Influencing Clinicians and Healthcare Managers: Can ROC be more persuasive?

S. Taylor-Phillips*[a], M. G. Wallis[b], A. Duncan[c], A.G.Gale[a]
[a]Applied Vision Research Centre, Loughborough University, Leicestershire, LE11 3TU
[b]Cambridge Breast Unit, Addenbrookes Hospital, Cambridge, CB2 0QQ
[c]Warwickshire, Solihull, and Coventry Breast Screening Service, University Hospital (Coventry), Warwickshire, CV2 2DX

ABSTRACT

Receiver Operating Characteristic analysis provides a reliable and cost effective performance measurement tool, without using full clinical trials. However, when ROC analysis shows that performance is statistically superior in one condition than another it is difficult to relate this result to effects in practice, or even to determine whether it is clinically significant. In this paper we present two concurrent analyses: using ROC methods alongside single threshold recall rate data, and suggest that reporting both provides complimentary data. Four mammographers read 160 difficult cases (41% malignant) twice, with and without prior mammograms. Lesion location and probability of malignancy was reported for each case and analyzed using JAFROC. Concurrently each participant chose recall or return to screen for each case. JAFROC analysis showed that the presence of prior mammograms improved performance ($p<.05$). Single threshold data showed a trend towards a 26% increase in the number of false positive recalls without prior mammograms ($p=.056$). If this trend were present throughout the NHS Breast Screening Programme then discarding prior mammograms would correspond to an increase in recall rate from 4.6% to 5.3%, and 12,414 extra women recalled annually for assessment. Whilst ROC methods account for all possible thresholds of recall and have higher power, providing a single threshold example of false positive, false negative, and recall rates when reporting results could be more influential for clinicians. This paper discusses whether this is a useful additional method of presenting data, or whether it is misleading and inaccurate.

Keywords: ROC Methodology, Observer Performance Evaluation, Prior mammograms

1. INTRODUCTION

When a breast screening centre first upgrades from film to digital mammography they must take a decision of whether to continue to present the prior mammograms, which are of course in film format. The research presented here was designed to provide information to these breast screening centers so that they can make a more informed choice about whether it is worth the effort of hanging these film prior mammograms. This raises a wider question of how evidence from ROC studies using enriched case sets should be presented. Providing single threshold sensitivity and specificity data makes the data easier for a clinician to apply to a local setting, but is this appropriate considering an enriched data set is being used and a single recall threshold must be chosen?

*s.taylor-phillips@warwick.ac.uk

Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment,
edited by David J. Manning, Craig K. Abbey, Proc. of SPIE Vol. 7627, 76270X

Context: Evidence Based Radiology

Clinicians and healthcare managers are increasingly being encouraged to take the approach of evidence based medicine in their decision making. An understanding of this process can help researchers produce data which is more useful and applicable in decision making. Evidence Based Medicine (EBM) is concerned with systematically analyzing the available evidence to make both clinical decisions for individual patients and policy decisions. Evidence Based Radiology (EBR) follows similar principles, but Sardanelli *et al.* [1] describe several ways in which it differs. Firstly the need for the decision maker to have an in depth knowledge of the physics of image generation and manipulation, as these can have a very large effect on performance measurements. Secondly the high speed of technical developments resulting in a lack of time to produce meaningful data concerning the efficacy, effectiveness and cost effectiveness of each technology. Finally the harmful effects of the radiation associated with many imaging techniques impacts both the data which can be collected, and the technologies which can be implemented.

There are two possible approaches for EBR; top down and bottom up. The bottom up approach involves a practitioner for example a radiologist or a manager of a hospital department or screening centre defining a problem that they face (either from an individual patient or a policy issue) and finding, appraising and applying the findings from the available evidence. This relies on the practitioner having the necessary skills to source, critically analyze, and apply the evidence available, but does allow local factors to be considered. Top down EBR relies on institutions with specific expertise to systematically review the evidence and produce guidelines for local application. This has the advantages of availability of specialized expertise, but may be not directly applicable to a range of clinical situations, and may both be adhered to by the practitioner due to lack of involvement in and understanding of the processes involved. Additionally this is a time consuming process and therefore not always possible in the fast moving field of radiology. These two approaches are shown systematically in figure 1.

