
Sensitivity with CAD (all polyps)	53.2 (43.9 to 61.4)	62.1 (54.1 to 70.3)	-8.9 (-16.6 to -1.9)
Specificity with CAD (all polyps)	88.0 (82.2 to 93.3)	88.3 (83.8 to 92.4)	-0.3 (-5.6 to 5.0)
Change in sensitivity with CAD (all polyps)	14.1 (6.8 to 21.4)	4.6 (-0.2 to 9.3)	9.6 (1.2 to 17.7)
Change in specificity with CAD (all polyps)	-6.1 (-12.0 to -0.2)	-2.7 (-6.3 to 0.8)	-3.4 (-9.6 to 3.0)
CAD net benefit measure (polyps ≥6mm)	9.9 (0.1 to 19.2)	3.7 (-2.6 to 10.1)	6.1 (-4.0 to 15.6)
Unassisted sensitivity (polyps ≥6mm)	49.5 (40.0 to 58.9)	65.9 (56.4 to 74.7)	-16.4 (-24.0 to -8.3)
Unassisted specificity (polyps ≥6mm)	92.6 (89.0 to 95.5)	93.5 (90.5 to 95.9)	-0.9 (-4.4 to 2.5)
Sensitivity with CAD (polyps ≥6mm)	61.1 (50.0 to 71.1)	70.1 (60.5 to 78.7)	-9.0 (-18.4 to -0.3)
Specificity with CAD (polyps ≥6mm)	89.2 (84.7 to 92.8)	92.7 (89.0 to 95.5)	-3.5 (-7.4 to 0.2)
Change in sensitivity with CAD (polyps ≥6mm)	11.6 (1.9 to 20.5)	4.2 (-2.0 to 10.5)	7.5 (-2.6 to 16.8)
Change in specificity with CAD (polyps ≥6mm)	-3.4 (-8.0 to 0.8)	-0.8 (-3.4 to 1.7)	-2.6 (-7.5 to 2.0)

With the analyses restricted to patients with polyps ≥6mm there was a significant mean rise in sensitivity with CAD of 11.6% for inexperienced readers (rising from 49.5% unassisted to 61.1%) compared with a mean rise of 4.2% for experienced readers (rising from 65.9% to 70.1%), which was not significant (Table 13). In this case, however, the fall in specificity with CAD was non-significant for both groups, with a mean fall of 3.4% for inexperienced readers and 0.8% for experienced readers. Thus,

in a series of 200 patients (100 with polyps) inexperienced readers using CAD would on average correctly identify 11 or 12 additional patients with polyps, at the cost of approximately 3 or 4 additional false-positives, whereas experienced readers would identify 4 or 5 additional patients with polyps at the cost of 1 additional false-positive. Net benefit was significant for inexperienced readers (9.9%, 95% CI 0.1% to 19.2%) but not experienced readers (3.7%, 95%CI -2.6 to 10.1) with a non-significant difference between groups (6.1%, 95%CI -4.0% to 15.6%) (Table 13).

Per-polyp analysis

Per-polyp sensitivity for readers when unassisted and when using CAD are shown in table 14. For all polyps there was a significant mean rise in sensitivity with CAD of 9.0% for inexperienced readers (rising from 15.4% unassisted to 24.4%) compared with a mean rise of 4.1% for experienced readers (rising from 30.3% to 34.4%), which was also significant. Restricting analysis to polyps ≥ 6 mm the mean rise of 10.0% (rising from 28.5% to 38.5%) for inexperienced readers was significant but the mean rise of 3.0% (rising from 51.0% to 54.0%) for experienced readers was not. When the analysis was restricted to polyps ≤ 5 mm the mean rise in sensitivity with CAD was significant for both groups, 8.3% (rising from 5.9% to 14.2%) for inexperienced readers and 4.8% (15.3% rising to 20.1%) for experienced readers. The magnitude of benefit with CAD was not significantly different between the two groups.

Table 14: Per-polyp sensitivity for CAD assistance when used in concurrent mode for interpretation of CT colonography by inexperienced and experienced readers. For all comparisons differences are calculated as performance with CAD assistance minus performance when unassisted. All data are percentages.

	Novice readers [mean % (95%CI)]	Experienced readers [mean % (95%CI)]	Difference (Novice – Experienced)
Unassisted sensitivity (all polyps)	15.4 (11.3 to 20.8)	30.3 (23.9 to 37.7)	-14.9 (-19.6 to -10.5)
Sensitivity with CAD concurrent (all polyps)	24.4 (18.8 to 31.3)	34.4 (27.4 to 42.5)	-10.0 (-14.7 to -5.4)
Change in sensitivity with CAD (all polyps)	9.0 (5.1 to 13.2)	4.1 (1.0 to 7.5)	4.9 (0.3 to 9.5)
Unassisted sensitivity (polyps ≥6mm)	28.5 (20.2 to 36.9)	51.0 (40.4 to 60.9)	-22.5 (-29.9 to -14.7)
Sensitivity with CAD (polyps ≥6mm)	38.5 (29.7 to 48.3)	54.0 (43.0 to 64.7)	-15.5 (-23.0 to -7.6)
Change in sensitivity with CAD (polyps ≥6mm)	10.0 (3.0 to 17.3)	3.0 (-2.1 to 8.7)	7.0 (-1.2 to 14.7)
Unassisted sensitivity (polyps ≤5mm)	5.9 (3.0 to 10.0)	15.3 (10.4 to 21.2)	-9.3 (-14.3 to -5.6)
Sensitivity with CAD (polyps ≤5mm)	14.2 (8.4 to 21.4)	20.1 (13.9 to 28.0)	-5.9 (-10.6 to -0.9)
Change in sensitivity with CAD (polyps ≤5mm)	8.3 (4.1 to 13.6)	4.8 (1.5 to 8.7)	3.5 (-1.2 to 8.9)

Second-read CAD

Data for second-read CAD were only available for experienced readers and are shown in table 15 with 95%CI. There was a significant rise in mean sensitivity of 6.9% for patients with all polyps (rising from 57.5% to 64.4%), with a non-significant fall in mean specificity of -2.0% (falling from 91.0% to 89.0%). Thus in a series of 200 patients (100 with polyps) experienced readers would identify 6 or 7 additional patients with polyps on average, at a cost of 2 additional false-positives. These data gave a significant CAD net benefit of 6.0 (95%CI 1.2 to 10.5). Mean per-patient sensitivity for patients with polyps ≥6mm rose significantly by 6.9% also, with a non-significant fall in specificity of -0.9%. Per-polyp sensitivity rose significantly by a mean of 7.2% for all polyps, with

significant gains in mean sensitivity of 9.1% for polyps $\geq 6\text{mm}$ and 5.8% for polyps $\leq 5\text{mm}$.

Table 15: Effect of CAD assistance when used in second-read mode for interpretation of CT colonography by experienced readers. For all comparisons differences are calculated as performance with CAD assistance minus performance when unassisted. All data are percentages.

	Per-Patient analysis [mean % (95%CI)]	Per-polyp analysis [mean % (95%CI)]
CAD net benefit measure (all polyps)	6.0 (1.2 to 10.5)	n/a
Unassisted sensitivity (all polyps)	57.5 (49.6 to 65.2)	30.3 (23.9 to 37.7)
Unassisted specificity (all polyps)	91.0 (87.0 to 94.8)	n/a
Sensitivity with CAD (all polyps)	64.4 (56.6 to 72.3)	37.5 (29.5 to 46.1)
Specificity with CAD (all polyps)	89.0 (84.1 to 93.3)	n/a
Change in sensitivity with CAD (all polyps)	6.9 (2.8 to 11.2)	7.2 (3.9 to 10.6)
Change in specificity with CAD (all polyps)	-2.0 (-6.2 to 1.6)	n/a
CAD net benefit measure (polyps $\geq 6\text{mm}$)	6.6 (1.2 to 11.9)	n/a
Unassisted sensitivity (polyps $\geq 6\text{mm}$)	65.9 (56.4 to 74.7)	51.0 (40.4 to 60.9)
Unassisted specificity (polyps $\geq 6\text{mm}$)	93.5 (90.5 to 95.9)	n/a
Sensitivity with CAD (polyps $\geq 6\text{mm}$)	72.8 (63.3 to 81.4)	60.1 (48.9 to 70.4)
Specificity with CAD (polyps $\geq 6\text{mm}$)	92.6 (89.0 to 95.6)	n/a
Change in sensitivity with CAD (polyps $\geq 6\text{mm}$)	6.9 (1.9 to 12.5)	9.1 (3.8 to 13.8)
Change in specificity with CAD (polyps $\geq 6\text{mm}$)	-0.9 (-3.7 to 1.8)	n/a
Unassisted sensitivity (polyps $\leq 5\text{mm}$)	n/a	15.3 (10.4 to 21.2)
Sensitivity with CAD (polyps $\leq 5\text{mm}$)	n/a	21.1 (14.3 to 29.7)
Change in sensitivity with CAD (polyps $\leq 5\text{mm}$)	n/a	5.8 (2.3 to 9.7)

Second-read CAD was not tested with inexperienced readers but we can expect at least a similar impact to that seen in experienced readers, with second-read CAD likely to confer increased net benefit. Using second-read CAD experienced readers achieved an average sensitivity 2.5% above that when using concurrent CAD (6.9% increase with second-read, 4.6% increase with concurrent read; tables 13 and 15). Specificity for experienced readers rose by 0.7% using second-read CAD compared to concurrent reading (-2.0% change for second-read versus -2.7% change for concurrent; table 13 and 15). Conservative estimates suggest a significant rise in sensitivity for inexperienced readers of 16.6% (14.1% plus 2.5%; table 13) with a potentially significant fall in specificity of approximately -5.5% (-6.1% plus +0.7%; table 13).

Other analyses

It is possible to fortuitously achieve a true-positive diagnosis for a patient when assigning a false-positive polyp diagnosis while simultaneously failing to identify a true polyp. The mean number of such patients was 4.3 for both experienced and inexperienced readers when unassisted, falling with CAD to 3.9 for experienced readers and rising to 5.0 for inexperienced readers. Thus the proportion of such patients was small and the effect on the analyses overall was negligible; i.e. increased sensitivity with CAD was not due to false-positive detections in patients with true polyps elsewhere.

When unassisted, mean reading time for inexperienced readers was 11.2 min (95%CI 10.7 to 11.7) compared with 7.9 min (7.4 to 8.2) for experienced readers. When using CAD concurrently, this fell to 8.9 (8.3 to 9.4) for inexperienced readers but rose to 8.7 (8.2 to 9.3) for experienced readers.

Discussion

This study aimed to quantify the incremental benefit of CAD for inexperienced versus experienced readers; both groups read the same CT colonography data using a concurrent CAD paradigm. Our primary outcome was a weighted combination of the change in sensitivity and specificity when using CAD to detect patients with polyps of any size. We found that inexperienced readers achieved a significant net benefit with CAD of 11.2% but experienced readers only achieved 3.2%; the net benefit for inexperienced readers was nearly four times that achieved by experienced readers. This occurred despite a significant fall in specificity with CAD for inexperienced readers (a phenomenon not seen with experienced readers), confirming that the rise in sensitivity outweighed correspondingly diminished specificity when using our weighted

analysis. For both groups the impact of CAD was spread across 83% of cases with polyps, indicating that benefit was not confined to a small number of pivotal cases and suggesting that our findings are generalisable. Analysis restricted to patients with polyps $\geq 6\text{mm}$ similarly found net benefit greatest for inexperienced readers. Per-polyp analyses found that inexperienced readers achieved gains in sensitivity when CAD-assisted for polyps of all sizes and also when restricted to polyps $\geq 6\text{mm}$ and $\leq 5\text{mm}$. Experienced readers also achieved significant gains in sensitivity for the “all polyps” and “ $\leq 5\text{mm}$ ” analyses but not polyps $\geq 6\text{mm}$.

Several studies have investigated the effect of CAD-assistance on inexperienced readers but direct comparisons with experienced readers are uncommon, possibly because experienced readers are difficult to recruit versus less experienced individuals who are often trainees and/or those who wishing to learn CT colonography. Mang and colleagues used a second-read paradigm (87), finding that CAD raised sensitivity for two inexperienced readers close to that achieved by two experienced readers. Our findings suggest that while CAD improves sensitivity for inexperienced readers, in isolation it cannot compensate for proper training and experience. For example, per-polyp sensitivity with CAD at the 6mm threshold was 38.5% for inexperienced readers versus 54% for experienced readers. Supporting this, a study of 6 inexperienced readers who had participated in a prior CAD study found that a single day of clinical training significantly increased subsequent sensitivity (89). Researchers have also investigated the role of CAD for training inexperienced readers (90).

Second-read CAD was restricted to experienced readers, as per Chapter 2, in whom we found a net benefit that was not seen for concurrent CAD, suggesting the second-read paradigm is more accurate. To recap, because the CAD marks made during each paradigm read were identical, the effect cannot be a consequence of fewer true-positive marks during concurrent reading. Decreased vigilance paid to the unannotated endoluminal surface (including unannotated polyps) during concurrent reading may be an explanation. Alternatively, correctly annotated polyps may have been dismissed inappropriately by the reader although it is unclear why this should apply more to the concurrent paradigm.

Other researchers have also found second-read CAD beneficial for experienced readers, using ROC AUC as the primary analysis (75). We did not test second-read CAD on inexperienced readers but it is plausible to expect at least a similar benefit to that observed in experienced readers. A conservative estimate might assume a similar magnitude of difference between concurrent and second-read paradigms to that observed for experienced readers. In reality, a larger net benefit is likely for inexperienced readers given that they achieved more benefit from concurrent CAD

than did experienced readers. This assumption suggests that with the second-read paradigm sensitivity for patients with any polyp would increase by approximately 16.6% with a decrease in specificity of approximately -5.5%.

Our primary outcome was based on detection of patients with any polyp because such patients are potential candidates for colonoscopy if a polyp reaches the diameter threshold for referral. We did not apply a threshold for our primary outcome because there is disagreement regarding which diameter should trigger referral (91). Moreover, 3 or more diminutive polyps alone attracts a higher CRADS score, also prompting colonoscopy (92). Also, since smaller polyps are more difficult to detect than larger polyps, the *a priori* expectation is that CAD will have most impact here.

We used the net benefit analysis outlined in Chapter 1 and used in Chapter 2, but with different values for W and p . For the present analysis our value of W was evidence-based, having been obtained from the experiment detailed in Chapter 3 rather than the overly conservative value of 3 used in Chapter 2; i.e. the value of 6 used here is twice the value based in conservative expert opinion. As per our findings in Chapter 3, weighting to favour sensitivity becomes much more pronounced when cancer is considered, rising to 2250. These data are similar to those found for mammographic screening where women will trade 500 false-positive diagnoses for a single additional true-positive cancer (12). The higher the weighting factor, the more falls in specificity are regarded as increasingly clinically irrelevant. The present experiment considered detection of polyps alone but, in reality, any colorectal cancer screening programme using CTC will inevitably encounter cancers as well as polyps (although the latter will be far more numerous). I believe that analysis of the present data would likely be more representative if W also included the chance of encountering cancer but it is unclear exactly how the two weightings should be combined to provide a single value for W (and it should be borne in mind that while CAD does indeed detect cancers, it is generally not intended/designed to do so). Perhaps a weighting that is based on the relative prevalence of polyps and cancer in the general population would be most appropriate? Given a polyp prevalence of 25% and cancer prevalence of 0.2%, with values for W of 6 and 2250, the overall value of W would rise to 24; i.e. $6 + 18$ since cancers are detected 125 times less frequently than polyps so $2250/125 = 18$. This issue requires more research.

Mean unassisted interpretation time was longer for inexperienced readers, by approximately three minutes. Although we might expect experienced readers to be quicker, it could be argued that accurate interpretation arises from slow, careful inspection. Concurrent CAD shortened interpretation time for inexperienced readers (by over two minutes) yet raised it for experienced readers (by just under a minute). It

is possible that when using CAD concurrently, inexperienced readers pay less attention to un-annotated endoluminal surface than when unassisted, suggesting an “over-reliance” on CAD. Experienced readers may be more aware that CAD can be inaccurate, although both groups were told in advance that CAD made both true-positive and false-positive marks, and may miss polyps.

This study does have limitations. Inexperienced participants read under “laboratory” conditions over a week whereas experienced readers’ interpretations occurred over a month, at their workplace. The difference in reading environment may have exerted an influence. The CAD algorithm was identical for both groups, with identical true-positive and false-positive marks for cases shared between studies, but the reading platform was different: inexperienced readers used an in-house interface whereas experienced readers used commercially-available workstations with the CAD algorithm integrated. Our assumption that second-read CAD would benefit inexperienced readers is based on direct comparison between groups using the concurrent paradigm coupled with the incremental benefit of second-read over concurrent for experienced readers. Although statistically plausible, our estimate remains speculative.

In summary, we found that concurrent CAD resulted in a significant beneficial net benefit when used by inexperienced readers to identify patients with any polyp by CT colonography. The net benefit was approximately three times the magnitude of that observed in experienced readers. Experienced readers derived more benefit from second-read CAD than concurrent CAD, suggesting that second-read CAD would also be more effective for inexperienced readers.

Chapter 5: Detection of extracolonic pathology by CT colonography: A discrete choice experiment of perceived benefits versus harms.

This Chapter has been published as:

Plumb AA, Boone D, Fitzke H, Helbren E, Mallett S, Zhu S, Yao GL, Bell N, Ghanouni A, von Wagner C, Taylor SA, Altman DG, Lilford R, Halligan S. Detection of extracolonic pathologic findings with CT colonography: A discrete choice experiment of perceived benefits versus harms. *Radiology* 2014;273:144-152.

Abstract

Purpose: To determine the maximum rate of false-positive diagnoses that patients and healthcare professionals were willing to accept in exchange for detection of extracolonic malignancy when using CT colonography (CTC) for colorectal cancer screening.

Materials and Methods: Following ethical approval and informed consent, 79 patients and 50 healthcare professionals undertook two discrete choice experiments where they chose between unrestricted CTC that examined intra- and extra-colonic organs or CTC restricted to the colon, across different scenarios. Test A detected one extracolonic malignancy per 600 cases with a false-positive rate varying across scenarios from 0% to 99.8%. One experiment examined radiological follow-up generated by false-positives while the other examined invasive follow-up. Intracolonic performance was identical for both tests. The median “tipping point” (maximum acceptable false-positive rate for extracolonic findings) was calculated overall and for both groups via bootstrapping.

Results: The median tipping-point for radiological follow-up occurred at a false-positive rate of >99.8% (IQR 10 to >99.8%); i.e. participants would tolerate at least a 99.8% rate of unnecessary radiological tests in order to detect an additional extracolonic malignancy. The median tipping-point for invasive follow-up occurred at a false-positive rate of 10% (IQR 2 to >99.8%). Tipping-points were significantly higher for patients (>99.8 vs 40 and >99.8 vs 5 respectively for the two experiments, both $p < 0.001$).

Conclusion: Patients and healthcare professionals are willing to tolerate high rates of false-positives by CTC in exchange for diagnosis of extracolonic malignancy. The actual specificity of screening CTC for extracolonic findings in clinical practice is likely to be highly acceptable to both patients and healthcare professionals.

Introduction

Diagnostic tests used for cancer screening programmes usually target a specific organ. However, when screening for colorectal cancer (CRC) with CT colonography (CTC) extracolonic abdominal and pelvic tissues are imaged unavoidably, potentially detecting disease in organs other than the primary target. For example, a systematic review of 24 studies estimated that approximately 20% of indeterminate renal masses detected by CTC ultimately prove malignant (93). While some of these findings will be clinically important, the majority will not. For example, the same systematic review estimated false-positive diagnoses of extracolonic malignancy by CTC in 4.6% men and 6.8% women (93). Clarification often requires further investigations including biopsy and even surgery, which may be worrisome, costly and occasionally harmful, all for no ultimate benefit in most patients. Balancing benefit and harm is important because further tests precipitated by extracolonic findings are common, occurring in 7 to 11% of all screenees following CTC (94-98). One series estimated average costs for additional tests at just under \$100 per patient screened (98). However, evidence suggests that more asymptomatic cancers are detected beyond the colon than within it: A retrospective study of 10,286 screenees found that CTC detected extracolonic malignancies in 0.35% versus colorectal cancer in 0.21% (99).

