Measuring performance in the interpretation of chest radiographs: a pilot study

Y. Chen a,*, J.J. James b, L. Dong a, A.G. Gale a

a Applied Vision Research Centre, Loughborough University, Loughborough, UK
b Nottingham Breast Institute, Nottingham University Hospitals, Nottingham, UK

AIM: To develop a system to assess the image interpretation performance of radiologists in identifying signs of malignancy on chest radiographs.

MATERIALS AND METHODS: A test set of 30 chest radiographs was chosen by an experienced radiologist consisting of 11 normal and 19 abnormal cases. The malignant cases all had biopsy-proven pathology; the normal and benign cases all had at least 2 years of imaging follow-up. Fourteen radiologists with a range of experiences were recruited. Participants individually read the test set displayed on a standard reporting workstation, with their findings entered directly into a laptop running specially designed reporting software. For each case, relevant clinical information was given and the reader was asked to mark any perceived abnormality and rate their level of suspicion on a five-point scale (normal, benign, indeterminate, suspicious, or malignant). On completion, participants were given instant feedback with performance parameters including sensitivity and specificity automatically calculated. An opportunity was then given to review the cases together with an expert opinion and pathology. The time each participant took to complete the test was recorded.

RESULTS: Six consultant radiologists who took part showed significantly better performance as determined by receiver operating characteristic (ROC) analysis compared to eight specialist registrars (area under the ROC curve [AUC]=0.9297 and 0.7648 respectively, p=0.003). There was a significant correlation with years of experience in the interpretation of chest radiographs and performance on the test set (r=0.573, p=0.032). Consultant radiologists completed the test significantly more quickly than the specialist registrars: mean time 19.65 minutes compared to 26.51 minutes (p=0.033).

CONCLUSION: It is possible to use a test set to measure individual differences in the interpretation of chest radiographs. This has the potential to be a useful tool in performance testing.

© 2016 Published by Elsevier Ltd on behalf of The Royal College of Radiologists.

Introduction

In recent years, there has been considerable interest in generating individual performance data in the UK National Health Service (NHS). An assessment of individual performance is part of the appraisal and revalidation process for all doctors in the UK and can be used to improve services, detect problems early, and inform the public.
Radiology is not immune to scrutiny and so consideration needs to be given to methods of assessing performance in the different subspecialties. Performance measures might include those derived from real-life data generated from individual clinical practice, a peer-review process where the clinical work of one radiologist is reviewed by another, and from batches of test cases. In the UK, the NHS Breast Screening Programme (NHSBSP) successfully uses a test-set-based self-assessment scheme to generate individual performance data that can be compared across the programme. The Personal Performance in Mammographic Screening (PERFORMS) scheme has been running for over 25 years and is currently undertaken by over 800 readers each year. A batch of 60 mammograms, consisting of a mixture of normal and abnormal studies are made available twice a year. Each case is viewed and interpreted by the reader with decisions entered into a password-protected website. At the end of the reading session, the individual receives immediate feedback on performance compared to an “expert opinion” generated from a panel of experienced breast radiologists. When all the readers have undertaken the test, detailed anonymised feedback is produced enabling an individual to see how their performance compares to their peers nationally.

Chest radiographs are one of the most commonly requested and reported imaging investigations worldwide and the first-line imaging test for patients with chest symptoms suspicious of malignancy. As with mammography, the ability to spot the early and potentially subtle signs of malignancy is crucial. Missed lung cancer on the chest radiograph is a very common source of error and cases are commonly encountered at error and discrepancy meetings.

The aim of this pilot study was to develop a similar system to the PERFORMS scheme for mammographic interpretation to assess the performance of radiologists in identifying signs of malignancy on chest radiographs.

Materials and methods

A test set of 30 digital chest radiographs was chosen by a radiologist (J.J.J.) with 13 years of experience of chest radiograph interpretation at consultant level, who also provided the expert opinion on each case. The case mix for the study consisted of 19 abnormal cases and 11 normal cases. The abnormal cases consisted of 16 biopsy-proven malignant lesions and three benign cases. The benign cases all had features that were judged to be abnormal on the chest radiograph at the time of clinical presentation and warranting further investigation: one was a case of multiple nodules (chicken pox pneumonia on CT and follow-up) and two were solitary masses (a biopsy-proven sclerosing haemangioma, and an area of inflammation that resolved on follow-up). All the malignant cases had biopsy-proven disease; the normal and benign cases also had at least 2 years of follow-up. All images were available as anonymised DICOM (digital imaging and communications in medicine) files.

