Eur Radiol (2015) 25:932–939
DOI 10.1007/s00330-014-3487-0

GASTROINTESTINAL

Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach

Steve Halligan · Douglas G. Altman · Susan Mallett

Received: 13 May 2014 / Revised: 16 September 2014 / Accepted: 3 November 2014 / Published online: 20 January 2015

Abstract
Objectives The objectives are to describe the disadvantages of the area under the receiver operating characteristic curve (ROC AUC) to measure diagnostic test performance and to propose an alternative based on net benefit.
Methods We use a 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 was 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 it 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.
Key points
- *The area under the receiver operating characteristic curve (ROC AUC) measures diagnostic accuracy.*
- *Confidence scores used to build ROC curves may be difficult to assign.*
- *False-positive and false-negative diagnoses have different misclassification costs.*
- *Excessive ROC curve extrapolation is undesirable.*
- *Net benefit methods may provide more meaningful and clinically interpretable results than ROC AUC.*

Keywords ROC curve · Sensitivity and specificity · Area under curve · Data interpretation · Statistical · CT colonography

Introduction

Radiologists interpret medical images in order to identify potentially harmful lesions. Test choice depends on many factors including availability and cost but is usually influenced most by how effectively the test resolves potential abnormalities. Sensitivity, how well a test identifies an abnormality, is a measure of diagnostic test accuracy very familiar to radiologists. Sensitivity is inextricably linked to specificity – how well a test identifies normal patients. Sensitivity and specificity usually move in different directions. Most obviously, if we reported every image as disease-positive, sensitivity would be 100 % but specificity would be 0 %, and normal patients would be subjected to unnecessary further investigation and possibly treatment, which would be inconvenient, illogical,

S. Halligan (✉)
Centre for Medical Imaging, University College Hospital, University College London, Podium Level 2, 235 Euston Road, London NW1 2BU, UK
e-mail: s.halligan@ucl.ac.uk

D. G. Altman
Centre for Statistics in Medicine, University of Oxford, Oxford, UK

S. Mallett
Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

precipitate anxiety, and be extremely costly. Conversely, if we reported every image as negative, specificity would be perfect but we would never diagnose any abnormality. Sensitivity and specificity are "two sides of the same coin" and should always be considered together, which can be difficult when comparing tests; if one test has high sensitivity and another high specificity, which is better? Combining sensitivity and specificity into a single measure of "diagnostic accuracy" facilitates comparisons. For radiologists, the most familiar combined measure is the area under the receiver operating characteristic curve (ROC AUC) [1].

The ROC curve

The ROC curve is a plot of the test true-positive rate (y-axis) against the corresponding false-positive rate (x-axis); i.e., sensitivity against 1-specificity (Fig. 1). The curve is built from test performance at different "diagnostic thresholds." For example, while urinalysis for glucose is "present/absent," blood sugar has a range of normal values. While diabetes is increasingly likely with higher values, the proportion of patients ultimately diagnosed depends on applying a diagnostic threshold that denotes a positive test. For imaging tests that depend on subjective interpretation, a threshold can be applied at a level that reflects diagnostic confidence, e.g., the mammographic BI-RADS scale: "negative", "benign", "probably benign", "suspicious", and "highly suggestive of malignancy" [2]. Broader scales use 0 (definitely no disease) to 100 (definitely disease) [3]. Scales amalgamate whether lesions are resolved by imaging, whether a radiologist perceived a lesion, and whether it was interpreted correctly. For example, consider a research study of a radiologist faced with CT colonography examinations from 100 patients, 50 of whom have colon cancer. Although competent radiologists will usually make the correct diagnosis, occasionally they will not because small cancers may be unresolved, missed, or misinterpreted as spasm; even "obvious" tumours are sometimes missed. Whereas in clinical practice we must apply a single diagnostic threshold at which (and above) the patient has an abnormality and below which they do not, in research studies we can calculate the proportion of correct (true-positive) and incorrect (false-negative) diagnoses at all thresholds by comparing the test result for each patient with the true diagnosis known via an independent reference test(s).

Figure 1 shows our CT colonography example. If a threshold of "definitely cancer" is required for diagnosis, then most patients so labeled will probably have cancer. However, patients labeled "probably cancer" and below who have cancer will be missed with such a high threshold. Dropping the threshold to "probably cancer" increases the proportion of cancers detected (sensitivity increases) but more normal patients are labeled positive; the false-positive fraction increases (decreased specificity). Plotting the proportion of true-positive against false-positive patients at each diagnostic threshold builds a ROC curve (Fig. 1) and different test (and readers) may have different curves (Fig. 2). The ROC plot, therefore, describes test performance measured using sensitivity and specificity at different thresholds and is a composite of two distributions, patients with and without abnormalities (Fig. 3).

