Accuracy of radiographer preliminary clinical evaluation of skeletal trauma radiographs, in clinical practice at a district general hospital

Verrier, W., Pittock, L. J., Bodoceanu, M. and Piper, K.

Accepted version of:

W. Verrier, L.J. Pittock, M. Bodoceanu, K. Piper,

Accuracy of radiographer preliminary clinical evaluation of skeletal trauma radiographs, in clinical practice at a district general hospital,

Radiography,

2022

https://doi.org/10.1016/j.radi.2021.12.010.

Introduction

The radiographic image reflects a pivotal diagnostic tool for the emergency patient and referrers' ability to accurately interpret pathology is paramount in determining appropriate clinical management[1]. The potential negative outcomes of misinterpretation of trauma images are multiple; increased patient morbidity and mortality, additional costs and detrimental impact upon Department of Health service quality indicators[2, 3]. Such errors frequently occur[4] and whilst the immediate formal reporting of images is seen to be the most effective means of minimising interpretive error[5], this is not always feasible secondary to staff shortages[6]. Initial evaluation of images by the radiographer has become a proven method in reducing emergency department (ED) error[7] and providing interim guidance until a report is produced[6].

Utilisation of radiographers has been commonplace in the United Kingdom (UK) since the 1980s[8] and the 'Red dot' system (RD) is the most commonly employed model[4]. This system, originated from the practice of the radiographer affixing a red dot sticker to an abnormal trauma film, has some well-documented limitations associated with ambiguity in the absence of RD, inability to specify the nature of an abnormality and incompatibility with quantitative audit systems[9,10]. Subsequently, the stakeholder professional bodies have encouraged departments to develop from RD into preliminary clinical evaluation (PCE)[11,12,13]. PCE enables the radiographer to unambiguously communicate both the nature and location of a traumatic finding in a short written comment visible to the referrer[13]. Whilst, a number of studies have been conducted which evaluated radiographers' ability to provide PCE[14], many of these used a test bank methodology where a purposive sample of images were selected[15]. Such studies are performed under test conditions which are not directly comparable to clinical practice[16] and may not reflect the true disease prevalence encountered in a typical ED caseload[8,17]. Indeed, it is suggested that such studies provide 'little or no direct evidence' of the impact of radiographer image interpretation in practice[15] unless a purposeful image bank that reflects local clinical workload is adopted[18]. Two studies have been performed which specifically evaluated radiographers' ability to provide PCE in clinical practice[19,20]. McConnell *et al.*[19] demonstrated PCE accuracy, sensitivity and specificity as 89%, 95% and 95% respectively, with results for the same measures in Brown's study[19] achieving 92%, 71% and 98%. However, both of these studies were based in Australia where radiographers participating in PCE, locally referred to as Preliminary Image Evaluation (PIE) are required to complete an in-house programme of study as a prerequisite to providing comments. Additionally, the study by McConnell *et al.* only included appendicular images[19]. These factors limit the application of findings to a UK setting where both appendicular and axial performance are under evaluation.

The aim of this study was to determine the appropriateness of transition from RD system to PCE for skeletal trauma radiographs at a local district general hospital (DGH). The objectives were 1) to calculate accuracy, sensitivity and specificity of radiographer RD and PCE; and, 2) to examine any statistically significant differences in radiographers' ability to provide PCE on appendicular and axial trauma radiographs.

Methods

Ethics

Appropriate approval for the study was obtained from the University Faculty of Health and Wellbeing Ethics Panel (Ref: 19-001) and Trust Research and Development Department. National Health Service (NHS) Research Ethics committee approval was deemed unnecessary by local guidance and the Health Research Authority decision tool[21] secondary to being categorised as a service evaluation.

Study setting and participants

The study was conducted in a DGH. All qualified diagnostic radiographers regularly performing radiographic trauma imaging at the main hospital site (except reporting practitioners) were invited to participate (n=40). Relevant information was provided to all participants and it was explicitly outlined that PCEs would be audited against a definitive radiological report, in line with Medical Research Council recommendations[22]. Formal consent was obtained which included notification of right to withdraw from the study. The only specific training for participants was a short presentation with commentary which outlined the practicalities and PCE structure in line with 'What, where, how?' methodology[23].

