Image Evaluation Performance of Diagnostic Radiographers: Benchmarking New Graduates

Tatsuhito Akimoto

A thesis submitted in partial fulfilment of the requirements of
Sheffield Hallam University
for the degree of Doctor of Philosophy

September 2019

Declaration

I hereby declare that:

1. I have not been enrolled for another award of University, or other academic or professional organisation, whilst undertaking my research degree.
2. None of the material contained in the thesis has been used in any other submission for an academic award.
3. I am aware of and understand the University's policy on plagiarism and certify that this is my own work. The use of all published or other sources of material consulted have been properly and fully acknowledged.
4. The work undertaken towards the thesis has been conducted in accordance with the SHU Principles of Integrity in Research and the SHU Research Ethics Policy.
5. The word count of the thesis is 36,997.

Name	Tatsuhito Akimoto
Date	27 September 2019
Award	PhD
Faculty	Faculty of Health and Wellbeing
Director of Studies	Dr Pauline Reeves

Acknowledgement

I would like to express my sincere appreciation to Dr Pauline Reeves and Dr Chris Wright for the meticulous supervision and support. I would also like to thank the advisor of my PhD study, James Harcus, for his immense knowledge and perceptive comments.

Besides my supervisors and advisor, my special gratitude goes to statistics specialists, Dr Karen Kilner, Dr Sammy Rashid and Sean Demack, as well as library information specialists. I am deeply indebted for their invaluable assistance. I am also grateful to the staff and my fellow doctoral students at the Centre for the Health and Social care of Sheffield Hallam University.

Finally, I would like to thank my family: my parents and sister for spiritually supporting me.

Abstract

Aim: Preliminary clinical evaluation (PCE) is a new clinical role of diagnostic radiographers in the United Kingdom (UK). Radiographers participating in PCE are now expected, not only to view radiographs and make reliable clinical decisions, but also to express the clinical findings in unambiguous written forms. The Society and College of Radiographers (SCoR) (2013) expects that newly qualified radiographers have the underpinning education and training to take part in PCE. However, the feasibility of PCE by radiographers, especially newly qualified radiographers, has not been empirically challenged. This research therefore set out to determine whether final year diagnostic radiography students at the point of graduation and qualification were capable of providing reliable PCE.

Method: An X-ray image evaluation test was conducted to assess PCE performance of the final year undergraduate diagnostic radiography students. An image bank, consisting of 30 appendicular radiographs, was developed for the test. A total of 87 students from nine universities in England and Wales took the test. The students provided their clinical decisions (normal or abnormal) and comments (PCE). Accuracy, sensitivity and specificity were calculated based on their decisions. A PCE taxonomy was developed to classify comments and identify types and frequencies of PCE errors. The comments were also systematically evaluated with a scoring system, which was developed to assess three essential components of skeletal trauma reports: type, location and displacement/dislocation of fractures. Comments were further analysed by the results of the scoring.

Results: The results demonstrated that mean sensitivity and specificity of the student group were 79.62 % (95% CI: 77 – 82%) and 67.13% (95% CI: 64 – 71%) respectively. Accuracy was 73.37% (95% CI: 72 – 75%). PCE error classification found that the students made more false positives than false negatives. A further analysis of the comments using the scoring system indicated that, although many commented on types and locations of abnormalities, very few described displacement/dislocations of fractures.

Conclusion: Low specificity with higher rate of false positive decisions suggests that education providers should collaborate in partnership with clinical placement sites to devote greater focus on evaluation of normal radiographs. A certain proportion of newly qualified radiographers may benefit from post qualification learning to provide more reliable PCE. Preceptorship, which is a transitional phase for newly qualified radiographers to become independent practitioners, could incorporate PCE training as one of its key educational components. The error classification system and scoring model are ideally suited for regular audits at any stage of image evaluation learning and practicing.

Keywords: audit, benchmarking, newly qualified radiographers, preceptorship, preliminary clinical evaluation (PCE), X-ray image evaluation

Table of contents

List of tables

List of figures

Glossary of terms

Term	Definition
Accuracy	A measure that incorporates sensitivity and specificity into a single index.
Anteroposterior (AP) radiographs	Radiographs taken from front-to-back.
Appendicular skeleton	Upper and lower limbs including shoulder and pelvic girdles.
Axial skeleton	Bones including the skull, vertebral column and thoracic cage.
False negative	Decision outcome that indicates the presence of abnormality is incorrectly identified (abnormality is missed).
False positive	Decision outcome that indicates the absence of abnormality is incorrectly identified (normal anatomy is judged as abnormal).
Image evaluation	Radiographers’ practice to determine the significance of a radiographic finding.
Image interpretation	Radiographers’ practice to assign a meaning to a radiographic finding.
Musculoskeletal system	Human body system that is made up of bones, joints and associated anatomical structures such as cartilages and tendons.
Posteroanterior (PA) radiographs	Radiographs taken from back-to-front.

Sensitivity	Ability to correctly identify abnormal appearances on radiographs.
Specificity	Ability to correctly identify normal appearances on radiographs.
True negative	Decision outcome that indicates the absence of abnormality is correctly identified.
True positive	Decision outcome that indicates the presence of abnormality is correctly identified.

List of abbreviations

Abbreviation	Explanation
A&E	Accident and emergency
AFROC	Alternative FROC
AP	Anteroposterior
BTEC	Business and Technology Education Council
CINAHL	Cumulative Index to Nursing and Allied Health Literature
CT	Computed tomography
DICOM	Digital Imaging and Communications in Medicine
ED	Emergency department
ENP	Emergency nurse practitioner
ERS	Extreme Response Style
ESP	Extended-scope physiotherapist
FN	False negative
FP	False positive
FRCR	The Fellow and the Royal College of Radiologists
FROC	Free-response receiver operating characteristics
GP	General practitioner
HCPC	Health and Care Professions Council
HEI	Higher education institute
HESA	Higher Education Statistics Agency
JAFROC	Jacknife AFROC
MeSH	Medical Subject Headings
MIU	Minor injury unit
MRI	Magnetic Resonance Imaging
MRPBA	Medical Radiation Practice Board of Australia
NHS	National Health Service
NLM	The National Library of Medicine

PA	Posteroanterior
PACS	Picture Archiving and Communication System
PCE	Preliminary Clinical Evaluation
PI	Principal Investigator
QUADAS2	Quality Assessment of Diagnostic Accuracy Studies
RADS	Radiographer abnormality detection scheme
ROR	The Royal College of Radiologists
SCoR	The Society and College of Radiographers
SIGRR	The Special Interest Group in Radiographic Reporting
SOS	Satisfaction of search
TN	True negative
TP	True positive
WHO	World Health Organization
WWH	WHAT, WHERE and HOW framework

Chapter 1. Introduction

Preliminary clinical evaluation (PCE) is a new clinical duty of diagnostic radiographers introduced by The Society and College of Radiographers (SCoR) (2013). The SCoR defines PCE as “the practice of radiographers whereby they assess imaging appearances, make informed clinical judgements and decisions and communicate these in unambiguous written forms to referrers”.

SCoR (2013) aspires that PCE will be a core practice of diagnostic radiographers in the United Kingdom (UK). The SCoR (2013) also postulates that successful implementation of PCE will bolster clinical imaging services and satisfy the needs of patients and referrers to allow faster admission to the appropriate clinical treatment, and thus ensure enhanced patient outcomes. However, has the feasibility of PCE by diagnostic radiographers been empirically proven?

Ever increasing emergency department (ED) admission rates have become a global problem. Prolonged patient waiting times and delay in treatment have been recognised as a public health problem (Pines et al. 2011). Delay and absence of clinical reports are known to adversely affect patient care and department management (Brealey et al. 2006), therefore prompt clinical reporting in EDs is vital to sustain clinical decision making. However, the medical imaging workforce is now in crisis. The Royal College of Radiologists (RCR) (2015) acknowledged the chronic shortage of radiologists in the United Kingdom (UK) and sustainability of future radiology services is now questioned (RCR, 2017a). The RCR (2015) estimated that 212,970 plain X-ray examinations were unreported for 30 days or more for all 155 National Health Service (NHS) acute trusts in England in 2015. Diagnostic delayed reports for plain X-rays could be alleviated by allocating more radiologists and at the expense of an increased delay in reports for Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). However, this portends a serious failure in patient care, since CT and MRI are more expensive and likely to provide conclusive diagnostic information than plain X-ray examination. The RCR (2015) identified only two possible remedies for relatively easier problem of unreported plain X-ray examinations: outsourcing or reporting by radiographers.

Radiographers were precluded from expressing clinical decisions in radiographic imaging until 1980s (Price, 2001). However, decision making for radiographs has since become the most predominant area of role expansion in diagnostic radiography (Snaith, 2013). Diagnostic radiographers in the UK participate in the Radiographer abnormality detection schemes (RADS) in order to mitigate the delayed diagnoses. Prior to the installation of digital systems, radiographers typically used the RADS, (often referred to the Red-dot system), by signalling the presence of abnormalities with a red dot to support other emergency staff. This terminology persisted, despite the current use of digital markers such as asterisks to denote abnormalities. Research evidence suggests that the RADS has been shown to reduce patient waiting time and radiologists' workload (Smith & Baird, 2007). RADS has been widely embraced as part of the extended role of radiographers since the introduction of its early form by Berman et al. (1981). The last national survey to investigate the UK RADS practice found that the majority of EDs and minor injuries units (MIUs) (n =

284/306; 92.8%) operated RADS, of which 77.8% (n = 221) adopted the Red-dot system (Snaith & Hardy, 2008). Radiographers with a postgraduate qualification in reporting now provide definitive medical reports to assist the imaging service. Moreover, an alternative form of the Red-dot system, The Traffic Light (TL) system has emerged (Higgins & Wright, 2016). In the TL system, radiographers are required to make a decision on every imaging examination: ‘Red = Abnormal’, ‘Green = Normal’ or ‘Amber = Unsure’. The TL system allows radiographers to make more explicit expression of their decisions and eliminate the ambiguity of the Red-dot system, but has yet to be widely adopted.

