RadBench : benchmarking image interpretation skills

WRIGHT, Chris and REEVES, Pauline

Published version

WRIGHT, Chris and REEVES, Pauline (2016). RadBench : benchmarking image interpretation skills. Radiography, 22 (2), e131-e136.

Introduction

The fact that radiographers have the ability to provide an accurate report on diagnostic images is well established [1-3]. The provision of a preliminary accurate opinion for all diagnostic images to the referring clinician ahead of the official report, offers the potential for rapid assessment of treatment requirements and optimisation of emergency department time [4,5.] Education and training can overcome the potential barriers to this approach[6,7], such as anxiety and transparency [3], and misconceptions or misunderstandings over medico-legal aspects [8]. The platform to underpin the move from radiographer abnormality detection systems (RADS) such as 'red dot' towards the provision of written comments (or preliminary clinical evaluation PCE) began by introducing image interpretation as an integral part of modern undergraduate education[9]. However, at least one study has concluded that this education, of itself, is insufficient [6].

The concept of accreditation (or benchmarking) has been applied to healthcare systems (particularly in the United States) for some time, but only recently has this included radiology [10] . Accreditation is said to promote professional development, amongst other benefits. The Society and College of Radiographers (SCoR) and the Australian Institute of Radiography (AIR) now offer formal accreditation of individual radiography advanced practitioners [11,12].

Over the past decade numerous authors have carried out a wide range of studies to investigate the image interpretation performance of different professions. Gold-standard accuracy of 95% is based on that of experienced consultant radiologists[13-15]. Image interpretation studies to date have broadly followed a similar quantitative methodology, either focusing on a single profession, or comparing one professional group with another. Many have been bespoke, relatively small scale studies, however there are examples of larger studies and systematic reviews [16-18]. Studies have also been carried out which investigated radiographers' abilities to provide a written comment after suitable further education[6,7.]

Method

The key aim of this pilot study was to develop an objective, accurate assessment tool with which to provide regular measurement and monitoring of image interpretation performance*. RadBench* is a software program which was conceived

as an approach to objectively measure image interpretation performance en masse and identify development needs. The research aimed to build and test a web based platform to enable benchmarking of image interpretation skills (with a view to its potential for testing across global populations).

Ethical approval was gained from the host university.

A participant sheet outlined the research and provided relevant information. In addition a registration form enabled the collection of demographic variables and written consent. Participants were assigned a unique software generated user code to provide anonymity. This code, along with the unique password, was required to enter the RadBench system.

As a starting point, two test banks were generated (Test 1 & Test 2), each containing twenty appendicular musculoskeletal images, half were normal, half contained fractures. The image banks were created to include variety of appendicular body parts of anticipated comparable difficulty: Ankle (3) Foot (4) Knee (3) Hand (3) Wrist (4) Elbow (3). Three images per test were paediatric, seventeen adult. All images were double reported by radiologists with consistent findings. They were then anonymised in accordance with ethical governance and data protection legislation. A response section was created within the software adjacent to each image presented. Images were presented sequentially, although the respondent had the option to go back and forth within the image set until the point of commitment. Each image could be maximised to full screen to optimise viewing. Certainty of decision making was assessed using a five point scale (definitely normal (1), probably normal (2), possibly abnormal (3), probably abnormal (4) or definitely abnormal (5)). A free response text box enabled the addition of clinical commentary.

A pilot study (n=42) was carried out within one calendar month to test the method and analysis approach. A convenience sample of volunteers included general radiographers (34), reporting radiographers (3), radiologists (2) (all from one UK NHS Trust), and medical imaging academics (3). Qualitative feedback on their experiences was also sought via *Survey Monkey* [19].

The benchmarking option within the software enabled the user to compare their score with the highest, lowest and mean score of others who had taken the same test. Feedback was provided in the form of a CPD certificate identifying accuracy, sensitivity, and specificity performance; a graphical display of decision making skills comparing ‘the ideal’ with their own performance; and an output table comparing the respondent's clinical commentary with the actual report highlighting any errors.