Sample size

Sample size was established to determine whether there was any significant difference in radiographers' ability to provide PCE between appendicular and axial trauma radiographs. No previous studies were available which investigated this comparison. The most relevant UK dataset was derived from a meta-analysis by Brealey *et al.*[24] on radiographer reporting in clinical practice. Upon re-analysis of the data, a postulated accuracy of 97% for both appendicular and axial cases was estimated and a two-tailed tail comparative test methodology was employed, as accuracy of PCE in either area could have been superior or inferior to the other[25]. Based upon these postulated accuracies it was appropriate to use an equivalence trial which demonstrates similarity or non-inferiority of outcomes between two groups[26].An appropriate sample size was calculated to allow any significant statistical difference in radiographer PCE accuracy between axial and appendicular cases to be established. Allowing an inferiority of no more than 5% to be acceptable, the sample size for each arm of the test was 240, resulting in a total sample of 480. In the course of performing the research, an increase in appendicular sample size was necessary secondary to the reduced incidence of axial cases and the need to collect both appendicular and axial data simultaneously thus avoiding seasonal variations in disease prevalence[15]. Data collection was monitored and terminated when 240 axial cases had been completed. This resulted in a final sample size of 762 (240 axial and 522 appendicular).

Research procedure

Participants were asked to provide RD and PCE on all available consecutive appendicular and axial skeletal trauma examinations, excluding images obtained following application of plaster cast. Upon completion of an examination, participants were asked to review the images and determine if an acute abnormality was detected. If so, the radiographer was asked to mark the image electronically with the wording 'red dot'. Images were then reviewed using *Insignia InSight* Picture Archiving and Communication System application on a portrait 20" Barco Digital Imaging and Communications in Medicine[28] compliant monitor. The required content and format of PCE (Table 1) were outlined to participants in a narrated PowerPoint presentation and participants were asked to describe all acute abnormalities if more than one was present. PCE were recorded using the *Insight User Action* function on the secure *Insignia InSight* server; only visible to the researcher. This methodology had potential for the introduction of film selection bias whereby a participant had the opportunity to only provide PCE on images upon which they felt confident to interpret, thus giving a potentially inflated estimate of performance[27].

Table 1 PCE format

PCE format	Only acute abnormalities such as: recent fracture, subluxation or dislocation, raised elbow fat pads or knee lipohaemarthrosis or foreign body should be described.
	The PCE should explicitly specify if the exam is normal or abnormal and if abnormal, the nature and site of the abnormality outlined.
	The individual participant code should be added to the PCE.

Individual participants were allotted a code by a research assistant, in order to maintain anonymity and minimise arbiter comparator bias[27]. This allowed individual performance measures to be calculated, as evidence of personal achievement[9]. Individual performance measures were not included in the study, as performance between participants cannot be compared due to observer comparator bias; where individuals report a different set of cases and thus differences in performance may be due to case mix rather than ability[29].

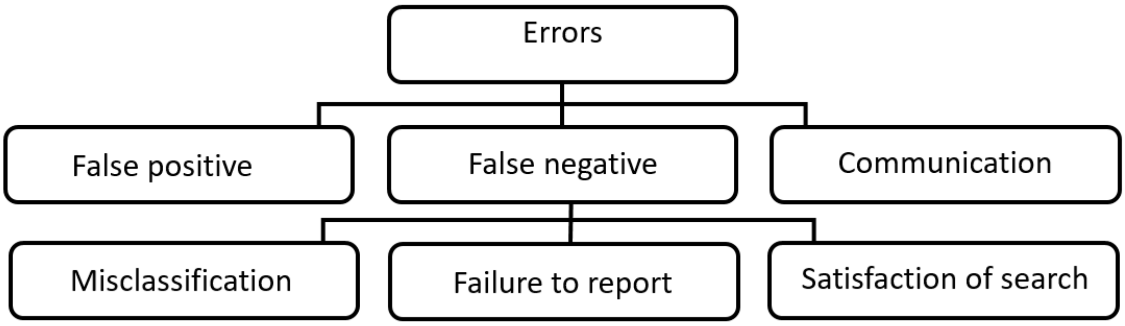
Accuracy of radiographer PCE and RD were determined by comparison against the formal report, reflecting the reference standard which is necessary to calculate diagnostic performance measures[15]. Although a double/triple blind report is thought to be the most robust reference standard, the single blind report used in the study meets the minimum standards for validity[30]. As not all cases were compared against the reports from the same individual, some difference in reference standard quality may have occurred, a phenomenon known as reference standard bias which can impact the validity of findings[30]. Upon comparison with the verified report; RD and PCEs were classified using the true positive (TP), true negative (TN), false positive (FP), false negative (FN) system[31] and the study marking assumptions (Table 2).