Since the discovery of X-rays by a German physicist, W. C. Roentgen in 1895, decision making in radiographic imaging by non-medically trained healthcare staff has been a subject of considerable debate. In the early years of medical radiation, the terms *radiographer* and *radiologist* were used interchangeably. In 1923, the SCoR clarified the differences of those occupational groups that *radiologists* were members of the medical profession who undertake medical diagnosis, while *radiographers* were trained non-medical assistants. In 1944, Furby, a radiographer, reiterated the difference in these professional groups that the clinical duty of radiologists was the provision of clinical reports for X-ray images, while the primary duty of radiographers was to be “the utmost service to radiologists” (Price, 2001). This distinction between radiologists’ and radiographers’ clinical duties remained unchallenged until 1971.

Swinburne (1971), a radiologist, argued that radiographers could accurately distinguish between abnormal and normal X-ray images. He suggested, in order to reduce the workload of radiologists and improve the service quality within the radiology department, trained radiographers could make accurate decisions for radiographs. Swinburne's paper became a seminal work as this was the first to point out the radiographers' clinical potential in X-ray image evaluation roles. Prime, Paterson & Henderson (1999) noted that Swinburne's paper also established several key themes that later researchers explored:

1. Shortage of radiologists whose workload could be alleviated by radiographers.
2. Reporting roles have the potential to improve radiographers' job satisfaction.
3. Pattern recognition is the main skill that radiographers will acquire.
4. Training and research are necessary for radiographers to gain skills for reporting.
5. Although not critical, medico-legal issues concerning radiographers' reporting should be considered.

Despite this, Swinburne's proposal did not receive immediate attention.

Four years later, the clinical potential of radiographers recaptured radiologists' attention. In 1975, Swinburne's idea was revitalised by an anonymous letter to the British Journal of Radiology (Anon., 1975) and responses to the letter from radiologists (Aberdour, 1976; Brindle, 1975; Cooper, 1976). The anonymous letter raised the problem of increasing workload of radiologists and subsequent increase in the number of unreported films, as well as the fact that decisions regarding patient management had been already made long before radiologists had a chance to view radiographs. The author of the letter considered that such situations could be perceived as failure of radiology service provision, and in order to increase the number of radiological reports, the author proposed that some elements of reporting roles of radiologists could be delegated to family doctors. Brindle (1975) acknowledged several problems in radiology service such as the shortage of radiologists, diminishing recruitment, increasing workload and expanding areas of practice. He then proceeded to question whether radiologists were physically capable of producing reports for every film whilst simultaneously maintain a desirable quality of radiology service. Aberdour (1976) also maintained that attempting to report on all X-ray examinations would lead to increased errors and distortion of work pattern. Aberdour considered reporting all X-ray films by radiologists was “foolish to try”. Cooper (1976), on the other hand, disagreed about the idea of delegation and argued that the quality of radiology service and patient management would become controllable only if radiologists viewed all the X-ray films.

Berman et al. (1985) conducted pioneering research to measure and compare abnormality detection accuracy of radiographers of all grades and junior casualty officers. This was the first attempt to place Swinburne’s proposal into the clinical context. In their study, both groups viewed 1496 plain X-ray films, of which 85% (n = 1272) consisted of suspected trauma cases. The results showed that abnormality detection accuracy of radiographers and junior casualty officers were 87.4% and 89.0%, respectively. One significant finding of this study was, although false negative (missed abnormality) rates of the two groups were similar (radiographers: 4.5% and junior casualty officers: 4.2%), nearly half of the clinically or medico-legally significant abnormalities missed by junior casualty officers were correctly found by the radiographers. The authors therefore proposed establishing a new system – subsequently known as RADS or the Red-dot system – that allowed radiographers to highlight abnormalities for casualty officers would reduce diagnostic errors.

Saxton (1992) reiterated several key issues relating to radiologists' reporting role:

1. Confining radiologists to reporting every film increased radiologists' workload and it had reached the point of inefficiency.
2. Limited radiological manpower was being used ineffectively.
3. Recruiting more radiologists for better radiological service provision was not a realistic solution owing to a lack of financial support from the NHS.
4. A lack of radiologists' time for plain X-ray reporting due to their new clinical responsibilities in other imaging modalities such as CT, MRI and ultrasound.

Saxton warned that these problems raised practical and medico-legal issues, and radiologists' medico-legal position would remain untenable unless these problems were acknowledged and directly addressed. One of Saxton's solutions to the problems was to spread certain areas of radiologists' practice to other non-medically trained staff, such as radiographers and nurses, and he argued that, under proper training and supervision, radiographers could undertake fracture reporting in trauma radiographs.

Loughran (1994) examined Saxton's proposal in the practical context. Loughran conducted a study to determine the effects of a six-month training programme on three radiographers' ability to make accurate decisions. The results showed the radiographers' sensitivity improved from 81.1% to 95.9% ($p < .001$) while specificity also improved from 94.4% to 96.6% ($p < .05$). The results also showed no statistically significant difference of sensitivity between radiologists and the radiographers after completing the training programme ($p < .001$), although the difference of specificity between these groups remained statistically significant ($p < .001$). Loughran concluded that experienced radiographers with a supplemental training programme in skeletal trauma film reporting could report plain X-ray films and had the potential to alleviate radiologists' workload. Robinson (1996) also conducted similar research and maintained that suitably trained radiographers could provide full text reports on trauma plain radiographs.

In parallel to the early research evidence from Berman et al. (1981), Loughran (1994) and Robinson (1996), the SCoR's vision of radiographers' decision making in radiographic imaging became more explicit in the 1990s. Radiographers who performed obstetric ultrasound scanning started to provide numerical data and clinical reports to doctors in the 1980s (Price, 2000). The SCoR acknowledged this newly expanding area of radiographers' responsibility in ultrasound studies and subsequently modified its Code of Professional Conduct in 1988 to allow radiographers to provide descriptions of images, measurements and numerical data in medical ultrasound. This was further amended in 1994 to suggest that radiographers could provide verbal comments to patients and written reports to medical staff. The RCR's Code of Conduct in 1994 also stated that radiographers could provide verbal and written reports and this formally supplanted the previously informal radiographers' reporting role (Freckleton, 2012). This was further modified to confirm that there was no statutory restriction on reporting of radiographic images performed by specially trained non-medical personnel (RCR, 1995).

As a measured response to the RCR's 1995 statement, the SCoR made an aspirational announcement that reporting would be a future requirement for radiographers (SCoR, 1997). Following this, a joint paper published by the SCoR and the RCR (1998) outlined certain tasks previously performed by radiologists, such as clinical reporting, and stated that they could be delegated to radiographers, although the need for appropriate training for radiographers before engaging in the tasks was also recognised. Concurrent with these policies developed by the SCoR and RCR, research into radiographers' decision-making accuracy burgeoned during this period. More evidence emerged to suggest that appropriately trained radiographers were beneficial additions to clinical reporting (Brealey et al., 2005; Carter & Manning, 1999; Piper, Paterson & Godfrey, 2005; Piper, Paterson & Ryan, 1999; Smith & Younger, 2002), with additional positive implications to patient management and cost-effective treatment (Friedenberg, 2000). Against these empirical backgrounds, the SCoR (2006) stated that radiographers (in an extension of RADS) should be able to provide written reports for trauma radiographs by 2010. However, Snaith and Hardy (2008) construed that the SCoR's statement did not suggest mandatory definitive reporting by radiographers. Instead, the SCoR introduced the "middle ground" of radiographers' role between the Red-dot system and clinical reporting, which would allow a more proactive role in image evaluation rather than simply signalling the presence of abnormalities. The "middle ground" sprang to life in 2013.

The SCoR (2013) solidified its position and introduced the concept of Preliminary Clinical Examination (PCE). The SCoR previously identified and distinguished two pillars of radiographers' practice: RADS and clinical reporting. RADS, generally referred as the Red-dot system, represents radiographers' "image evaluation" duties, in which general radiogaphers use basic image viewing strategies to judge or determine the significance of a finding. On the other hand, clinical reporting is the highest degree of decision making in diagnostic imaging. The role of clinical reporting is led by reporting radiographers who hold a post-graduate qualification in medical image interpretation. "Image interpretation" is the core practice of the reporting radiographers where they assign clinical meanings to radiographic findings in a written form. Research has consistently found that competencies of reporting radiographers in musculoskeletal imaging are favourably comparable with medically trained radiologists (Blakeley et al., 2008; Buskov et al., 2013; Carter & Manning, 1999; Piper et al., 1999; Piper et al., 2005; Robinson, 1996). Despite the beneficial implications, limitations of the Red-dot system have been known for many years. Robinson (1996) pointed out that the RADS system is characterised with three limitations:

1. It only distinguishes normalities and abnormalities – severity or significance of abnormalities are ignored.
2. It is a precursor to the clinical reports of radiologists and Accident and Emergency (A&E) physicians rather than replacing them.
3. It is informal – practice standard cannot be established.