Results

Upon submission of the completed test, the *RadBench* software generated a calculation of sensitivity, specificity, and accuracy in addition to a decision making map. Early findings highlighted a 5% mean difference between image banks, confirming that benchmarking must be related to a specific test. This was despite the fact that the tests were designed to be (in principle) of equal difficulty. Half the candidates sat test 1 before test 2 and vice versa. Test 2 proved consistently more difficult regardless of the order taken. On average respondents took around twenty minutes to complete each test. All respondents completed both tests as requested with a short break between each one to reduce eye strain and relaxation time

Reporting radiographers (n=3), radiologists (n=2) and medical imaging academics (n=3) all scored 95 -100% with accurate anatomical identification in both tests. With education and experience, confidence in decision making improves. The image banks contained no equivocal cases and so, as expected, the experts made confident decisions each time, although did make the occasional error. Table 1 shows comparative data between this expert group and the group of general radiographers.

Insert table 1

The remainder of the results section will now focus on the general radiographer respondents (n=34) since these are the population of interest with regard to the proposed move from RADS to written commenting. The mean age of the general radiographer respondents was 37, with a span from 21 to 59. Of these, 18 were male and 24 female. Post graduate experience ranged from 4 to 26 years with a mean of 7.5 years. All were recruited from the same UK NHS Trust and were active

participants of a red dot abnormality detection scheme (RADS) at the time of testing. Mean accuracy was 84% for Test 1 and 79% for Test 2. Sensitivity was 92% and 86%, specificity was 77% and 73%, respectively as shown in Table 2. These results demonstrate how the content of a test may affect performance, confirming the need to benchmark by specific test. The mean scores of the two tests were calculated per respondent in order to provide a fairer reflection of performance, evening out the inter-test variation.

Insert table 2

The general radiographer population gained their radiography qualifying degree at eight different English Universities (see figure 1).

Insert figure 1

Figures 2,3 and 4 demonstrate the range of score for the combined test performance of the radiographers in terms of percentage accuracy, sensitivity and specificity.

Insert figures 2,3,4

Analysis of variance (ANOVA) between groups at a 95% confidence level demonstrated a statistically significant difference in Accuracy (P=0.019) and Sensitivity (P=0.001) although not in Specificity (P=0.340). Post hoc tests were not possible because at least one group had fewer than two members; however it is clear from figures 2-4 that University 8 is the outlier.

35% (n=12) of the general radiographer population had accuracy less than 80%. Of the 65% (n=22) who scored greater than this, 38% (n=13) scored 95% and over, producing decision making reliability consistent with reporting personnel. Figure 5 shows the level of accuracy plotted against years of experience of the radiographers. Interestingly, the mean level of accuracy drops the longer the experience of the participants. This suggests that, whilst decision-making may be more confident in those with more experience, unless this is backed up by continued training and development, the ability to make a correct decision may deteriorate.

Insert figure 5

Overall, decisions made by the 65% (n=22) of general radiographers scoring >80% accuracy correlated closely with comments suggesting that respondents were correctly identifying the anatomical regions of interest. Confidence in decision making is particularly useful in mapping further training and leads to an alternative approach to signalling; which may be termed the *traffic light* system. The 'definitely abnormal' decisions effectively equate to *red dot.* The 'definitely normal' decisions may be termed *green dot*. All other decisions are designated *amber* because the respondent is unsure. Their written comments did not change the computed accuracy, sensitivity and specificity score. The on-line software does not currently correct for text response however in this pilot study there were no examples of 'right for the wrong reason' or the reverse.

ROC was calculated using JROCFIT [20]. Receiver operating curves (ROC) were calculated retrospectively from the downloaded data using MS Excel [21] (See figure 6). Surprisingly, user feedback strongly recommended this function be available as a research output only, hence was excluded from the general results output to the user; however the raw data is available within RadBench for research purposes. In addition, upon completion of the test the user was provided with the actual reports versus their PCE enabling a qualitative interpretation designed to provide a positive impact on learning and development for the user.