ROC AUC

ROC AUC, the area under the ROC curve, is often used to summarise test performance across all thresholds in the plot. Simplistically, AUC represents how likely it is that the test will rank two patients; one with a lesion and one without, in the correct order, across all possible thresholds [4, 5]. More intuitively, AUC is the chance that a randomly selected patient with a lesion will be ranked above a randomly selected normal patient [1]. A perfect test would have 100 % sensitivity with zero false-positives (100 % specificity), across all thresholds. This point lies at the extreme top left-hand corner of the ROC plot; AUC=1·0. Such tests don't exist in real life, and we expect some failure to separate normal and abnormal patients. A straight line connecting the extreme bottom-left (sensitivity, FPR: 0,0) and top-right (1,1) corners (the "chance diagonal") describes a test with no discrimination; AUC=0.5.

MRMC studies

"Multi-reader, multi-case" studies (MRMC) employ multiple readers interpreting multiple cases, to maximise statistical power and enhance generalisability of results [6, 7]. Different radiologists have different ROC curves. Some are more experienced, some are more competent, and all have different internal thresholds for reporting abnormalities ("over-callers" vs. "under-callers"). Because diagnosis for the *same* image may differ depending on the radiologist, no single ROC curve can really describe any imaging investigation incorporating human interpretation. As noted already, the curve combines the ability of the test to resolve lesions with the ability of observers to detect them.

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. MRMC studies introduce "clustering" because multiple radiologists view the same cases. For example, small lesions are generally seen less frequently than larger lesions, i.e., reader observations are clustered within cases. Similarly, more experienced