The SCoR (2013) also acknowledged that the Red-dot system was ambiguous owing to its informal and voluntary nature, resulting in inconsistent outcomes for patients and referrers. The SCoR therefore introduced PCE to mitigate the limitations of the Red-dot system. Furthermore, the SCoR no longer holds a view about mandatory reporting by radiographers. Instead, they have introduced PCE ("image evaluation" with commenting) which acts as the middle ground between the Red-dot system and clinical reporting. The SCoR (2013) believes that PCE should become a core competence of radiographers and replace the ambiguous Red-dot system in the future.

The SCoR (2006) recommended that all the undergraduate programmes of diagnostic radiography in the UK incorporate image evaluation into their education. Evidence suggests this is in place. Hardy and Snaith (2009) conducted a survey questionnaire across the UK to elicit information regarding image evaluation education. 19 Higher Education Institutions (HEIs) indicated they had embedded image evaluation in their undergraduate programmes (n = 19/25; 76.0%), although educational contents and timing of their delivery varied. The authors concluded that the participating HEIs had offered appropriate education to satisfy the aspiration of the SCoR.

The SCoR (2013) argued that new graduates of diagnostic radiography at the point of registration with the Health and Care Professions Council (HCPC) should now have the necessary education and training to initiate PCE, despite acknowledging that they must continue to advance their competencies through preceptorship. Preceptorship is a period of adaptation into a new role. This is the period for newly qualified radiographers to "consolidate knowledge (educative), to be induced into the policies and procedures of the workplace (normative) and to reflect on their practice, especially on challenging experience (restorative)" (SCoR, 2003). A determined effort has gone into preceptorship in both medical and non-medical professional fields (Billay & Myrick, 2008). For example, the nursing profession has accepted the preceptorship as an effective education model that promotes a successful transition from students to more competent practitioners (Marks-Maran et al., 2012; Nielsen et al., 2017; Quek & Shorey, 2018). Notwithstanding the SCoR's anticipation, there is a severe lack of research evidence to evaluate the effect of the preceptorship on newly qualified radiographers in the UK. Literature suggests that there is only one publicly recorded preceptorship scheme by Nisbet (2008) without a practical evaluation of the programme. Tan, Feuz, Bolderson and Palmer (2011) also pointed out poor documentation of the preceptorship in radiography in the North American context. The benefits of the preceptorship known in other professional groups may not be directly transferable to diagnostic radiography. However, similar benefits are conceivable. A recent study (Stevens & Thompson, 2018) found that a focused training during the preceptorship could improve newly qualified radiographers' ability to detect and describe abnormalities. However, research has also identified difficulties that radiographers may encounter in their early career (Harvey-Lloid, Morris & Stew, 2019; Hyde, 2015; Naylor, Ferris & Burton, 2015). The research evidence encourages well documented preceptorship with the aim of alleviating newly qualified radiographers' difficulties and improve their clinical competencies.

Two professional bodies that represent diagnostic radiography force in the UK, (SCoR and HCPC), have not established performance standards for PCE. This may be a possible barrier to its implementation. The SCoR (2013) indicated that they expected the Red-dot system to phase out and be superseded by PCE, although a successful and widespread implementation requires that radiographers must demonstrate PCE accuracy that is at least equivalent to red-dot accuracy. No improvement to imaging service can be expected when the reliability of PCE falls behind the Red-dot system. What constitutes clinically reliable decisions in image evaluation is perhaps open to debate. However, in quantitative terms, "95% accuracy" is widely perceived as the performance standard for qualified reporting radiographers (Brealey, 2001a; Paterson, Price, Thomas & Nuttall, 2004; Stephenson et al., 2012). 80% accuracy has been suggested as the minimum performance standard for A&E skeletal decision making (Brealey, 2001b). Wright and Reeves (2017) reflected on the changes in radiography education and maintained that radiographers are now reasonably expected to achieve 90% accuracy in any form of decision making. However, the reliability of PCE has not been rigorously considered by the SCoR. The SCoR's current standards of practice states that radiographers must be "demonstrably competent", while simultaneously conceding that PCE performance standards are difficult to quantitatively define. Moreover, the SCoR and HCPC, a regulatory body which defines radiographers' standards of practice, have not worked towards an amicable agreement regarding performance standard and means of communication (The Red-dot, PCE or clinical reporting). In *Standards of proficiency*, HCPC merely acknowledged that "Registrant radiographers must be able to distinguish normal and abnormal appearances evident on images" (HCPC, 2013).

Lancaster and Hardy (2012) argued that the lack of evidence is one possible barrier for the implementation of a radiographer comment scheme. PCE is more cognitively demanding than the Red-dot system. Radiographers need to provide, not only red-dot style decisions, but also written comments to effectively communicate with referrers. Since the introduction of PCE by the SCoR in 2013, is there research evidence to underpin the extended image evaluation practice without eroding the professional autonomy and accountability of diagnostic radiogaphers? This research therefore set out to determine whether newly qualified radiographers were capable of providing reliable PCE. This research was conducted with the following aim and objectives:

Aim:

To benchmark new graduate radiographers' competencies in evaluation of plain appendicular X-ray images.

Objectives:

1. To measure accuracy, sensitivity and specificity of new graduate radiographers by conducting an image evaluation test with a test bank consisting 30 images of appendicular skeleton.
2. To evaluate quality of radiographic descriptions (comments) of PCE by using a new scoring system.
3. To understand and classify types and frequencies of PCE errors by using a PCE taxonomy.
4. To differentiate between, and critically review, the bodies of literature as they relate specifically to RADS and PCE.

Chapter 2. Literature review

2.1. Introduction

Healthcare research places immense value on public health. Research evidence has promoted new medical findings, development of treatments and improved quality of healthcare delivery to society (Institute of Medicine, 2009). However, an unmanageable deluge of healthcare evidence now confronts healthcare personnel, patients, researchers and policy makers. A systematic literature review grapples with this dilemma. A systematic literature review aims to identify and appraise research evidence, then summarise and synthesise a body of knowledge of a specific academic field (Higgins & Green, 2011). It also serves to determine methodological flaws in studies and identify a gap of knowledge where a lack of empirical studies is found (Knopf, 2006).

The need for a research base that underpinned development and expansion of radiographers' practice started to prevail in the mid-1990s in the UK. However, some perceived radiography to be "semi-professional" due to the use of research knowledge of other disciplines in practice, rather than its own (Nixon, 2001). Despite a broad range of possible research areas in radiography (SCoR, 2015), the research culture has not yet fully evolved in the profession (Harris & Paterson, 2016; Nightingale, 2016). A true healthcare profession requires to establish its own knowledge foundation which in turn allows autonomous management of clinical practice (Manning & Hogg, 2006). Failure to develop such a knowledge base will, arguably, continue to hold the profession back (McKenna, O'Neil & McIntyre, 1995).

Implementation of medical image evaluation by radiographers needs to be justified by research evidence. Research into radiographers' ability in X-ray image evaluation emerged in early 1990s in the UK, and it is now gradually spreading on an international scale (Buskov et al., 2013; Hazel, Motto & Chipeya, 2015; McConnell et al., 2012). Systematic literature reviews are an ideal means of synthesising research evidence and providing a penetrating insight into research areas of interest in radiography (Marshall & Sykes, 2010). Chapter 1 questioned whether there is research evidence that supports the feasibility of PCE by radiographers. First, a literature search was conducted to determine whether there were sufficient image evaluation studies to allow formulation of dependable research-based knowledge. The result indicated that research has been continuously conducted to investigate radiographers' image evaluation skills. However, the academic effort to synthesise research evidence appeared to have been discontinued since the mid-2000s (Brealey & Scally et al., 2005; Brealey et al., 2006). Importantly, the search found no PCE studies conducted since the SCoR's announcement in 2013. A literature review was therefore conducted to differentiate between, and critically review, the bodies of literature as they relate specifically to RADS and PCE.

2.2. Literature search strategies

Literature search in a systematic literature review must be as broad as possible to retrieve all the relevant studies with minimum effects of reporting biases (Smith, Devane, Begley & Clarke, 2011). Errors in a literature search may reduce sensitivity or precision, resulting in a biased and incomplete evidence base (Sampson & McGowan, 2006). Literature search strategies were therefore developed before the search to allow optimised results.

2.2.1. Eligibility criteria

The literature search applied eligibility criteria to ensure that all the relevant studies were included (Meline, 2006). Inclusion criteria established the standards for systematically searching relevant studies. Irrelevant studies were rejected when their titles and abstract clearly satisfied the exclusion criteria. The inclusion and exclusion criteria were defined prior to the literature search (Table 2.1). The SCoR (2013) does not specify areas of examinations that PCE must include. However, the survey results of Snaith and Hardy (2009) implied that radiographers do not have formal education for chest and abdominal radiograph evaluation. Robinson (1996) also maintained that advanced level of education is necessary for chest and abdomen, owing to the complex anatomy and a diverse range of abnormalities; therefore, the criteria were arranged so that the literature search would include skeletal image studies (with possible additional anatomical areas) but exclude studies that solely investigated radiographers' evaluation skills for chest or abdominal radiographs. Chapter 1 pointed out the difference between image evaluation (Red-dot system and PCE) and interpretation (clinical reporting). Investigating clinical reporting studies was outside the focus of this literature review. Studies that investigated reporting radiogaphers' competencies in image interpretation were therefore excluded.