Insert figure 6

Qualitative feedback via Survey Monkey regarding the *RadBench* platform and concept was extremely positive. Net promoter score was 100% with all participants recommending the product to their peers. Some minor design modifications were suggested, some to improve ergonomics and others to widen the scope of the application to include axial skeleton, chest, and other imaging modalities. Promotion requires focus in order to develop a global brand. Pricing raised some interesting points; whilst 90% would be willing to pay an annual fee, 55% of the general radiographers felt that this could be integrated into the SCoR or HCPC membership. Site licensing for NHS Trusts was recommended by 60%; interestingly this suggestion was also mirrored by the academics for Universities, but with the added request for greater access such that the product could be used assessment as part

of degree programmes. Suggested market extensions include other allied health professionals, particularly physiotherapy and nursing. A small change to the image bank format would provide consistency with the Fellowship of the Royal College of Radiologists (FRCR) rapid reporting assessment for radiologists and also open the platform to medical education.

Discussion

The evidence from this initial pilot study has confirmed that image interpretation performance varies with difficulty of the test, highlighting the importance of benchmarking to specific image banks. The authors propose the adoption of the following standard set of criteria ; 95% ideal, 90% optimal, and 80% minimal accuracy as an approach to categorising decision-making performance. The special interest group in radiography reporting (SIGRR) guidance also suggested a 95% standard[13,15]. Achieving a 95% performance standard in binary decision making (normal or abnormal) could be a credible goal for most general radiographers, particularly if they exit University education close to this performance level, although improving the quality of commenting to this level may be a tougher challenge in the short term. Reporting radiographers remain clearly differentiated at a higher level where 95% accuracy with confident decision making will continue to be expected, consistent with the postgraduate education required to produce a full accurate written report.

The results from both pilot image bank tests demonstrated higher sensitivity scores compared to specificity, indicating that the ability to identify fractures was better than the ability to identify normal variants. The overall effect was to reduce the accuracy score. This is consistent with other research [6,7]. This becomes the first point of focus for developing the non-reporting radiographer population. Is this due to insufficient education as part of undergraduate studies [6] or does knowledge lapse with increasing years? Further research will probe any link between undergraduate education, performance once qualified and post graduate educational interventions.

Most respondents appeared to take a professional, reasoned approach to decision making. This is consistent with recent findings reinforcing the need to scaffold learning[15]; first make the decision (RADS/ red dot) before progressing to commentary (PCE). A minority took the 'maverick' approach of making very confident

decisions (ranking 1 or 5) but without the underlying skills to underpin them, all too often making bad judgements which could have a negative impact on patient care and the radiographic profession. This tendency will form the basis of further research

The key aim was to develop an objective, accurate assessment tool with which to provide regular screening of performance, including measurement and monitoring; a function for which *RadBench* was specifically designed. Assessing decision making skills in terms of accuracy, sensitivity and specificity is the first stage in developing image interpretation skills. Only when a credible level of accuracy is achieved is it appropriate to make written comments or ultimately (with further postgraduate education) progress to reporting. The College of Radiographers (2013) stated that such levels of accuracy were difficult to define in quantitative terms but that radiographers participating in such schemes must demonstrate continuous professional development in respect of their participation. The *Radbench* software is designed to address both of these issues.

RadBench testing could be phased to complement local clinical audit for reporting radiographers. This is recommended by SIGRR on a monthly or two monthly basis, or more practically four monthly combined with annual summary and appraisal;[16] the latter being also more viable for en masse review. Delivering *RadBench* from a web based platform facilitates the benchmarking of anyone who has taken the same test in any geographical location, within any defined organisation or institution, and with any defined personal characteristics.