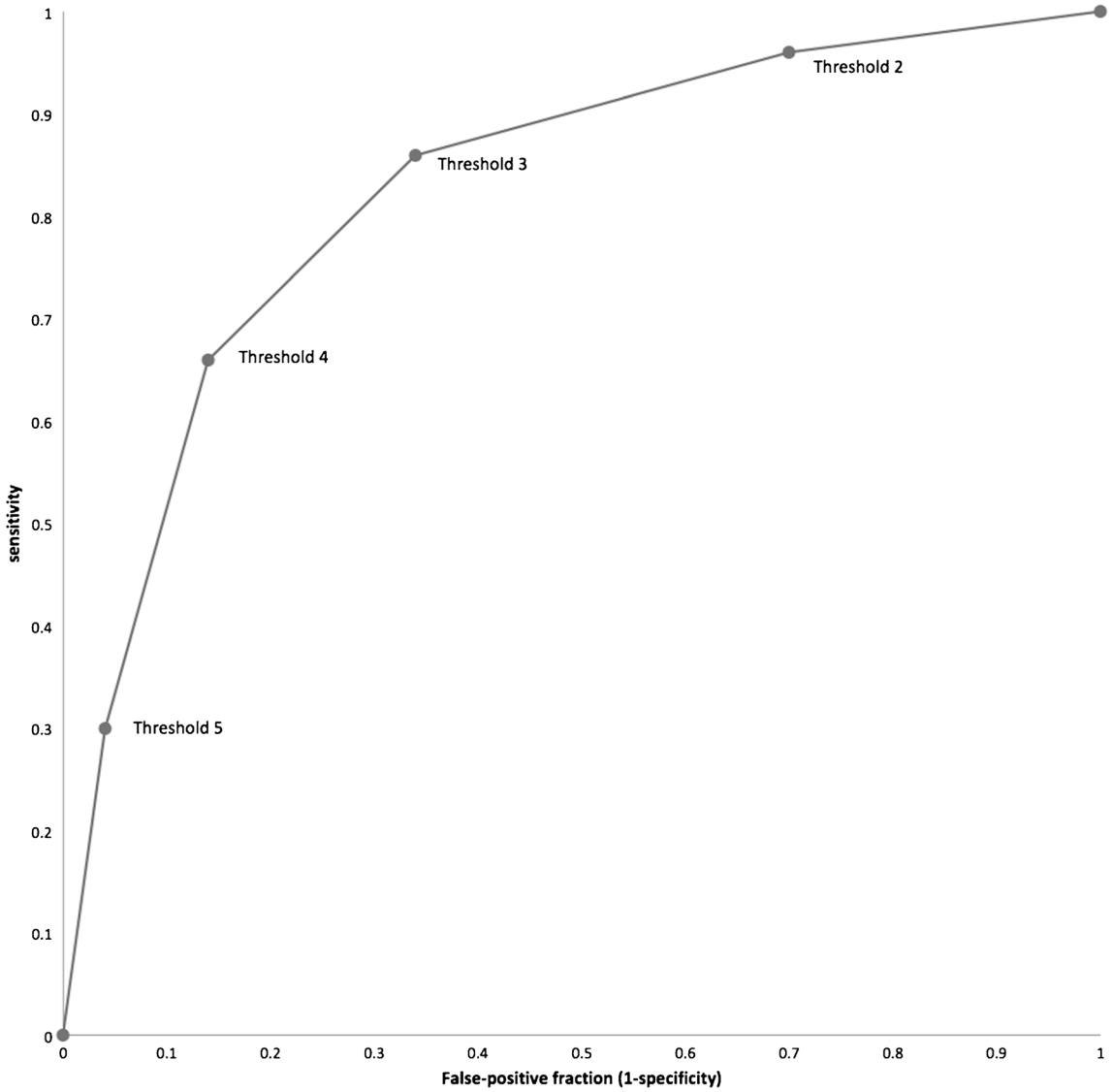

Fig. 1 Data from a hypothetical study of 100 patients who underwent CT colonography, 50 of whom have colorectal cancer. A radiologist uses the following rating scale to indicate their belief that the CT shows cancer in each individual patient: 1 – Definitely normal; 2 – Probably normal; 3 – Equivocal; 4 – Probably has cancer; 5 – Definitely has cancer. The following table details the rating when compared to the reference diagnosis in each case:

	Rating					
Reference diagnosis	1	2	3	4	5	Total
Cancer	2	5	10	18	15	50
No cancer	15	18	10	5	2	50

The ROC curve for these data is shown above. Assuming a threshold of 5 is required to diagnose cancer then 15 patients with cancer are correctly identified (30 % sensitivity) with only two false-positive diagnoses (96 % specificity). By lowering the threshold needed for diagnosis to 3 (i.e., all patients allocated a rating of 3 and above are considered to have cancer), then seven positive cases are missed (86 % sensitivity) with 17 false-positive diagnoses (specificity 66 %). Dropping the diagnostic threshold to include "definitely normal" (a small proportion of whom may actually have cancer) results in 100 % sensitivity but 0 % specificity since all patients are called positive. The empiric ROC AUC is 0.83 (calculated using a web-based calculator for ROC curves. Baltimore: Johns Hopkins University 2006 and drawn using GraphPad Prism 6·0, La Jolla, USA)

readers are likely to perform better across a series of cases than less experienced readers, i.e., results are correlated within readers. Analysis methods must account for clustering or the 95 % confidence intervals will be too narrow. Bootstrap resampling and multilevel modeling can account for clustering, linking results from the same observers and cases [8].

Advantages of ROC AUC

ROC AUC is a single metric facilitating comparison between tests without needing to "juggle" sensitivity and specificity. Proponents claim it measures "overall diagnostic performance" since ROC AUC is averaged across all possible

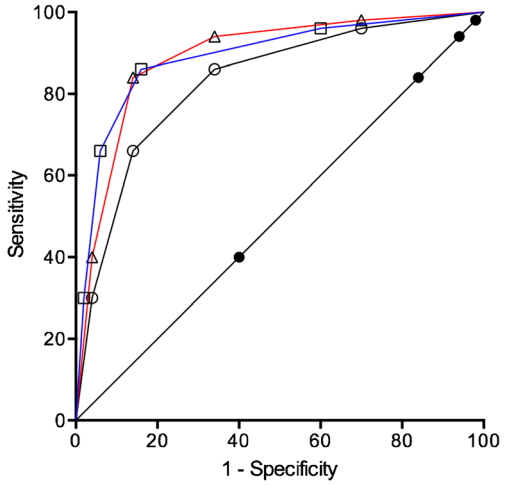

Fig. 2 The black ROC curve with circles at diagnostic thresholds shows the data from Fig. 1, with AUC 0.83. The red curve with triangles describes the following data where test specificity is the same as in Fig. 1, but sensitivity has been increased. Here, assuming a rating of 3 or more conveys a diagnosis of cancer, then three positive cases are missed (sensitivity 94 %) and 17 negative cases are labeled positive (specificity 66 %); the AUC is 0.89:

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

The blue ROC curve with squares describes the following data where test sensitivity is the same as Fig. 1, but specificity has been increased. Here, assuming a rating of 3 or more conveys a diagnosis of cancer, then seven positive cases are missed (sensitivity 86 %) and eight negative cases are labeled positive (specificity 84 %); the AUC is 0.89:

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

The black ROC curve with solid circles describes the following data where test sensitivity is increased to the same level as the red curve but with specificity dropped by a corresponding amount. Assuming a rating of 3 or more conveys a diagnosis of cancer, then just three cases are missed (sensitivity 94 %) but 47 negative cases are labelled positive (specificity 6 %). The empirical ROC AUC is 0.50 and the "curve" is actually a straight line:

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

diagnostic thresholds [9]. ROC AUC is constant across prevalence of abnormality [3] and authors argue that issues of prevalence and misclassifications costs (see Section 2 below) should only be considered once the "intrinsic" performance of a test is known [10]. ROC AUC is claimed to account for

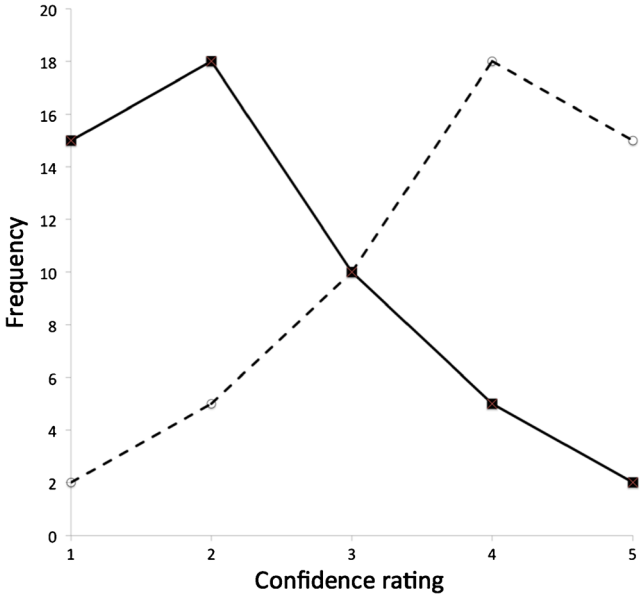

Fig. 3 Distribution of confidence scores (data from Fig. 1). Patients with cancer are represented by the dotted line and patients without cancer by the solid line. In this example using a threshold at greater or equal to a confidence rating of 3 provides the best separation between patients with and without cancer and is optimally balanced between sensitivity versus specificity. Moving this boundary to the right or left of the graph corresponds to raising or lowering the diagnostic threshold, respectively. If CT were perfect at discriminating between patient groups, the two distributions would not overlap. The less good a test at discriminating between patients with and without an abnormality, the more the two distributions will overlap

different thresholds between readers ("response criteria") and different curves are compared easily.

Disadvantages of ROC AUC

Clinical comprehension and relevance

Sensitivity and specificity are familiar concepts to clinicians, who are used to interpreting the results of diagnostic tests in these terms. In contrast, ROC AUC means little to clinicians (especially non-radiologists), patients, or health care providers. While a test whose AUC is 0.9 is considered "better" than one of 0.8, what does this mean for patients and what is clinically important? It is well established that diagnostic tests are understood best when presented in terms of gains and losses to individual patients [11]. AUC lacks clinical interpretability because it does not reflect this. Clinicians are uninterested in performance across all thresholds - they focus on clinically relevant thresholds. However, because AUC measures performance over all thresholds, it includes both those clinically relevant and clinically illogical. Moreover, different tests can have identical AUC but different performance at clinically important thresholds. Narrowing the range

of thresholds via "partial" AUC (pAUC) is possible [12] but choosing a single clinically important threshold is usually more practical.

Are sensitivity and specificity equally important?

ROC AUC treats sensitivity and specificity as equally important overall when averaged across all thresholds. But what if the clinical consequences of changes in sensitivity and specificity are very different? Consider our CT colonography example: Poor sensitivity could mean missed cancer and delayed treatment or even death, whereas poor specificity just means unnecessary colonoscopy. A recent study of colorectal cancer screening found that patients and healthcare professionals were willing to accept 2250 false-positive diagnoses in exchange for one additional true-positive cancer [13]. Similarly, for mammography, women will exchange 500 false-positives for one additional cancer [14].