The benefits or otherwise of extracolonic imaging has been debated widely, and neither clinicians (100) nor policy-makers (101) are clear whether it is helpful or not for population screening. Furthermore, little research has investigated whether clinicians or their patients regard extracolonic imaging as desirable. Specifically, it is unclear how individual patients and healthcare professionals balance the possibility of detecting life-threatening extracolonic pathology against the larger chance of fruitless (or even harmful) testing precipitated by extracolonic findings. While qualitative studies have found that screening-age patients generally view the ability of CTC to visualise extracolonic organs as advantageous (55), we do not know at what point (if at all) perceived benefit is outweighed by the inconvenience, worry and risks of unnecessary further investigation. This point is important and the discrete choice methodology used in Chapter 3 could be used to elicit the trade-offs between extracolonic detections and potential harms. We therefore aimed to determine the maximum rate of false-positive diagnoses that patients and healthcare professionals were willing to accept in exchange for detection of extracolonic malignancy when using CTC for colorectal cancer screening.

Methods

Ethical committee approval was granted. All participants gave written informed consent. Opinions were elicited using a discrete choice experiment where participants chose between two alternatives, as per the methodology used in Chapter 3. Again, since discrete choice experiments are difficult to administer via postal questionnaires (63), we used face-to-face interviews primarily, providing background information for participants via an interactive laptop presentation.

Recruitment

Consecutive adults of CRC screening age (55 to 69 years), scheduled to attend for unrelated ultrasound investigations, were identified via the hospital booking system (UCLH). A research assistant mailed information and consent forms beforehand and responders were interviewed on their appointment day. Individuals with a history of CRC or undergoing investigations for suspected CRC were excluded to avoid bias(102). Additionally, radiologists, colorectal surgeons, gastroenterologists, specialist nurses and radiographers who requested, performed, or interpreted colorectal imaging were recruited via hospital e-mail. Participants were offered a £10 voucher. Because we had previously found no difference for healthcare professionals between online and face-to-face completion (Chapter 3), this group could complete the experiment online. All patient participants were interviewed face-to-face by either radiology research fellows (AP, DB, EH), or research assistants (HF, NB); see Acknowledgements for full names.

Attributes

The experiment focus was extracolonic false-positive diagnosis by CTC at CRC screening, described to participants as “false-alarms”. While some extracolonic findings are fully characterised via further imaging (e.g. ultrasound for indeterminate renal lesions), others require invasive tests (endoscopy, biopsy, even surgery). To address both situations, participants undertook two separate experiments. In the first (the “radiological testing” experiment), participants were told that false-positive extracolonic diagnosis would precipitate unnecessary further imaging. Participants were instructed to assume the rates of such imaging to be 50% ultrasound, 45% CT and 5% MRI, per published literature (98). Disadvantages of imaging were explained: Ultrasound and MRI were described as safe but carrying inconvenience and potentially anxiety. Noise and claustrophobia were described for MRI. CT was described as carrying a very small chance of cancer induction several years afterwards (103). In the second (the “invasive testing” experiment), false-positives led to biopsy/endoscopy/surgery. Participants were

instructed to assume that approximately 50% of “invasive tests” would be operative, 25% needle biopsy (either FNA or core) and 25% endoscopy (98). Pain, bleeding, perforation and a small risk of death were mentioned as possible complications. Participants were told that tests for “false-alarms” were unnecessary, i.e. the screenee derived no benefit from the ultimate diagnosis; that the hypothetical population being tested was asymptomatic; and that the focus was purely on findings outside the colon.

We presented two hypothetical tests: “Unrestricted CTC” evaluated both the colon and extracolonic organs whereas “restricted CTC” was confined to the colon. Participants were told that diagnostic accuracy for colonic neoplasia was identical for both tests (95). Unrestricted CTC was assumed to detect extracolonic malignancy at an early/curable stage in 1 in 600 cases (a conservative estimate derived from literature (99)), whereas restricted CTC did not. Participants were informed that the impact of early diagnosis on overall survival was unknown, and that even using unrestricted CTC, many extracolonic malignancies would remain undetected by a one-off screening examination; i.e. that detection did not necessarily result in cure. Across the scenarios presented, the specificity (false-positive rate) of unrestricted CTC was varied from 100% to 0.17%, corresponding to a rate of additional testing ranging between 0 and 599 extra tests to diagnose 1 extracolonic malignancy per 600 screenees (table 16). The “invasive testing” experiment included a scenario where the chance of death was directly equivalent to the chance of diagnosing extracolonic malignancy and carried the additional disadvantage of the near-certainty of an unnecessary invasive procedure.

Table 16: Attributes and levels presented in the “radiological testing” (1A) and “invasive testing” (1B) experiments.

Table 16A: “Radiological testing” experiment

“Radiological testing” question number	Unrestricted CTC – looks inside and outside the colon	Restricted CTC – only looks inside the colon		
	False positive rate	False positive rate	Curable extracolonic cancers detected per 600 screening examinations	Additional false positive detections per 600 screening examinations
1r	0	0	1	0
2r	0.17	0	1	1
3r*	4	0	1	24
4r*	4	0	1	24
5r	10	0	1	60
6r	20	0	1	120
7r	40	0	1	240
8r	60	0	1	360
9r	80	0	1	480
10r**	99.8	0	1	599

*Questions are identical to test for internal consistency

**Participants choosing unrestricted CTC in response to question 10 were considered non-traders

Table 16B: “Invasive testing” experiment

“Invasive testing” question number	Unrestricted CTC – looks inside and outside the colon	Restricted CTC – only looks inside the colon		
	False positive rate	False positive rate	Curable extracolonic cancers detected per 600 screening examinations	Additional false positive detections per 600 screening examinations
1i	0	0	1	0
2i	0.17	0	1	1
3i*	1	0	1	6
4i*	1	0	1	6
5i	2	0	1	12
6i	4	0	1	24
7i	10	0	1	60
8i	20	0	1	120
9i	40	0	1	240
10i	60	0	1	360
11i	80	0	1	480
12i	99.8	0	1	599
13i**	99.8	0	1	599 + 1 death

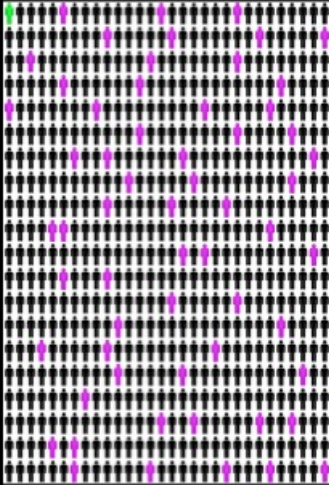
*Questions are identical to test for internal consistency

**Participants choosing unrestricted CTC in response to question 13 were considered non-traders

Experiment format

Background information regarding CRC, screening, CTC, and risk were presented via a multimedia presentation. Consequences of false-positive results were described in terms of need for additional imaging or biopsy (e.g. resection of indeterminate ovarian cysts). The nature of needle, endoscopic and surgical excision biopsy was explained. The chance of requiring an extra test was presented graphically and via text, using both absolute risks and natural frequencies to maximize understanding (104) (figure 11). Patients were told to assume they were asymptomatic and at average risk of both intracolonic and extracolonic pathology, and to pick the test they would choose for themselves or a close friend or relative, with no opt-out. Healthcare professionals were asked to pick the test they felt was best suited to population screening (rather than for their own care). The “radiological testing” experiment was performed first, followed by the “invasive testing” experiment. For each experiment, the differing false-positive rates for unrestricted CTC were presented in a non-sequential order (i.e. the rate did not rise or fall incrementally). One choice was repeated to test consistency. If inconsistent, the response was clarified with the participant; the second response was used for online participants. Unrestricted CTC was also presented with a zero false-positive rate and those choosing restricted CTC in this scenario labelled as “irrational” responders. If this occurred, the reason for choosing restricted CTC was recorded after qualitative exploration, but the response was retained for the subsequent analysis. Participants who preferred unrestricted CTC despite a false-positive rate of 599/600 in the “invasive testing” scenarios were presented with additional information emphasising potential harms (including death) arising from this situation. The risk of death was stated as 1 in 600 (105); participants were therefore choosing between a 1 in 600 chance of early extracolonic malignancy diagnosis versus an equivalent risk of death plus the near-certainty of unnecessary additional invasive procedures. Some data rounding was used in the scenarios when events were rare.

Figure 11: Example question from the “invasive testing” experiment. A hypothetical screening population of 600 individuals were presented and participants invited to choose between a test generating a variable rate of false-positives (pink) but a 1 in 600 chance of finding an early stage extracolonic malignancy (green) or a test that generated no false-positives but no chance of finding an extracolonic malignancy (yellow). Participants were informed that their chance of receiving any particular result could not be predicted in advance and was essentially random. Relative and absolute percentages were presented. This image corresponds to the median “tipping point” for patients and professionals combined: On average, unrestricted CTC (Test A on the Figure) was preferred to this level of false-positive invasive tests, but not beyond.

<p>Test A: Looks inside and outside the bowel</p> <p>For every 600 patients going for bowel cancer screening, one will have cancer outside the bowel which will be picked up at an early stage by this test.</p> <p>But 60 patients get worrying test results and need to go for extra tests such as biopsy or surgery</p> 	<p>Test B: Does not look outside the bowel.</p> <p>For every 600 patients going for bowel cancer screening, one will have an early cancer outside the bowel which cannot be picked up by this test.</p> <p>But no patient gets worrying test results or needs to go for extra tests such as biopsy or surgery.</p> 
--	---

Pilot testing

To confirm comprehensibility, estimate completion time, and inform sample size, 15 individuals (10 professionals, 5 patients) were piloted (data not included in the final analysis). This confirmed that attributes and levels, and the concept of choosing between two scenarios (“trading” test benefits versus harms) were comprehensible. However, simultaneous consideration of false-positives leading to both radiological follow-up and invasive testing was judged too confusing, explaining why these were presented ultimately as separate experiments.

Statistical analysis

The primary outcome measure was the maximum false-positive rate that the average participant was willing to accept in exchange for a 1 in 600 chance of diagnosing an extracolonic malignancy; the “tipping point”. The pilot suggested a mean acceptable

false-positive rate of 11% for invasive testing with a standard deviation of 0.23. To determine the median tipping point \pm 5% at two-sided alpha 0.05 and 90% power required 81 participants ($N=4\sigma^2(z_{crit})^2/D^2$ where $D=0.10$, $\sigma=0.23$ and $z_{crit}=1.960$) (72). Since pilot data were non-normal, we aimed to recruit a further 15%. A pre-specified secondary outcome to compare subgroups of patients versus healthcare professionals required 56 participants each for 90% power to detect a 20% difference in the maximum false-positive rate.

Participants' overall median tipping-point was calculated for each experiment, and for patients/professionals individually. Since numbers of patients and professionals differed, the overall tipping-point was calculated using 2000 bootstrap estimates of medians and interquartile ranges using equally sized samples from each group ($n=50$). Tipping-points were non-normal and were therefore summarised with medians and interquartile ranges. "Non-traders" were defined as participants who consistently chose one test over the alternative across all of the scenarios presented. Non-traders were therefore regarded as requiring higher tipping-points than were offered in the experiment but responses were still included. The Mann-Whitney U test and Wilcoxon signed rank sum test were used for unpaired and paired comparisons respectively. Data were collated using Microsoft Excel 2011 for Mac v14.3.4 (Microsoft Corp, Redmond, WA) and analysed with R version 2.15.1 (R Foundation for Statistical Computing, Wirtschaftsuniversität Wien, Welthandelsplatz 1, 1020 Vienna, Austria) by Drs. Susan Mallett and Andrew Plumb.

Results

318 patients and 96 healthcare professionals were invited. 79 patients (25%) responded positively but only 52 were interviewed due to scheduling conflicts. 50 professionals participated, a response rate of 52%; 21 (42%) were interviewed face-to-face and 29 (58%) responded online. On average, patients were older than professionals (median 64.5 vs. 29.5 years; $p<0.001$) and had discontinued education earlier (50% educated to degree level vs. 94%; $p<0.001$; Table 17). There were no significant differences in gender ratio ($p=0.54$) or ethnicity ($p=0.14$).

Table 17: Demographic characteristics of patient and professional participants.

Characteristic	Patients (n=52)	Professionals* (n=50)	Total (n=102)
	N (%)	N (%)	N (%)
Gender			
Male	27 (52)	29 (58)	56 (55)
Female	25 (48)	21 (42)	46 (45)
Age (years)			
<25	0 (0)	1 (2)	1 (1)
25-34	0 (0)	31 (62)	31 (30)
35-54	0 (0)	18 (36)	18 (18)
55-59	1 (2)	0 (0)	1 (1)
60-69	51 (98)	0 (0)	51 (50)
Ethnicity			
White	42 (81)	34 (68)	76 (75)
Other	10 (19)	16 (32)	26 (25)
Education			
Degree-level	26 (50)	47 (94)	73 (72)
Other	26 (50)	3 (6)	29 (28)

*Comprising 10 radiologists, 5 gastroenterologists, 4 surgeons, 19 registrars in these specialties, 2 specialist colorectal nurses and 10 radiographers.

Non-traders

For the “radiological testing” experiment, 61 participants of the 102 interviewed (60%; 41 patients, 20 professionals) were deemed non-traders (i.e. they always chose unrestricted CTC), 24 of whom (24% overall; 23 patients, 1 professional) were also non-traders for the “invasive testing” experiment. These 24 non-traders felt unable to choose restricted CTC, despite one scenario for unrestricted CTC presenting a risk of death equivalent to the chance of detecting an extracolonic malignancy. Conversely, a single patient participant never chose unrestricted CTC (even for the zero false-positive scenario), stating a firm opinion that colorectal cancer screening should examine the colon alone. On average, non-traders were significantly older (median age 64.5 vs. 29.5; $p < 0.001$), significantly more likely to be patients (41 [79%] vs. 20 [40%]; $p < 0.001$), and were less educated than traders (38 [62%] degree-level education vs. 35 [85%]; $p < 0.01$). There was no significant difference in gender (52% vs. 59% male; $p = 0.55$) or ethnicity (79% vs. 68% white; $p = 0.24$).

Radiological testing discrete choice experiment

When the consequence of extracolonic findings was radiological testing, the median tipping-point occurred at a false-positive rate of >99.8% (IQR 10 to >99.8%; Table 18). Thus, the average participant was prepared to tolerate at least a 99.8% rate of unnecessary additional radiological tests in order to diagnose a single additional extracolonic malignancy. The median tipping point was significantly higher for patients than professionals at >99.8 (IQR >99.8 to >99.8) versus 40 (IQR 10 to >99.8, $p < 0.001$; Table 18). Overall, at a prevalence of 1 in 600 for potentially curable extracolonic malignancy, this corresponds to >599 (IQR 60 to >599) unnecessary additional radiological tests to find one curable extracolonic malignancy. Patients were prepared to accept a significantly higher number of false-positives (median >599, IQR >599 to >599; Table 18, figure 12A) than professionals (median 240, IQR 60 to >599, $p < 0.001$; Table 18, figure 12B).

Table 18: Tipping points and number of FP deemed acceptable in each scenario, for patients, professionals, and the two groups combined.

	Tipping point (Maximum false positive rate acceptable to find one extracolonic cancer)		Average number of additional FP tolerated per additional extracolonic cancer found	
	Median	IQR	Median	IQR
PATIENTS				
Scans	>99.8	>99.8 to >99.8	>599	>599 to >599
Invasive tests	>99.8	20 to >99.8	>599	120 to >599
PROFESSIONALS				
Scans	40	10 to >99.8	240	60 to >599
Invasive tests	5	2 to 10	30	12 to 60
COMBINED				
Scans	>99.8	10 to >99.8	>599	60 to >599
Invasive tests	10	2 to >99.8	60	12 to >599

Figure 12: Cumulative plot of tipping-points expressed as absolute numbers of additional unnecessary tests for patients (A) and professionals (B) in the “radiological testing” experiment. Each grey dot shows an individual’s tipping-point. Large red square shows the median value, corresponding to “an average participant”. Blue squares show 25 and 75 percentage points.

Figure 12A:

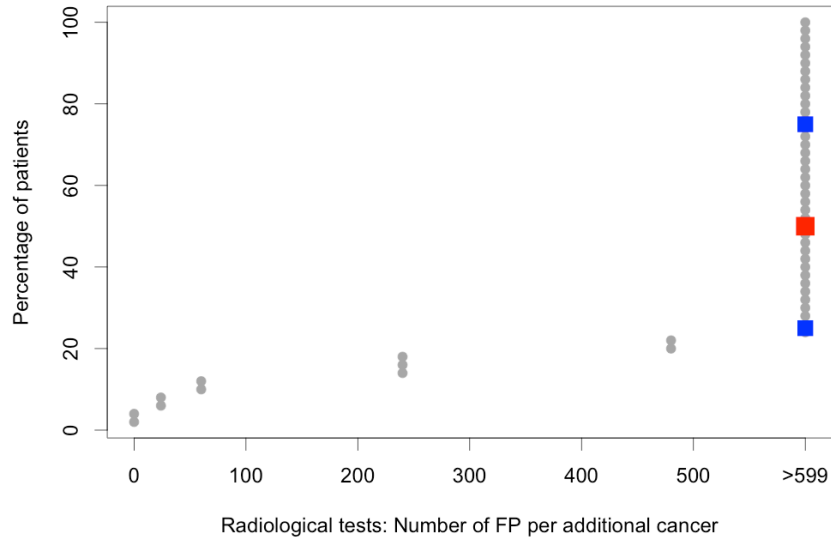
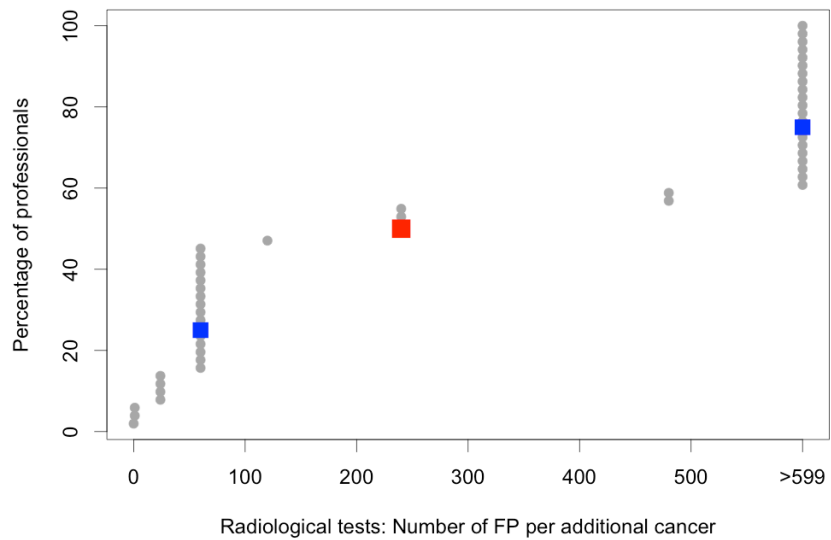


Figure 12B:



Invasive testing discrete choice experiment

When the consequence of extracolonic findings was invasive, the median tipping-point occurred at a false-positive rate of 10% (IQR 2 to >99.8%). Thus, the average participant was prepared to tolerate a 10% rate of unnecessary additional invasive tests in exchange for diagnosis of a single extracolonic malignancy. The median tipping point was significantly higher for patients than professionals at >99.8 (IQR 20 to >99.8) versus 5 (IQR 2 to 10, $p < 0.001$; Table 18). Overall, at population prevalence of 1 in 600, this corresponds to 60 (IQR 12 to >599) additional invasive tests per extracolonic malignancy. Again, patients were prepared to tolerate higher numbers of false-positives (median >599, IQR 120 to >599 plus risk of death; Table 18, figure 13A) than professionals (median 30, IQR 12 to 60, $p < 0.001$; Table 18, figure 13B).

Figure 13: Cumulative plot of tipping-points expressed as numbers of additional unnecessary tests for (A) patients and (B) professionals in the “invasive testing” experiment: Color-codes are the same as in figure 12.