The distinction between top down and bottom up approaches is critically important when deciding how to disseminate results from a research study, as it determines who is the target recipient, their level of expertise in data analysis and appraisal, and their objectives. A large number of decisions in radiology are made at the level of the individual institution, where best practice would be use of bottom up evidence based radiology. This paper investigates whether we could be presenting research in a format which is of more use to the clinicians and managers making decisions in a bottom up manner for their individual institutions.

Clinicians and managers taking the bottom up approach to evidence based radiology must consider not only the efficacy of a diagnostic test (performance under ideal conditions), but also its effectiveness (performance in real world conditions) and its efficiency (whether it is cost effective versus other treatments). The Standards for Reporting Diagnostic Accuracy Studies (STARD)[2] provide an excellent reference to ensure the information is present to enable the reader to critically analyse both internal and external validity of the study. However, if results of a diagnostic test are presented using ROC curves, and there is a statistically significant difference between conditions, how can a clinician determine whether this difference is clinically significant in the context of their institution? The area under an ROC curve is equal to the probability that a radiologist will rank a randomly chosen abnormal case higher than a randomly chosen normal case. This is, of course, independent of decision threshold. When abnormality location information within the image is important, and tools such JAFROC are used, how can the reader interpret the magnitude in the difference in figures of merit between different conditions? Should results be extrapolated to 'real world' consequences in a specific setting i.e. provide false positive and false negative rates at a set threshold? This would render clinical significance easier to interpret, but has several associated issues. Should this be extrapolated further to include analysis of costs? There may be a trade off between pure scientific integrity and producing results which clinicians and managers taking a bottom up approach to EBR would find easier and less time consuming to interpret.