ROC AUC ignores clinical differentials in "misclassification cost" and, therefore, risks finding a new test worthless when patients and physicians would consider otherwise.

Strictly speaking, ROC AUC weighs changes in sensitivity and specificity equally only where the curve slope equals one [5]. Other points assign different weights, determined by curve shape and without considering clinically meaningful information; e.g., a 5 % improvement in sensitivity contributes less to AUC at high specificity than at low specificity. Thus, AUC can consider a test that increases sensitivity at low specificity superior to one that increases sensitivity at high specificity. However, when screening, better tests must increase sensitivity at high specificity to avoid numerous false-positives [15].

Confidence scales may be inconsistent and unreliable

While confidence scales are used to construct ROC curves in radiology, there is little evidence they are assigned consistently and reliably. Confidence scales should ideally be ordinal, with a meaningful order and constant difference between points. However, a study asking what is "high confidence", found radiologists gave ten different interpretations including, "image quality is good", "the finding is obvious", and "the finding is familiar" [16]. Consistent scales are perturbed further by the multifaceted nature of radiological interpretation. For example, a rating may describe whether a pulmonary nodule is present or absent, whether it is benign or malignant, and also its location. Potentially, there are three tasks – detection, characterisation, and localisation. For the simplest analyses, empirical methods can be used to calculate and compare ROC AUC. However, multi-reader multi-case analyses usually require ratings be distributed normally (or transformed to normal distribution) for valid comparison of reader and test performance. However, having perceived an abnormality, readers are unlikely to state then they did so with low confidence. For example, the authors, undertaking a research study to seek US Food and Drug Administration approval for computer-assisted-detection (CAD) software for diagnosis of colorectal polyps [17], were obliged to use ROC AUC as the primary outcome for licensing. Following guidance [6, 18], readers rated the presence/absence of polyps using a 100 point (continuous) scale; 60 of 107 patients had polyps [17]. Confidence ratings were influenced strongly by polyp size, with larger polyps attracting higher scores. By definition, observers do not see false-negative polyps, so in true-negative patients' only false-positives attract ratings. True-negative patients may, therefore, attract no per-polyp score. While zero scores can be imposed when data coding, this scoring introduces a parallel binary rating method inconsistent with the continuous scale used by readers. We found confidence ratings highly non-normal because, in effect, there were two distributions, one continuous, and one binary (Fig. 4). While some suggest that extensive scales and encouragement to use the whole range will broaden distributions [3], this contradicts clinical practice where binary decisions are usual. Gur et al. [19] state, "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."

True-negative scores can potentially apply to normal patients or those with benign abnormalities. Lewin [20] describes 4,945 screening mammograms where zero scores were given to both cases classified as no abnormality and cases classified with benign abnormalities. We would expect better

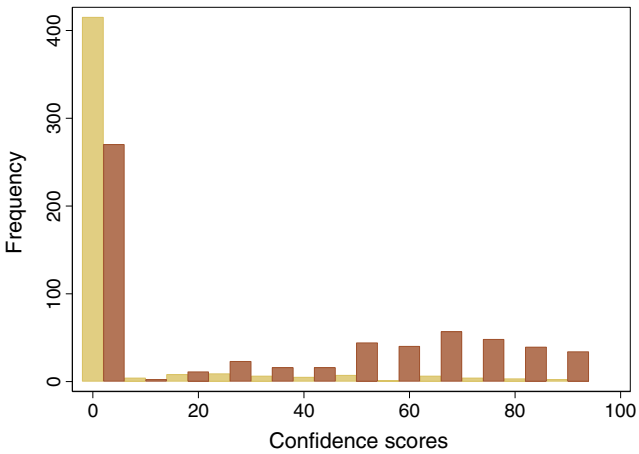

Fig. 4 Histogram of confidence ratings ascribed by ten radiologists in a prior study of CT colonography [17]. The dark brown bars represent ratings for 107 patients (of whom 60 had colon polyps) when using computer-assisted detection (CAD) whereas the light brown bars represent ratings when unassisted. 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

tests to improve confidence scores, but a zero score for normal cases cannot be improved whereas scores for benign lesions may improve if better imaging switches equivocal findings to benign. AUC summarises only a subset of study data as patients with zero or equivalent lowest scores do not contribute to AUC. In our colonography study [17] only 15 % to 47 % (depending on reader) of the 107 patients actually contributed to the curve and, hence, AUC. Harrington states, "The radiologist reports confidence levels only for a finding actually seen, or for a finding of normality. 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].