Further research is required in order create a viable strategy for the NHS. Budgets are limited and so measuring the return on investment in training is critical to quality and success. Protectionism is also an important factor as this can have a negative impact on otherwise logical strategic decisions within organisations. Additional work and consultation is required to develop the research of previous authors[17,18] , exploring the perceptions of the radiologist's viewpoint on radiographer role extension and re-educate the minority (18%) who do not support the concept of radiographer commenting [18]. There is also work to do within the radiography profession to avert the concerns of those who do not feel commenting should be part of the general role by routine [4].

By benchmarking performance across all radiography staff, managers can plan post graduate training activities as part of continuous professional development to maximum benefit. Where necessary or desired, individual development plans can be designed. Annual reassessment provides an auditable measure of performance appraisal. More frequent testing, for example before and after training interventions, may be implemented as desired. Performance is certificated through *RadBench* and forms a useful part of the CPD portfolio required to be maintained by the HCPC.

Once radiographers are determined to be reliably making the correct decisions, the progression is to improve their decision making confidence, and then they may decide, after further postgraduate education, to move on to writing reports. This may take some time however the *RadBench* tool again allows progress to be monitored. In parallel, the difficulty of the staging tests can be increased, as can the proportion type and /or of abnormal cases presented in order to reflect clinical practice. Image interpretation skills of this standard have the potential to radically change the NHS.

Limitations and next steps

A limitation of this pilot study was the size of the population and it is therefore inappropriate at this time to benchmark performance according to the number of years post registration experience, and geographical considerations; although early indications suggest that the training University (with associated clinical placements) may be a more significant factor than the actual number of years post qualification. Whilst it may be true to say that the more respondents you have from the same University, the greater the potential for variation in performance but this misses the point; Universities are supposed to produce graduates able to meet the CoR 2013 policy of delivering reliable PCE. The results presented are too small a sample to draw significant conclusions but the trend is obvious from the wide range of scores gained. This is a subject of further research, although at least one other study has concluded that undergraduate education is insufficient to meet the accuracy targets proposed [6].

Currently the software only detects the normal/ abnormal decision making for a single abnormality. Research is ongoing which may result in the ability to analyse the written comments; at the time of writing these required separate analysis.

Now that the tool has been tested on fracture recognition, future test banks will be developed to take account of other injuries and pathologies which may be encountered such as dislocations and osteomyelitis. There is also the potential to adjust or develop test banks which account for local disease prevalence; for example Paget's disease in the North West of England.

Conclusion

Benchmarking image interpretation performance is important because it provides the quantitative evidence to assess current status of potential or actual participants and develop training plans for the future. Utilising identical image bank tests across large scale populations increases the power of research. In 2014, 18,647 RadBench tests were taken by healthcare professions across the world; development and testing continues.

Decision making in musculoskeletal imaging, particularly through Accident and Emergency, is slowly moving away from the radiologist and into the domain of the reporting radiographer [19], although this elite group alone are unlikely to be able to sustain the service improvements modern healthcare demands. Taking a new approach, the general radiographer population (largely using the traditional ‘red-dot’ flagging schemes (RADS) at present) have the potential to be trained to provide initial written comments prior to the provision of a formal report; a process which should begin at the point of entry to undergraduate education. Benchmarking image interpretation performance at the point of UCAS application is the subject of further ongoing research. The *RadBench* platform potentially enables the hosting of an infinite number of image test banks, with instantaneous analysis, to facilitate the development of image interpretation skills by radiographers and other healthcare professionals (potentially on a global basis), in addition to providing a rich source of research data for future studies. Tomorrow's general radiographer could have an individual image interpretation performance rating which follows their career, also forming one of the component parts of HCPC registration.

A scaffolding approach to continuous professional development allows training to be tailored to specific needs, providing a rich talent pool capable of delivering accurate image interpretation decisions en masse, which in turn offers the modern NHS a whole new level of efficiency improvement potential and justified auditable return on

investment, and also creating a natural pool of talent ready to develop into the formal reporting role.