Figure 13A

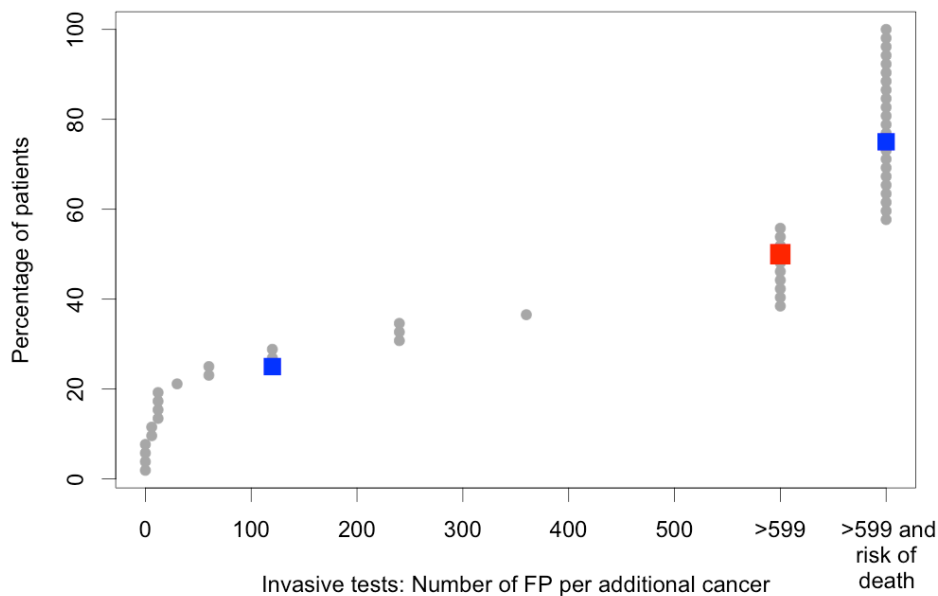
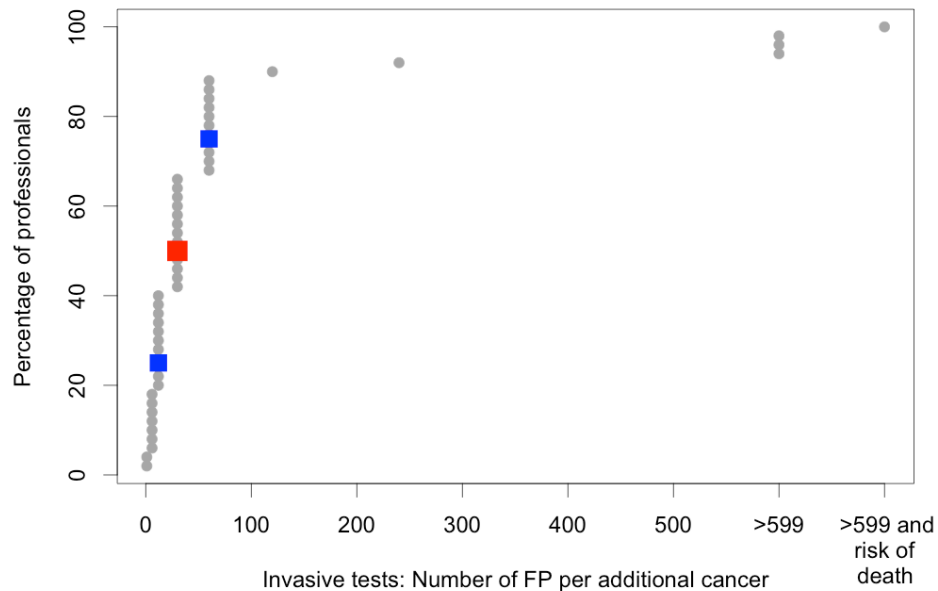


Figure 13B



The median number of false-positives tolerated per extracolonic malignancy was significantly higher for the “radiological testing” than the “invasive testing” experiment ($p < 0.001$), demonstrating that additional imaging were deemed more acceptable than additional invasive tests.

There was no significant difference in the median tipping-points for patients interviewed by either radiologists or research assistants ($p = 0.57$), or between professionals who gave their responses online as opposed to face-to-face ($p = 0.81$).

Discussion

Extracolonic findings at CTC present a clinical dilemma: Early diagnosis of important pathology might be curative but unnecessary investigation of ultimately irrelevant findings has physical, psychological and financial costs. How patients and healthcare professionals balance these costs is not known with precision, hence our decision to conduct these experiments. We found that patients were prepared to tolerate an extremely high rate (>99.8%) of unnecessary additional imaging or invasive tests subsequent to screening CTC in order to reap the potential benefits of finding early-stage extracolonic malignancy. While healthcare professionals were less tolerant of

unnecessary follow-up testing than patients, nevertheless we were surprised by the very high rates they accepted; on average 40% was deemed acceptable for radiological follow-up. Invasive tests were deemed less acceptable by healthcare professionals, with the median “tipping point” at 5%. Nevertheless, these rates are substantially greater than occur in published series, where approximately 7-11% of individuals required further diagnostic testing following CTC with only 1-2% requiring an invasive test (98). Our data therefore suggest that the specificity of screening CTC for extracolonic malignancy in current clinical practice is likely to be highly acceptable to both patients and healthcare professionals.

Previous studies quantifying patients’ perceptions of false-positives have shown an overwhelming preference for improved sensitivity despite the costs of diminished specificity. When considering breast cancer screening with mammography, one study (12) found that 63% of women were prepared to accept 500 or more additional false-positive mammograms to save a single life by early breast cancer diagnosis. Similarly, in Chapter 3 we found patients would tolerate over 4000 false-positive diagnoses to avoid a single missed colorectal cancer. We found patients equally tolerant in the context of extracolonic malignancy, confirming that potentially lifesaving detection of pathology outweighs the perceived disadvantages of subsequent testing irrespective of the organ being evaluated.

Our data, while framed in the context of colorectal cancer screening, potentially have wider implications for incidental findings discovered by other imaging modalities. Such “incidentalomas” are common in CT urography, CT coronary angiography, thoracic CT when screening for lung cancer, and abdominopelvic CT, provoking the American College of Radiology white paper for management guidance (106). Our results suggest that both patients and healthcare professionals are likely to tolerate additional work-up in these different clinical settings. We chose a prevalence of incidentally-detected extracolonic malignancy of 1 in 600, based on available data for CTC screening populations (99); we would expect tipping-points to differ if this prevalence changed. For example, our 10% invasive test rate deemed acceptable overall might be unacceptably high if the chance of uncovering an unexpected malignancy fell to 1 in 2,000 (as was found in one large lung cancer screening trial (107)).

As explained in Chapter 3, when completing simple rank-order preferences, patients choose the most accurate, comfortable, convenient and cheapest test (80). In contrast, discrete choice experiments reflects “real-world” choices, where different attributes must be traded against each other (63, 80). Such experiments are complex to comprehend, so we interviewed patient participants face-to-face, provided extensive background information using laptops and graphical aids, and were available to answer

questions. Despite this, many patients made the (apparently irrational) decision to choose unrestricted CTC for the scenario where it was presented alongside a risk of death from unnecessary further investigations equal to the chance of detecting extracolonic malignancy, and the near-certainty of an unnecessary invasive procedure. Such counter-intuitive behavior may arise from prior belief (108); many patients could not deviate from their conviction that early cancer diagnosis is always preferable, irrespective of associated risks. Others stated an unwillingness to live with the uncertainty of not knowing whether an equivocal extracolonic lesion was significant or not, apparently overlooking the fact that such uncertainty could be avoided entirely by choosing restricted CTC, which was unable to image beyond the colon.

Limitations of our study include the necessary simplifications required to render discrete choice experiments comprehensible, just as for the experiments detailed in Chapter 3. For example, attributes of intracolonic performance, bowel preparation, discomfort during the examination, and radiation dose were not included other than as background information; increasing the number of attributes and their levels renders the experiment impracticable in terms of both cognitive complexity and the number of individual scenarios required to achieve statistical power. Our piloting confirmed this, so we examined radiological and invasive follow-up via separate experiments whereas, in reality, a false-positive result might generate both. Also, participants' responses may not reflect their real-life behavior, a limitation of all stated-preference methodologies. We could not estimate an upper boundary to patients' tolerance of false-positives because the non-trade rate was so high; higher than suggested by the pilot. Consequently, estimates of the median tipping point had a broader IQR than anticipated.

Future researchers should consider treating patients and professionals as separate groups from the outset, and power accordingly. For example, healthcare workers requesting screening tests will tend to be younger than the patients receiving them; Indeed, Table 17 shows that all professionals were aged 54 years or younger while all patients were older (because we wished to reflect screening age). It is possible that as professionals age and become more susceptible to cancer themselves, then their opinions might converge with those of patients. However, we believe the reason that professionals were less tolerant of false-positives was more influenced by their superior medical knowledge than by their younger age. Our current sample reflects the demographics of the two groups but it would be interesting in future to solicit the views of retired healthcare professionals who are themselves of screening age, to see if their opinion has changed. It may also be informative to test community physicians since our healthcare professionals were all hospital based. The response rate for patients was

also lower than expected, perhaps because of cognitive complexity. Costs arising from additional testing were not investigated and it is possible that responses may have been different if patients were responsible for these costs; the experiments were carried out in the English National Health Service where patients do not bear costs directly. Finally, we studied outpatients attending hospital for unrelated investigations, who may not fully represent an unselected screening population.

In summary, via discrete choice experiment we found that both patients and healthcare professionals believe diagnosis of extracolonic malignancy by screening CTC greatly outweighs the potential disadvantages of subsequent radiological or invasive investigation precipitated by false-positives. This belief was held more strongly by patients than healthcare professionals. The specificity of CTC for extracolonic malignancy in clinical practice is likely to be highly acceptable to both patients and healthcare professionals.

Chapter 6: Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: Systematic review with a focus on quality of data reporting.

This Chapter has been published as:

Dendumrongsup T, Plumb AA, Halligan S, Fanshawe TR, Altman DG, Mallett S. Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: Systematic review with a focus on the quality of data reporting. PLoS One. 2014 Dec 26;9(12):e116018.

Abstract

Introduction: We examined the design, analysis and reporting in multi-reader multi-case (MRMC) research studies using the area under the receiver-operating curve (ROC AUC) as a measure of diagnostic performance.

Methods: We performed a systematic literature review from 2005 to 2013 inclusive to identify a minimum 50 studies. Articles of diagnostic test accuracy in humans were identified via their citation of key methodological articles dealing with MRMC ROC AUC. Two researchers in consensus then extracted information from primary articles relating to study characteristics and design, methods for reporting study outcomes, model fitting, model assumptions, presentation of results, and interpretation of findings. Results were summarized and presented with a descriptive analysis.

Results: Sixty-four full papers were retrieved from 475 identified citations and ultimately 49 articles describing 51 studies were reviewed and extracted. Radiological imaging was the index test in all. Most studies focused on lesion detection vs. characterization and used less than 10 readers. Only 6 (12%) studies trained readers in advance to use the confidence scale used to build the ROC curve. Overall, description of confidence scores, the ROC curve and its analysis was often incomplete. For example, 21 (41%) studies presented no ROC curve and only 3 (6%) described the distribution of confidence scores. Of 30 studies presenting curves, only 4 (13%) presented the data points underlying the curve, thereby allowing assessment of extrapolation. The mean change in AUC was 0.05 (-0.05 to 0.28). Non-significant change in AUC was attributed to underpowering rather than the diagnostic test failing to improve diagnostic accuracy.

Conclusions: Data reporting in MRMC studies using ROC AUC as an outcome measure is frequently incomplete, hampering understanding of methods and the reliability of results and study conclusions. Authors using this analysis should be encouraged to provide a full description of their methods and results.

Introduction

Chapter 1 of this thesis describes the several difficulties we encountered when trying to implement ROC AUC as the primary outcome measure in a prior study of CT colonography. Many of these difficulties were related to issues implementing confidence scores in a transparent and reliable fashion, which led ultimately to a flawed analysis. We considered, therefore, that for ROC AUC to be a valid measure there are methodological components that need addressing in study design, data collection and analysis, and interpretation. Based on our attempts to implement the MRMC ROC AUC analysis, we were interested in whether other researchers have encountered similar hurdles and, if so, how they had tackled these. Therefore, in order to investigate how often other studies have addressed and reported on issues with ROC AUC, we performed a systematic review of MRMC studies using this analysis as an outcome measure. We searched and investigated the available literature to determine the statistical methods used, the completeness of data presentation, and whether any problems with analysis were encountered and reported.

Methods

Search strategy, inclusion and exclusion criteria

This systematic review was performed guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses (109). We developed an extraction sheet for the systematic review, broken down into different sections (used as subheadings for the Results section of this report), with notes relating to each individual item extracted. In consensus we considered approximately 50 articles would provide a sufficiently representative overview of current reporting practice. Based on our prior experience of performing systematic reviews we believed that searching for additional articles beyond 50 would be unlikely to yield valuable additional data (i.e. we believed we would reach “saturation” by 50 articles) yet would present a very considerable extraction burden.

In order to achieve this, potentially eligible primary articles published between 2005 and February 2013 inclusive were identified by a radiologist researcher (TD) using PUBMED via their citation of one or more of 8 key methodological articles relating to MRMC ROC AUC analysis (110-117). To achieve this, the Authors' names (combined using “AND”) were entered in the PUBMED search field and the specific article identified and clicked in the results list. The abstract was then accessed and the “Cited

By # PubMed Central Articles” link and “Related Citations” link used to identify those articles in the PubMed Central database that have cited the original article. There was no language restriction. Online abstracts were examined in reverse chronological order, the full text of potentially eligible papers then retrieved, and selection stopped once the threshold of 50 studies fulfilling inclusion criteria had been passed.

To be eligible, primary studies had to be diagnostic test accuracy studies of human observers interpreting medical image data from real patients, and attempting to use a MRMC ROC AUC analysis as a study outcome, based on the following methodological approaches (110-117); Reviews, solely methodological papers, and those using simulated imaging data were excluded.

Data extraction

An initial pilot sample of 5 full-paper articles were extracted and the data checked by a subgroup of investigators in consensus, to both confirm the process was feasible and to identify potential problems. A further 10 full-papers were extracted by two radiologist researchers working independently (Thaworn Dendomrungsup, Andrew Plumb) to check agreement further. The remaining articles were extracted predominantly by TD, who discussed any concerns/uncertainty with AP. Any disagreement following their discussion was arbitrated by the author and/or Susan Mallett where necessary.

The extraction covered the following broad topics: Study characteristics, methods to record study outcomes, model assumptions, model fitting, data presentation.

We extracted data relating to the organ and disease studied, the nature of the diagnostic task (e.g. characterisation vs. localisation vs. presence/absence), test methods, patient source and characteristics, study design (e.g. prospective/retrospective, secondary analysis, single/multicenter) and reference standard. We extracted the number of readers, their prior experience, specific interpretation training for the study (e.g. use of CAD software), blinding to clinical data and/or reference results, the number of times they read each case and the presence of any washout period to diminish recall bias, case ordering, and whether all readers read all cases (i.e. a fully-crossed design). We extracted the unit of analysis (e.g. patient vs. organ vs. segment), and sample size for patients with and without pathology.

We noted whether study imaging reflected normal daily clinical practice or was modified for study purposes (e.g. restricted to limited images). We noted the confidence scores used for the ROC curve and their scale, and whether training was provided for scoring. We noted if there were multiple lesions per unit of analysis. We

noted if scoring differed for positive and negative patient cases, whether score distribution was reported, and whether transformation to a normal distribution was performed.

We extracted if ROC curves were presented in the published article and, if so, whether for individual readers, whether the curve was smoothed, and if underlying data points were shown. We defined unreasonable extrapolation as an absence of data in the right-hand 25% of the plot space. We noted the method for curve fitting and whether any problems with fitting were reported, and the method used to compare AUC or pAUC. We extracted the primary outcome, the accuracy measures reported, and whether these were overall or for individual readers. We noted the size of any change in AUC, whether this was significant, and made a subjective assessment of whether significance could be attributed to a single reader or case. We noted how the study authors interpreted change in AUC, if any, and whether any change was reported in terms of effect on individual patients. We also noted if a ROC researcher was named as an author or acknowledged, defined as an individual who had published indexed research papers dealing with ROC methodology.

Analysis

Data were summarised in an Excel worksheet (Excel For Mac 14.3.9, Microsoft Corporation) with additional cells for explanatory free text. The author then compiled the data and extracted frequencies, consulting the two radiologists who performed the extraction (TD and AP) for clarification when necessary. The investigator group discussed the implication of the data subsequently, to guide interpretation.

Results

Four-hundred and seventy-five citations of the 8 key methodological papers were identified and 64 full papers retrieved subsequently. Fifteen (118-132) of these were rejected after reading the full text (Table 19) leaving 49 (75, 133-180) for extraction and analysis. Two papers (161, 175) contributed two separate studies each, meaning that 51 studies were extracted in total.

Table 19: Citations for the 49 papers (contributing 51 studies) included in the systematic review. Details are also provided for the 15 articles excluded from the systematic review after reading the full-text, along with primary reasons for their exclusion (multiple reasons for exclusion were possible).

Included Articles (first author, year)	
Aoki, 2011 (133)	
Aoki, 2012 (134)	
Berg, 2012 (135)	
Bilello, 2010 (136)	
Choi, 2012 (137)	
Cole, 2012 (138)	
Collettini, 2012 (139)	
Dachman, 2010 (75)	
Dromain, 2012 (140)	
Gennaro, 2010 (141)	
Hupse, 2013 (142)	
Kelly, 2010 (143)	
Kim, 2012 (144)	
Kim, 2010 (145)	
Li, 2012 (146)	
Li, 2011a (147)	
Li, 2011b (148)	
Matsushima, 2010 (149)	
McNulty, 2012 (150)	
Medved, 2011 (151)	
Mermuys, 2010 (152)	
Moin, 2010 (153)	
Muramatsu, 2010 (154)	
Noroozian, 2012 (155)	
Ohqiya, 2012 (156)	
Otani, 2012 (157)	
Padilla, 2013 (158)	
Pollard, 2012 (159)	
Purysko, 2012 (160)	
Rafferty, 2013 (161)	Contributed two studies
Saade, 2013 (162)	
Salazar, 2011 (163)	
Shimauchi, 2011 (164)	
Shiraishi, 2010 (165)	
Subhas, 2011 (166)	
Sung, 2010 (167)	
Svahn, 2012 (168)	
Takahashi, 2010 (169)	
Tan, 2012 (170)	
Timp, 2010 (171)	
Toomey, 2010 (172)	
Uchiyama, 2012 (173)	
Visser, 2012 (174)	
Wallis, 2012 (175)	Contributed two studies
Wardlaw, 2010 (176)	
Way, 2010 (177)	
Yamada, 2011 (178)	
Yamada, 2012 (179)	
Yoshida, 2013 (180)	

Excluded Articles (first author, year)	Primary reason for exclusion
Warren, 2012 (118)	Simulated imaging
Berbaum, 2012 (119)	Simulated imaging
Destounis, 2011 (120)	No MRMC ROC analysis
Jinzaki, 2011 (121)	No MRMC ROC analysis
Krupnski, 2012 (122)	Simulated imaging
Leong, 2012 (123)	Simulated imaging
Nishida, 2011 (124)	No MRMC ROC analysis
Obuchowski, 2010 (125)	Non sequential design
Okamoto, 2011 (126)	No human readers
Reed, 2011 (127)	Simulated imaging
Svane, 2011 (128)	No MRMC ROC analysis
Szucs-Farkas, 2010 (129)	No MRMC ROC analysis
Webb, 2011 (130)	Simulated imaging
Yakabe, 2010 (131)	Simulated imaging
Zanca, 2012 (132)	Not a primary study with original data

Study characteristics

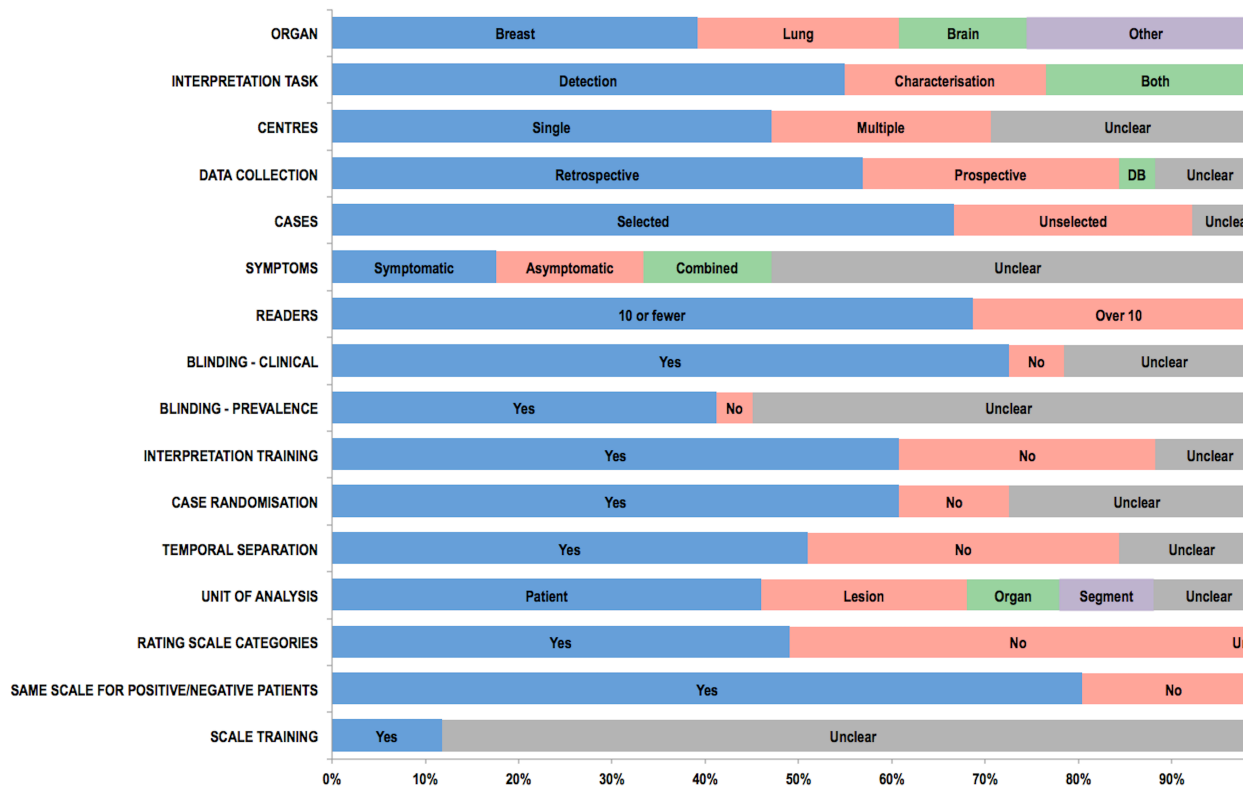
The index test was imaging in all studies. Breast was the commonest organ studied (20 studies), followed by lung (11 studies) and brain (7 studies). Mammography (15 studies) was the commonest individual modality investigated, followed by plain film (12 studies), CT and MRI (11 studies each), tomosynthesis (six studies), ultrasound (two studies) and PET (one study); nine studies investigated multiple modalities. In most studies (28 studies) the prime interpretation task was lesion detection. Eleven studies focused on lesion characterisation and 12 combined detection and characterisation. Most studies compared 2 tests/conditions (i.e. a single test but implemented in different ways) to a reference standard (41 studies), with 2 studies comparing 1 test/condition, 7 studies comparing 3 tests/conditions, and 1 study comparing 4 tests/conditions. Twenty-five studies combined data to create a reference standard while the reference was a single finding in 24 (14 imaging, 5 histology, 5 other – e.g. endoscopy). The reference method was unclear in 2 studies (154, 155).

