Radiography 22 (2016) e131–e136

Contents lists available at ScienceDirect

Radiography

journal homepage: www.elsevier.com/locate/radi

RadBench: Benchmarking image interpretation skills

C. Wright, P. Reeves*

Sheffield Hallam University, Robert Winston Building, Broomhall Rd, Sheffield S10 2BP, UK

ARTICLE INFO

Article history:
Received 9 February 2015
Received in revised form
28 December 2015
Accepted 30 December 2015
Available online 20 January 2016

Keywords:
Abnormality detection
Accreditation
Accuracy
Audit
CPD
Development needs

ABSTRACT

Purpose: The key aim of this research was to develop an objective, accurate assessment tool with which to provide regular measurement and monitoring of image interpretation performance. The tool was a specially developed software program (*RadBench*) by which to objectively measure image interpretation performance en masse and identify development needs.

Method: Two test banks were generated (Test 1 & Test 2), each containing twenty appendicular musculoskeletal images, half were normal, half contained fractures. All images were double reported by radiologists and anonymised. A study (n = 42) was carried out within one calendar month to test the method and analysis approach. The participants included general radiographers (34), reporting radiographers (3), radiologists (2) (all from one UK NHS Trust) and medical imaging academics (3).

Results: The *RadBench* software generated calculations of sensitivity, specificity, and accuracy in addition to a decision making map for each respondent. Early findings highlighted a 5% mean difference between image banks, confirming that benchmarking must be related to a specific test. 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. Reporting radiographers and radiologists all scored 95% or above accuracy in both tests. The general radiographer population scored between 60 and 95%.

Conclusions: The evidence from this research indicates that the *Radbench* tool is capable of providing benchmark measures of image interpretation accuracy, with the potential for comparison across populations.

Introduction

Image interpretation by radiographers now has three potential levels; firstly radiographer abnormality detection systems (RADS) such as 'red dot'; Secondly preliminary clinical evaluation (PCE) whereby the 'red dot' or similar annotation is replaced by a written comment which may, or may not, be in a structured format[1–3] and finally definitive clinical reporting performed by radiographers who have undergone postgraduate education.[4,5]

The fact that radiographers have the ability to provide an accurate report on diagnostic images is well established.[4,5] The provision of a preliminary accurate opinion for all diagnostic images to the referring clinician ahead of the definitive report, offers the potential for rapid assessment of treatment requirements and optimisation of emergency department time.[1–3] 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

* Corresponding author. Tel.: +44 1978 313244.
E-mail address: p.reeves@shu.ac.uk (P. Reeves).

http://dx.doi.org/10.1016/j.radi.2015.12.010

interpretation studies to date have broadly followed a similar quantitative methodology, either focussing 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.

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.

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

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

Table 1
Comparative test scores between groups.

	Mean Accuracy		Mean sensitivity		Mean specificity		ROC	
	Radiologists/Reporting Radiographers/Academics	General radiographers	Radiologists/Reporting Radiographers/Academics	General radiographers	Radiologists/Reporting Radiographers/Academic	General radiographers	Radiologists/Reporting Radiographers/Academics	General radiographers
Test 1	100%	84%	100%	92%	100%	77%	1.000	0.930
Test 2	98%	79%	95%	86%	100%	73%	0.985	0.920
Two tests mean	99%	82%	98%	89%	100%	75%	0.995	0.925