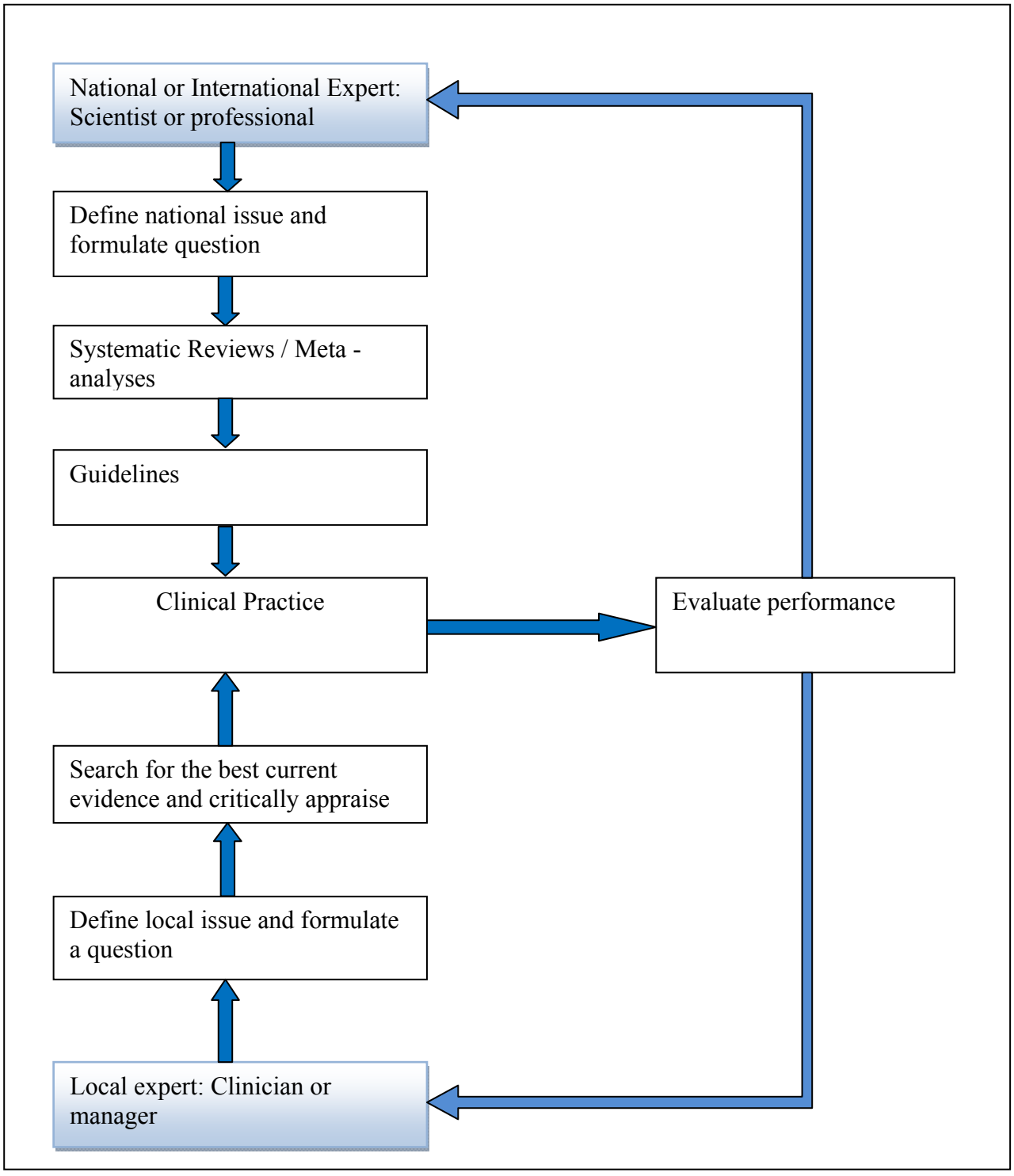

Figure 1 - Top down vs bottom up approaches to evidence based medicine. Adapted from Sardanelli (2010)

Application of Enriched Case ROC studies to Clinical Practice

There are several potential issues specifically with the application of ROC performance data from human observers reading enriched case sets to clinical practice using a set threshold:

1. The threshold for recall should be the same as that used in practice. Using a significantly different threshold for recall than used in practice could significantly affect measures of sensitivity and specificity.
2. The case set should mirror clinical practice. To ensure ROC results with a small case set it can be enriched with difficult cases and an increased number of abnormal cases. However, the test can still mirror clinical practice in the proportions of the different types of abnormalities.
3. The test environment should mirror the clinical environment as closely as possible. The equipment specifications and time pressures should be similar.

2. METHODS

Participants and Cases

Four participants from one breast screening centre in the UK took part in the study, two breast specialist radiologists and two radiography advanced practitioners. Radiography Advanced Practitioners are radiographers who have been trained and qualified to read mammograms in the NHS Breast Screening Programme (NHSBSP).

Each participant read an enriched case set of 160cases. These cases can be classified as follows: 66 abnormal cases proven by biopsy; 60 normal cases which had been recalled for further tests in the NHSBSP; 28 normal cases which had been arbitrated (i.e. one reader thought it should be recalled and the other did not, so it was referred to a third reader for arbitration) in the NHSBSP but not recalled for further tests; and 6 normal cases which were not recalled by either reader in the NHSBSP. Of the abnormal cases two had more than one lesion; they two lesions each.

Each participant was asked to mark the locations of any suspicious lesions, rate the probability of malignancy of each lesion identified on a scale of 1 to 100%, and state whether they would have recalled the case if they had encountered in whilst working in the NHSBSP.

Workstation

The workstation used in this study was digital, mammograms were obtained from the MicroDose Mammography system (Sectra, Sweden) and were displayed using Sectra mammography PACS on twin five megapixel LCD screens (EIZO, Japan). The previous mammograms were acquired using a Mammomat 3000 Nova (Siemens, Germany), with Kodak MIN-R2000 mammography film, and developed using a Kodak X-OMAT Multiloader 7000 (Carestream Health, Toronto, Canada). They were displayed on a Mammolux XL multi-viewer (Planilux, Germany), which was positioned adjacent and perpendicular to the digital workstation, as shown in figure 2.