Twenty-four studies were single center, 12 multicenter, with the number of centers unclear in 15 (29%) studies. Nine studies recruited symptomatic patients, 8 asymptomatic, and 7 a combination, but the majority (53%; 27 studies) did not state whether patients were symptomatic or not. 42 (82%) studies described the origin of patients with half of these stating a precise geographical region or hospital name. However, 9 (18%) studies did not sufficiently describe the source of patients and 21 (41%) did not describe patients' age and/or gender distribution.

Study design

Extracted data relating to study design and readers are presented graphically in Figure 14. Most studies (29; 57%) used patient data collected retrospectively. Fourteen (28%) were prospective while 2 used an existing database. Whether prospective/retrospective data was used was unstated/unclear in a further 6 (12%). While 13 studies (26%) used cases unselected other than for the disease in question, the majority (34; 67%) applied further criteria, for example to preselect “difficult” cases (11 studies), or to enrich disease prevalence (4 studies). How this selection bias was applied was stated explicitly in 18 (53%) of these 34. Whether selection bias was used was unclear in 4 studies.

Figure 14: Bar chart showing data extracted by the systematic review relating to study readers, design, and the confidence scales used to build ROC curves.



The number of readers per study ranged from 2 (156) to 258 (176). The mean number was 13, median 6. The large majority of studies (35; 69%) used fewer than 10 readers. Reader experience was described in 40 (78%) studies but not in 11. Specific reader training for image interpretation was described in 31 (61%) studies. Readers were not trained specifically in 14 studies and in 6 it was unclear whether readers were trained

specifically or not. Readers were blind to clinical information for individual patients in 37 (73%) studies, unblind in 3, and this information was unrecorded or uncertain in 11 (22%). Readers were blind to prevalence in the dataset in 21 (41%) studies, unblind in 2, but this information was unsure/unrecorded or uncertain in the majority (28, 55%).

Observers read the same patient case on more than one occasion in 50 studies; this information was unclear in the single further study (170). A fully crossed design (i.e. all readers read all patients with all modalities) was used in 47 (92%) studies, but not stated explicitly in 23 of these. A single study (172) did not use a fully crossed design and the design was unclear or unrecorded in 3 (135, 170, 176). Case ordering was randomised (either a different random order across all readers or a different random order for each individual reader) between consecutive readings in 31 (61%) studies, unchanged in 6, and unclear/unrecorded in 14 (27%). The ordering of the index test being compared varied between consecutive readings in 20 (39%) studies, was unchanged in 17 (33%), and was unclear/unrecorded in 14 (27%). 26 (51%) studies employed a temporal separation between readings that ranged from 3 hours(150) to 2 months (163), with a median of 4 weeks. There was no separation (i.e. reading of cases in all conditions occurred at the same sitting) in 17 (33%) studies, and temporal separation was unclear/unrecorded in 8 (16%).

Methods of reporting study outcomes

The unit of analysis for the ROC AUC analysis was the patient in 23 (45%) studies, an organ in 5, an organ segment in 5, a lesion in 11 (22%), other in 2, and unclear or unrecorded in 6 (12%); one study (135) examined both organ and lesion so there were 52 extractions for this item. Analysis was based on multiple images in 33 (65%) studies, a single image in 16 (31%), multiple modalities in a single study(140), and unclear in a single study (157); no study used videos.

The number of disease positive patients per study ranged between 10 (179) and 100 (153) (mean 42, median 48) in 46 studies, and was unclear/unrecorded in 5 studies. The number of disease positive units of outcome for the primary ROC AUC analysis ranged between 10 (179) and 240 (141) (mean 59, median 50) in 43 studies, and was unclear/unrecorded in 8 studies. The number of disease negative patients per study ranged between 3 (169) and 352 (135) (mean 66, median 38) in 44 studies, was zero in 1 study (180), and was unclear/unrecorded in 6 studies. The number of disease negative units of analysis for the primary outcome for the ROC AUC analysis ranged between 10 (151) and 535 (75) (mean 99, median 68) in 42 studies, and was unclear/unrecorded in the remaining 9 studies. The large majority of studies (41, 80%)

presented readers with an image or set of images reflecting normal clinical practice whereas 10 presented specific lesions or regions of interest to readers.

Calculation of ROC AUC requires the use of confidence scores, where readers rate their confidence in the presence of a lesion or its characterisation. In our previous study (17) we identified the assignment of confidence scores to be potentially on separate scales for disease positive and negative cases (181). For rating scores used to calculate ROC AUC, 25 (49%) studies used a relatively small number of categories (defined as up to 10) and 25 (49%) used larger scales or a continuous measurement (e.g. visual analogue scale). One study did not specify the scale used (176). Only 6 (12%) studies stated explicitly that readers were trained in advance to use the scoring system, for example being encouraged to use the full range available. In 15 (29%) studies there was the potential for multiple abnormalities in each unit of analysis (stated explicitly by 12 of these). This situation was dealt with by asking readers to assess the most advanced or largest lesion (e.g.(143)), by an analysis using the highest score attributed (e.g.(142)), or by adopting a per-lesion analysis (e.g.(152)). For 23 studies only a single abnormality per unit of analysis was possible, whereas this issue was unclear in 13 studies.

Model assumptions

The majority of studies (41, 80%) asked readers to ascribe the same scoring system to both disease-positive and disease-negative patients. Another 9 studies asked that different scoring systems be used, depending on whether the case was perceived as positive or negative (e.g.(161)), or depending on the nature of the lesion perceived (e.g.(166)). Scoring was unclear in a single study (176). No study stated that two types of true-negative classifications were possible (i.e. where a lesion was seen but misclassified vs. not being seen at all), a situation that potentially applied to 22 (43%) of the 51 studies. Another concern occurs when more than one observation for each patient is included in the analysis, violating the assumption that data are independent. This could occur if multiple diseased segments were analysed for each patient without using a statistical method that treats these as clustered data. An even more flawed approach occurs when analysis includes one segment for patients without disease but multiple segments for patients with disease.

When publically available MRMC software is used for ROC AUC modelling, this requires assumptions of normality for confidence scores or their transformations if the standard parametric ROC curve fitting methods are used. When scores are not normally distributed, even if non parametric approaches are used to estimate ROC

AUC, this lack of normality may indicate additional problems with obtaining reliable estimates of ROC AUC (19, 21-23, 26). While 17 studies stated explicitly that the data fulfilled the assumptions necessary for modelling, none described whether confidence scores were transformed to a normal distribution for analysis. Indeed, only 3 studies (154, 173, 176) described the distribution of confidence scores, which was non-normal in each case.

Model fitting

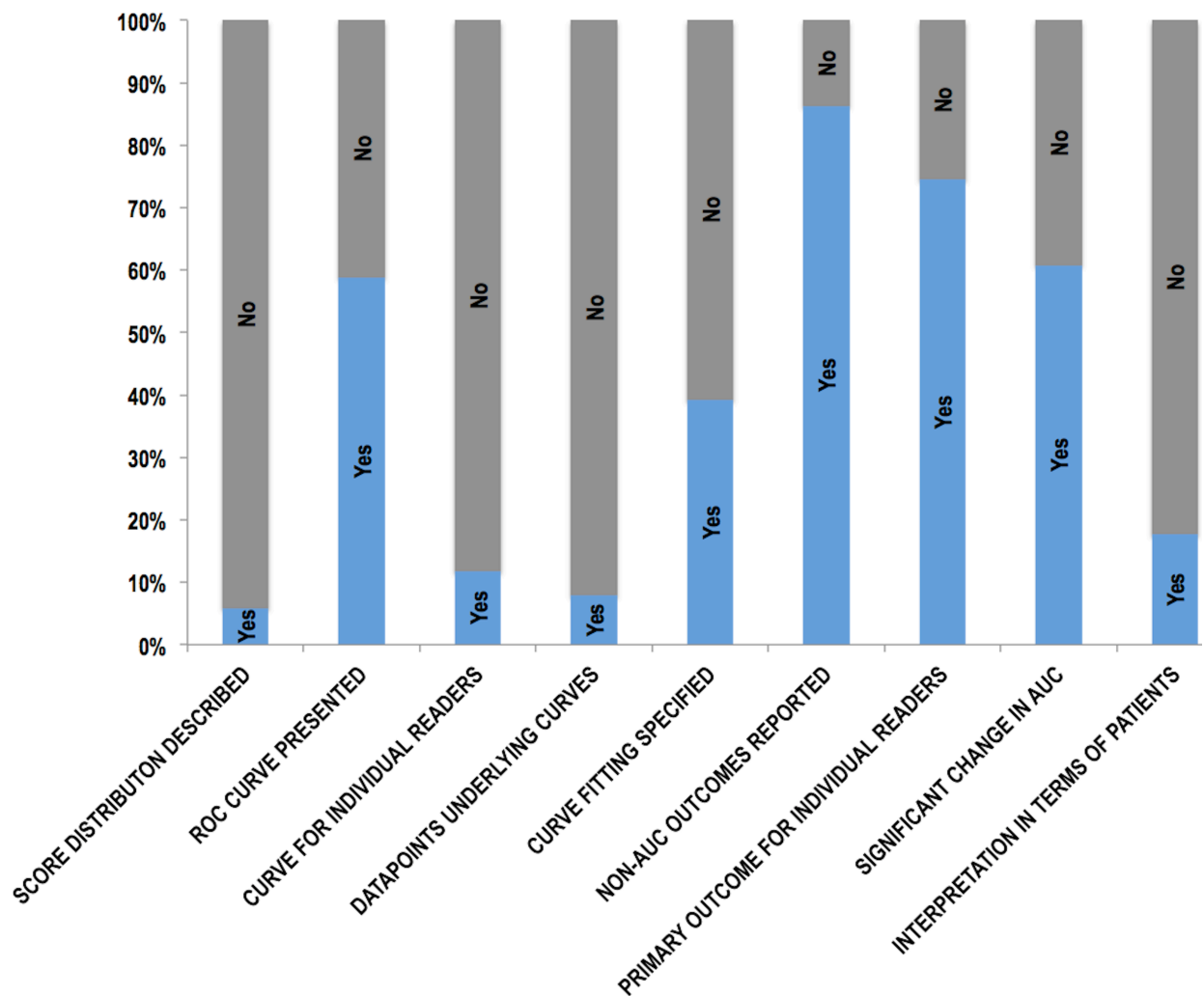
Thirty (59%) studies presented ROC curves based on confidence scores; i.e. 21 (41%) studies showed no ROC curve. Of the 30 with curves, only 5 presented a curve for each reader whereas 24 presented curves averaged over all readers; a further study presented both. Of the 30 studies presenting ROC curves, 26 (87%) showed only smoothed curves, with the data points underlying the ROC curve presented in only 4 (13%) (143, 151, 163, 178). Thus, a ROC curve with underlying data points was presented in only 4 of 51 (8%) studies overall. The degree of extrapolation is critical in understanding the reliability of the ROC AUC result (181). However, extrapolation could only be assessed in these four articles, with unreasonable extrapolation, by our definition, occurring in two (143, 163).

The majority of studies (31, 61%) did not specify the method used for curve fitting. Of the 20 that did, 7 used non-parametric methods (Trapezoidal/Wilcoxon), 8 used parametric methods (7 of which used Proproc), 3 used other methods, and 2 used a combination. Previous research (22, 181) has demonstrated considerable problems fitting ROC curves due to degenerate data where the fitted ROC curve corresponds to vertical and horizontal lines, e.g. there are no FP data. Only 2 articles described problems with curve fitting (155, 161). Two studies stated that data were degenerate: Subhas and co-workers (166) stated that, "data were not well dispersed over the five confidence level scores". Moin and co-workers (153) stated that, "If we were to recode categories 1 and 2, and discard BI-RADS 0 in the ROC analysis, it would yield degenerative results because the total number of cases collected would not be adequate". While all studies used MRMC AUC methods to compare AUC outcomes, 5 studies also used other methods (e.g. t-testing) (138, 152, 160, 167, 177). Only 3 studies described using a partial AUC (142, 155, 177). 44 studies reported additional non-AUC outcomes (e.g. McNemar's test to compare test performance at a specified diagnostic threshold (158), Wilcoxon signed rank test to compare changes in patient management decisions (164)). Eight (16%) of the studies included a ROC researcher as an author (75, 147, 148, 154, 160, 165, 166, 172).

Presentation of results

Extracted data relating to the presentation of individual study results is presented graphically in Figure 15. All studies presented ROC AUC as an accuracy measure with 49 (96%) presenting the change in AUC for the conditions tested. Thirty-five (69%) studies presented additional measures such as change in sensitivity/specificity (24 studies), positive/negative predictive values (5 studies), or other measures (e.g. changes in clinical management decisions (164), intraobserver agreement (137)). Change in AUC was the primary outcome in 45 (88%) studies. Others used sensitivity (135, 140), accuracy (136, 169), the absolute AUC (144) or JAFROC figure of merit (168). All studies presented an average of the primary outcome over all readers, with individual reader results presented in 38 (75%) studies but not in 13 (25%). The mean change/difference in AUC was 0.051 (range -0.052 to 0.280) across the extracted studies and was stated as “significant” in 31 and “non-significant” in the remaining 20. No study failed to comment on significance of the stated change/difference in AUC. In 22 studies we considered that a significant change in AUC was unlikely to be due to results from a single reader/patient but we could not determine whether this was possible in 11 studies, and judged this not-applicable in a further 18 studies. One study appeared to report an advantage for a test when the AUC increased, but not significantly (165). There were 5 (10%) studies where there appeared to be discrepancies between the data presented in the abstract/text/ROC curve (137, 139, 169, 177, 180).

Figure 15: Bar chart showing data extracted by the systematic review relating to the presentation of individual study results.



While the majority of studies (42, 82%) did not present an interpretation of their data framed in terms of changes to individual patient diagnoses, 9 (18%) did so, using outcomes in addition to ROC AUC: For example, as a false-positive to true-positive ratio(136) or the proportion of additional biopsies precipitated and disease detected(164), or effect on callback rate(143). The change in AUC was non-significant in 22 studies and in 12 of these the authors speculated why, for example stating that the number of cases was likely to be inadequate(165, 170), that the observer task was insufficiently taxing(137), or that the difference was too subtle to be resolved(145). For studies where a non-significant change in AUC was observed, authors sometimes framed this as demonstrating equivalence (16 studies, e.g.(155, 174)), stated that there were other benefits (3 studies), or adopted other interpretations. For example, one study stated that there were “beneficial” effects on many cases despite a non-significant change in AUC(154) and one study stated that the intervention “improved

visibility” of microcalcifications noting that the lack of any statistically significant difference warranted further investigation (165).

Discussion

While many studies have used ROC AUC as an outcome measure, very little research has investigated how these studies are conducted, analysed and presented. We could find only a single existing systematic review that has investigated this question (182). The authors stated in their Introduction, “we are not aware of any attempt to provide an overview of the kinds of ROC analyses that have been most commonly published in radiologic research.” They investigated articles published in the journal “Radiology” between 1997 and 2006, identifying 295 studies (182). The authors concluded that “ROC analysis is widely used in radiologic research, confirming its fundamental role in assessing diagnostic performance”. For the present review, we wished to focus on MRMC studies specifically, since these are most complex and are often used as the basis for technology licensing. We also wished to broaden our search criteria beyond a single journal. Our systematic review found that the quality of data reporting was frequently incomplete in those MRMC studies using ROC AUC as an outcome measure. We would therefore agree with Shiraishi et al. who concluded from their review of articles in “Radiology” that studies, “were not always adequate to support clear and clinically relevant conclusions” (182).

Many omissions we identified were those related to general study design and execution, and are well-covered by the STARD initiative (183) as factors that should be reported in studies of diagnostic test accuracy in general. For example, we found that the number of participating research centres was unclear in approximately one-third of studies, that most studies did not describe whether patients were symptomatic or asymptomatic, that criteria applied to case selection were sometimes unclear, and that observer blinding was not mentioned in one-fifth of studies. Regarding statistical methods, STARD states that studies should, “describe methods for calculating or comparing measures of diagnostic accuracy” (183); this systematic review aimed to focus on description of methods for MRMC studies using ROC AUC as an outcome measure.

The large majority of studies used less than 10 observers, some did not describe reader experience, and the majority did not mention whether observers were aware of prevalence of abnormality, a factor that may influence diagnostic vigilance. Most studies required readers to detect lesions while a minority asked for characterisation, and others were a combination of the two. We believe it is important for readers to

understand the precise nature of the interpretative task since this will influence the rating scale used to build the ROC curve. A variety of units of analysis were adopted, with just under half being the patient case. We were surprised that some studies failed to record the number of disease-positive and disease-negative patients in their dataset. Concerning the confidence scales used to construct the ROC curve, only a small minority (12%) of studies stated that readers were trained to use these in advance of scoring. We believe such training is important so that readers can appreciate exactly how the interpretative task relates to the scale; there is evidence that radiologists score in different ways when asked to perform the same scoring task because of differences in how they interpret the task (16). For example, readers should appreciate how the scale reflects lesion detection and/or characterisation, especially if both are required, and how multiple abnormalities per unit of analysis are handled. Encouragement to use the full range of the scale is required for normal rating distributions. Whether readers must use the same scale for patients with and without pathology is also important to know.

Despite their importance for understanding the validity of study results, we found that description of the confidence scores, the ROC curve, and its analysis was often incomplete. Strikingly, only three studies described the distribution of confidence scores and none stated whether transformation to a normal distribution was needed. When publically available DBM MRMC software (e.g. http://www-radiology.uchicago.edu/kr/KRL_ROC/software_index6.htm) is used for ROC AUC modelling, this requires assumptions of normality for confidence scores or their transformations when ROC curve fitting methods are used. Where confidence scores are not normally distributed these software methods are not recommended (21-24). Although Hanley shows that ROC curves can be reasonable under some distributions of non normal data (25), concerns have been raised particularly for imaging detection studies measuring clinically useful tests with good performance to distinguish well defined abnormalities. In tests with good performance two factors make estimation of ROC AUC unreliable. Firstly readers' scores are by definition often at the ends of the confidence scale so that the confidence score distributions for normal and abnormal cases have very little overlap (19, 21-23, 26). Secondly tests with good performance also have few false positives making ROC AUC estimation highly dependent on confidence scores assigned to possibly fewer than 5% to 10% of cases in the study (21).

Most studies did not describe the method used for curve fitting. Over 40% of studies presented no ROC curve in the published article. When present, the large majority were smoothed and averaged over all readers. Only four articles presented data points

underlying the curve meaning that the degree of any extrapolation could not be assessed despite this being an important factor regarding interpretation of results(28). While, by definition, all studies used MRMC AUC methods, most reported additional non-AUC outcomes. Approximately one-quarter of studies did not present AUC data for individual readers. Because of this, variability between readers and/or the effect of individual readers on the ultimate statistical analysis could not be assessed.