Table 2 Marking assumptions

Assumptions	Only text relating the acute abnormalities outlined in Figure 1 was analysed. Documentation of pathology outside this scope and chronic abnormalities were not analysed.
	All terms of equivocation were disregarded for the purposes of analysis and scoring, and if these terms proceeded a diagnosis, the PCE should be interpreted as positive.
	As PCE primarily reflects an abnormality flagging system, 0.5 TP was awarded for correctly identifying an abnormal exam, the remaining 0.5 TP for abnormality type and site. If multiple abnormalities were present, all abnormalities had to be described to award a full 1 TP. All fractional marks had to total 1.

Fractional scores were utilised where some element of the PCE was incorrect, as applied in previous studies[20,31]. Scores were summed and accuracy, sensitivity and specificity determined. Performance measures, overall agreement and confidence intervals were calculated for RD and PCE and Two Sample Z-tests to Compare Two Independent Proportions[32] used to determine any statistical difference between RD versus PCE and appendicular versus axial performance measures. Error analysis was undertaken and classified as FP, FN and communication errors. FN errors were then sub-divided into misclassification, failure to report and satisfaction of search (SOS) errors, based on the work of Renfrew *et al.*[33] (Figure 1). Textual analysis was also undertaken on the wording of erroneous cases in order to determine if any further themes arose.

Figure 1 Error classification system



Comparison and scoring were performed by the researcher and as such, any cases reported by the researcher were excluded in order to avoid arbiter review bias[29]. In order to evaluate inter-observer variability, any cases where there was disagreement between the PCE and report, were reviewed by a moderator to ensure consistency of marking. A proportion of accordant cases were also reviewed and Two Sample Z-tests utilised to determine inter-observer agreement.

Results

Participation and prevalence

There were 23 consented participants. One of these withdrew and one participant failed to contribute any data. A total number of 813 exams were performed, 51 exams were rejected as they did not meet the inclusion criteria. The disease prevalence was 27% (34% and 12% respectively for the appendicular and axial tails of the study).

Performance

The overall RD and PCE accuracy, sensitivity and specificity for the study were 90%, 72% and 97% (RD), and 92%, 80% and 97% (PCE) respectively (Table 3).

Table 3 Raw data, performance measures and confidence intervals for radiographer RD and PCE

Element of study	Red dot or PCE	Raw data				Accuracy	Sensitivity	Specificity
		TP	TN	FP	FN	(Confidence Intervals)		
Whole study	Red dot	156.50	530.50	14.00	61.00	0.9 (0.88-0.92)	0.72 (0.65-0.78)	0.97 (0.96-0.99)
	PCE	178.00	522.50	17.75	43.75	0.92 (0.90- 0.94)	0.80 (0.74-0.85)	0.97 (0.95-0.98)
Appendicular	Red dot	145.50	329.50	12.00	35.00	0.91 (0.88-0.93)	0.81 (0.74-0.86)	0.96 (0.94-0.98)
	PCE	156.25	325.50	11.00	29.25	0.92 (0.90-0.94)	0.84 (0.78-0.89)	0.97 (0.94-0.98)
Axial	Red dot	11.00	200.50	2.00	26.50	0.88 (0.83-0.92)	0.29 (0.16-0.46)	0.99 (0.96-1.00)
	PCE	21.75	197.00	6.75	14.50	0.91 (0.87-0.94)	0.60 (0.43-0.75)	0.97 (0.93-0.99)

Two Sample Z-tests demonstrated a significant statistical difference for sensitivity between RD versus PCE (Table 4), and appendicular vs axial (Table 5)[34].

Table 4 Comparison between overall RD and PCE performance with a Two Sample Z-Test*

Performance measure	Red dot result	95% CI	PCE result	95% CI	*Z*-score	*p*-value
Accuracy	0.90	0.88-0.92	0.92	0.90-0.94	-1.26	0.21
Sensitivity	0.72	0.65-0.78	0.80	0.74-0.85	-2.11	0.03
Specificity	0.97	0.96-0.99	0.97	0.96-0.99	0.57	0.57

***assuming a two-tail test where a *p*-value < .05 reflects a true statistical difference**

Table 5 Comparison between Appendicular and Axial performance with a Two Sample Z-Test*