Table 2.1.

Eligibility criteria for the literature review.

Inclusion criteria	Exclusion criteria
Study was included when following criteria are met:	Study was excluded when at least one of following criteria is met:
• It involved diagnostic radiographers or diagnostic radiographers and other healthcare personnel. • Imaging modality was plain X-rays. • Skeletal X-ray images (with possible additional anatomical areas) were used. • It measured diagnostic radiographers' performance of X-ray image evaluation (Red-dot or PCE) in clinical practice (test bank or audit) • Research result was presented with at least two types of outcome variables: Sensitivity and specificity.	• It did not involve diagnostic radiographers. • It did not measure radiographers' performance of X-ray image evaluation. • Imaging modality was not plain X-rays (e.g., CT and MRI). • It solely investigated radiographers' evaluation skills on chest or abdominal radiographs. • Clinical reporting studies that measured interpretation skill of reporting radiographers. • Research result was not presented with sensitivity and specificity. • The research methods and results are unique and a comparison with other study results is unachievable.

2.2.2. Databases and keywords

Electronic searches provide the most up-to-date information and relevant information in sources other than traditional books and journals (Knopf, 2006). The use of multiple electronic databases enables the capture of all the pertinent studies although it also increases the time and effort (Stevinson & Lawlor, 2004). The literature search of this review therefore included five electronic databases to retrieve relevant studies: PubMed, Cumulative Index to Nursing and Allied Health Literature (CINAHL), ScienceDirect, Web of

Science and ProQuest. The reasoning for including each of the databases is summarised in Table 2.2.

Table 2.2.

Reasoning for including the databases used for the literature search.

Database names	Reasoning for inclusion
PubMed	PubMed accesses mainly MEDLINE which contains over 23 million biomedical literature, but also other life science journals and online books. MeSH (Medical Subject Headings) is the controlled vocabulary thesaurus developed by The National Library of Medicine (NLM). The use of MeSH terms in PubMed allows uniform indexing of literature.
CINAHL complete	CINAHL complete provides a wide range of literature from nursing and allied health professions, including diagnostic radiography. The database can be searched by using CINAHL subject headings, similar to MeSH but headings reflect terms commonly used in nursing and allied health professions.
ScienceDirect	ScienceDirect is a database of literature from medical research as well as other scientific subjects including health and social care research.
Web of Science	Web of Science offers a multidisciplinary and comprehensive index of scientific, technical, social sciences, arts and humanities journal articles and conference papers.
ProQuest	ProQuest is a platform that provides documents from a wide range of sources (Arts, Business, Health & Medicine, History, Literature & Language, Science & Technology and Social Sciences) from different sources.

Preliminary literature search using a set of simple keywords and Boolean logic was first conducted in each database to explore literature that could provide a quick insight into the area of interest. Author keywords and common terms used in the titles and abstracts of the identified studies were assembled to develop three sets of themed keywords for the literature search (Table 2.3).

Table 2.3.

Categorised keywords used for free-text search (combined with a Boolean operator: "OR").

Keywords	Rationale
Radiographer* OR Radiography	The primary themes of the study.
Accuracy OR Competemc* OR Education* OR Program* OR Sensitivity OR Specificity OR Training	Keywords related to diagnostic radiographers' skills in X-ray image evaluation.
Comment* OR Interpret* OR PCE OR Preliminary clinical evaluation OR Red dot OR Red-dot OR Report*	Keywords related to radiographers' clinical roles in X-ray image evaluation. "Interpret*" was included because old literature used two terms "evaluation" and "interpretation" interchangeably. "Report*" was also used since the term traditionally indicated informal reporting (or commenting) by radiographers.

Several databases, such as PubMed and CINAHL, offer controlled vocabulary searching to alleviate the limitations of free-text searches. Controlled vocabularies are standardised indexing terms that flag relevant literature irrespective of author-supplied terminology. Controlled vocabularies ensure that articles with the same concepts are indexed uniformly (Brusco, 2010) and the use of such vocabularies in literature searches enables a coherent way to locate literature that may use different terminology for the same concept. Therefore, when searching in PubMed and CINAHL, free-text keywords and controlled vocabularies were used in conjunction as this technique enhances search quality

(Jain & Raut, 2011). Only free-text keyword searches were conducted for other databases that did not offer a controlled vocabulary search function.

Search filters were not used in order to maintain the breadth and balance of literature sources. A "Help" section of each database was reviewed before searching and the search strategies and keywords were modified accordingly. Appendix A summarises the rationale and results of the literature search in each database.

2.3. Results of the literature search

The literature search found 75 potentially relevant studies from five databases (Table 2.4).

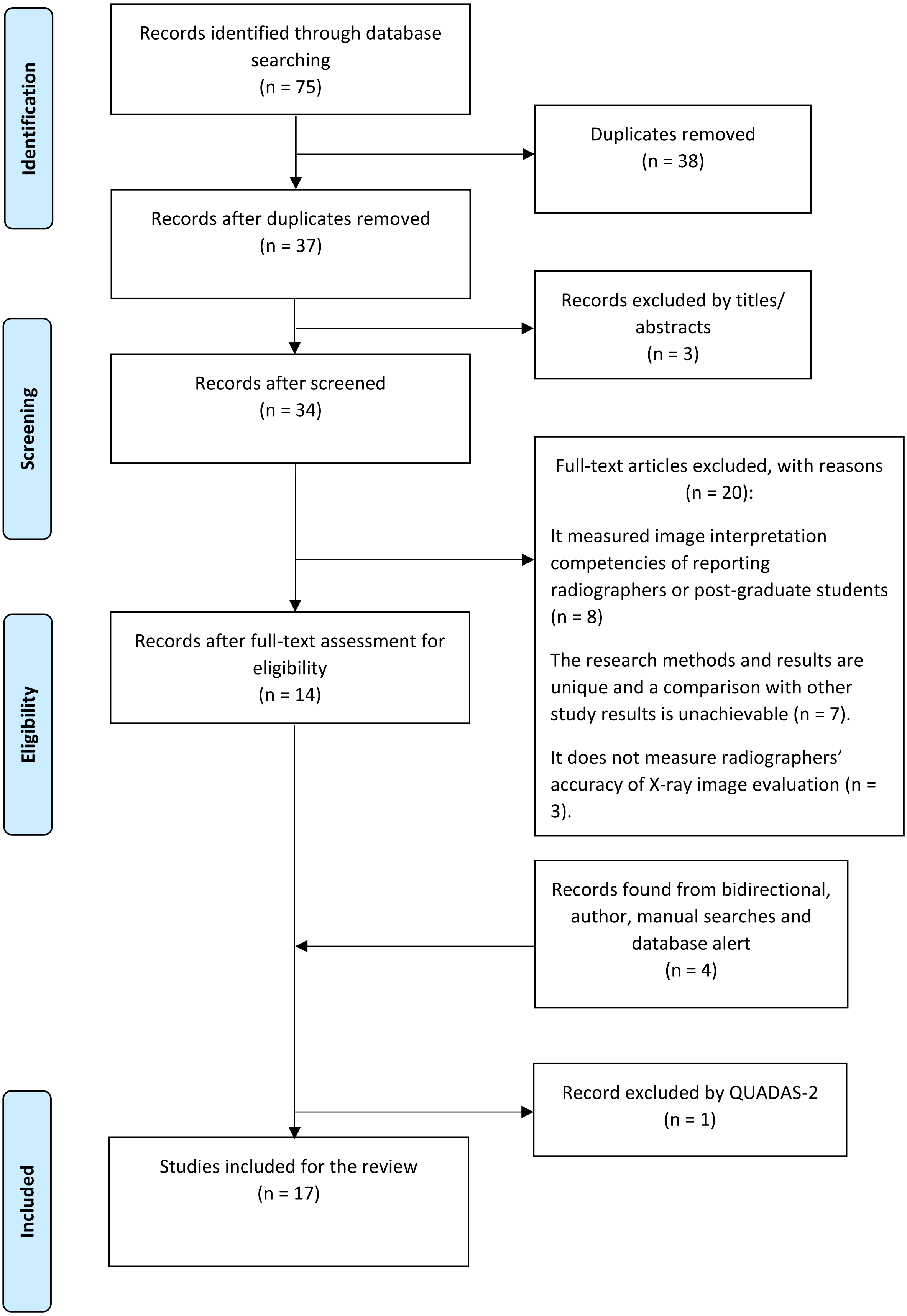
Table 2.4.

Results of the literature search in five electronic databases.

Databases	Results	Retrieved
PubMed	1,723	15
CINAHL	743	17
ScienceDirect	215	14
Web of Science	654	14
ProQuest	1,259	15
Total	4,594	75

Duplicates (n = 38) were first excluded by screening titles and abstracts of the extracted studies. Studies (n = 3) that clearly met at least one of the exclusion criteria were excluded. The remaining 34 full-text articles were scrutinised for eligibility. 20 studies were excluded with reasons, of which eight were clinical reporting studies. Bidirectional searches (citation and reference searches) and author search of the included studies were also conducted in Scopus to minimise publication and location biases (Stevinson & Lawlor, 2004; Hinde & Spackman, 2015). An additional manual search was also performed in key journal catalogues. After the completion of the literature search, database alert was set in each database to regularly update the search results. These searches added four eligible studies. The remaining literature (n = 18) was further scrutinised by using QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies) (discussed in the next section). One study was rejected owing to low methodological quality. The literature retrieval process identified that 17 studies were eligible for the review. (Figure 2.1).