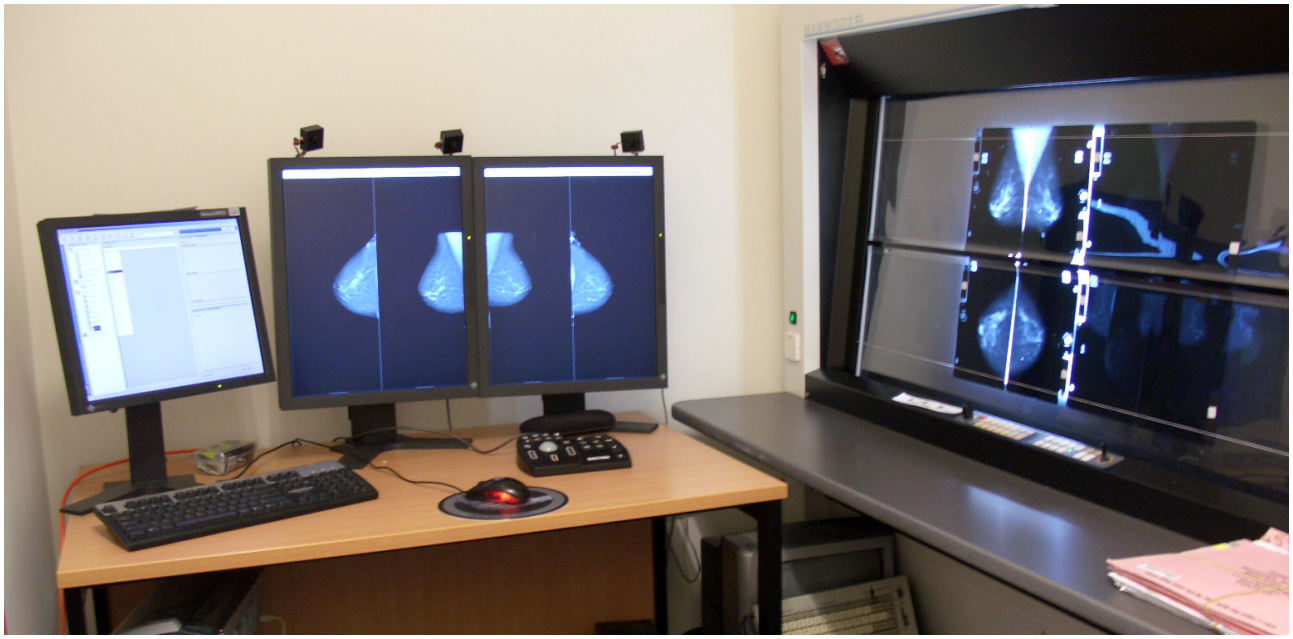

Figure 2 – The workstation used in the study. The current digital mammograms are displayed to the left on LCD screens and the film prior mammograms (when present) are displayed to the right on a multi-viewer.

Analysis

Overall performance was compared in the conditions with prior mammograms and without prior mammograms using JAFROC software [3], as is typical for observer performance studies with location information, and cases with more than one lesion. An additional analysis of whether the number of normal cases recalled differed between conditions was conducted using a within subjects students t test. This was considered a metric which would be of greater interest to clinicians. These results were then applied to an example breast screening centre to give a measure of clinical significance in one clinical setting.

3. RESULTS

JAFROC Analysis

JAFROC analysis showed that performance in digital mammography was greater with the film prior mammograms available and mounted on an adjacent multi-viewer ($p<.05$). These are illustrated using a FROC curve as shown in figure 3. This analysis may be considered sufficient for publication, and demonstrates that using the film prior mammograms produces a statistically significant improvement. However it may not advise a clinician or manager whether the effort of hanging all of the film prior mammograms resulted in a clinically significant improvement.