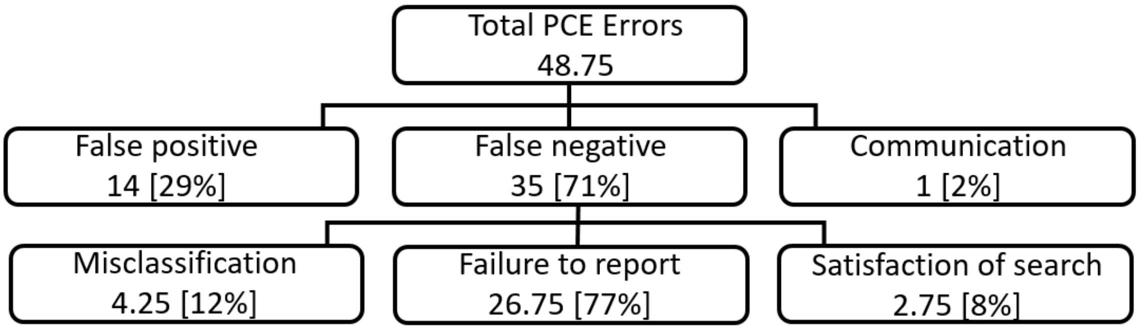
Red dot or PCE	Performance measure	Appendicular result	Axial result	*Z*-score	*p*-value
Red dot	Accuracy	0.91	0.88	1.15	0.25
	Sensitivity	0.81	0.29	6.43	<0. 02
	Specificity	0.96	0.99	1.80	0.07
PCE	Accuracy	0.92	0.91	0.51	0.61
	Sensitivity	0.84	0.6	3.24	0.00
	Specificity	0.96	0.97	0.11	0.92

***assuming a two-tail test where a *p*-value < .05 reflects a true statistical difference**

Error analysis

Of the 762 studies reviewed, 694 showed correct use of RD and 682 had completely correct PCE. The overall error rate for RD was 10% and PCE was 8%. 48.75 PCE errors were identified and analysed using the classification algorithm. Error analysis showed that the majority of PCE errors were classified FN (Figure 2).

Figure 2 PCE error classification



Textual analysis of the PCEs was performed and it was identified that 46% of PCE FN errors occurred in cases where there was equivocation in the reference standard report. These included the following terms: possible, suspicious, suggestive, may represent, difficult to/cannot be exclude(d), spurious, indeterminate.

Inter-observer variability

80 cases with discrepancy between PCE and reference standard and a random 48 accordant cases were reviewed by a moderator in order to evaluate inter-observer variability in marking criteria application.

Of 128 reviewed cases, the researcher deemed 48 PCE in complete accordance with the reference standard and the moderator deemed 40 in complete agreement (Table 6).

Table 6 Scoring for Researcher and Moderator for 128 sample cases

Role	TP	TN	FP	FN	Complete agreement	Accuracy	Sensitivity	Specificity
Researcher	29.75	43.00	13.75	41.50	48.0	0.57	0.42	0.76
Moderator	28.00	38.80	15.00	46.25	40.0	0.52	0.38	0.72

Two Sample Z-tests demonstrated no significant statistical difference between observers (Table 7) [34].

Table 7 Comparison between Researcher and Moderator scoring on 128 sample cases with a Two Sample Z-Test*

Performance measure	Researcher sample	95% CI	Moderator sample	95% CI	*Z*-score	*p*-value
Accuracy	0.57	0.48-0.66	0.52	0.43-0.61	0.75	0.45
Sensitivity	0.42	0.30-0.54	0.38	0.26-0.50	0.44	0.66
Specificity	0.75	0.62-0.85	0.72	0.58-0.83	-0.28	0.78

***assuming a two-tail test where a p-value < .05 reflects a true statistical difference**

Discussion

Overall performance

Compared to other RD studies conducted in clinical practice without specific training, the 90% accuracy level demonstrated in this study was similar to Hargreaves and Mackay[35] and Renwick, Butt and Steele[36] at 90% but slightly inferior to Brown and Leschke[37] and Hlongwane and Pitcher[38] who both reported radiographer accuracy at 94%. It should be noted that Brown and Leschke's study only included the appendicular skeleton which may explain their slightly improved accuracy levels[37]. Hlongwane and Pitcher's[38] study was based on data from an afterhours South African trauma hospital, making comparisons to a UK DGH general hospital difficult.

Compared with other studies, PCE performance measures were similar to the pre-education arm of Loughran's reporting training study[39], 81% and 94% (sensitivity/specificity), and Brown *et al.*'s research[20]; 92%,71% and 98% (accuracy/sensitivity/specificity). Interestingly, radiographers in the latter had a pre-requisite image interpretation education prior to study participation or provision of PIE in practice. Such, pre-training was not provided in Younger and Smith' study[40] which demonstrated performance measures of 93%, 95% and 92% (accuracy/sensitivity/specificity). It can be seen that radiographer sensitivity in the current study is much lower at 80% but reasons for this discrepancy are unclear; however, inclusion of chest and abdominal cases make the studies difficult to compare and the study did not include data in sufficient detail to enable re-calculation of the statistics[30]. Some PCE studies also included an option of ‘Not sure’ for cases where the radiographer was unsure of findings[20]; in the context of this study where 46% of FN errors occurred where there was equivocation in the reference, the inclusion of a similar option would have the potential to improve overall accuracy of the results.