Extrapolation

In our study [17], few false-positives were reported, so data was clustered in the lower left portion of the ROC plot (Fig. 5). Completing a curve across all thresholds necessitates extrapolation beyond the last available data point. AUC is then dominated by a region containing no data and no clinically practical thresholds. Furthermore, the statistical method used for curve extrapolation also influences the calculated AUC [8] (Fig. 5). Gur states "selection of a specific analysis approach could affect the study conclusion" [21], noting problems with extrapolation, "when observers tend to be more decisive". Frequently, ROC curves cannot be fitted using standard methods due to "degenerate" data distributions. In our study this occurred in half the readers due to low false-positives. Also, when false-positive diagnoses are infrequent, those present exert disproportionate influence on curve shape versus more numerous true-positive scores; i.e., AUC is dominated by a small portion of observed data. In our study, without CAD-assistance the median number of patients with false-positive scores was just 2 of 107 [17].

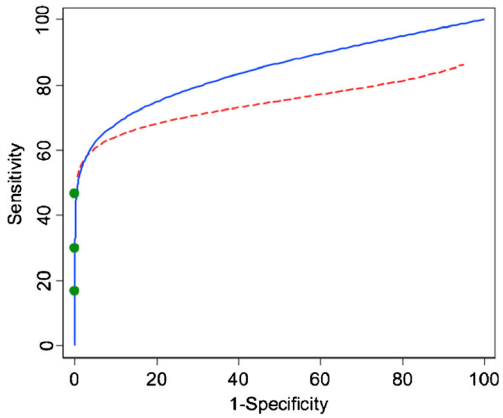

Fig. 5 Data extrapolation: ROC plots for an individual reader of CT colonography (without CAD) using data from a prior study [17]. Green dots indicate real data points underlying curve fitting. ROC curves are shown extrapolated from these data using two software methods, LabMRMC (red dashed line) and Proproc (blue solid line). It can be seen that the AUC depends on the method used for curve fitting, and that almost all of the AUC is determined by the extrapolated curve where there is no patient data

Prevalence of abnormality

A stated advantage of ROC AUC is its independence from prevalence of abnormality [10]; AUC is unchanged at differing prevalence. AUC corresponds to the probability of correctly ranking pairs of patients, one abnormal, one normal, implicitly suggesting 50 % prevalence. In clinical practice, the number of patients classified accurately by a test changes with prevalence. In high-prevalence situations the number of test-positive patients increases greatly for a given increase in sensitivity compared with low-prevalence situations (e.g., screening). AUC itself cannot account for how changing prevalence impacts on results for individual patients, so instead sensitivity and specificity are used, with the operating point directed by prevalence. While sensitivity and specificity are prevalence-independent, these measures separate positive and negative patients so prevalence can be incorporated by users as part of their interpretation.

Alternatives to ROC AUC

We have described problems with ROC AUC that encompass conceptual issues (confidence scores may be meaningless), statistical issues (non-normal distributions, extrapolation), practical issues (some patients do not contribute to AUC), and ethical issues (patients' and doctors' values cannot be incorporated easily). An alternative should be easy to comprehend and express, incorporate explicit weightings regarding gain in sensitivity vs loss of specificity, and account for prevalence. In particular, "costs" should be ascribed to the misclassification of true-positive and true-negative patients that account for the different clinical consequences of such misdiagnoses.

The need for alternatives to ROC AUC is well recognised, with several methods proposed. These have not penetrated the radiological literature, probably because ROC AUC is so predominant [22, 23]. Some methods move the operating point from one that optimises separation of events and non-events towards one or the other, depending on the relative misclassification costs [24]. "Net Reclassification Improvement", "Weighted Net Reclassification Improvement", and "Relative Utility" all account for differing consequences of correct and incorrect diagnosis [24]. Many measures, such as weighted-comparison [25] and Net Reclassification Index with two categories [26], are based directly on the difference in sensitivity and specificity between the tests assessed. We used a "net benefit" method in a second study of CAD for CT colonography [27]. Correct and

incorrect classification costs were expressed directly and adjustment for prevalence incorporated. A net benefit formula may be expressed as:

$$Net\ benefit = \Delta sensitivity + [\Delta specificity x\ (1/W)\ x\ (1-p)/p]$$