Interpretation of study results was variable. Notably, when no significant change in AUC was demonstrated, authors stated that the number of cases was either insufficient or that the difference could not be resolved by the study, appearing to claim that their studies were underpowered rather than conclude that the intervention was ineffective when required to improve diagnostic accuracy. Indeed some studies claimed an advantage for a new test in the face of a non-significant increase in AUC, or turned to other outcomes as proof of benefit. Some interpreted no significant difference in AUC as implying equivalence.

Our review does have limitations. Indexing of the statistical methods used to analyse studies is not common so we used a proxy to identify studies; their citation of “key” references related to MRMC ROC methodology. While it is possible we missed some studies, our aim was not to identify all studies using such analyses. Rather, we aimed to gather a representative sample that would provide a generalisable picture of how such studies are reported. It is also possible that by their citation of methodological papers (and on occasion including a ROC researcher as an author), our review was biased towards papers likely to be of higher methodological quality than average. This systematic review was cross-disciplinary and two radiological researchers (TD and AP) performed the bulk of the extraction rather than statisticians. This proved challenging since the depth of statistical knowledge required was demanding, especially when details of the analysis was being considered. We anticipated this and piloted extraction on a sample of five papers to determine if the process was feasible, deciding that it was. Advice from experienced statisticians (SM, DGA and TF) was also available when uncertainty arose.

In summary, via systematic review we found that MRMC studies using ROC AUC as the primary outcome measure often omit important information from both the study design and analysis, and presentation of results is frequently not comprehensive. Authors using MRMC ROC analyses should be encouraged to provide a full description of their methods and results so as to increase interpretability. The following Chapter of this thesis provides guidelines regarding how to do so.

Chapter 7: Guidelines for the reporting of multi-reader multi-case imaging studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy.

At the time of writing, this Chapter is in preparation for submission to an indexed journal.

Abstract

Introduction: Chapter 6 found that data reporting is frequently incomplete in multi-reader multi-case studies of medical imaging that use the area under the receiver operator curve (ROC AUC) as an outcome measure. We aimed to develop guidelines to facilitate fuller reporting of such studies.

Methods: The findings of the systematic review of 51 multi-reader multi-case studies of medical imaging that used ROC AUC as an outcome measure were consulted (Chapter 6). The lead author noted where studies were deficient in reporting individual items and drafted a set of reporting guidelines that were then considered by the investigator group as a whole, and modified subsequently in the light of these discussions.

Results: We present guidelines for the reporting of studies grouped under 11 separate items that deal with the imaging test, study cases, study readers, their diagnostic task, confidence rating scales used to build the ROC curve, study design, units of analysis, confidence rating score distribution and curve fitting, data analysis, data presentation, and study interpretation.

Conclusions: Authors reporting multi-reader multi-case studies using ROC AUC as an outcome measure are encouraged to follow the guidelines presented in this article so as to increase interpretability of their study results.

Introduction

Because we were interested in the quality of their data reporting, the previous Chapter (Chapter 6) describes a systematic review of MRMC studies using ROC AUC as an outcome measure. Our review identified 49 primary research articles describing 51 individual studies from which we extracted data regarding the quality of study reporting, with a particular emphasis on how ROC AUC was calculated and presented. In brief, we concluded that such studies often omit important information when describing both study design and analysis, and that presentation of results is frequently not comprehensive. We encourage authors using MRMC ROC analyses to provide a full description of their methods and results so as to increase interpretability. In order to facilitate this aim, in this Chapter we now present detailed guidelines suggesting how to report MRMC studies that use ROC AUC as a measure of diagnostic accuracy.

Methods

The guidelines presented here apply predominantly to studies of diagnostic test accuracy that use human observers to interpret medical image data from real patients, and that attempt to use a MRMC ROC AUC analysis as a study outcome. Although the guidelines could potentially apply to such studies of other technologies, the systematic review described in Chapter 6 identified no non-imaging studies. We developed the following guidelines by considering the results from Chapter 6, noting especially where studies were particularly deficient in reporting the individual items considered. For ease of comprehension, we grouped the guidelines into 11 individual “items”, each dealing with a specific aspect of study reporting. Under each item we also detail the rationale for the guideline where this is not clear from reading of the item. The guidelines were drafted by Steve Halligan, then considered by the investigator group as a whole, and modified by Steve Halligan subsequently in the light of these discussions.

Results: The guidelines.

Item #1: The imaging test

The test method(s) should be described in detail including the nature of the test (e.g. CT scan, MR imaging), the technical parameters used to acquire imaging data (including patient preparation if relevant), and the exact nature of any different conditions under which the test was assessed (e.g. with and without computer-assisted detection). Was the data presented to study readers representative of the test when used in daily clinical practice or not? For example, were only selected images, portions of the imaging data, or pre-specified regions of interest presented to readers? Deviations from daily clinical practice will impact on the generalisability of study findings.

Item #2: Patients (cases) used for the study

The source of patients (e.g. local hospital, online database, multi-centre or single-centre) and their demographics (male/female proportion, age and range) should be stated. The organ being studied and the pathology/pathologies (and any subcategories of pathology, e.g. benign mass/malignant mass), if any, should be stated, as should whether patients were symptomatic and/or asymptomatic, and the exact proportions of each. The exact proportions of patients with and without lesions/pathology should be clear, as should the number of lesions per patient when multiple so that prevalence can be calculated. All of these data are necessary in order to assess whether study patients reflect those encountered in everyday clinical practice. For the same reason it is also necessary to describe any specific selection criteria that were applied to cases included in the study, for example whether only specific categories of pathology were included (e.g. more severe cases), or whether criteria were applied to select cases that are deemed particularly “difficult” to interpret. Whether patient data were obtained prospectively and specifically for the purposes of the study, whether the study is a prospective comparison of retrospectively acquired data, or whether the study was wholly retrospective (i.e. observational) should be clear. The nature of any reference test(s) and/or panels used to establish a ground-truth reference standard result for each case should be described along with how this was achieved.

Item #3 Study readers

The number and source of study readers should be described along with their profession, status, and prior experience (both in terms of number of similar clinical cases interpreted previously and the time-span over which this occurred). Whether

readers underwent specific training for the study (e.g. use of CAD software) should be noted and, if so, the nature and duration of this.

Item #4 Readers' diagnostic task(s)

The exact nature of the diagnostic task for study readers should be described. For example, whether they were required to consider the presence or absence of lesions and/or lesion characterisation and/or lesion localisation. The exact instructions given to study readers should be available in the article or available as an appendix (e.g. online). The degree to which study readers are blind to clinical data, reference results, and their knowledge of prevalence of abnormality in the study dataset should be clear.

Item #5 Confidence rating scales

The exact nature of the confidence rating scale(s) completed by readers (and therefore used subsequently to build the ROC curve) should be clear, e.g. the exact number of categories, and exactly how these categories relate to the diagnostic task. In order to build a ROC curve, we would expect categories to be ordinal (i.e. to have a logical order), but this is sometimes not the case. For example, the BI-RADS scoring system has been used to create ROC curves but a rating of 2, defined as benign abnormality, does not imply a greater suspicion of cancer than a BI-RADS rating of 1, which is defined as no abnormality; both are confident diagnoses of non-malignancy (3). If an ordinal scale is used, it is also important to know whether the intervals between individual categories are constant. If not, then the scale may not be used equivalently by different readers in a multi-reader study. This is especially the case when a confidence rating score is actually a composite of assessments of different characteristics, for example size, shape, morphology and other visual information (16). For the same reason it must be stated whether different scales were used for disease-positive and disease-negative cases (if any). Were study readers trained to use the rating categories correctly in advance of their interpretation? Were the study readers encouraged to use the full range of scoring categories available to them, a procedure that is sometimes recommended to facilitate subsequent ROC curve analysis (6, 40)? If so, how was this done and the effect assessed?

Item #6 Study design

A description of the number of times each reader interprets each individual case should be described so that the study design is clear. For example is the design fully-crossed, i.e. every reader interprets every case? Case-ordering should be described since this may introduce potential biases, e.g. is the order of case interpretation the same for all readers and is the ordering the same for consecutive interpretations of the same case

by each individual reader? How was case ordering arrived at and if this was changed between readers and/or interpretations, how was this achieved (e.g. by randomisation)? The presence of any washout period between interpretations to diminish recall bias should be stated as should the exact conditions under which the interpretation were made and their duration; for example, were interpretations performed in a situation representative of daily clinical practice or in a “laboratory” environment, with batches of consecutive cases reported rapidly?

Item #7 Unit of analysis

The unit of analysis to build the ROC curve (or multiple curves where present) should be apparent. For example, is the unit of analysis a single image, multiple images, or portion of an image? Is the unit a patient, an organ, or portion of an organ (e.g. liver segment, colonic segment)? It should be clear whether single or multiple lesions are possible per individual unit, as should whether it is possible for individual units to be negative for lesions. It should also be clear if two types of true-negative classifications were possible by readers, i.e. where a lesion was seen but misclassified by a study reader vs. not being seen at all. The total number of lesion-positive and lesion-negative units of analysis should be stated in the final article. If other study outcomes are used, then their unit of analysis should be stated also.

Item #8 Distribution of rating scores and data fitting

The distribution(s) of reader rating scores should be presented in the published article, both overall and for readers individually. It should be stated whether the distribution(s) is normal or not. The method used for curve fitting should be stated and referenced. Whether groups of scoring categories were collapsed/amalgamated to facilitate building the ROC curve should be clear. Whether the authors encountered any problems with fitting ROC curves to reader rating scores should be reported. In particular, whether the data distribution satisfied any assumptions of normality demanded by the fitting method or whether transformation of the data was necessary in order to achieve this. If so, then the method of transformation should be stated. This information is important so that readers can assess whether ROC analysis is likely to be appropriate.

Item #9 Data analysis

The primary study outcome should be stated clearly and unambiguously. For example, was improved test accuracy defined in advance by a significant increase in ROC AUC or by significantly improved sensitivity and specificity, the latter together or individually? If there are multiple outcomes, the distinction between primary and secondary should

be clear. In particular, it should be stated exactly which primary outcome was decided upon in advance of the analysis so that readers can balance knowledge of this against the emphasis attributed to other outcomes in the published manuscript. If the intention was to have co-primary outcomes (for example change in ROC AUC and sensitivity/specificity) then the authors should describe how they intended in advance to account for this statistically. Where ROC AUC and/or pAUC is compared between tests (or between the same test under different conditions), the statistical method used to compare AUC or pAUC should be stated.

Item #10 Data presentation

Each statistical measure should be reported both averaged over all readers and for each reader individually so that the impact of individual readers on the outcome overall can be assessed. Where ROC AUC is compared between tests (or between the same test under different conditions), the AUC and change in AUC should be clear, and its significance stated, again averaged over all readers and for each reader individually. ROC curves for each interpretation condition should be presented in the published article, both overall and individually for each reader, preferably unsmoothed. The data points underlying the ROC curve should be shown so that the influence of individual readers on the curve and its area can be assessed. Visible data points are also necessary so that the degree of curve extrapolation, if any, can be assessed by readers so that they can come to a decision as to whether extrapolation exerted a significant effect on the AUC (28).

Item #11 Study interpretation

Interpretation of study results should focus on the primary outcome and not be diverted towards other outcomes, especially where any change in the primary outcome was found to be non-significant. Because lack of statistical power was often blamed for a non-significant change in AUC, the authors should describe their rationale for powering and its calculation for sample size. To increase clinical interpretability of the study results, an attempt should also be made to frame any change in AUC in terms of its effect on individual patient diagnoses at relevant clinical thresholds. For example, when comparing two test conditions, the number of additional true-positive and false-positive patients per 100 patients at a clinically sensible diagnostic threshold could be stated. It will also be necessary to state the prevalence of abnormality since that will influence the numbers detected. The stated prevalence should reflect the prevalence encountered in daily clinical practice so that the interpretation is generalisable.

Discussion

We have developed and presented a set of guidelines recommended by us for the reporting of MRMC research studies that use ROC AUC as an outcome measure. These guidelines arose from difficulties encountered by us when attempting to understand how authors had deployed this analysis in studies of CT colonography available in the published literature. These difficulties precipitated a systematic review, described in Chapter 6, and the present guidelines arose directly from the findings of that review.

These guidelines focus on aspects of study reporting that deal with design and analysis, and presentation of results. These guidelines are not intended to replace the STARD guidelines for the reporting of studies of diagnostic test accuracy (183). The systematic review detailed in Chapter 6 found that many studies omitted information related to general study design and execution, which are well-covered by the STARD initiative (183). The present guideline is intended to complement STARD but necessarily duplicates some of the information that is necessary when attempting to understand how the primary research was performed, the origin of patients for example. With specific reference to statistical methods, STARD states that studies should, “describe methods for calculating or comparing measures of diagnostic accuracy” (183). The present guideline therefore builds on this recommendation and goes beyond STARD by requiring more detailed, specific information regarding the methods and analyses performed in MRMC studies using ROC AUC as an outcome measure.

Our guideline comprises 11 separate headings, ranging from the imaging test being investigated through to interpretation of the study findings. A clear description of the test and the patients on which it is used is necessary so that readers can assess whether study findings are likely to be generalisable, i.e. do the test and patients reflect those encountered in daily clinical practice? A description of readers is also important since studies often recruit readers who do not reflect the majority likely to use the test in normal practice, for example using either inexperienced trainees (since they are easier to recruit) or highly expert radiologists working in a specialist centre (since these are more likely to undertake research).

Our guidelines have taken a particularly detailed focus on the diagnostic task required of study readers and the confidence rating scale that they must complete in order for the ROC curve(s) to be built. Our personal experience of the literature has been that it is frequently very difficult to be clear regarding exactly what was required of readers and how the confidence rating scale related to the diagnostic task, especially when there is some confusion around the unit of analysis and whether this can be

positive/negative for disease. For example, it may be unclear how readers handle multiple abnormalities, or whether this is even possible. Our systematic review found that only 12% of studies stated that readers were trained to use the confidence rating scale in advance of scoring, despite the fact that this procedure is recommended(6).

Our systematic review found that details relating to fitting the ROC curve, its presentation, and methods for comparing the AUC were often lacking, and so our guidelines have emphasised these aspects of study reporting. Because we found that curves were usually smoothed, averaged, and showed no underlying data points we have emphasised presentation of data points for individual readers so that their influence on the curve can be assessed along with the degree of any extrapolation. For example, our review found that a ROC curve with underlying data points was presented in only 4 (8%) of 51 studies and extrapolation could be determined in these 4 only. Most studies did not specify the method used for curve fitting and only 3 described the distribution of confidence scores and none stated whether transformation to a normal distribution was needed.

Finally, because our review found that interpretation of studies in the literature often deviated from the primary outcome (usually change in ROC AUC), especially if this was found to be non-significant, we have asked that researchers focus their interpretation on their primary outcome, stated in advance. Because changes in diagnostic test accuracy are most comprehensible when presented in terms of gains and losses to individual patients (10), we suggest that authors frame their findings in this way.

It occurred to us that publication of this set of guidelines might by inference indicate a tacit support for MRMC studies using ROC AUC as an outcome measure. While it is clear that we believe there are considerable issues regarding the correct application of ROC AUC to these studies, we do believe that the data described in the guidelines need to be presented when studies using this analysis are reported. It is also possible that with relatively little modification these guidelines could also apply to MRMC studies that use net benefit as the primary outcome measure. Doing so would offer the opportunity for us to combine proponents of both ROC AUC and net benefit as authors of the same guideline, which would likely serve to ameliorate fundamental differences of approach between the two. This would broaden the appeal of the guidelines which, at present, could be accused of bias since they have been drafted by workers perceived as opponents of ROC AUC. A combined approach would likely pre-empt criticism of the guidelines arising from perceptions that the guideline authors are biased towards one analysis or another. A combined approach would facilitate guideline generalisability.

In summary, we have developed guidelines for the reporting of MRMC research studies that use ROC AUC as an outcome measure. The guidelines are intended to provide clarity around the research methodology used and encourage a full description of both the analysis and presentation of findings so as to increase the interpretability of study results.

Chapter 8: Thesis summary and recommendations for future research

The belief that ROC curve analysis is fundamental to understanding the diagnostic accuracy of an imaging test is deeply imbedded in radiological thinking. For example, the most highly cited paper in the entire clinical medical imaging literature is Hanley and McNeil's 1982 article, "The meaning and use of the area under a receiver operating characteristic curve" (2). Accordingly, when designing a MRMC study in order to achieve licensing approval for CAD software for CT colonography, the USA FDA obliged us to use ROC AUC as our primary outcome measure (78). My statistical collaborators (Susan Mallet and Douglas Altman) soon realised that this approach was problematic. In particular, the primary study question turned on whether radiologists found it easier to detect polyps when using CAD software. Confidence scores used to build the ROC curve therefore reflected radiologists' opinions regarding whether a polyp was present or not. We developed and wrote a study protocol that was accepted subsequently by the FDA. It was not until reader scores had been obtained and we were performing the intended analysis that we realised the data were far from normally distributed (see Chapter 1); we could not build ROC curves. We began to examine the reasons underpinning this, which, in turn, led us to unearth more and more problematic issues with ROC AUC as an outcome measure for studies of radiological technologies. This thesis gathers together our thinking on ROC AUC, the search for an alternative based on net benefit, and further research that was precipitated by both our discontent with ROC AUC and the need for additional data necessary to implement our proposed alternative. At the same time, this thesis also demonstrates that the development of an alternative method will present another set of problems, best exemplified here by the need to perform experiments in order to arrive at a value for W .

This thesis details non-trivial issues related to the use of ROC AUC in MRMC studies and presents an alternative based on net benefit (Chapter 1). Chapter 2 uses the net benefit method to analyse a MRMC study of CAD for CT colonography. The analysis requires an estimate of relative misclassification costs for false-negative versus false-positive diagnoses; " W ". This study used a conservative value for W , arrived at via consensus between the investigators; while we were confident that false-negative diagnoses "cost" more than false-positive diagnoses, we were concerned that by using too high a figure, we would lay ourselves open to claims that we were biasing the analysis to such an extent that CAD could not fail to be found beneficial. In any event, despite the conservative (to us) figure of 3 used originally, the FDA rejected the analysis as "unproven and unvalidated", despite it being based on a combination of

sensitivity and specificity, two measures for which it would be hard to imagine better established validation. It is informative to note that when submitting the contents of Chapter 2 for peer-reviewed, indexed publication, the Editor of the journal *Radiology* asked that we change our primary outcome measure in retrospect: We were obliged to present the net benefit analysis in an online Appendix (78). Seeking to understand why the analysis was “unvalidated” we could only assume that this was because the value of W had been arrived at by consensus discussion, since the other components of the equation (namely change in sensitivity, change in specificity, and prevalence) are known with certainty. Chapter 3 therefore describes an experiment used to obtain a precise value for W . Exactly as we predicted, the value for W ascribed by our study participants was far higher than our overly conservative value of 3, being 2250 for cancers and 6 for polyps. It was natural, then, to re-analyse the study performed in Chapter 2, this time using the evidence-based values of 2250 and 6 for cancers and polyps respectively, for W . At the same time, we elected to compare the performance of experienced and inexperienced readers of CT colonography using the data from Chapter 2 and a prior study of CT colonography performed by us (17); these analyses are described in Chapter 4. We then decided to extend the work described in Chapter 3, seeking to discover if the seemingly high value for W extended to other clinical scenarios, using as an example the detection of extra-colonic pathology by CT colonography. I believe this is a very germane situation since the ability of CT colonography to depict extracolonic pathology is often promulgated as a distinct advantage over other methods to examine the colon, yet false-positive diagnoses have considerable capacity to cause morbidity (both physical and psychological) and even mortality. Nevertheless, again we found that the value of W was relatively high; 599 for non-invasive follow-up testing and 60 for invasive testing.

Given the problems we encountered when attempting to use ROC AUC as an outcome measure, it was natural to seek evidence that other researchers had come across similar hurdles. The systematic review described in Chapter 6 revealed that if other researchers have encountered similar problems, then they have not declared these in the literature in any substantial numbers. While we found a paucity of problems described, at the same time we found that comprehensive presentation of study methods, analysis, and data presentation was very unusual. Accordingly, the final Chapter describes a set of guidelines (i.e. a minimum dataset) that we believe researchers should adhere to when reporting such studies. As Noted in Chapter 7, this does not necessarily imply that we “approve” of the MRMCC ROC AUC analysis but we do believe that these data need to be presented when such studies are reported.