Figure 2.1. PRISMA flow diagram describing the result of the literature retrieval.


Identification
Records identified through database searching (n = 75)
Duplicates removed (n = 38)
Records after duplicates removed (n = 37)
Screening
Records excluded by titles/ abstracts (n = 3)
Records after screened (n = 34)
Full-text articles excluded, with reasons (n = 20):
It measured image interpretation competencies of reporting radiographers or post-graduate students (n = 8)
The research methods and results are unique and a comparison with other study results is unachievable (n = 7).
It does not measure radiographers' accuracy of X-ray image evaluation (n = 3).
Eligibility
Records after full-text assessment for eligibility (n = 14)
Records found from bidirectional, author, manual searches and database alert (n = 4)
Record excluded by QUADAS-2 (n = 1)
Included
Studies included for the review (n = 17)

2.4. Study quality assessment

QUADAS-2 defines quality as: both the risk of bias and applicability of a study; 1) the degree to which estimates of diagnostic accuracy avoided risk of bias, and 2) the extent to which primary studies are applicable to the review's research question (Whiting et al., 2011). QUADAS-2 consists of four domains (patient/participant selection, index test, reference standard and flow and timing) to assess risk of bias and concerns regarding applicability (Appendix B). Risk of bias was assessed by signalling questions. As recommended by the developers, the original signalling questions that did not apply to the aim and objectives of the review were tailored to adequately address specific aspects of the literature review. These questions were then answered either yes, no or unclear: "yes" indicating low risk of bias. Concerns regarding applicability were judged by the information gathered while performing full-text analysis of the literature. Concerns regarding applicability for each domain were then rated as low, high or unclear. "Unclear" options for both risk of bias and concerns regarding applicability were used when the articles presented insufficient data to determine the quality. The results of the quality assessment are summarised and discussed in the discussion section of this chapter. This quality assessment using QUADAS-2 excluded one study from the literature review. This study was an audit poster and eight of 13 signalling questions remained "Unclear" due to insufficient research information. The results of QUADAS-2 assessment are presented and discussed in Chapter 2.7.1.

2.5. Collection of the literature data and synthesis of evidence

The literature review first elicited information regarding radiographers' sensitivity, specificity and/or accuracy in image evaluation. All studies in the literature review assessed radiographers' performance in image evaluation by first categorising their clinical decisions into four outcomes: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Then, sensitivity, specificity and accuracy were determined by the resulting TPs, FPs, TNs and FNs. TP, FP, TN and FN are defined as:

TP = Abnormal radiographs (positive) correctly identified.

FP = Normal radiographs (negative) incorrectly identified.

TN = Normal radiographs (negative) correctly identified.

FN = Abnormal radiographs (positive) incorrectly identified.

Sensitivity of radiographers in image evaluation denotes their ability to correctly classify abnormal radiographs as being abnormal and is estimated by the proportion of TP / (TP+FN), while specificity expresses their ability to correctly classify normal radiographs as being normal and is estimated by the proportion of TN / (FP + TN). Accuracy is the radiographers' ability to differentiate normal and abnormal radiographs and is estimated by the proportion of (TP + TN) / (TP + FP + TN + FN) (Baratloo, Hosseini, Negida & Ashal, 2015). The collected data and commonly discussed themes in the included studies were then critically evaluated and summarised. Chapter 3.2 discusses the calculation and use of sensitivity, specificity and accuracy in more detail.

2.6. Literature review

This literature review aimed to differentiate between, and critically review, the bodies of literature concerning the practice of image evaluation by radiographers. The review directed a primary attention to PCE studies (n = 6). However, PCE is a conceptual extension of the Red-dot system. This study therefore added Red-dot studies (n = 9) to the review. Two exceptional studies that investigated radiographers' Red-dot and PCE performance independently using the same sample population were also included (discussed separately in the following sections).

Red-dot studies employ a simple research method to determine radiographers' performance in image evaluation. Radiographers' binary decisions (red-dot or no red-dot / presence or absence of abnormality) are compared to the gold standard (radiological reports) and classified as correct or incorrect. Radiographers' performance is then expressed in the form of sensitivity specificity and accuracy, and/or the fraction of TPs, FPs, TNs and FNs.

PCE studies determine radiographers' skills in detecting abnormality (Red-dot) as well as their descriptive performance. Participants of PCE studies are typically general radiographers with or without a short training programme. Participating radiographers may be asked to provide Red-dot style decisions but comments are used to verify their clinical decisions. Radiographers' comments are classified as either concordant or discordant with the gold standard in most studies. Some studies that were conducted and considered as "reporting studies" prior to the definitions by the SCoR (2013) now fall into this category.

The studies also varied in the research conditions. Seven studies were conducted under clinical practice (audit) and 10 were carried out in a controlled condition (use of an X-ray image bank). Eight studies adopted a pre-post study design to evaluate the impact of education or training programme on radiographers' image evaluation competencies.

2.6.1. Radiographer abnormality detection schemes (RADS aka Red-dot system)

Nine studies measured radiographers' performance in Red-dot. Four studies were conducted within clinical practice (audit) while five used an image bank. Three Red-dot studies used a pre-post training design. Additionally, there are two studies that measured Red-dot and descriptive skills independently using the same groups of radiographers with a pre-post training design and image banks. These studies are discussed with other Red-dot studies. Table 2.5 summarises the Red-dot studies assessed in the review.

Table 2.5.

Summary of Red-dot studies.

Study	Study description	Study type	Education/training	Measurement of performance	Reference standard
Brown & Leschke (2012)	No information about the participants. A total of 3638 appendicular trauma radiographs from a hospital over a period of four months were retrospectively assessed for radiographer applied red-dots.	Audit.	None.	Each response was categorised as either true positive (TP), false positive (FP), true negative (TN) or false negative (FN). Sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	Validated radiologist report.
du Plessis & Pitcher (2015)	Nine radiographers with a minimum of 10 years of experience. The radiographers assessed the presence or absence of abnormality for a bank of 40 appendicular trauma radiographs.	Image bank.	None.	Each response was categorised as either TP, FP, TN or FN. Accuracy, sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	Consensus reports of three consultant radiologists.

Hardy & Culpan (2007)	115 radiographers with 1 to more than 25 years of experience. The radiographers assessed the presence or absence of abnormality and provided comments for a bank of 20 skeletal radiographs (12 appendicular and 8 axial) twice: before and after a short course.	Image bank.	A short course on musculoskeletal trauma.	Sensitivity and specificity were calculated for Red-dot and comments (PCE) independently (method of calculation is not fully described).	Unknown or insufficient description.
Hargreaves & Mackay (2003)	Seven radiographers with less than 1 to 35 years of experience. The radiographers assessed the presence or absence of abnormality twice: 8 weeks before (493 appendicular and axial skeleton radiographs) and 8 weeks after (546 appendicular, axial skeleton and chest radiographs) a training programme.	Audit.	A training was delivered by a clinical radiographer over a period of four months (three times a week).	Sensitivity and specificity were calculated (method of calculation is not fully described).	A reporting radiographer with a postgraduate qualification in diagnostic reporting.

Hazel, Motto & Chipeya (2015)	Nine radiographers with unknown clinical experience and educational background. The radiographers assessed the presence or absence of trauma and other pathological changes twice: before and after a training programme for image banks, each consisting of 100 appendicular and axial skeleton (without skull) radiographs. The radiographers also provided comments.	Image bank.	Six lectures and tutorials approximately 2 hours in duration. Each tutorial was offered on two separate occasions over a period of four months.	Each response was categorised as either TP, FP, TN or FN. Accuracy, sensitivity and specificity were calculated based the fraction of TP, FP TN or FN. The comments were categorised as incorrect, partially correct and correct. Accuracy, sensitivity and specificity were not calculated.	A single consultant radiologist.
Hlongwane & Pitcher (2013)	No information about the participants. A total of 369 appendicular and axial trauma radiographs from a hospital over a period of two months were retrospectively	Audit.	None.	Sensitivity and specificity were calculated (method of calculation is not fully described).	Consultant radiologist's report.

	assessed for radiographer applied red-dots.				
Mackay (2006)	133 Radiographers with 1 to 36 years of experience (n = 133, 132 and 39 for pre-, post and 6-month following a training programme respectively). The radiographers assessed the presence or absence of abnormality three times: before, after and six months following a training programme for an image bank of 30 appendicular and axial skeleton radiographs.	Image bank.	A two-day training programme with short keynote lectures and small group tutorials delivered by radiologists and reporting radiographer.	Each response was categorised as either TP, FP, TN or FN. Accuracy, sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	Consensus reports of a consultant radiologist in practice and a consultant and reporting radiographer who developed the image bank.
McConnell & Baird (2017)	16 radiographers (over 2 years of experience). The radiographers assessed the presence or absence of skeletal trauma for a bank of 209 radiographs.	Image bank.	None.	The radiographers used electronic worksheets to indicate presence or absence of abnormality (red-dot) and written description of abnormality (comments). Whether the comments influenced the allocation of TP, FP, TN and FN is unknown. Accuracy, sensitivity and	Consensus reports of three radiologists.

				specificity were calculated based the fraction of TP, FP TN or FN.	
McConnell & Webster (2000)	22 radiographers with 2 to 33 years of experience. The radiographers assessed the presence or absence of skeletal trauma and provide a short comment to clarify their decision three times: before, after and 6-8 weeks following a short course of training for the same series of an image bank, consisting of 42 unknown body parts of trauma radiographs.	Image bank.	A short course of training aimed at improving radiographers' red-dot accuracy.	Each response was categorised as either TP, FP, TN or FN. Accuracy, sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	Unknown or insufficient description.
Renwick, Butt & Steel (1991)	Unknown number of unselected radiographers of all grades. The radiographers assessed 3994 A&E radiographs of all body parts (including soft tissue) by using a choice of four categories, normal, abnormal, insignificantly	Audit.	None.	Radiographers' choice was compared with an assessment of radiologists who had a choice of three categories: normal, abnormal or insignificantly abnormal. False positives and false negatives were calculated based on the radiologists' assessment.	Reporting radiologists of all grade with a minimum of 18 months of experience.

	abnormal or further advice required, over a period of six weeks.				
Wright & Reeves (2016)	34 general radiographers with 4 to 26 years of experience. The radiographers assessed the presence or absence of abnormality by using a choice of five categories (definitely normal, probably normal, possibly abnormal, probably abnormal and definitely abnormal) for two sets of image bank, each consisting of 20 appendicular skeleton radiographs (50% prevalence of fractures).	Image bank.	None.	Accuracy, sensitivity, specificity were calculated based on the radiographer's decisions.	Double reported images by radiologists with consistent findings.