Fourteen radiologists took part in the pilot study. Written consent was obtained from all participating radiologists (ethical approval was not deemed necessary following discussion with the local ethics committee). There were six consultant radiologists and eight radiology specialist registrars. The consultant radiologists experience ranged from 11–26 years (mean 17.8 years) and the registrars experience ranged from 1–5 years (mean 2.5 years).

Participants individually examined each case on a standard clinical workstation in a darkened reporting room with controlled ambient lighting. The usual image manipulation tools were available and relevant clinical information was provided. Findings were entered directly into a laptop running specially designed reporting software adapted from the PERFORMS scheme. The participant was asked to mark any perceived abnormality on a low-resolution version of the image on the laptop and rate their level of suspicion on a five-point scale: normal, benign, indeterminate, suspicious, or malignant. The participants were instructed that a rating of indeterminate, suspicious, or malignant meant that referral for further investigation was required. A normal or benign rating meant no further action was needed. Each participant read the test set in a random order, which was generated by the reporting software. Once the reporting software had identified the next case for reading, this was selected from the worklist displayed on the reporting workstation. The time taken for each participant to complete the test was recorded.

On completion of the test set, each participant was given instant feedback on their individual performance. Five personal performance parameters were calculated automatically by the software and were immediately available to view. There were two measures of sensitivity: cancer-detection rate and a correct recommendation for further investigation. This latter measure was used to take account of the cases that have an abnormality on the chest radiograph judged to require further assessment, which were subsequently found to be benign. Specificity, defined as a correct recommendation for no further investigation, was also calculated as was a positive and negative predictive value (PPV and NPV). If an individual’s result disagreed with the expert radiology opinion, this was highlighted. An opportunity was given to review any of the cases again, this time with the expert radiology opinion available to view and histopathology provided where applicable. After completing the test set and reviewing the results, participants were asked to fill in a questionnaire to provide feedback on the test itself and the educational potential.

Participants’ overall performance was examined by performing receiver operating characteristic (ROC) curve analysis, with areas under the curve (AUC) calculated. The results were analysed for the two groups of readers: consultant radiologists and specialist registrars. Analysis of variance (ANOVA), Bonferroni post hoc tests, and Pearson product–moment correlation coefficients were used to determine the significance of any difference in the performance parameters between the groups with a p-value of <0.05 considered significant.
Results

The performance of the consultant radiologists and specialist registrars is shown in Fig 1. The performance of the consultant group as determined by ROC analysis was significantly better than the specialist registrars (AUC=0.9297 and 0.7648, respectively, p=0.003).

Fig 2 shows the performance attributes for consultant radiologists and specialist registrars. There was a trend for superior performance by consultant radiologists across all the measures. The PPVs and NPVs of the consultant radiologists was significantly better than those of the specialist registrars (p=0.021 and p=0.045, respectively).

There was a significant correlation with years of experience in the interpretation of chest radiographs and test-set performance. This was determined by a positive correlation between years of experience and the AUC (r=0.573, p=0.032) for each of the 14 individuals. A scatterplot summarises the results (Fig 3). For specialist registrars as a group, there was a significant association between experience and performance (p=0.011). Consultant radiologists as a group showed no significant association between experience and performance.

Most participants finished the test within 30 minutes. Fig 4 shows the mean completion times of the test set for both groups, along with the standard errors. Consultant radiologists completed the test significantly more quickly than the specialist registrars: mean time 19.42 minutes (95% confidence interval [CI]: 15.94–22.9 compared to 27.71 minutes (95% CI:18.32–37.1; p=0.0397).

The post-test questionnaire revealed that all participants found the exercise useful for training purposes. Suggestions for improvement included making prior images available and supplying more demographic data. All the participants found the ability to review the cases along with an expert opinion once the test set had been completed useful.