Appendicular vs Axial

The results of this study showed radiographer sensitivity on appendicular examinations compared to axial was statistically superior for both RD (81% versus 29%) and PCE (84% versus 60%). There is minimal literature comparing radiographer’s ability to interpret appendicular versus axial cases. Some work has been produced based upon test bank methodology. Whilst Hardy and Culpan's study[9] comparing RD and PCE found no significant distinction in results of appendicular and axial cases, Renwick, Butt and Steele[36] found that the radiographers demonstrated lower RD sensitivity and specificity for the axial skeleton compared with appendicular (89% and 91% vs 90% and 94%).

Studies conducted within clinical practice mainly concur with Renwick, Butt and Steele[36] and the findings of this study, in the trend of lower sensitivity for axial skeleton. Hlongwane and Pitcher[38] reported lower axial sensitivity (57.1% vs 76) whilst a smaller disparity was seen in Brown *et al.*’s work[20] with axial sensitivity only 3% lower than appendicular. However, these radiographers had completed in-house training which included specific axial modules.

Interestingly, the low axial sensitivity found in this study correlates with some authors’ work on radiographer reporting. Both Brealey *et al.*[41] and Piper *et al.*[42] found that trained radiographers were able to report abnormal radiographs of the appendicular skeleton significantly more accurately than for the axial skeleton.

It is difficult to determine why this theme appears consistently across the majority of the available literature. Lancaster and Hardy's study[43] on barriers to participation in radiographer comment schemes suggested that whilst 77% of respondents were confident to comment on radiographs of the appendicular skeleton, only 53% were confident to comment on axial. The work of Neep[44] *et al.* mirrors these findings with radiographers reporting much lower self-perceived accuracy in both detecting and describing traumatic axial abnormalities compared to appendicular. The authors further speculated that the lower levels of confidence and perceived accuracy in describing radiographs of the axial skeleton may be attributable to the more complex anatomy and potentially more complex pathology encountered in the axial skeleton[44].

The recognised significant potential impact of disease prevalence on diagnostic performance outcomes[29] should also not be overlooked. The study showed a marked difference in abnormality prevalence between appendicular and axial cases (34% vs 12%). Even minimal disparities in prevalence can lead to different values of accuracy for a test[45] and some research has shown that sensitivity may be lower with a lower disease prevalence[46].

Benchmark diagnostic performance

In order to determine appropriateness of local PCE implementation, it is necessary to consider what constitutes an appropriate performance benchmark. Whilst 95% accuracy is perceived as the required performance standard for qualified reporting radiographers[47] a benchmark performance level for RD and PCE is less clear cut. The SCoR standards of professional practice[13] suggest that radiographers should be demonstrably competent, the Health and Care Professions Council Standards of Proficiency[48] state that they must be able to distinguish abnormal appearances evident on images. This is quantified by Brealey[47] who indicates that any professional involved in the clinical reporting of emergency skeletal radiographs should demonstrate a minimum of 80% accuracy, although 90% is optimal and 95% ideal. Wright and Reeves[49] concur suggesting that radiographers should be expected to achieve 90% accuracy in any form of decision making. Based on these benchmark figures and allowing for bias challenges and performing in the contemporaneous rather than exam based environment, it can be seen that accuracy levels for both RD and PCE in this study are within the acceptable threshold.

Limitations

Whilst the inter-observer variability component of the data analysis suggests that the reproducibility element of this research appears statistically robust, there are a few areas where biases may have occurred.

The risk of bias occurs when the reference standard report is from a single reporting practitioner; what Brealey and Scally refer to as reference standard bias[29]. Such a reference standard does not take into account inherent reporting practitioner error which has been estimated at 4%[49] nor does it allow for differential verification bias which derives from different reporting practitioners providing what is effectively different reference standards[50]. This variable quality is reflected in the prevalence of terms of reference standard equivocation noted in a high proportion of the study's FN errors.

The study recruitment method, referred to by McConnell and Baird[51] as a self-selecting sample population is also a potential source of bias. This, observer-cohort bias[29] may limit the generalisability of the research findings and overestimate radiographer performance secondary to lack of participation by radiographers with less confidence or interest in image interpretation. Furthermore, the proportional contribution of each participant has not been factored into the data analysis which may skew the study's internal validity, and question whether the overall results sufficiently represent the population's performance[52].