It is important to understand that we do not object to ROC AUC per se. I believe that the analysis is especially useful at the early stages of test development, when it is important to understand whether a diagnostic test (that uses different diagnostic thresholds) works at all. However, ROC AUC appears to be promulgated widely as the only analysis applicable when it is necessary to know the performance of a diagnostic imaging test. This seems, to me, to be the case irrespective of the context of the assessment and this is where my major objection lies. It is clear and undeniable that misclassification costs for false-negative and false-positive diagnoses may differ depending on the clinical context, and often differ very considerably. However ROC AUC almost always ignores these and also ignores the effect that prevalence might have on the result. It is also worth considering the fact that “good” diagnostic tests will be better able to sort patients into “diseased” and “not diseased” categories. Thus, true-positive and true-negative diagnoses tend to be confident with the result that readers’ scores are often at each end of the confidence scale. Accordingly, the confidence score distributions for normal and abnormal cases have very little overlap (19, 21-23, 26). Secondly, tests with good performance also have few false positives making ROC AUC estimation highly dependent on confidence scores assigned to possibly fewer than 5% or 10% of cases in the study (21). Thus, a wide range of confidence scores would be unexpected in a test that performed well, for example one that was ready for clinical implementation. However, a wide range of scores is necessary to build the ROC curve. Ultimately, ROC AUC also lacks clinical interpretability because it is not immediately obvious how an AUC translates into numbers of patients correctly and incorrectly diagnosed. This seems, to me, to be a major disadvantage because, in clinical radiology at least (where ROC AUC is the most dominant analysis), the clinical context is well-established in most studies. It seems illogical, then, to ignore this.

ROC proponents state that it is possible to address these issues but the solution does not seem, to me, to be simple. For example, Zweig and Campbell (9) state that to use ROC AUC, “for patient management” two “major elements” of selecting the appropriate sensitivity/specificity pair threshold are required. These are, “the relative cost or undesirability of errors, i.e. false-positive and false-negative classifications; the value or benefits of correct classifications”, and the need to know, “the relative proportions of the two states of health that the test is intended to discriminate between. This is related to prevalence”. The authors therefore acknowledge explicitly that knowledge of both misclassification costs and disease prevalence is necessary to interpret ROC AUC in clinical practice. They then proceed to describe a method to incorporate these elements within ROC AUC that appears, to a non-statistician such as myself, to be complex and difficult to implement, at least when compared to net benefit. They derive

a slope, m , that is the product of false-positive costs divided by false-negative costs and one minus prevalence divided by prevalence (9). It is then necessary to calculate where a line with the slope m first contacts the ROC curve; this provides the appropriate operating point. Dwyer (5) proposes a similar method, again one that requires an estimate of relative misclassification costs. It is interesting to note that while our net benefit equation was criticised as being “unvalidated” because misclassification costs were unknown, the solution proposed by well-recognised ROC methodologists requires exactly the same data.

Similarly, a paper from McClish (184) published subsequent to my starting work on this thesis recognises many of the issues that we raise. For example, McClish states, “One definition of the optimal threshold is the test value that minimises the Bayes cost or, equivalently, maximises the generalized Youden index (GYI). This definition incorporates both the cost or utility of misclassification and the prevalence of disease” (184) (the more familiar Youden index [sensitivity + specificity – 1] maximises correct classification but does not account for prevalence and costs). However, while recognising these issues fully, the author then proceeds to describe a complex method to constrict the AUC into a region around the optimal point (184). Again, this seems to me to be an overly complex solution and implies a desire to stay wedded to ROC AUC when to discard the analysis altogether seems, to me, to be a simpler solution in many situations.

So, should net benefit or similar indices replace ROC AUC? As already stated, I believe ROC AUC has particular value early in the assessment process, i.e. when deciding if a test should progress to clinical testing. Also, ROC AUC has particular validity when human interpretation is unnecessary, as is the case with many laboratory tests, since issues related to ascribing confidence scores do not apply. However, further research regarding net benefit is also required before it can be recommended with absolute confidence. For example, I am aware that statisticians arguing against ROC AUC do not regard indices similar to net benefit as an acceptable alternative. For example, Hilden objects to ROC AUC in general (185) but notes that there are some circumstances where it is preferable to other measures of maximisation of expected utility (MEU), specifically net reclassification index (NRI) and integrated discrimination improvement (IDI) (186-188). My reading of this is that the arguments against these analyses is in the context of prediction models where multiple measures are scored and considered simultaneously to generate a prediction index. The fact that several measures are combined serves to normalise the distributions of scores from normal/abnormal patients, which favours ROC AUC (here known as the c-index).

Taking an example from our own work, we were surprised by the results of the analysis described in Chapter 4. Specifically, the net benefit for experienced readers was non-significant at 3.2% (95%CI -1.9% to 8.3%), and we did not expect this from our inspection of the raw sensitivity and specificity data. This set of circumstances arose because the prevalence of disease (polyps in this case) was relatively low and the value of W was far less than that obtained for diagnosis of cancer, i.e. 6 versus 2250. I believe we need to consider in greater detail the interaction between prevalence and W , and the effect they exert on the outcome (akin to a “sensitivity analysis”) to be confident that the equation truly provides the “right” result. As is obvious from the content of this thesis, while the value of W has been established in Chapters 3 and 5 for specific clinical situations, more work is needed to define the precise value of W in other, multiple clinical situations. We found that patients and healthcare practitioners held different values of W , with the latter, as might be expected, holding more conservative views. Quoting Chapter 5, a very recent piece of work has found that “overdiagnosis” in several different contexts for cancer screening (i.e. detection of cancers that are destined *not* to harm the patient) is well tolerated by potential screenees (189).

It must also be noted that while Chapters 3 and 5 used a DCE methodology to arrive at a value for W , this analysis has attracted criticism. As noted in the Discussion related to those Chapters, DCEs require considerable cognitive ability to complete, relative to some alternatives. Also, both Chapters considered the diagnosis of cancer, which is likely to elicit a fear response, particularly in patients versus healthcare professionals, which may result in irrational answers underpinned by a fear of cancer and death. DCE may also be difficult where the event being studied is rare, as is usually the case for screening. Patients in particular may have difficulty appreciating that the event is, on average, very unlikely to happen. This may mean that the trade-off being asked of participants is complex and involves the weighing-up of many different attributes simultaneously, all with different consequences and high levels of uncertainty. Decision tree analysis is an alternative to discrete choice analysis that allows multiple attributes to be considered via a series of binary yes/no decisions. Their disadvantage is that it is less easy to vary the balance between attributes.

The question then arises, whose values for W should be used: Patients? Healthcare workers? Health economists? Politicians? A combination? In what proportions? I am inclined to suggest that medical doctors are most likely to provide the “correct” value of W for a given context since they have the broadest understanding of the clinical situation and consequences for the patient, and act on behalf of their patients’ best interests. However, it has been argued to me that, ultimately, patients are the most

appropriate group to provide W since they are both the consumers of the test, its consequences, and also, ultimately, pay for the test (whether indirectly via taxation or directly via insurance). I have some concerns around that approach, not least because of the issues around irrational and fearful responses noted in the preceding paragraph. It should also be noted that W is also likely to vary in the same group of patients, for example depending on their sex and social status. Indeed, W might change value in the same patient depending on their age since this will likely influence their perception of disease likelihood, and how consequences of the disease in question are perceived by the patient. Clearly, this issue requires more research as does the possibility of using a more complex net benefit equation to incorporate a multi-layered value for W that could be dependent on multiple factors such as age and prior probability of disease in an individual (for example, by accounting for family history).

It is natural that further work flowing from this thesis should compare both ROC AUC and net benefit analyses when applied directly to datasets arising from the same study. Doing so would allow us to more closely examine the advantages and disadvantages inherent to both analyses. A range of studies should be investigated since the applicability of ROC AUC is likely to vary across these contingent on fluctuating difficulties with applying confidence scores reliably, as described in previous Chapters. Chapter 7 presents a set of guidelines for the comprehensive reporting of MRMC studies using ROC AUC as an outcome measure but it is probable that, with relatively little modification, these guidelines could also apply to the reporting of studies using net benefit measures. Future work should attempt to develop these guidelines further so that they are applicable to both types of analysis in MRMC studies, an approach that would also serve to ameliorate apparently fundamental differences between researchers who are advocates of one particular approach.

In summary, in this thesis I have argued that there are considerable difficulties when using ROC AUC as an outcome measure for clinical studies of radiological technologies. These difficulties are non-trivial and are related in part to problems assigning confidence scores, non-normal score distributions and curve extrapolation, and an inability to easily ascribe misclassification costs. Further, the AUC lacks clinical interpretability. I have argued for an alternative analysis based on net benefit and have shown how this can be used to analyse clinical imaging study data. I have also shown how values for relative misclassification costs can be arrived at by discrete choice experiment and then incorporated into the net benefit analysis. I argue that an analysis based on net benefit is statistically transparent since it is based on changes in sensitivity and specificity, incorporates estimates of prevalence and misclassification costs, and is clinically interpretable since it reflects changes in correct and incorrect

patient diagnoses when using a new diagnostic test. Net benefit analyses do however require further consideration before they can be recommended unreservedly. In this respect the concept of “validation”, applied in the title of this thesis, is still in its infancy. The majority of this thesis has concentrated on the development of a net benefit alternative to ROC AUC. Future work should concentrate on further validation of the net benefit method.

Contribution and acknowledgements

This PhD is unusual in that it was executed at a relatively late stage in my career. The experience has been quite different to that of my first thesis – an MD submitted in 1996 – since I am now a team leader rather than a junior researcher. While for my MD I completed the bulk of the work at “grass-roots” level, i.e. recruiting patients, collecting data, and preparing first drafts of putative manuscripts for my supervisor, this time around (happily!) I have been living at the other end of the food chain and it is right that I acknowledge here my precise contribution and that of others. So, I would very much like to thank my own research fellows, especially those who did much of the “spade-work” for this thesis: Darren Boone, Andrew Plumb, Emma Helbren, and Thaworn Dendumrongsup. Darren Boone deserves special mention as his was the original suggestion that I might assemble my work around ROC AUC/net benefit into a credible personal PhD thesis after I had mentioned that I wished I had completed a PhD “first time around”. Heather Fitzke (also my PA) and Nichola Bell are also due thanks for their help recruiting and interviewing subjects for the second discrete choice experiment. Heather also deserves further thanks for managing my NIHR Programme Grant for Applied Research, which funded most of the work described in this thesis.

This thesis is an indisputable example of “other men’s flowers” (Lord Wavell). It is obvious that I am no statistician yet this thesis pivots on statistics. So, many, many thanks are due to my collaborators. I would particularly like to thank Douglas Altman and Susan Mallett, without whose statistical insights none of this work would have happened. So, what part did I play in this thesis? As Chief Investigator for the NIHR Programme Grant that forms a substantial part of this thesis, I was responsible for the overall strategy and coordination of the work, for supervising my research fellows, for redrafting and finalising their articles, and for final approval for submission (performing submission myself in the vast majority of cases) after my collaborators had due input. In particular, I had considerable intellectual input regarding research strategy, the methods used, their execution, data interpretation, and drafting of all final articles. To reflect this, I am corresponding author for *all* of the work presented in this thesis. In some cases I produced the initial article draft myself, with little input from my research fellows. My precise role for the work presented in this thesis was as follows:

- I was first and corresponding author for the published work arising from Chapters 1, 2 and 7 (i.e. I produced the initial draft myself).
- I was last and corresponding author for the published work arising from Chapters 3, 4, and 5 (i.e. I revised first drafts produced by my research fellows).

- I was third and corresponding author for Chapter 6. In this case I believed Dr. Susan Mallett should be placed last since she devised the extraction table for the systematic review. As Drs. Dendumrongsup and Plumb did the extraction, I believed they should go first and second respectively. However, I produced the first draft of the article and then revised it following input from my senior collaborators.

Many thanks are also due to Medicsight PLC (sadly now long gone) who supported the studies of CAD for CT colonography that precipitated our concerns re analysis using ROC AUC. In particular Justine McQuillan, Gareth Beddoe, Mary Roddie, Leslie Honeyfield and Greg Slaubaugh should be singled out. Thanks are also due to:

Professor Stuart Taylor (Professor of medical Imaging, UCL); Dr Christian von Wagner (Senior research associate, UCL), Mr Alex Ghanouni (health psychology research fellow), and Professor Jane Wardle (Professor of clinical psychology, UCL); Professor Richard Lilford (Professor of clinical epidemiology, University of Birmingham), Dr Shihua Zhu (postgraduate fellow) and Dr Lily Yao (University of Southampton); Dr. Tom Fanshawe (senior statistician, Oxford). Thanks are also due to Professor Wendy Atkin (Professor of gastrointestinal epidemiology, Imperial College) for facilitating my original introduction to Professor Altman – Professor Atkin has been a much-valued collaborator of mine for many years.

UCLH/UCL is a great place to work, with an ethos aimed squarely at facilitating research. Thanks are due to them, notably Professor Raymond MacAllister and my clinical colleagues in radiology, for allowing me to undertake a sabbatical from January to June 2013 that afforded me time to work on many component articles of this thesis. Of course, I am very grateful to the NIHR who funded my Programme Grant for Applied Research (RP-PG-0407-10338), which funded the research.

I must thank my “supervisors”, primary Professor Stuart Taylor and secondary, Professor David Hawkes. Thanks also to Dr Shonit Punwani for his help with my upgrade viva.

Thanks to Jon Deeks (Professor of biostatistics, University of Birmingham) and Adoni Toms (Professor of radiology, University of East Anglia) who conducted my PhD viva in September 2015. Being examined is never easy but the professional manner with which they conducted the viva is appreciated and their comments have undoubtedly improved this thesis, and focussed my thoughts on this topic further.

Of course my family deserve thanks, Alison, Sarah and Julia, but while my wife was devastated at the thought I would attempt a second thesis, I hope my assertion that it would be “much easier the second time around” has proven true?

Last but far from least, I would like to thank Professor Clive Bartram for recognising that my desire to pursue a research career was not entirely doomed to failure. His support was pivotal to my decision to become an academic radiologist and his mentoring meant that I had the right tools to do so: Thanks Clive - enjoy your well-deserved retirement.

Appendix 1: Indexed, peer-reviewed journal articles arising from this thesis.

1. Halligan S, Mallett S, Altman DG, McQuillan J, Proud M, Beddoe G, Honeyfield L, Taylor SA. Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: Multiobserver study. *Radiology* 2011;258:469-76.
2. Mallett S, Halligan S, Thompson M, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. *BMJ* 2012 Jul 2;345:e3999.
3. Boone D, Mallett S, Zhu S, Yao GL, Bell N, Ghanouni A, von Wagner C, Taylor SA, Altman DG, Lilford R, Halligan S. Patients' & healthcare professionals' values regarding true- & false-positive diagnosis when screening by CT colonography: Discrete choice experiment. *PLoS One* 2013 Dec 9;8:e80767.
4. Plumb AA, Boone D, Fitzke H, Helbren E, Mallett S, Zhu S, Yao GL, Bell N, Ghanouni A, von Wagner C, Taylor SA, Altman DG, Lilford R, Halligan S. Detection of extracolonic pathologic findings with CT colonography: A discrete choice experiment of perceived benefits versus harms. *Radiology* 2014; 273:144-152.
5. Mallett S, Halligan S, Collins GS, Altman DG. Exploration of analysis methods for diagnostic imaging tests: Problems with ROC AUC and confidence scores in CT colonography. *PLoS One*. 2014 Oct 29;9(10):e107633.
6. Dendumrongsup T, Plumb AA, Halligan S, Fanshawe TR, Altman DG, Mallett S. Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: Systematic review with a focus on the quality of data reporting. *PLoS One*. 2014 Dec 26;9(12):e116018.
7. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology* 2015;25:932-9.
Winner, Gold Medal for best GI paper published in European Radiology.
8. Boone D, Mallett S, McQuillan J, Taylor SA, Altman DG, Halligan S. Assessment of the incremental benefit of computer-aided detection (CAD) for interpretation of CT colonography by experienced and inexperienced readers. *PLoS ONE* 2015;10: e0136624.

Although arising directly from work presented in this thesis, articles nos. 2 and 5 above do not appear here as Chapters because they were chiefly the work of Dr. Susan Mallett, and drafted primarily by her.

References

1. Fawcett T. An Introduction to ROC analysis. *Pattern recognition Letters*. 2006;27:861-74.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982 Apr;143(1):29-36. PubMed PMID: 7063747.
3. Boyer B, Canale S, Arfi-Rouche J, Monzani Q, Khaled W, Balleyguier C. Variability and errors when applying the BIRADS mammography classification. *European journal of radiology*. 2013 Mar;82(3):388-97. PubMed PMID: 22483607.
4. Obuchowski NA. ROC analysis. *AJR American journal of roentgenology*. 2005 Feb;184(2):364-72. PubMed PMID: 15671347.
5. Dwyer AJ. In pursuit of a piece of the ROC. *Radiology*. 1996 Dec;201(3):621-5. PubMed PMID: 8939207.
6. Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM. Assessment of medical imaging and computer-assist systems: lessons from recent experience. *Academic radiology*. 2002 Nov;9(11):1264-77. PubMed PMID: 12449359.
7. Obuchowski NA. New methodological tools for multiple-reader ROC studies. *Radiology*. 2007 Apr;243(1):10-2. PubMed PMID: 17392244.
8. Obuchowski NA, Beiden SV, Berbaum KS, Hillis SL, Ishwaran H, Song HH, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Academic radiology*. 2004 Sep;11(9):980-95. PubMed PMID: 15350579.
9. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*. 1993;39(4):561-77.
10. Spiegelhalter D, Pearson M, Short I. Visualizing uncertainty about the future. *Science*. 2011 Sep 9;333(6048):1393-400. PubMed PMID: 21903802.
11. McClish DK. Analyzing a portion of the ROC curve. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1989 Jul-Sep;9(3):190-5. PubMed PMID: 2668680.
12. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG. US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. *Bmj*. 2000 Jun 17;320(7250):1635-40. PubMed PMID: 10856064. Pubmed Central PMCID: 27408.
13. Lusted LB. Decision-making studies in patient management. *The New England journal of medicine*. 1971 Feb 25;284(8):416-24. PubMed PMID: 4395661.
14. Adams NM, Hand DJ. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*. 1999;32:1139-47.
15. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute*. 2003 Apr 2;95(7):511-5. PubMed PMID: 12671018.
16. Harrington MB. Some methodological questions concerning receiver operating characteristic (ROC) analysis as a method for assessing image quality in radiology. *Journal of digital imaging*. 1990 Nov;3(4):211-8. PubMed PMID: 2085557.
17. Halligan S, Altman DG, Mallett S, Taylor SA, Burling D, Roddie M, et al. Computed tomographic colonography: assessment of radiologist performance with and without computer-aided detection. *Gastroenterology*. 2006 Dec;131(6):1690-9. PubMed PMID: 17087934.
18. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Academic radiology*. 2007 Jun;14(6):723-48. PubMed PMID: 17502262.
19. Gur D, Rockette HE, Bandos AI. "Binary" and "non-binary" detection tasks: are current performance measures optimal? *Academic radiology*. 2007 Jul;14(7):871-6. PubMed PMID: 17626312.