where Δsensitivity is the change in sensitivity and Δspecificity is the change in specificity when using CAD [28]. A net benefit will be positive if CAD is beneficial, zero indicates no benefit, and a negative value means a net loss. We would expect CAD to increase sensitivity but decrease specificity. As explained, increased sensitivity may be particularly desirable and outweigh the negative consequences of lowered specificity. To account for this, a weighting factor "W" is used to diminish the effect of reduced specificity via multiplying Δspecificity by 1/W (i.e., the larger W is, the less effect exerted by a given fall in specificity). The p is prevalence of abnormality in the population for which we calculate the benefit. At low prevalence, true-negative diagnosis is easier to achieve since most subjects are normal. The 1-p gives the proportion of normal subjects and dividing this by p gives the odds of having normal patients diagnosed over and above those with lesions. In our MRMC study we calculated average net benefit using a multilevel approach similar to meta-analysis, treating each reader as if in an individual "study." Bootstrap methods obtained 95 % confidence intervals empirically, accounting for the correct clustering of results within readers and cases. A significant benefit for a new test is defined as a positive net effect whose 95 % confidence interval does not include zero.

In our second study 16 radiologists read 112 cases with and without CAD assistance; 56 patients had 132 polyps. A challenging requirement is the need for assumptions regarding the relative costs of false-negative and false-positive classifications, reflected by W. While the precise value of W may be unknown, some insight will usually be available. We used W=3, a conservative estimate based on discussion with clinical colleagues; i.e., an additional true-positive was judged equal to the cost of three additional false-positives. The mean net benefit measure for second-read CAD overall was 6·2 % (95 %CI 3·1 % to 9·3 %) indicating significant benefit versus unassisted interpretation [27].

Advantages of net benefit methods

Net benefit combines sensitivity and specificity in a single metric, facilitating comparisons between tests. It provides advantages over ROC AUC as misclassification costs are transparent and incorporated explicitly (see worked example). Further, where W is unknown or known imprecisely, a range of weightings can be assigned via sensitivity analysis to examine the effect of different values. Errors induced by interpretation of confidence scores are avoided, prevalence is incorporated, and there is no need to fit curves or extrapolate beyond the data. Ultimately, net benefit is clinically relevant and interpreted easily since study data are expressed in terms of false-negative and false-positive patient diagnoses (specified as difference in sensitivity and difference in specificity).

Disadvantages of net benefit methods

As noted in the paragraph above, net benefit methods allow the effect of disease prevalence and misclassification costs to be incorporated explicitly into the analysis. However, a potential disadvantage is that these values must be known for them to be incorporated. Because most radiological research investigates applications that are ready for everyday use, the clinical context is usually established and estimates of disease prevalence in the population of interest should be relatively easy to obtain. Relative misclassification costs are more difficult to assess, especially with precision, because little research has been carried out in this area. However, such research in both mammographic [13] and colorectal cancer screening [14] has shown that patients and healthcare professionals greatly value gains in sensitivity over and above loss of specificity. Where the precise value of W has not been established, then it should be possible to arrive at a value by expert consensus. We used consensus to arrive at a value of 3 for our prior study [27], but subsequently found the precise value to be far higher [14], meaning that the initial analysis had underestimated the benefit of the new imaging test.

Summary: ROC AUC or net benefit?

Arguing for ROC AUC, Zweig and Campbell [10] 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". We agree, and believe ROC AUC to be most useful in the early stages of diagnostic test assessment, especially for tests not requiring subjective interpretation. However, most radiological research investigates tests or applications that are ready for clinical use, so the context *is* established. Because of this, meaningful evaluation must incorporate how the test influences results for individual patients, at a 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, net benefit methods provide meaningful and clinically interpretable results.

Acknowledgments The authors are very grateful to Professor Adrian Dixon, Emeritus Professor of Radiology Cambridge University, for his advice and suggestions regarding the manuscript.

This article presents independent research funded by the UK National Institute for Health Research (NIHR) under its Programme Grants for Applied Research funding scheme (RP-PG-0407-10338). The views expressed are those of the authors and not necessarily those of the UK NHS, the NIHR, or the Department of Health. A proportion of this work was undertaken at UCL and UCLH. They received a proportion of funding from the NIHR Biomedical Research Centre funding scheme. DGA is supported by a Cancer Research UK programme grant (C5529).