Finally, the number of cases which were in the radiographer scope, but where no RD/PCE was provided, has not been factored into the analysis. This would give a more accurate representation of the overall 'service accuracy' during the research period[20].

Conclusion and recommendations

Overall, the cohort of radiographers achieved acceptable accuracy in RD and PCE when compared to the available literature and recommended benchmarks. Diagnostic performance measures are uniformly better for PCE than RD suggesting that implementation of PCE has the potential to assist referring clinicians in the interpretation of skeletal trauma radiographs by overcoming the ambiguities associated with RD.

The study found that study participants were less able to correctly interpret abnormal axial images than appendicular. On the backdrop of reported reduced radiographer confidence in axial image interpretation and low axial sensitivity, improvement in undergraduate provision and local focused continuing professional development in this area is recommended.

Further research with a more robust reference standard, use of a wider range of statistics and evaluation of overall service accuracy should be considered.

References

1. Lee, G.A., Chou, K., Jennings, N., O'Reilly, G., McKeown, E., Bystrzycki, A. and Varma D. (2014) 'The accuracy of adult limb radiograph interpretation by emergency nurse practitioners: A prospective comparative study'. *International Journal of Nursing Studies*. Apr;51(4) pp.549-54. doi: 10.1016/j.ijnurstu.2013.08.001.
2. Department of Health (2010) A *& E Clinical Quality Indicators Data definitions*. London: Department of Health. doi: 10.1258/acb.2010.010234.

3. Snaith, B., Hardy, M. and Lewis, E. F. (2014) ‘Reducing image interpretation errors - Do communication strategies undermine this?’, *Radiography,* 20(3), pp. 230–234. doi: 10.1016/j.radi.2014.03.006.
4. Kranz, R. and Cosson, P. (2015) ‘Anatomical and/or pathological predictors for the “incorrect” classification of red dot markers on wrist radiographs taken following trauma’, *British Journal of Radiology,* 88 (1046). pp.1-9. doi: 10.1259/bjr.20140503.
5. Hardy, M., Snaith, B., & Scally, A. (2013). The impact of immediate reporting on interpretive discrepancies and patient referral pathways within the emergency department: a randomised controlled trial. *The British journal of radiology*, *86*(1021), 20120112-20120112.
6. Bain, P., Wareing, A. and Henderson, I. (2017) ‘A review of peer-assisted learning to deliver interprofessional supplementary image interpretation skills’, *Radiography,* 23, s65-s69, doi: 10.1016/j.radi.2017.05.002.
7. Coelho, J. M. and Rodrigues, P. P. (2011) ‘THE RED DOT SYSTEM - Emergency Diagnosis Impact and Digital Radiology Implementation: A Review’, *HEALTHINF 2011 - International Conference on Health Informatics*. BIOSTEC, Rome International Conference on Health Informatics, 26-29 January pp. 508–511. doi: 10.5220/0003135305080511.
8. Snaith, B. and Hardy, M. (2008) ‘Radiographer abnormality detection schemes in the trauma environment-An assessment of current practice’, *Radiography,* 14(4), pp. 277–281. doi: 10.1016/j.radi.2007.09.001.
9. Hardy, M. and Culpan, G. (2007) ‘Accident and emergency radiography: A comparison of radiographer commenting and “red dotting”’, *Radiography,* 13, pp. 65-71. doi: 10.1016/j.radi.2005.09.009.
10. Cosson, P. and Dash, R. (2015) ‘A taxonomy of anatomical and pathological entities to support commenting on radiographs (preliminary clinical evaluation)’, *Radiography,* 21, pp. 47-53. doi: 10.1016/j.radi.2014.06.013.
11. The College of Radiographers (2006) *Medical Image Interpretation and Clinical Reporting by Non-Radiologists: The Role of the Radiographer.* London: The College Of Radiographers.