Discussion

Errors will occur in any visual inspection task, however expert the reader. The measurement of individual performance is potentially an important tool in radiology. It has the potential to educate and improve the quality of interpretations, to spot problems early reducing potential harm, and provide evidence for the individual, institution, and public of a safe and effective service. In an ideal world, suitable performance measures would be extracted from an individual’s everyday reporting practice. When it can be generated, such information even is useful, but it does not allow comparisons between individuals and institutions where the case mix is different. Obtaining a true reflection of individual performance is also difficult when the incidence of abnormalities is relatively low and patients are not followed up. For instance, it may take several years before interpretation errors become apparent when a patient presents with more advanced disease. The use of test sets

![Figure 1](image-url)
Figure 2 Mean performance measures shown with standard error.

Figure 3 Scatterplot demonstrating the relationship between AUC and years of experience.
provides a solution, and this pilot study demonstrates that it would be possible to develop a performance testing scheme for the interpretation of chest radiographs along the lines of that used successfully for mammography interpretation for the last 25 years. The results show that it is possible to measure performance differences between individuals. Consultant radiologists showed significantly better performance compared to specialist registrars. The use of test sets to measure performance in the interpretation of chest radiographs is not new and others have also found that improved performance is associated with experience. Similarly, in the PERFORMS scheme for mammographic interpretation, performance is significantly related to years of reading experience. In the present study, only radiologist performance was measured, but others have used test sets to measure performance between different professional groups. For instance, Monnier-Cholley et al. showed similar performance between radiologists and chest physicians in the interpretation of chest radiographs, with both groups being significantly better than anaesthetists.

It is important for any performance-testing system to measure sensitivity and specificity. There are two potential sources of interpretation error that the participant could make. The first is under-reporting or the failure to spot the abnormality on the image. The second is over-reporting leading to the patient being referred for unnecessary additional tests, which adds to costs and also increases patient anxiety. For this pilot study, the software used in PERFORMS was adapted, which has evolved over the last 25 years as the mammographic interpretation testing scheme has matured. One of its strengths is the ability of each individual participant to obtain instant feedback, with key performance parameters including sensitivity and specificity instantly available. The software used in this test also requires the participant to mark the area that is considered abnormal. This avoids the situation where the participant may correctly score the case as abnormal, but have actually misclassified a normal feature. Once the test is completed the software allows the participant to review either all the cases along with the expert opinion and pathology available or just elect to review discrepant cases. Learning from the discrepancies can improve performance and the ability to review cases provides educational opportunities for the participant similar to those provided by error and discrepancy meetings and peer-review processes. The results of the post-test questionnaire complete by all the participants also confirmed the learning opportunities such schemes provide.

If assessing performance with tests sets is to be acceptable, it is important that taking part does not require an excessive time commitment. In this study, participants were able to work at their own rate with no set time limit; most completed the test within 30 minutes. Interestingly, the consultant radiologists who showed the better performance also finished significantly more quickly. Similar observations have also been noted in participants undertaking the PERFORMS scheme. Most readers will complete the PERFORMS set of 60 mammography studies in <2 hours. Those who completed the test within 1.5 hours showed no difference in cancer detection, but did show significantly better specificity.

One of the potential criticisms of using tests set to assess an individual is whether it is an accurate reflection of real-life performance. By its very nature, a test set is enriched with abnormal cases and the incidence of cancers is significantly higher than it would be in everyday practice. In addition, if a study is flagged as being abnormal requiring further investigation, the reader knows that no additional tests or investigations will be performed on the patient as they would be in real life. Consequently, the participant is viewing and interpreting studies in an artificial environment, which has the potential to interfere with image interpretation. Data from the PERFORMS scheme suggest that there are correlations with real-life performance, for instance, there is a strong correlation between the detection of small breast cancers in real life and the overall cancer detection rate in the PERFORMS set. Work is needed to...
see if a similar relationship can be demonstrated between the real-life interpretation of chest radiographs and performance in a test set.

In conclusion, this pilot study suggests that it is possible to measure individual performance in the interpretation of chest radiographs in a similar fashion to the PERFORMS scheme undertaken by readers of screening mammography in the UK. Clearly, these findings need to be validated in a much larger group of participants. There is already much interest in performance testing of individuals in the NHS. In radiology, the use of test sets to assess performance in a commonly requested and interpreted investigation, such as the chest radiograph, has the potential to be a useful tool.

References