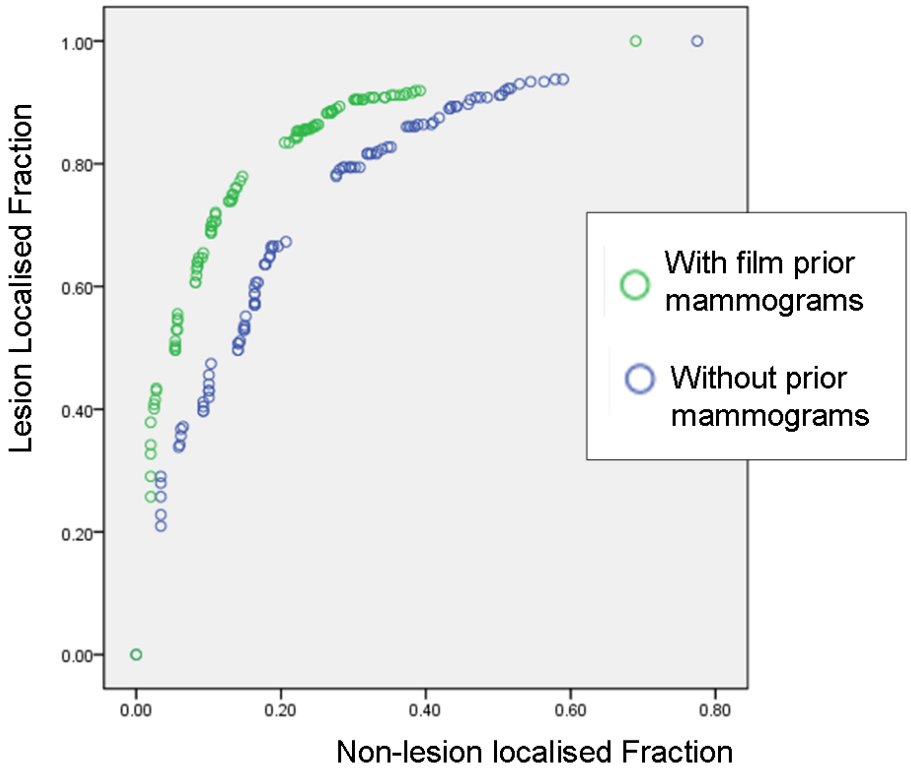

Figure 3 – A FROC curve of performance in digital mammography both with and without film prior mammograms.

Recall Decision Analysis

The number true positive responses did not differ between conditions i.e. the number of cancers detected did not differ between conditions. There was a trend towards a greater number of false positives (26% increase, p=.056) when the prior mammograms were not available in comparison to when they were displayed in film format (i.e. a trend towards recalling more women who do not have cancer when the prior mammograms are not available). If this trend were present throughout the NHS Breast Screening Programme then discarding prior mammograms would correspond to an increase in recall rate from 4.6% to 5.3%, and 12,414 extra women recalled annually for assessment.