12. The Royal College of Radiologists and The Society and College of Radiographers (2012) *Team working in clinical imaging*. London: The Royal College of Radiologists and The Society and College of Radiographers.
13. The Society and College of Radiographers (2013) *Preliminary Clinical Evaluation and Clinical Reporting by Radiographers : Policy and Practice Guidance.* London: The Society and College of Radiographers doi: 10.1021/jo501817m.
14. Lockwood, P. and Pittock, L. (2019) 'Radiography Multi-professional image interpretation : Performance in preliminary clinical evaluation of appendicular radiographs', *Radiography*. https://doi.org/10.1016/j.radi.2019.04.013
15. Brealey, S. and Scally, A. J. (2008) 'Methodological approaches to evaluating the practice of radiographers' interpretation of images: A review', *Radiography,* 14, pp.e46-e54. doi: 10.1016/j.radi.2008.01.001.
16. Panacek, E. A. and Thompson, C. B. (2007) 'Sampling methods: Selecting your subjects', *Air Medical Journal,* March-April, pp. 75-78. doi: 10.1016/j.amj.2007.01.001.
17. Neep, M. J, Steffens, T., Riley, V., Eastgate, P. and McPhail, S. M. (2017) 'Development of a valid and reliable test to assess trauma radiograph interpretation performance', *Radiography,* 23(2), pp. 153–158. doi: 10.1016/j.radi.2017.01.004.
18. Hardy, M., Flintham, K., Snaith, B., & Lewis, E. F. (2016). The impact of image test bank construction on radiographic interpretation outcomes: A comparison study. Radiography, 22(2), 166-170.
19. McConnell, J., Devaney, C. and Gordon, M. (2013) 'Queensland radiographer clinical descriptions of adult appendicular musculo-skeletal trauma following a condensed education programme', *Radiography,* 19, pp.48-55. doi: 10.1016/j.radi.2012.09.002.
20. Brown, C., Neep, M. J., Pozzias, E., & McPhail, S. M. (2019). 'Reducing risk in the emergency department: a 12-month prospective longitudinal study of radiographer preliminary image evaluations' *Journal of Medical Radiation Sciences,* 66 (3) pp. 154-162 https://doi.org/10.1002/jmrs.341

21. Health Research Authority (2018) *HRA Decision Tool, Medical Research Council*. Medical Research Council. Available at: http://www.hra-decisiontools.org.uk/research/.(Accessed 25 April 2019).
22. Medical Research Council (2018) 'General Data Protection Regulation (GDPR): Consent in Research and Confidentiality', *MRC Regulatory Support Centre*, 2018(March), pp. 1–11. Available at: https://mrc.ukri.org/documents/pdf/gdpr-guidance-note-3-consent-in-research-and-confidentiality/. (Accessed 5 May 2019)
23. Harcus, J., and Wright, C. (2014). *What, where, and how: a proposal for structuring preliminary clinical evaluations* [Poster] Exhibited at UKRC, Manchester, 9-11th June 2014.
24. Brealey, S., Scally, A., Hahn, S., Thomas, N., Godfrey, C. and Coomarasamy, A. (2005) 'Accuracy of radiographer plain radiograph reporting in clinical practice: A meta-analysis', *Clinical Radiology,* 60(2), pp. 232–241. doi: 10.1016/j.crad.2004.07.012.
25. Walker, E. and Nowacki, A. S. (2011) 'Understanding equivalence and noninferiority testing', *Journal of General Internal Medicine*, 26(2), pp. 192–196. doi: 10.1007/s11606-010-1513-8.
26. Ahn, S., Park, S. H. and Lee, K. H. (2013) 'How to Demonstrate Similarity by Using Non-inferiority and Equivalence Statistical Testing in Radiology Research', *Radiology,* 267(2), pp. 328–338.
27. Scally, A. and Brealey, S. (2002) 'Confidence Intervals and Sample Size Calculations for Studies of Film-reading Performance', *Clinical Radiology*, 58, pp. 238–246.
28. Rubin, D. L. (2011). Informatics in radiology: measuring and improving quality in radiology: meeting the challenge with informatics. *Radiographics*, 31(6), 1511-1527.
29. Brealey, S. and Scally, A. J. (2001) 'Bias in plain film reading performance studies', *British Journal of Radiology*, 74, pp. 307-316 doi: 10.1259/bjr.74.880.740307.
30. Brealey, S., Scally, A. J. and Thomas, N. B. (2002) 'Methodological standards in radiographer plain film reading performance studies', *British Journal of Radiology*, 75(890) pp. 107–113. doi: 10.1259/bjr.75.890.750107.

31. Piper, K. J., Paterson, A. M. and Godfrey, R. C. (2005) ‘Accuracy of radiographers’ reports in the interpretation of radiographic examinations of the skeletal system: A review of 6796 cases’, *Radiography*, 11(1), pp. 27–34. doi: 10.1016/j.radi.2004.05.004.