The scientific guarantor of this publication is Professor Steve Halligan. The authors of this manuscript declare relationships with the following companies: Professor Halligan has a non-remunerated research agreement with iCAD Inc. Two of the authors are statisticians. Institutional Review Board approval was not required for this work because it describes differing statistical approaches towards data analysis. No patients were approached for the purposes of the present study: We used existing data obtained from prior, published studies. The observer data used for the statistical examples in the present study was published previously as: Halligan S, Altman DG, Mallett S, Taylor SA, Burling D, Roddie M, Honeyfield L, McQuillan J, Amin H, Dehmeshki J. Computed tomographic colonography: Assessment of radiologist performance with and without computer-aided detection. **Gastroenterology** 2006;131:1690-9, and 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. Methodology: Retrospective analysis of existing data, Authors are from multiple institutions.

References

1. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
2. Boyer B, Canale S, Arfi-Rouche J, Monzani Q, Khaled W, Balleyguier C (2013) Variability and errors when applying the BIRADS mammography classification. Eur J Radiol 82:388–397
3. Obuchowski NA (2005) ROC analysis. AJR Am J Roentgenol 184: 364–372
4. Dwyer AJ (1996) In pursuit of a piece of the ROC. Radiology 201: 621–625
5. Fawcett T (2006) An Introduction to ROC analysis. Pattern Recogn Lett 27:861–874
6. Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM (2002) Assessment of medical imaging and computer-assist systems: lessons from recent experience. Acad Radiol 9:1264–1277
7. Obuchowski NA (2007) New methodological tools for multiple-reader ROC studies. Radiology 243:10–12
8. Obuchowski NA, Beiden SV, Berbaum KS et al (2004) Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. Acad Radiol 11:980–995
9. Park SH, Goo JM, Jo CH (2004) Receiver operating characteristic (ROC) curve: practical review for radiologists. Korean J Radiol 5: 11–18
10. Zweig MHCG (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577
11. Spiegelhalter D, Pearson M, Short I (2011) Visualizing uncertainty about the future. Science 333:1393–1400
12. McClish DK (1989) Analyzing a portion of the ROC curve. Med Decis Making 9:190–195
13. Boone D, Mallett S, Zhu S et al (2013) Patients' & healthcare professionals' values regarding true- & false-positive diagnosis when colorectal cancer screening by CT colonography: discrete choice experiment. PLoS One 8:e80767
14. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG (2000) US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. BMJ 320:1635–1640
15. Baker SG (2003) The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. J Natl Cancer Inst 95:511–515
16. Harrington MB (1990) Some methodological questions concerning receiver operating characteristic (ROC) analysis as a method for assessing image quality in radiology. J Digit Imaging 3:211–218
17. Halligan S, Altman DG, Mallett S et al (2006) Computed tomographic colonography: assessment of radiologist performance with and without computer-aided detection. Gastroenterology 131: 1690–1699
18. Wagner RF, Metz CE, Campbell G (2007) Assessment of medical imaging systems and computer aids: a tutorial review. Acad Radiol 14:723–748
19. Gur D, Rockette HE, Bandos AI (2007) "Binary" and "non-binary" detection tasks: are current performance measures optimal? Acad Radiol 14:871–876
20. Lewin JM, Hendrick RE, D'Orsi CJ et al (2001) Comparison of full-field digital mammography with screen-film mammography for cancer detection: results of 4,945 paired examinations. Radiology 218: 873–880
21. Gur D, Bandos AI, Rockette HE (2008) Comparing areas under receiver operating characteristic curves: potential impact of the "Last" experimentally measured operating point. Radiology 247: 12–15
22. Alemayehu D, Zou KH (2012) Applications of ROC analysis in medical research: recent developments and future directions. Acad Radiol 19:1457–1464
23. Zou KH (2012) Professor Charles E. Metz leaves profound legacy in ROC methodology: an introduction to the two Metz Memorial Issues. Acad Radiol 19:1447–1448
24. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW (2013) Evaluation of Markers and Risk Prediction Models: Overview of Relationships between NRI and Decision-Analytic Measures. Med Decis Making 33:490–501
25. Moons KG, Stijnen T, Michel BC et al (1997) Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves. Med Decis Making 17:447–454
26. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27: 157–172, discussion 207-112
27. Halligan S, Mallett S, Altman DG et al (2011) Incremental benefit of computer-aided detection when used as a second and concurrent reader of CT colonographic data: multiobserver study. Radiology 258:469–47628
28. Mallett S, Halligan S, Thompson M, Collins GS, Altman DG (2012) Interpreting diagnostic accuracy studies for patient care. BMJ 345: e3999