20. Krzanowski WJ, Hand DJ. ROC curves for continuous data. . 1 ed. Boca Raton, FL: Chapman and Hall/CRC; 2009 2009.
21. Mallett S, Halligan S, Collins GS, Altman DG. Exploration of analysis methods for diagnostic imaging tests: problems with ROC AUC and confidence scores in CT colonography. *PLoS one*. 2014;9(10):e107633. PubMed PMID: 25353643. Pubmed Central PMCID: 4212964.
22. Baker ME, Bogoni L, Obuchowski NA, Dass C, Kendzierski RM, Remer EM, et al. Computer-aided detection of colorectal polyps: can it improve sensitivity of less-experienced readers? Preliminary findings. *Radiology*. 2007 Oct;245(1):140-9. PubMed PMID: 17885187.
23. Zhou XH, Obuchowski N, McClish DK. Statistical methods in diagnostic medicine. New York NY: Wiley; 2002.
24. Petrick N, Haider M, Summers RM, Yeshwant SC, Brown L, Iuliano EM, et al. CT colonography with computer-aided detection as a second reader: observer performance study. *Radiology*. 2008 Jan;246(1):148-56. PubMed PMID: 18096536.
25. Hanley JA. The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1988 Jul-Sep;8(3):197-203. PubMed PMID: 3398748.
26. Dorfman DD, Berbaum KS, Brandser EA. A contaminated binormal model for ROC data: Part I. Some interesting examples of binormal degeneracy. *Academic radiology*. 2000 Jun;7(6):420-6. PubMed PMID: 10845401.
27. Lewin JM, Hendrick RE, D'Orsi CJ, Isaacs PK, Moss LJ, Karellas A, et al. Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. *Radiology*. 2001 Mar;218(3):873-80. PubMed PMID: 11230669.
28. Gur D, Bandos AI, Rockette HE. Comparing areas under receiver operating characteristic curves: potential impact of the "Last" experimentally measured operating point. *Radiology*. 2008 Apr;247(1):12-5. PubMed PMID: 18258813. Pubmed Central PMCID: 2637106.
29. Alemayehu D, Zou KH. Applications of ROC analysis in medical research: recent developments and future directions. *Academic radiology*. 2012 Dec;19(12):1457-64. PubMed PMID: 23122565.
30. Zou KH. Professor Charles E. Metz leaves profound legacy in ROC methodology: an introduction to the two Metz Memorial Issues. *Academic radiology*. 2012 Dec;19(12):1447-8. PubMed PMID: 23122563.
31. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of Markers and Risk Prediction Models: Overview of Relationships between NRI and Decision-Analytic Measures. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2013 May;33(4):490-501. PubMed PMID: 23313931.
32. Moons KG, Stijnen T, Michel BC, Buller HR, Van Es GA, Grobbee DE, et al. Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1997 Oct-Dec;17(4):447-54. PubMed PMID: 9343803.
33. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*. 2008 Jan 30;27(2):157-72; discussion 207-12. PubMed PMID: 17569110.
34. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical decision making : an international journal of the Society for Medical Decision Making*. 2006 Nov-Dec;26(6):565-74. PubMed PMID: 17099194. Pubmed Central PMCID: 2577036.
35. Summers RM, Yao J, Pickhardt PJ, Franaszek M, Bitter I, Brickman D, et al. Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. *Gastroenterology*. 2005 Dec;129(6):1832-44. PubMed PMID: 16344052. Pubmed Central PMCID: 1576342.

36. Taylor SA, Halligan S, Burling D, Roddie ME, Honeyfield L, McQuillan J, et al. Computer-assisted reader software versus expert reviewers for polyp detection on CT colonography. *AJR American journal of roentgenology*. 2006 Mar;186(3):696-702. PubMed PMID: 16498097.
37. Taylor SA, Charman SC, Lefere P, McFarland EG, Paulson EK, Yee J, et al. CT colonography: investigation of the optimum reader paradigm by using computer-aided detection software. *Radiology*. 2008 Feb;246(2):463-71. PubMed PMID: 18094263.
38. Taylor SA, Laghi A, Lefere P, Halligan S, Stoker J. European Society of Gastrointestinal and Abdominal Radiology (ESGAR): consensus statement on CT colonography. *European radiology*. 2007 Feb;17(2):575-9. PubMed PMID: 16967260.
39. Halligan S, Taylor SA, Dehmeshki J, Amin H, Ye X, Tsang J, et al. Computer-assisted detection for CT colonography: external validation. *Clinical radiology*. 2006 Sep;61(9):758-63; discussion 64-5. PubMed PMID: 16905382.
40. Dodd LE, Wagner RF, Armato SG, 3rd, McNitt-Gray MF, Beiden S, Chan HP, et al. Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the lung image database consortium. *Academic radiology*. 2004 Apr;11(4):462-75. PubMed PMID: 15109018.
41. Punwani S, Halligan S, Irving P, Bloom S, Bungay A, Greenhalgh R, et al. Measurement of colonic polyps by radiologists and endoscopists: who is most accurate? *European radiology*. 2008 May;18(5):874-81. PubMed PMID: 18176807.
42. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG. US women's attitudes to false-positive mammography results and detection of ductal carcinoma in situ: cross-sectional survey. *The Western journal of medicine*. 2000 Nov;173(5):307-12. PubMed PMID: 11069862. Pubmed Central PMCID: 1071147.
43. Mallett S, Deeks JJ, Halligan S, Hopewell S, Cornelius V, Altman DG. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. *BMJ*. 2006 Aug 26;333(7565):413. PubMed PMID: 16849365. Epub 2006/07/20. eng.
44. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol*. 2010 Aug;63(8):854-61. PubMed PMID: 20056381. Epub 2010/01/09. eng.
45. Salz T, Richman AR, Brewer NT. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psycho-oncology*. 2010 Oct;19(10):1026-34. PubMed PMID: 20882572.
46. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*. 2007 Apr 5;356(14):1399-409. PubMed PMID: 17409321. Pubmed Central PMCID: 3182841. Epub 2007/04/06. eng.
47. Skaane P, Hofvind S, Skjennald A. Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: follow-up and final results of Oslo II study. *Radiology*. 2007 Sep;244(3):708-17. PubMed PMID: 17709826. Epub 2007/08/22. eng.
48. Yankaskas BC, Taplin SH, Ichikawa L, Geller BM, Rosenberg RD, Carney PA, et al. Association between mammography timing and measures of screening performance in the United States. *Radiology*. 2005 Feb;234(2):363-73. PubMed PMID: 15670994. Epub 2005/01/27. eng.
49. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol*. 1989 Mar;24(3):234-45. PubMed PMID: 2753640. Epub 1989/03/01. eng.
50. Metz CE. Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems. *Journal of the American College of Radiology*. 2006;3(6):413-22.
51. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *Bmj*. 2009;338:b605. PubMed PMID: 19477892.

52. Shiraishi J, Pesce LL, Metz CE, Doi K. Experimental Design and Data Analysis in Receiver Operating Characteristic Studies: Lessons Learned from Reports in Radiology from 1997 to 2006. *Radiology*. 2009 December 1, 2009;253(3):822-30.
53. Mallett S, Halligan S, Thompson M, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. *Bmj*. 2012;345:e3999. PubMed PMID: 22750423.
54. von Wagner C, Halligan S, Atkin WS, Lilford RJ, Morton D, Wardle J. Choosing between CT colonography and colonoscopy in the diagnostic context: a qualitative study of influences on patient preferences. *Health expectations : an international journal of public participation in health care and health policy*. 2009 Mar;12(1):18-26. PubMed PMID: 19250149.
55. Ghanouni A, Smith SG, Halligan S, Plumb A, Boone D, Magee MS, et al. Public perceptions and preferences for CT colonography or colonoscopy in colorectal cancer screening. *Patient education and counseling*. 2012 Oct;89(1):116-21. PubMed PMID: 22705250.
56. Dhruva SS, Phurrough SE, Salive ME, Redberg RF. CMS's landmark decision on CT colonography--examining the relevant data. *N Engl J Med*. 2009 Jun 25;360(26):2699-701. PubMed PMID: 19474419. Epub 2009/05/29. eng.
57. Ryan M. Discrete choice experiments in health care. *BMJ*. 2004 Feb 14;328(7436):360-1. PubMed PMID: 14962852. Pubmed Central PMCID: 341374. Epub 2004/02/14. eng.
58. Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. *BMJ*. 2000 June 3, 2000;320(7248):1530-3.
59. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA, et al. Conjoint analysis applications in health--a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*. 2011 Jun;14(4):403-13. PubMed PMID: 21669364.
60. Pisani P, Bray F, Parkin DM. Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. *International journal of cancer Journal international du cancer*. 2002 Jan 1;97(1):72-81. PubMed PMID: 11774246.
61. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. *CA: a cancer journal for clinicians*. 2009 Jul-Aug;59(4):225-49. PubMed PMID: 19474385.
62. Schoenfeld P, Cash B, Flood A, Dobhan R, Eastone J, Coyle W, et al. Colonoscopic screening of average-risk women for colorectal neoplasia. *The New England journal of medicine*. 2005 May 19;352(20):2061-8. PubMed PMID: 15901859.
63. Marshall D, Bridges JF, Hauber B, Cameron R, Donnalley L, Fyie K, et al. Conjoint Analysis Applications in Health - How are Studies being Designed and Reported?: An Update on Current Practice in the Published Literature between 2005 and 2008. *The patient*. 2010 Dec 1;3(4):249-56. PubMed PMID: 22273432.
64. Boynton PM, Wood GW, Greenhalgh T. Reaching beyond the white middle classes. *BMJ*. 2004 June 12, 2004;328(7453):1433-6.
65. Gatta G, Capocaccia R, Sant M, Bell CM, Coebergh JW, Damhuis RA, et al. Understanding variations in survival for colorectal cancer in Europe: a EURO CARE high resolution study. *Gut*. 2000;47(4):533-8.
66. Robinson MH, Hardcastle JD, Moss SM, Amar SS, Chamberlain JO, Armitage NC, et al. The risks of screening: data from the Nottingham randomised controlled trial of faecal occult blood screening for colorectal cancer. *Gut*. 1999 Oct;45(4):588-92. PubMed PMID: 10486370. Epub 1999/09/16. eng.
67. Winawer SJ. Colorectal cancer screening. *Best Practice & Research Clinical Gastroenterology*. 2007;21(6):1031-48.
68. Spiegelhalter D, Pearson M, Short I. Visualizing Uncertainty About the Future. *Science*. 2011 September 9, 2011;333(6048):1393-400.
69. Group UCCSP. Results of the first round of a demonstration pilot of screening for colorectal cancer in the United Kingdom. *BMJ*. 2004;329:133.
70. Boynton PM, Greenhalgh T. Selecting, designing, and developing your questionnaire. *BMJ*. 2004 May 29, 2004;328(7451):1312-5.

71. Von Wagner C, Knight K, Halligan S, Atkin W, Lilford R, Morton D, et al. Patient experiences of colonoscopy, barium enema and CT colonography: a qualitative study. *Br J Radiol.* 2009 January 1, 2009;82(973):13-9.
72. Eng J. Sample Size Estimation: How Many Individuals Should Be Studied? *1. Radiology.* 2003 May 1, 2003;227(2):309-13.
73. Marshall DA, Johnson FR, Phillips KA, Marshall JK, Thabane L, Kulin NA. Measuring patient preferences for colorectal cancer screening using a choice-format survey. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research.* 2007 Sep-Oct;10(5):415-30. PubMed PMID: 17888107.
74. Nayaradou M, Berchi C, Dejardin O, Launoy G. Eliciting population preferences for mass colorectal cancer screening organization. *Medical decision making : an international journal of the Society for Medical Decision Making.* 2010 Mar-Apr;30(2):224-33. PubMed PMID: 19692710.
75. Dachman AH, Obuchowski NA, Hoffmeister JW, Hinshaw JL, Frew MI, Winter TC, et al. Effect of computer-aided detection for CT colonography in a multireader, multicase trial. *Radiology.* 2010 Sep;256(3):827-35. PubMed PMID: 20663975. Pubmed Central PMCID: 2923726.
76. Summers RM, Yao J, Pickhardt PJ, Franaszek M, Bitter I, Brickman D, et al. Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. *Gastroenterology.* 2005;129(6):1832-44.
77. Halligan S, Mallett S, Altman DG, McQuillan J, Proud M, Beddoe G, et al. Incremental Benefit of Computer-aided Detection when Used as a Second and Concurrent Reader of CT Colonographic Data: Multiobserver Study. *Radiology.* 2010.
78. Halligan S, Mallett S, Altman DG, McQuillan J, Proud M, Beddoe G, et al. Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: multiobserver study. *Radiology.* 2011 Feb;258(2):469-76. PubMed PMID: 21084409.
79. Ryan M, Bate A, Eastmond CJ, Ludbrook A. Use of discrete choice experiments to elicit preferences. *Qual Health Care.* 2001 Sep;10 Suppl 1:i55-60. PubMed PMID: 11533440. Pubmed Central PMCID: 1765744. Epub 2001/09/05. eng.
80. Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, et al. Eliciting public preferences for healthcare: a systematic review of techniques. *Health Technol Assess.* 2001;5(5):1-186. PubMed PMID: 11262422.
81. Yi D, Ryan M, Campbell S, Elliott A, Torrance N, Chambers A, et al. Using discrete choice experiments to inform randomised controlled trials: an application to chronic low back pain management in primary care. *Eur J Pain.* 2011 May;15(5):531 e1-10. PubMed PMID: 21075024. Epub 2010/11/16. eng.
82. Watson V, Carnon A, Ryan M, Cox D. Involving the public in priority setting: a case study using discrete choice experiments. *J Public Health (Oxf).* 2011 Dec 15. PubMed PMID: 22173912. Epub 2011/12/17. Eng.
83. Ozdemir S, Mohamed AF, Johnson FR, Hauber AB. Who pays attention in stated-choice surveys? *Health Econ.* 2010 Jan;19(1):111-8. PubMed PMID: 19267358. Epub 2009/03/10. eng.
84. de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. *Health Econ.* 2012 Feb;21(2):145-72. PubMed PMID: 22223558. Epub 2012/01/10. eng.
85. Arnold D, Girling A, Stevens A, Lilford R. Comparison of direct and indirect methods of estimating health state utilities for resource allocation: review and empirical analysis. *BMJ.* 2009 July 22, 2009;339(jul20_3):b2688-.
86. Obuchowski NA, Hillis SL. Sample size tables for computer-aided detection studies. *AJR American journal of roentgenology.* 2011 Nov;197(5):W821-8. PubMed PMID: 22021528. Pubmed Central PMCID: 3494304.
87. Mang T, Peloschek P, Plank C, Maier A, Graser A, Weber M, et al. Effect of computer-aided detection as a second reader in multidetector-row CT colonography. *European radiology.* 2007 Oct;17(10):2598-607. PubMed PMID: 17351780.

88. Halligan S, Altman DG, Taylor SA, Mallett S, Deeks JJ, Bartram CI, et al. CT colonography in the detection of colorectal polyps and cancer: systematic review, meta-analysis, and proposed minimum data set for study level reporting. *Radiology*. 2005 Dec;237(3):893-904. PubMed PMID: 16304111.
89. Taylor SA, Burling D, Roddie M, Honeyfield L, McQuillan J, Bassett P, et al. Computer-aided detection for CT colonography: incremental benefit of observer training. *The British journal of radiology*. 2008 Mar;81(963):180-6. PubMed PMID: 18180260.
90. Neri E, Faggioni L, Regge D, Vagli P, Turini F, Cerri F, et al. CT colonography: role of a second reader CAD paradigm in the initial training of radiologists. *European journal of radiology*. 2011 Nov;80(2):303-9. PubMed PMID: 20832219.
91. Lieberman D. Debate: small (6-9 mm) and diminutive (1-5 mm) polyps noted on CTC: how should they be managed? *Gastrointestinal endoscopy clinics of North America*. 2010 Apr;20(2):239-43. PubMed PMID: 20451813.
92. Zalis ME, Barish MA, Choi JR, Dachman AH, Fenlon HM, Ferrucci JT, et al. CT colonography reporting and data system: a consensus proposal. *Radiology*. 2005 Jul;236(1):3-9. PubMed PMID: 15987959.
93. Wernli KJ, Rutter CM, Dachman AH, Zafar HM. Suspected Extracolonic Neoplasms Detected on CT Colonography: Literature Review and Possible Outcomes. *Academic radiology*. 2013 Mar 1. PubMed PMID: 23465379.
94. Yee J, Kumar NN, Godara S, Casamina JA, Hom R, Galdino G, et al. Extracolonic abnormalities discovered incidentally at CT colonography in a male population. *Radiology*. 2005 Aug;236(2):519-26. PubMed PMID: 16040909.
95. Kim DH, Pickhardt PJ, Taylor AJ, Leung WK, Winter TC, Hinshaw JL, et al. CT colonography versus colonoscopy for the detection of advanced neoplasia. *The New England journal of medicine*. 2007 Oct 4;357(14):1403-12. PubMed PMID: 17914041.
96. Veerappan GR, Ally MR, Choi JH, Pak JS, Maydonovitch C, Wong RK. Extracolonic findings on CT colonography increases yield of colorectal cancer screening. *AJR American journal of roentgenology*. 2010 Sep;195(3):677-86. PubMed PMID: 20729446.
97. Stoop EM, de Haan MC, de Wijkerslooth TR, Bossuyt PM, van Ballegooijen M, Nio CY, et al. Participation and yield of colonoscopy versus non-cathartic CT colonography in population-based screening for colorectal cancer: a randomised controlled trial. *The lancet oncology*. 2012 Jan;13(1):55-64. PubMed PMID: 22088831.
98. Pickhardt PJ, Hanson ME, Vanness DJ, Lo JY, Kim DH, Taylor AJ, et al. Unsuspected extracolonic findings at screening CT colonography: clinical and economic impact. *Radiology*. 2008 Oct;249(1):151-9. PubMed PMID: 18796673.
99. Pickhardt PJ, Kim DH, Meiners RJ, Wyatt KS, Hanson ME, Barlow DS, et al. Colorectal and extracolonic cancers detected at screening CT colonography in 10,286 asymptomatic adults. *Radiology*. 2010 Apr;255(1):83-8. PubMed PMID: 20308446.
100. Yee J, Sadda S, Aslam R, Yeh B. Extracolonic findings at CT colonography. *Gastrointestinal endoscopy clinics of North America*. 2010 Apr;20(2):305-22. PubMed PMID: 20451819.
101. Whitlock EP, Lin J, Liles E, Beil T, Fu R, O'Connor E, et al. Screening for Colorectal Cancer: An Updated Systematic Review. U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Rockville (MD)2008.
102. von Wagner C, Ghanouni A, Halligan S, Smith S, Dadswell E, Lilford RJ, et al. Patient acceptability and psychologic consequences of CT colonography compared with those of colonoscopy: results from a multicenter randomized controlled trial of symptomatic patients. *Radiology*. 2012 Jun;263(3):723-31. PubMed PMID: 22438366.
103. Brenner DJ, Georgsson MA. Mass screening with CT colonography: should the radiation exposure be of concern? *Gastroenterology*. 2005 Jul;129(1):328-37. PubMed PMID: 16012958.
104. Ahmed H, Naik G, Willoughby H, Edwards AG. Communicating risk. *Bmj*. 2012;344:e3996. PubMed PMID: 22709962.

105. Howlett DC, Drinkwater KJ, Lawrence D, Barter S, Nicholson T. Findings of the UK national audit evaluating image-guided or image-assisted liver biopsy. Part II. Minor and major complications and procedure-related mortality. *Radiology*. 2013 Jan;266(1):226-35. PubMed PMID: 23143026.
106. Berland LL. Incidental extracolonic findings on CT colonography: the impending deluge and its implications. *Journal of the American College of Radiology : JACR*. 2009 Jan;6(1):14-20. PubMed PMID: 19111266.
107. van de Wiel JC, Wang Y, Xu DM, van der Zaag-Loonen HJ, van der Jagt EJ, van Klaveren RJ, et al. Neglectable benefit of searching for incidental findings in the Dutch-Belgian lung cancer screening trial (NELSON) using low-dose multidetector CT. *European radiology*. 2007 Jun;17(6):1474-82. PubMed PMID: 17206426.
108. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. *Science*. 1974 Sep 27;185(4157):1124-31. PubMed PMID: 17835457.
109. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*. 2009 Jul 21;6(7):e1000097. PubMed PMID: 19621072. Pubmed Central PMCID: 2707599.
110. Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple observers and multiple tests: an ANOVA approach with dependent observations. *Comm Stat*. 1995;24:934-6.
111. Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. *Academic radiology*. 1995 Aug;2(8):709-16. PubMed PMID: 9419629.
112. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Investigative radiology*. 1992 Sep;27(9):723-31. PubMed PMID: 1399456.
113. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. *Academic radiology*. 1998 Sep;5(9):591-602. PubMed PMID: 9750888.
114. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. *Academic radiology*. 2005 Dec;12(12):1534-41. PubMed PMID: 16321742. Pubmed Central PMCID: 1550352.
115. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. *Academic radiology*. 2004 Nov;11(11):1260-73. PubMed PMID: 15561573.
116. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Statistics in medicine*. 2005 May 30;24(10):1579-607. PubMed PMID: 15685718.
117. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. *Statistics in medicine*. 2007 Feb 10;26(3):596-619. PubMed PMID: 16538699. Pubmed Central PMCID: 1769324.
118. Warren LM, Mackenzie A, Cooke J, Given-Wilson RM, Wallis MG, Chakraborty DP, et al. Effect of image quality on calcification detection in digital mammography. *Medical physics*. 2012 Jun;39(6):3202-13. PubMed PMID: 22755704.
119. Berbaum KS, Schartz KM, Caldwell RT, El-Khoury GY, Ohashi K, Madsen M, et al. Satisfaction of search for subtle skeletal fractures may not be induced by more serious skeletal injury. *Journal of the American College of Radiology : JACR*. 2012 May;9(5):344-51. PubMed PMID: 22554633. Pubmed Central PMCID: 3345121.
120. Destounis S, Somerville P, Murphy P, Seifert P. Perceived sufficiency of full-field digital mammograms with and without irreversible image data compression for comparison with next-year mammograms. *Journal of digital imaging*. 2011 Feb;24(1):66-74. PubMed PMID: 20162439. Pubmed Central PMCID: 3025124.
121. Jinzaki M, Matsumoto K, Kikuchi E, Sato K, Horiguchi Y, Nishiwaki Y, et al. Comparison of CT urography and excretory urography in the detection and localization of urothelial carcinoma of the upper urinary tract. *AJR American journal of roentgenology*. 2011 May;196(5):1102-9. PubMed PMID: 21512076.