32. Zou, K.H., Fielding, J.R., Silverman, S.G. and Tempany C.M. (2003) ‘Hypothesis testing I: proportions’ *Radiology*. 226(3):99. pp.609-613. doi: 10.1148/radiol.2263011500

33. Renfrew, D. L., Franken Jr, E. A., Berbaum, K. S., Weigelt, F. H., & Abu-Yousef, M. M. (1992). ‘Error in radiology: classification and lessons in 182 cases presented at a problem case conference’. *Radiology*, *183*(1), pp. 145-150.

34. Lowry, R. (2021) *VassarStats: Website for Statistical Computation*. Available at http://vassarstats.net/index.html. (Accessed 10 Jan 2021).

35. Hargreaves, J., & Mackay, S. (2003). ‘The accuracy of the red dot system: can it improve with training?’ Radiography, 9(4), 283-289.

36. Renwick, I. G. H., Butt, W. P. and Steele, B. (1991) ‘How well can radiographers triage x-ray films in accident and emergency departments?’ *British Medical Journal,* 302 pp. 568-569

37. Brown, N., and Leschke, P. (2012). ‘Evaluating the true clinical utility of the red dot system in radiograph interpretation’ Jou*rnal of Medical Imaging and Radiation Oncology*, 56(5), pp. 510-513.

38. Hlongwane, S. T. and Pitcher, R. D. (2013) ‘Accuracy of after-hour “red dot” trauma radiograph triage by radiographers in a South African regional hospital’, *South African Medical Journal*, 103(9), pp. 638–640. doi: 10.7196/SAMJ.6267.

39. Loughran, C.F. (1994) ‘Reporting of fracture radiographs by radiographers: the impact of a training programme’. *The British Journal of Radiology*; 67: pp. 945-950.

40. Younger, C., and Smith, T. (2002). ‘Accident and emergency radiological interpretation using the radiographer opinion form (ROF)’. *The Radiographer: The Official Journal of the Australian Institute of Radiography*, 49(1), pp. 27-31.

41. Brealey, S., Scally, A., Hahn, S., Thomas, N., Godfrey, C., & Coomarasamy, A. (2005). Accuracy of radiographer plain radiograph reporting in clinical practice: a meta-analysis. *Clinical radiology*, *60*(2), 232-241.

42. Piper, K. J., Paterson, A. M., & Godfrey, R. C. (2005). Accuracy of radiographers' reports in the interpretation of radiographic examinations of the skeletal system: a review of 6796 cases. *Radiography, 11*(1), 27-34.

43. Lancaster, A. and Hardy, M. (2012) 'An investigation into the opportunities and barriers to participation in a radiographer comment scheme, in a multi-centre NHS trust', *Radiography*, 18, pp. 105-108. doi: 10.1016/j.radi.2011.08.003.

44. Neep, M. J., Steffens, T., Owen, R. and McPhail, S. *M.* (2014) 'A survey of radiographers' confidence and self-perceived accuracy in frontline image interpretation and their continuing educational preferences', *Journal of Medical Radiation Sciences*, 61 pp.69-77. doi: 10.1002/jmrs.48.

45. Obuchowski, N. A., Blackmore, C. C., Karlik, S., & Reinhold, C. (2005). Fundamentals of clinical research for radiologists. *American Journal of Roentgenology*, 184(2), pp.364-372.

46. Leeflang, M. M., Rutjes, A. W., Reitsma, J. B., Hooft, L., & Bossuyt, P. M. (2013). Variation of a test's sensitivity and specificity with disease prevalence. *Canadian Medical Association Journal.* 185(11), E537–E544. https://doi.org/10.1503/cmaj.121286

47. Brealey, S. (2001). 'Measuring the effects of image interpretation: an evaluative framework'. *Clinical Radiology*, 56(5), pp. 341-347.

48. Health and Care Professions Council. (2013). *Standards of Proficiency – Radiographers*. London: Health and Care Professions Council.

49. Wright, C., and Reeves, P. (2017). 'Image interpretation performance: A longitudinal study from novice to professional' *Radiography*, 23(1), e1-e7. doi:10.1016/j.radi.2016.08.006

50. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., and Bossuyt, P. M. M. (2011). 'QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies'. *Annals of Internal Medicine*, 155(8), pp. 529-536. doi:10.7326/0003-4819-155-8-201110180-00009

51. McConnell, J. R., and Baird, M. A. (2017). 'Could musculo-skeletal radiograph interpretation by radiographers be a source of support to Australian medical interns: a quantitative evaluation.' *Radiography*, 23(4), pp. 321-329.

52. Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D., and Newman, T. B. (2013). *Designing clinical research*. Philadelphia : Wolters Kluwer Health/Lippincott Williams & Wilkins.