Including the study results of post-training assessment, Red-dot sensitivity ranged from 72.10% to 100.00%, while specificity ranged from 50.10% to 99.60%. Accuracy ranged from 65.47% to 93.7% (two studies did not report accuracy) (Table 2.6). Figure 2.2 compares the results of the Red-dot studies. From pre-post training studies, only the results of pre-training assessment were used to ensure the comparability with other non-training studies in the figure. In some Red-dot studies, radiographers provided comments to verify their decisions, but the comments had no influence in determining the outcome. These studies were therefore considered as Red-dot studies.

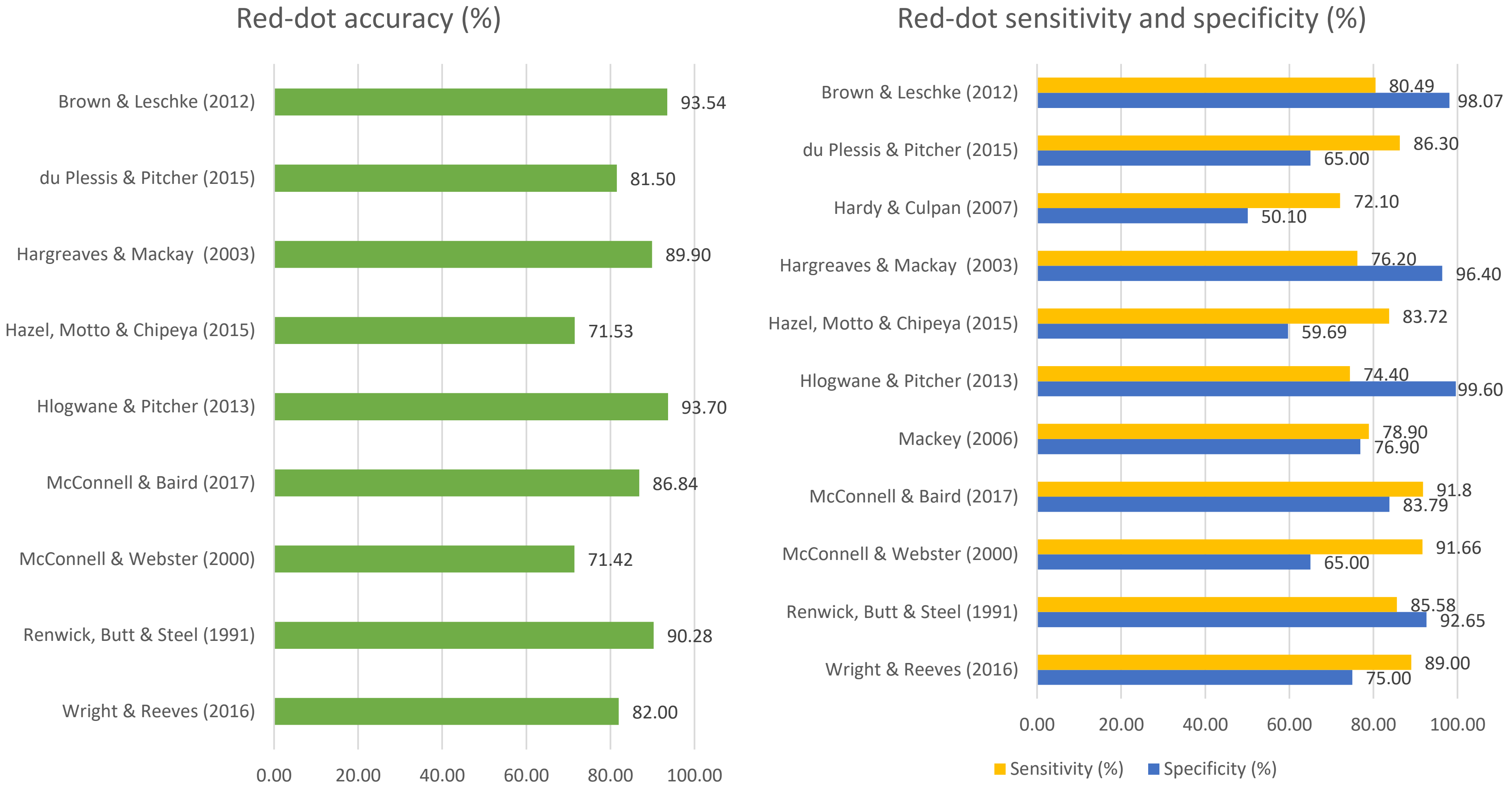
Table 2.6.

Results of the Red-dot studies.

Red-dot studies	Training	TP	FP	FN	TN	Accuracy (%)	Sensitivity (%)	Specificity (%)
Brown & Leschke (2012)	None	755	52	183	2648	93.54	80.49	98.07
du Plessis & Pitcher (2015)	None	-	-	-	-	81.50	86.30	65.00
Hardy & Culpan (2007)	Pre	-	-	-	-	-	72.10	50.10
	Post	-	-	-	-	-	88.50	53.40
Hargreaves & Mackay (2003)	Pre	-	-	-	-	89.90	76.20	96.40
	Post	-	-	-	-	93.00	81.30	96.10
Hazel, Motto & Chipeya (2015)	Pre	365	181	71	268	71.53	83.72	59.69
	Post	392	131	57	310	78.80	87.31	70.29
Hlongwane & Pitcher (2013)	None	-	-	-	-	93.70	74.40	99.60
Mackey (2006)	Pre	-	-	-	-	-	78.90	76.90
	Post	-	-	-	-	-	88.20	76.90
	6 months	-	-	-	-	-	76.50	79.90
McConnell & Baird (2017)	None	56	24	5	124	86.84	91.8	83.79
McConnell & Webster (2000)	Pre	-	-	-	-	71.42	91.66	65.00
	Post	-	-	-	-	65.47	100.00	53.33
	6 - 8 weeks	-	-	-	-	80.95	95.83	75.00
Renwick, Butt & Steel (1991)	None	1110	189	187	2383	90.28	85.58	92.65

Wright & Reeves (2016)	None	-	-	-	-	82.00	89.00	75.00

Figure 2.2. Accuracy, sensitivity and specificity of the Red-dot studies.



Hlongwane and Pitcher conducted the first Red-dot study in South Africa in 2013. This retrospective audit study analysed 369 trauma radiographs for the presence or absence of a red dot. The result demonstrated that the radiographers' accuracy was 93.7% with 74.4% sensitivity and 99.6% specificity. Further analysis showed that the sensitivity of experienced radiographers (82.0%) was better than inexperienced radiographers (63.9%), and therefore the authors concluded that years of clinical experience positively influence fracture detection rate. Possible improved validity is one methodological advantage of this study and other audit studies. In image bank studies, selection of radiographs is often arbitrary. The number of radiographs used for image bank studies is also limited (typically 20 to 30 images) when short test duration is desirable. On the other hand, audit studies can utilise a large number of X-ray images that reasonably mirror actual clinical load. As a result of this methodological strength, radiographers' performance determined in audit studies is likely to resemble their daily practice. One limitation is that audit studies assume a 100% participation rate in the Red-dot system. However, the Red-dot system is an informal and voluntary practice. Radiographers who decide not to participate in this practice exhibit 0% sensitivity and 100% specificity; thus, a low participation rate in a Red-dot audit will severely skew the results (unforeseen decreased sensitivity and increased specificity). Hlongwane and Pitcher (2013) did not report the radiographers' participation rate in their retrospective audit. The results of their study therefore need to be interpreted with caution.