122. Krupinski EA, Silverstein LD, Hashmi SF, Graham AR, Weinstein RS, Roehrig H. Observer performance using virtual pathology slides: impact of LCD color reproduction accuracy. *Journal of digital imaging*. 2012 Dec;25(6):738-43. PubMed PMID: 22546982. Pubmed Central PMCID: 3491161.
123. Leong DL, Rainford L, Haygood TM, Whitman GJ, Tchou PM, Geiser WR, et al. Verification of DICOM GSDF in complex backgrounds. *Journal of digital imaging*. 2012 Oct;25(5):662-9. PubMed PMID: 22535193. Pubmed Central PMCID: 3447097.
124. Nishida K, Yuen S, Kamoi K, Yamada K, Akazawa K, Ito H, et al. Incremental value of T2-weighted and diffusion-weighted MRI for prediction of biochemical recurrence after radical prostatectomy in clinically localized prostate cancer. *Acta radiologica*. 2011 Feb 1;52(1):120-6. PubMed PMID: 21498337.
125. Obuchowski NA, Meziane M, Dachman AH, Lieber ML, Mazzone PJ. What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance? *Academic radiology*. 2010 Jun;17(6):761-7. PubMed PMID: 20457419.
126. Okamoto S, Shiga T, Hattori N, Kubo N, Takei T, Katoh N, et al. Semiquantitative analysis of C-11 methionine PET may distinguish brain tumor recurrence from radiation necrosis even in small lesions. *Annals of nuclear medicine*. 2011 Apr;25(3):213-20. PubMed PMID: 21188660.
127. Reed WM, Ryan JT, McEntee MF, Evanoff MG, Brennan PC. The effect of abnormality-prevalence expectation on expert observer performance and visual search. *Radiology*. 2011 Mar;258(3):938-43. PubMed PMID: 21248231.
128. Svane G, Azavedo E, Lindman K, Urech M, Nilsson J, Weber N, et al. Clinical experience of photon counting breast tomosynthesis: comparison with traditional mammography. *Acta radiologica*. 2011 Mar 1;52(2):134-42. PubMed PMID: 21498340.
129. Szucs-Farkas Z, Kaelin I, Flach PM, Roszkopf A, Ruder TD, Triantafyllou M, et al. Detection of chest trauma with whole-body low-dose linear slit digital radiography: a multireader study. *AJR American journal of roentgenology*. 2010 May;194(5):W388-95. PubMed PMID: 20410383.
130. Webb LJ, Samei E, Lo JY, Baker JA, Ghate SV, Kim C, et al. Comparative performance of multiview stereoscopic and mammographic display modalities for breast lesion detection. *Medical physics*. 2011 Apr;38(4):1972-80. PubMed PMID: 21626930.
131. Yakabe M, Sakai S, Yabuuchi H, Matsuo Y, Kamitani T, Setoguchi T, et al. Effect of dose reduction on the ability of digital mammography to detect simulated microcalcifications. *Journal of digital imaging*. 2010 Oct;23(5):520-6. PubMed PMID: 19415382. Pubmed Central PMCID: 3046683.
132. Zanca F, Hillis SL, Claus F, Van Ongeval C, Celis V, Provoost V, et al. Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: results from independently conducted FROC studies in mammography. *Medical physics*. 2012 Oct;39(10):5917-29. PubMed PMID: 23039631. Pubmed Central PMCID: 3461051.
133. Aoki T, Oda N, Yamashita Y, Yamamoto K, Korogi Y. Usefulness of computerized method for lung nodule detection in digital chest radiographs using temporal subtraction images. *Academic radiology*. 2011 Aug;18(8):1000-5. PubMed PMID: 21718956.
134. Aoki T, Oda N, Yamashita Y, Yamamoto K, Korogi Y. Usefulness of computerized method for lung nodule detection on digital chest radiographs using similar subtraction images from different patients. *European journal of radiology*. 2012 May;81(5):1062-7. PubMed PMID: 21382681.
135. Berg WA, Madsen KS, Schilling K, Tartar M, Pisano ED, Larsen LH, et al. Comparative effectiveness of positron emission mammography and MRI in the contralateral breast of women with newly diagnosed breast cancer. *AJR American journal of roentgenology*. 2012 Jan;198(1):219-32. PubMed PMID: 22194501.
136. Bilello M, Suri N, Krejza J, Woo JH, Bagley LJ, Mamourian AC, et al. An approach to comparing accuracies of two FLAIR MR sequences in the detection of multiple sclerosis lesions in the brain in the absence of gold standard. *Academic radiology*. 2010 Jun;17(6):686-95. PubMed PMID: 20457413.

137. Choi HJ, Lee JH, Kang BS. Remote CT reading using an ultramobile PC and web-based remote viewing over a wireless network. *Journal of telemedicine and telecare*. 2012 Jan;18(1):26-31. PubMed PMID: 22067287.
138. Cole EB, Toledano AY, Lundqvist M, Pisano ED. Comparison of radiologist performance with photon-counting full-field digital mammography to conventional full-field digital mammography. *Academic radiology*. 2012 Aug;19(8):916-22. PubMed PMID: 22537503.
139. Collettini F, Martin JC, Diekmann F, Fallenberg E, Engelken F, Ponder S, et al. Diagnostic performance of a Near-Infrared Breast Imaging system as adjunct to mammography versus X-ray mammography alone. *European radiology*. 2012 Feb;22(2):350-7. PubMed PMID: 21947512.
140. Dromain C, Thibault F, Diekmann F, Fallenberg EM, Jong RA, Koomen M, et al. Dual-energy contrast-enhanced digital mammography: initial clinical results of a multireader, multicase study. *Breast cancer research : BCR*. 2012;14(3):R94. PubMed PMID: 22697607. Pubmed Central PMCID: 3446357.
141. Gennaro G, Toledano A, di Maggio C, Baldan E, Bezzon E, La Grassa M, et al. Digital breast tomosynthesis versus digital mammography: a clinical performance study. *European radiology*. 2010 Jul;20(7):1545-53. PubMed PMID: 20033175.
142. Hupse R, Samulski M, Lobbes MB, Mann RM, Mus R, den Heeten GJ, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. *Radiology*. 2013 Jan;266(1):123-9. PubMed PMID: 23091171.
143. Kelly KM, Dean J, Lee SJ, Comulada WS. Breast cancer detection: radiologists' performance using mammography with and without automated whole-breast ultrasound. *European radiology*. 2010 Nov;20(11):2557-64. PubMed PMID: 20632009. Pubmed Central PMCID: 2948156.
144. Kim H, Choi D, Lee JH, Lee SJ, Jo H, Gwak GY, et al. High-risk esophageal varices in patients treated with locoregional therapy for hepatocellular carcinoma: assessment with liver computed tomography. *World journal of gastroenterology : WJG*. 2012 Sep 21;18(35):4905-11. PubMed PMID: 23002363. Pubmed Central PMCID: 3447273.
145. Kim S, Yoon CS, Ryu JA, Lee S, Park YS, Kim SS, et al. A comparison of the diagnostic performances of visceral organ-targeted versus spine-targeted protocols for the evaluation of spinal fractures using sixteen-channel multidetector row computed tomography: is additional spine-targeted computed tomography necessary to evaluate thoracolumbar spinal fractures in blunt trauma victims? *The Journal of trauma*. 2010 Aug;69(2):437-46. PubMed PMID: 20699755.
146. Li F, Engelmann R, Pesce L, Armato SG, 3rd, Macmahon H. Improved detection of focal pneumonia by chest radiography with bone suppression imaging. *European radiology*. 2012 Dec;22(12):2729-35. PubMed PMID: 22763504.
147. Li F, Engelmann R, Pesce LL, Doi K, Metz CE, Macmahon H. Small lung cancers: improved detection by use of bone suppression imaging--comparison with dual-energy subtraction chest radiography. *Radiology*. 2011 Dec;261(3):937-49. PubMed PMID: 21946054.
148. Li F, Hara T, Shiraishi J, Engelmann R, MacMahon H, Doi K. Improved detection of subtle lung nodules by use of chest radiographs with bone suppression imaging: receiver operating characteristic analysis with and without localization. *AJR American journal of roentgenology*. 2011 May;196(5):W535-41. PubMed PMID: 21512042.
149. Matsushima M, Naganawa S, Ikeda M, Itoh S, Ogawa H, Komada T, et al. Diagnostic value of SPIO-mediated breath-hold, black-blood, fluid-attenuated, inversion recovery (BH-BB-FLAIR) imaging in patients with hepatocellular carcinomas. *Magnetic resonance in medical sciences : MRMS : an official journal of Japan Society of Magnetic Resonance in Medicine*. 2010;9(2):49-58. PubMed PMID: 20585194.
150. McNulty JP, Ryan JT, Evanoff MG, Rainford LA. Flexible image evaluation: iPad versus secondary-class monitors for review of MR spinal emergency cases, a comparative study. *Academic radiology*. 2012 Aug;19(8):1023-8. PubMed PMID: 22503894.
151. Medved M, Fan X, Abe H, Newstead GM, Wood AM, Shimauchi A, et al. Non-contrast enhanced MRI for evaluation of breast lesions: comparison of non-contrast enhanced high

- spectral and spatial resolution (HiSS) images versus contrast enhanced fat-suppressed images. *Academic radiology*. 2011 Dec;18(12):1467-74. PubMed PMID: 21962476. Pubmed Central PMCID: 3210583.
152. Mermuys K, De Geeter F, Bacher K, Van De Moortele K, Coenegrachts K, Steyaert L, et al. Digital tomosynthesis in the detection of urolithiasis: Diagnostic performance and dosimetry compared with digital radiography with MDCT as the reference standard. *AJR American journal of roentgenology*. 2010 Jul;195(1):161-7. PubMed PMID: 20566811.
153. Moin P, Deshpande R, Sayre J, Messer E, Gupte S, Romsdahl H, et al. An observer study for a computer-aided reading protocol (CARP) in the screening environment for digital mammography. *Academic radiology*. 2011 Nov;18(11):1420-9. PubMed PMID: 21971259.
154. Muramatsu C, Schmidt RA, Shiraishi J, Li Q, Doi K. Presentation of similar images as a reference for distinction between benign and malignant masses on mammograms: analysis of initial observer study. *Journal of digital imaging*. 2010 Oct;23(5):592-602. PubMed PMID: 20054606. Pubmed Central PMCID: 3046675.
155. Noroozian M, Hadjiiski L, Rahnema-Moghadam S, Klein KA, Jeffries DO, Pinsky RW, et al. Digital breast tomosynthesis is comparable to mammographic spot views for mass characterization. *Radiology*. 2012 Jan;262(1):61-8. PubMed PMID: 21998048. Pubmed Central PMCID: 3244671.
156. Ohgiya Y, Suyama J, Seino N, Hashizume T, Kawahara M, Sai S, et al. Diagnostic accuracy of ultra-high-b-value 3.0-T diffusion-weighted MR imaging for detection of prostate cancer. *Clinical imaging*. 2012 Sep-Oct;36(5):526-31. PubMed PMID: 22920357.
157. Otani H, Nitta N, Ikeda M, Nagatani Y, Tanaka T, Kitahara H, et al. Flat-panel detector computed tomography imaging: observer performance in detecting pulmonary nodules in comparison with conventional chest radiography and multidetector computed tomography. *Journal of thoracic imaging*. 2012 Jan;27(1):51-7. PubMed PMID: 21307781.
158. Padilla F, Roubidoux MA, Paramagul C, Sinha SP, Goodsitt MM, Le Carpentier GL, et al. Breast mass characterization using 3-dimensional automated ultrasound as an adjunct to digital breast tomosynthesis: a pilot study. *Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine*. 2013 Jan;32(1):93-104. PubMed PMID: 23269714. Pubmed Central PMCID: 3556642.
159. Pollard BJ, Samei E, Chawla AS, Beam C, Heyneman LE, Koweek LM, et al. The effects of ambient lighting in chest radiology reading rooms. *Journal of digital imaging*. 2012 Aug;25(4):520-6. PubMed PMID: 22349990. Pubmed Central PMCID: 3389094.
160. Puryrsko AS, Remer EM, Coppa CP, Obuchowski NA, Schneider E, Veniero JC. Characteristics and distinguishing features of hepatocellular adenoma and focal nodular hyperplasia on gadoxetate disodium-enhanced MRI. *AJR American journal of roentgenology*. 2012 Jan;198(1):115-23. PubMed PMID: 22194486.
161. Rafferty EA, Park JM, Philpotts LE, Poplack SP, Sumkin JH, Halpern EF, et al. Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial. *Radiology*. 2013 Jan;266(1):104-13. PubMed PMID: 23169790.
162. Saade C, Bourne R, Wilkinson M, Evanoff M, Brennan P. A reduced contrast volume acquisition regimen based on cardiovascular dynamics improves visualisation of head and neck vasculature with carotid MDCT angiography. *European journal of radiology*. 2013 Feb;82(2):e64-9. PubMed PMID: 23088881.
163. Salazar AJ, Camacho JC, Aguirre DA. Comparison between differently priced devices for digital capture of X-ray films using computed tomography as a gold standard: a multireader-multicase receiver operating characteristic curve study. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*. 2011 May;17(4):275-82. PubMed PMID: 21457011. Pubmed Central PMCID: 3099052.
164. Shimauchi A, Giger ML, Bhooshan N, Lan L, Pesce LL, Lee JK, et al. Evaluation of clinical breast MR imaging performed with prototype computer-aided diagnosis breast MR imaging workstation: reader study. *Radiology*. 2011 Mar;258(3):696-704. PubMed PMID: 21212365.

165. Shiraishi J, Abe H, Ichikawa K, Schmidt RA, Doi K. Observer study for evaluating potential utility of a super-high-resolution LCD in the detection of clustered microcalcifications on digital mammograms. *Journal of digital imaging*. 2010 Apr;23(2):161-9. PubMed PMID: 19277785. Pubmed Central PMCID: 3043779.
166. Subhas N, Kao A, Freire M, Polster JM, Obuchowski NA, Winalski CS. MRI of the knee ligaments and menisci: comparison of isotropic-resolution 3D and conventional 2D fast spin-echo sequences at 3 T. *AJR American journal of roentgenology*. 2011 Aug;197(2):442-50. PubMed PMID: 21785092.
167. Sung YM, Chung MJ, Lee KS, Choe BK. The influence of liquid crystal display monitors on observer performance for the detection of interstitial lung markings on both storage phosphor and flat-panel-detector chest radiography. *European journal of radiology*. 2010 Apr;74(1):275-9. PubMed PMID: 19304429.
168. Svahn TM, Chakraborty DP, Ikeda D, Zackrisson S, Do Y, Mattsson S, et al. Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy. *The British journal of radiology*. 2012 Nov;85(1019):e1074-82. PubMed PMID: 22674710. Pubmed Central PMCID: 3500806.
169. Takahashi N, Tsai DY, Lee Y, Kinoshita T, Ishii K, Tamura H, et al. Usefulness of z-score mapping for quantification of extent of hypoattenuation regions of hyperacute stroke in unenhanced computed tomography: analysis of radiologists' performance. *Journal of computer assisted tomography*. 2010 Sep-Oct;34(5):751-6. PubMed PMID: 20861780.
170. Tan T, Platel B, Huisman H, Sanchez CI, Mus R, Karssemeijer N. Computer-aided lesion diagnosis in automated 3-D breast ultrasound using coronal spiculation. *IEEE transactions on medical imaging*. 2012 May;31(5):1034-42. PubMed PMID: 22271831.
171. Timp S, Varela C, Karssemeijer N. Computer-aided diagnosis with temporal analysis to improve radiologists' interpretation of mammographic mass lesions. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*. 2010 May;14(3):803-8. PubMed PMID: 20403792.
172. Toomey RJ, Ryan JT, McEntee MF, Evanoff MG, Chakraborty DP, McNulty JP, et al. Diagnostic efficacy of handheld devices for emergency radiologic consultation. *AJR American journal of roentgenology*. 2010 Feb;194(2):469-74. PubMed PMID: 20093611. Pubmed Central PMCID: 2826276.
173. Uchiyama Y, Asano T, Kato H, Hara T, Kanematsu M, Hoshi H, et al. Computer-aided diagnosis for detection of lacunar infarcts on MR images: ROC analysis of radiologists' performance. *Journal of digital imaging*. 2012 Aug;25(4):497-503. PubMed PMID: 22215250. Pubmed Central PMCID: 3389095.
174. Visser R, Veldkamp WJ, Beijerinck D, Bun PA, Deurenberg JJ, Imhof-Tas MW, et al. Increase in perceived case suspiciousness due to local contrast optimisation in digital screening mammography. *European radiology*. 2012 Apr;22(4):908-14. PubMed PMID: 22071778. Pubmed Central PMCID: 3297744.
175. Wallis MG, Moa E, Zanca F, Leifland K, Danielsson M. Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution X-ray imaging observer study. *Radiology*. 2012 Mar;262(3):788-96. PubMed PMID: 22274840.
176. Wardlaw JM, von Kummer R, Farrall AJ, Chappell FM, Hill M, Perry D. A large web-based observer reliability study of early ischaemic signs on computed tomography. *The Acute Cerebral CT Evaluation of Stroke Study (ACCESS)*. *PloS one*. 2010;5(12):e15757. PubMed PMID: 21209901. Pubmed Central PMCID: 3012713.
177. Way T, Chan HP, Hadjiiski L, Sahiner B, Chughtai A, Song TK, et al. Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance. *Academic radiology*. 2010 Mar;17(3):323-32. PubMed PMID: 20152726. Pubmed Central PMCID: 3767437.
178. Yamada Y, Jinzaki M, Hasegawa I, Shiomi E, Sugiura H, Abe T, et al. Fast scanning tomosynthesis for the detection of pulmonary nodules: diagnostic performance compared with chest radiography, using multidetector-row computed tomography as the reference. *Investigative radiology*. 2011 Aug;46(8):471-7. PubMed PMID: 21487302.

179. Yamada Y, Mori H, Hijiya N, Matsumoto S, Takaji R, Kiyonaga M, et al. Extrahepatic bile duct cancer: invasion of the posterior hepatic plexuses--evaluation using multidetector CT. *Radiology*. 2012 May;263(2):419-28. PubMed PMID: 22447852.
180. Yoshida A, Tha KK, Fujima N, Zaitzu Y, Yoshida D, Tsukahara A, et al. Detection of brain metastases by 3-dimensional magnetic resonance imaging at 3 T: comparison between T1-weighted volume isotropic turbo spin echo acquisition and 3-dimensional T1-weighted fluid-attenuated inversion recovery imaging. *Journal of computer assisted tomography*. 2013 Jan-Feb;37(1):84-90. PubMed PMID: 23321838.
181. Mallett S, Halligan S, Collins GS, Altman DG. Exploration of analysis methods for diagnostic imaging tests: Problems with ROC AUC and confidence scores in CT colonography. *PloS one*. 2014;(in press).
182. Shiraishi J, Pesce LL, Metz CE, Doi K. Experimental design and data analysis in receiver operating characteristic studies: lessons learned from reports in radiology from 1997 to 2006. *Radiology*. 2009 Dec;253(3):822-30. PubMed PMID: 19864510. Pubmed Central PMCID: 2786192.
183. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Radiology*. 2003 Jan;226(1):24-8. PubMed PMID: 12511664.
184. McClish DK. Evaluation of the accuracy of medical tests in a region around the optimal point. *Academic radiology*. 2012 Dec;19(12):1484-90. PubMed PMID: 23122567.
185. Hilden J. The area under the ROC curve and its competitors. *Medical decision making : an international journal of the Society for Medical Decision Making*. 1991 Apr-Jun;11(2):95-101. PubMed PMID: 1865785.
186. Gerds TA, Hilden J. Calibration of models is not sufficient to justify NRI. *Statistics in medicine*. 2014 Aug 30;33(19):3419-20. PubMed PMID: 25042216.
187. Hilden J. Commentary: On NRI, IDI, and "good-looking" statistics with nothing underneath. *Epidemiology*. 2014 Mar;25(2):265-7. PubMed PMID: 24487208.
188. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in medicine*. 2014 Aug 30;33(19):3405-14. PubMed PMID: 23553436.
189. Van den Bruel A, Jones C, Yang Y, Oke J, Hewitson P. People's willingness to accept over-detection in cancer screening: population survey. *Bmj*. 2015;350:h980. PubMed PMID: 25736617.