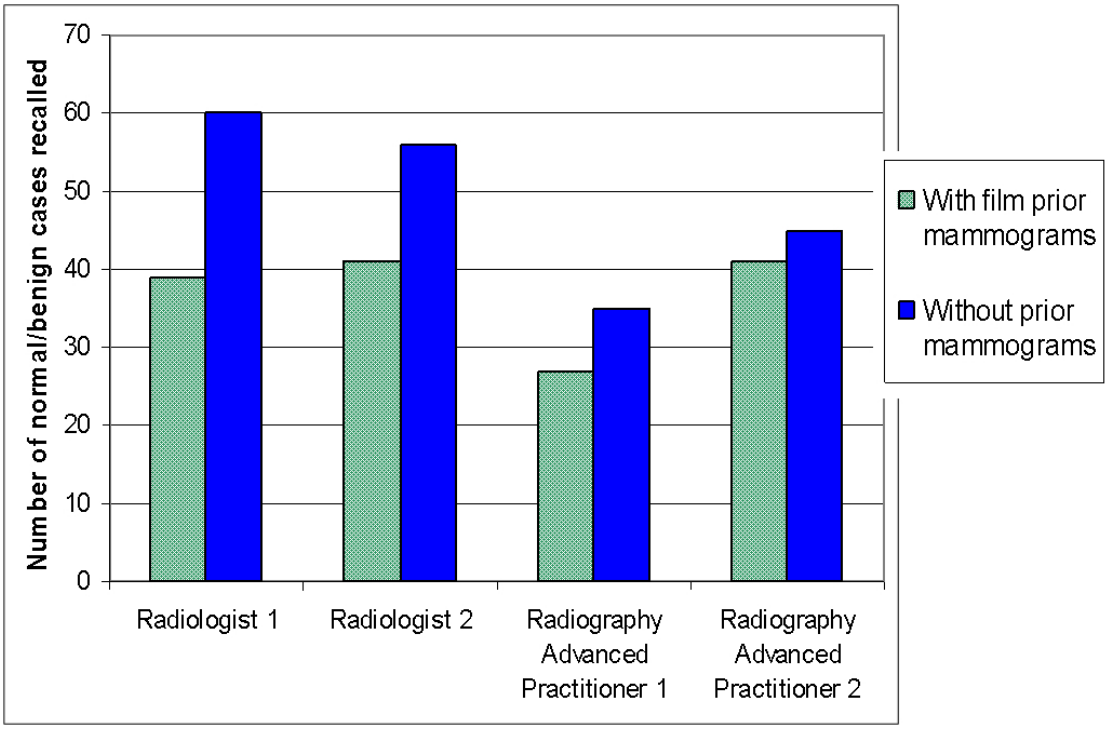

Figure 4 – The number of normal cases recalled using digital mammography both with and without film prior mammograms.

4. DISCUSSION

The three requirements outlined in the introduction for applying ROC studies to clinical practice using a single threshold were: using the same threshold as in clinical practice; the case set mirroring clinical practice; and the test environment mirroring clinical practice.

Using the same threshold as clinical practice appeared straightforward to implement. An extra question was simply added for each case asking the participant whether they would have recalled the case in clinical practice. However this may not have been an accurate representation of clinical practice as Gur *et. al.* [4] found that increasing the prevalence of disease in the case set was associated with decreases in the readers' levels of confidence. Therefore even though the experimental reporting is identical to that of clinical practice, simply enriching the case set for the test may change the radiologists' threshold for recall.

The case set approximately mirrored that of clinical practice. The proportions of the different types of abnormalities were not matched to the proportions experienced in a breast screening programme (i.e. the sampling was not stratified by abnormality type), but they were selected at random from the programme. This appears a reasonable approximation.

The test environment closely mirrored clinical practice as it was a workstation in a UK breast screening centre, which all participants were familiar with and used during the course of their work. This is as close to mirroring clinical practice as is possible, but the fact that the cases are not live screening cases, and that performance is being measured may encourage participants to adapt their behavior.

With the rapid introduction of technological advancements in radiology causing time pressures, and the economic downturn causing funding pressures, study designs using small numbers of participants reading enriched case sets for ROC analysis are increasingly attractive. However, these data are not directly applicable to clinicians taking an EBR approach. Extension of reporting to include single threshold sensitivity and specificity data could provide a richer source of data more applicable to clinical practice, but further research is needed to assess whether this approach is valid.

5. CONCLUSIONS

Presenting results of ROC studies in terms of single threshold specificity and sensitivity, or false positive and false negative rates could help clinicians and healthcare managers determine the clinical significance of the results in addition to the statistical significance. However further research is required to determine whether this is an appropriate strategy and produces results of sufficient internal and external validity.

REFERENCES

[1] Sardanelli, F., Hunink, M. G., Gilbert, F.J., Leo, G. D., and Krestin, G. P., "Evidence-based radiology: why and how?" Eur Radiol 20, 1-15 (2010).

[2] Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., and de Vet, H.C.W., "Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative," Radiol 226, 24-28 (2003).

[3] Chakraborty, D.P., Jackknife Free-Response Receiver Operating Characteristic Analysis Software, [computer software], Available at: www.devchakraborty.com [accessed on 1st July 2009].

[4] Gur, D., Bandos, A. I., Fuhrman, C. R., Klym, A. H., King, J. L., and Rockette, H. E., "The prevalence effect in a laboratory environment: changing the confidence ratings," Acad Radiol 14(1), 49-53 (2007).