Following Hlongwane and Pitcher (2013), du Plessis and Pitcher (2015) compared Red-dot performance of senior radiographers (n = 9) and medical officers (n = 8) by using an image test bank (n = 40). The images in the test bank were selected so that they represented the image profile of the Emergency unit where the study was undertaken. The result demonstrated that the accuracy and sensitivity of radiographers were higher than medical officers (81.5% vs 67.8% for accuracy and 86.3% vs 68.7% for sensitivity), although the two groups showed similar specificity (65%). Hlongwane & Pitcher (2013) and du Plessis & Pitcher (2015) conducted their studies in the same area (the Western Cape Province of South Africa) within a relatively short frame (approximately three years) but their findings appeared inconsistent (accuracy: 93.7% vs 81.5%, sensitivity: 74.4% vs 86.3%, specificity: 99.6% vs 65%). Figure 2.2 suggests that the results of this study are relatively comparable with other image bank studies. However, this study also shares many methodological limitations with other studies that used X-ray image banks. In quantitative research, a large and randomly selected sample is ideal for improved precision of estimate, generalisability and statistical power. However, image evaluation studies generally depend on a small group of radiographers. The participants in research are also volunteering or self-selected radiographers owing to a non-probability sampling method. Generalisation of the research results to a larger context is therefore often inappropriate. In addition to the sampling method, the development of image banks needs careful selection of X-ray images in order to reduce possible prevalence bias. Prevalence of abnormality in image banks is generally high (around or above 50%), compared to lower rate of abnormality in clinical practice. The prevalence of abnormality in the test bank of du Plessis and Pitcher (2015) was particularly high (75%, n = 30). Hardy, Flintham, Snaith and Lewis (2015) recommended the use of image banks that reflect clinical practice. However, the influence of a high prevalence of abnormality on image evaluation ability is still poorly understood. Critical scrutiny is therefore necessary while interpreting results of image evaluation studies with image banks.

Wright and Reeves (2016) devised an assessment tool (RadBench software) that allows objective measurement of image evaluation performance. Their pilot study primarily aimed to determine the technical feasibility of RadBench software to measure image evaluation skills of radiographers and other healthcare personnel. A total of 34 general radiographers participated in two sessions of image evaluation tests. In order to assess the radiographers' levels of confidence, they were asked to provide answers using a five-point scale (1: definitely normal, 2: probably normal, 3: possibly abnormal, 4: probably abnormal and 5: definitely abnormal). The results demonstrated that the average sensitivity and specificity were 89% and 75% respectively. Researchers generally conduct audits or image evaluation tests at one or two study sites. A geographical barrier is one possible reason for limiting the scale of research and sample size. A larger sample size is expected when research is free from the geographical barrier. The study by Wright and Reeves (2016) provided evidence that an online platform can remove the barrier. In 2014, 18, 647 online image evaluation tests were taken by various healthcare professions across the world Wright and Reeves (2016), indicating that research using a larger sample size is now technically feasible. However, this method also highlights a limitation that researchers cannot directly control image viewing conditions (e.g., monitor size, resolution and luminance etc) and this may affect the ability to evaluate X-ray images.

McConnell and Baird (2017) measured and compared image evaluation performance of final year medical students (n = 16) and radiographers (n = 16). They pointed out that the radiographers' potential to support other emergency department staff has been recognised since the 1980's in the UK. In Australia, however, radiographers are still an underused healthcare group even though appropriately educated radiographers in emergency departments could provide support for medical interns while they develop evaluation skills in musculoskeletal trauma radiographs. The measurement of the performance of two groups was carried out by using a test bank, consisting of 209 musculoskeletal radiographs with injury prevalence of 16.13%. The participants were provided with electronic worksheets to record their responses and the returned worksheets were compared against radiological reports with consensus on the diagnosis. Overall, the radiographers performed better. Accuracy, sensitivity and specificity of the radiographers and medical students were 86.84% vs 81.34%, 91.80% vs 86.07% and 83.79% vs 77.70% respectively. Medical Radiation Practice Board of Australia (MRPBA) (2013) requires that radiographers must be able to identify clinically significant radiographic appearances. The evidence from this study suggested that radiographers with appropriate education can and should assist junior doctors in early practice years, the interns and other ED multidisciplinary groups. One advantage of this study was the image bank used for the evaluation test. The bank correctly reflected examination types and trauma prevalence (16.13%) at the emergency department of the study site. The test bank could have improved the validity of the test. However, the large size of the test bank (n = 209) might have been contrarily affected. Fatigue is a known factor to influence image evaluation performance (Stec, Arje, Moody, Krupinski & Tyrrell, 2018). Although image reading time is likely to vary with examination types, if we accept that average reading time to process one examination of plain radiograph is 1.4 minutes (84 seconds) (Fleishon, Bhargavan & Meghea, 2006), the test required a total of 292.6 minutes (4.88 hours) to complete. The participants might not have viewed 209 images consecutively. However, fatigue with a possible reduction in performance should be acknowledged. The participants received their tests in USB memory sticks via Australian Post. Distributing tests by post could be another strategy to remove the geographical barrier and increase a sample size if the cost is justifiable. Despite this sampling potential, the radiographer response rate of this study remained low (4%) owing to the time required to complete the test. One disadvantage of this method is similar to Wright and Reeves (2016) that the viewing conditions cannot be controlled or supervised.

Brown and Leschke (2012) conducted a Red-dot study by retrospectively auditing 3638 appendicular musculoskeletal trauma radiographs at the Department of Emergency Medicine at an Australian metropolitan hospital. This study was conducted with a particular emphasis on radiographers' ability in detecting subtle fractures. Retrospectively audited radiographs contained 938 abnormal images (25.8%), of which 338 (9.3%) were considered as "subtle" (displacement or distraction < 1mm). Overall, mean sensitivity and specificity were 80.4% and 98.0% respectively. However, subgroup analysis of subtle fractures found that sensitivity dropped to 45.8%, indicating that 54.2% of the subtle fractures were not red-dotted by the radiographers compared to 20.5% of all fractures. The authors argued that the Red-dot system only alerts emergency physicians on an intermittent basis due to the radiographers' high under-calling (missing) rate for subtle fractures and advised a cautious approach to the introduction of radiographer reporting in Australia. They also asserted that the shortage of the radiologists in Australia was alleviated between 2000 and 2010 and therefore radiologist reports should remain as the gold standard.

Renwick, Butt and Steele (1991) conducted a prospective audit study to assess radiographers' ability to identify abnormal radiographs at A&E departments. Unselected radiographers of all grades (n = unknown) were asked to assess the radiographs with a choice of four categories: normal, abnormal, insignificantly abnormal, or further advice required. This study included a total of 3994 radiographs of all body parts (including soft tissues and sinuses) for analysis. The result demonstrated 7% false positive and 14% false negative rates. Despite acknowledging the radiographers' potential to assist casualty officers' clinical decision making, the authors maintained that a false positive rate of 7% was too high for the radiologist's reporting duty to be delegated to radiographers. However, the results must be interpreted with caution because of some methodological limitations. First, this study was conducted in 1991, before Radiography became a graduate profession in 1993. It is very likely that the participating radiographers had not have received formal education and training for image evaluation. Second, the radiographers' participation rate was not recorded. The radiographers did not know when the audit commenced and completed although they were informed about the research procedure. Some of the radiographers might have opted not to participate in the practice when they were uncertain or simply busy. Third, the authors noted that most missed fractures occurred for X-ray images of skull, facial bones, chest, abdomen and soft tissues. However, further breakdown of the results indicated that the radiographers demonstrated 90.14% sensitivity and 94% specificity for the appendicular skeleton with 5.47% false negative rate and 11.40% false negative rate. Sensitivity and specificity for the axial skeleton were slightly lower: 88.94% and 91.14% with 9.80% with 14.29% of false positive and false negative rates. Although the possible impacts of the radiographers who did not participate in the audit must be considered, the radiographers' image evaluation skills for musculoskeletal system appear reasonably acceptable.

There were three studies that evaluated educational impacts on radiographers' Red-dot competencies. McConnell and Webster (2000) devised a two-day training programme with a focus on improving Red-dot accuracy. The radiographers (n = 22) were assessed three times (before, after and six to ten weeks following the completion of the programme) in order to explore the effect of the programme on their image evaluation skills. The same series of 42 trauma radiographs were used for the three tests. The second test (at the end of the programme) showed a considerable increase in false positives, resulting in the median sensitivity of 100.00% with a concomitant decrease in the median specificity of 53.33% and accuracy of 65.47%. However, the third test, conducted at six to ten weeks after the programme showed a decreased rate of false positives (median sensitivity of 95.83%) with an improved specificity (75.00%) and accuracy (80.95%). The authors also posed a question whether the improved evaluation skills by training programmes can be retained for a longer time span. McConnell and Webster (2000) was the first study to investigate into the impacts of education on image evaluation skills and they attributed the improvement of performance to the training programme. However, this study highlights common limitations of educational intervention studies. The intervention studies have been conducted with a hypothesis that education and training have positive impacts on image evaluation skills. The authors of educational intervention studies have ascribed the improved performance to education and training. A positive link between education and skills is conceivable. However, evidence is needed to support the hypothesis. The influence of education has not yet been explicitly clarified owing to the absence of control groups in the research design. Moreover, there is a dearth of discussion to pinpoint what leads to improvement in image evaluation performance. McConnell and Webster (2000) observed 8.34% of improvement in sensitivity after the training. If the hypothesis is proven, a more pedagogically valuable question is what causes the improvement. Evidence appears fragmentary unless research answers this question. Therefore, the results of educational intervention studies must be construed with caution.

Hargreaves and Mackay (2003) measured educational impacts on radiographers' Red-dot skills with a longer length of training programme than McConnell and Webster (2000). Seven self-selected radiographers were first audited for their Red-dot performance before the commencement of an education programme. The radiographers were then provided with tutorials, three times a week (each restricted to 30 minutes) over a period of four months. The radiographers were re-audited after the completion of the programme over the same period (eight weeks). The radiographers' mean sensitivity improved from 76.2% to 81.3% with a negligible decrease in specificity (96.4% to 96.1%).

Mackay (2006) conducted a study to determine the impact of a short course on radiographers' Red-dot performance. In this study, a short course (two days) was developed. The participating radiographers (n = 133) took tests three times (before, after and six months after the completion of the course) by using the test bank consisting of 30 radiographs. The median sensitivity increased from 78.9% to 88.2% after the short course, but it then decreased to below the initial score after 6 months. The median specificity on the other hand showed no fluctuation throughout the study and remained at 76.9%. McConnell and Webster (2000) and Mackay (2006) reported the median values of sensitivity and specificity instead of the mean values seen in other image evaluation studies. The authors did not explain the reason for using the median values. Caution must be applied to interpret the results because radiographers' performance expressed by the mean and median values may be incomparable.

There are two studies that evaluated radiographers' Red-dot and comment skills independently. Hardy and Culpan (2007) developed a research design to compare radiographers' ability to Red-dot and comment on A&E radiographs. The radiographers (n = 115) undertook an assessment and the authors measured their ability to recognise (Red-dot) and describe (comment) abnormal appearances of radiographs at the start and end of a short course on musculoskeletal trauma by using a test bank consisting of 12 appendicular and eight axial skeletal radiographs. The results demonstrated that the radiographers' mean sensitivity after the short course improved from 72.1% to 88.5%. The mean specificity also improved from 50.1% to 53.4%, although this was considerably lower than the results from other studies. The radiographers' commenting performance demonstrated a similar pattern to red-dotting. After the short course, their commenting sensitivity and specificity improved from 47.8% to 74.4% and 50.7% to 51.4% respectively. This study was the first to compare radiographers' ability to red-dot in conjunction with descriptive skills. The authors found that radiographers' comment sensitivity curtailed when compared with their Red-dot sensitivity. This reduced comment sensitivity occurred when abnormal images were correctly classified as abnormal (a red-dot) but the reasoning behind their decisions were wrong (incorrect identification of abnormality). The findings carry crucial implications for clinical practice that radiographers' correct decisions could be made based on wrong reasons. The authors warned that such erroneous decisions would result in reduced service quality in A&E departments. The sample size of this study was a methodological advantage. This study recruited a larger group of radiographers (n = 115) than other image evaluation studies, which potentially resulted in improved generalisability and statistical power. The study primarily aimed to compare the radiographers' abilities to identify abnormalities and comment. One criticism is that they appeared to have directed a rapt attention to sensitivity. The radiogaphers demonstrated considerably low mean specificity, 50.1% (pre-training) and 53.4% (post-training), compared to the radiographers in other studies. It is possible that the study truly reflected the specificity of this particular group of radiographers. However, a reasoned discussion on specificity might have enriched the study findings. Piper and Paterson (2009) hypothesised that the inclusion of the axial skeleton radiographs in the bank resulted in the low specificity of the radiographers. Sample radiographs presented in the article also imply the possibility that the authors intentionally included many normal images that mimic fractures (normal variants) in the image bank. This could be the reason for the low specificity. It is probable, therefore, that the tests underestimated the radiographers' specificity. This accentuates the importance in careful selection of X-ray images for test banks.

Hazel et al. (2015) evaluated the impact of a training programme on radiographers' pattern recognition (Red-dot) ability and descriptive comments on musculoskeletal images.

In this pre-post training study, the training programme included a tutorial aimed to assist the radiographers in a systematic analysis of radiographs and how to compose descriptive comments. The radiographers (n = 9) were first asked to identify if the image was normal or abnormal (Red-dot), then provide comments on the images that they identified as abnormal. This study did not quantify the comments to allow calculations of sensitivity and specificity. Instead, the comments were classified into three categories: correct, partially correct and incorrect. The result demonstrated that accuracy, sensitivity and specificity of Red-dot improved. The analysis of comments also showed that incorrect comments decreased (from 24.11% to 17.78% of all the comments made in pre- and post- training tests) after the training programme, while partially correct and correct comment increased (from 16.78% to 21.78% and 7.78% to 10.33 respectively). A qualitative analysis of the comments also indicated that the radiographers used more acceptable medical terms to describe the pathology after the training programme. The authors did not make a direct comparison of Red-dot and comment sensitivity. However, the figures presented in their study showed that the mean pre- and post- sensitivity for Red-dot were 83.73% and 87.28% respectively, while correct comments only accounted for 7.78% for pre- and 10.33% for post- of all the comments made for abnormal images, indicating that some of the radiographers correctly classified abnormal images as abnormal with incorrect or partially correct reasoning. The finding therefore supports the conclusion drawn by Hardy and Culpan (2007) that radiographers' Red-dot sensitivity is not always concordant with comment sensitivity. Hardy and Culpan (2007) and Hazel et al. (2015) developed a similar research method. One characteristic methodological approach of Hazel et al. (2015) was that their tutorial focused on abnormality identification as well as a systematic method for structured commenting. The results indicated that the radiographers improved their descriptive skills after the tutorial. However, what constituted “good comments” seemed to have been subjectively determined. The development of a more scientific method to measure comment quality is desirable.

2.6.2. Preliminary Clinical Evaluation (PCE)

Six studies evaluated radiographers’ performance in PCE. Three studies were conducted within clinical practice while another three used an image bank. Four PCE studies used a pre-post training design. Table 2.7 summarises the PCE studies assessed in the review. Including the study results of post-training assessment and also PCE results from Hardy and Culpan (2007), PCE sensitivity ranged from 47.80% to 95.90%, while specificity ranged from 50.70% to 97.30%. Accuracy ranged from 64.17% to 95.70% (four studies did not report accuracy) (Table 2.8). Figure 2.3 compares the results of the PCE studies. From pre-post training studies, only the results of pre-training assessment were used to ensure the comparability with other non-training studies in the figure.

Table 2.7.

Summary of PCE studies.

Study	Study description	Study type	Education/training	Measurement of performance	Reference standard
Coleman & Piper (2009)	18 radiographers with band 5 and 6. The radiographers assessed the presence or absence of abnormality by using a choice of five categories (definitely normal, probably normal, possibly abnormal, probably abnormal and definitely abnormal) for a bank of 20 appendicular radiographs. They also provided reports for answers considered to be possibly abnormal, probably abnormal or definitely abnormal.	Image bank.	None.	A maximum score of two marks were awarded when the image was correctly classified, and location and description of the abnormalities were correct. One mark was recorded when the answer was partially correct. Partially correct answers were awarded fractional marks (e.g., ½ TP and ½ FP). Sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	A consensus on diagnosis was reached by original anonymised reports, a consultant radiologist with many years of skeletal reporting, a senior radiology registrar and an advanced practitioner radiographer with five years of plain film reporting experience for each radiograph.
Loughran (1994)	Three radiographers with a minimum of 5 years of experience.	Audit.	A regular series of weekly X-ray tutorials during the study period (6 months).	Unknown or insufficient description.	No formal gold standard. The author reassessed when there were discrepancies

	The radiographers provided 3595 reports of skeletal trauma radiographs over a 6-month period.				between radiologists' and radiographers' reports.
McConnell et al. (2012)	10 radiographers with unknown clinical experience and educational background. Each radiographer provided reports three times: before, after and 8-10 weeks following an educational programme for a bank of 102 randomly selected images of A&E appendicular skeleton.	Image bank.	An educational programme was delivered by a senior lecture in medical imaging with radiographer reporting training.	The radiographers used opinion worksheets to indicate presence or absence of abnormality (red-dot) and written description of abnormality (report). Each response for red-dot was verified by the report from the opinion sheets and categorised as either TP, FP, TN or FN. Accuracy, sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	A consensus on diagnosis was reached by at least three radiologists for each radiograph.
McConnell, Devaney & Gordon (2013)	10 radiographers with unknown clinical experience and educational background.	Audit.	An education programme offered in the study of McConnell et al. (2012).	The radiographers used opinion worksheets to indicate presence or absence of abnormality (red-dot) and written description of abnormality (report). Each	Radiologist report.

	The radiographers previously completed the educational programme in McConnell et al. (2012). The radiographers were audited for a total of 655 appendicular radiographs over 22-day period.			response for red-dot was verified by the report from the opinion sheets and categorised as either TP, FP, TN or FN. Accuracy, sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	
Piper & Paterson (2009)	18 radiographers with six months to over 20 years of experience. The radiographers assessed the presence or absence of abnormality by using a choice of five categories (definitely normal, probably normal, possibly abnormal, probably abnormal and definitely abnormal) for a bank of 20 appendicular radiographs (before and after a short course). They also provided reports for answers considered to be possibly abnormal, probably	Image bank.	A short course (six of two-hour sessions) of image evaluation.	A maximum score of two marks were awarded when the image was correctly classified, and location and description of the abnormalities were correct. One mark was recorded when the answer was partially correct. Partially correct answers were awarded fractional marks (e.g., ½ TP and ½ FP). Sensitivity and specificity were calculated based the fraction of TP, FP TN or FN.	Unknown or insufficient description.

	abnormal or definitely abnormal.				
Smith & Younger (2002)	26 radiographers with 1 to 27 years of experience. The radiographers provided 820 reports of all body parts (including chest and abdomen) over a period of three months.	Audit.	None.	Opinion sheets were used to indicate presence or absence of abnormality. Reports were used to verify the radiographers' clinical decision, and then categorised as true positive/negative or false positive/negative based on the reference standard.	Radiologists' reports.

Table 2.8.

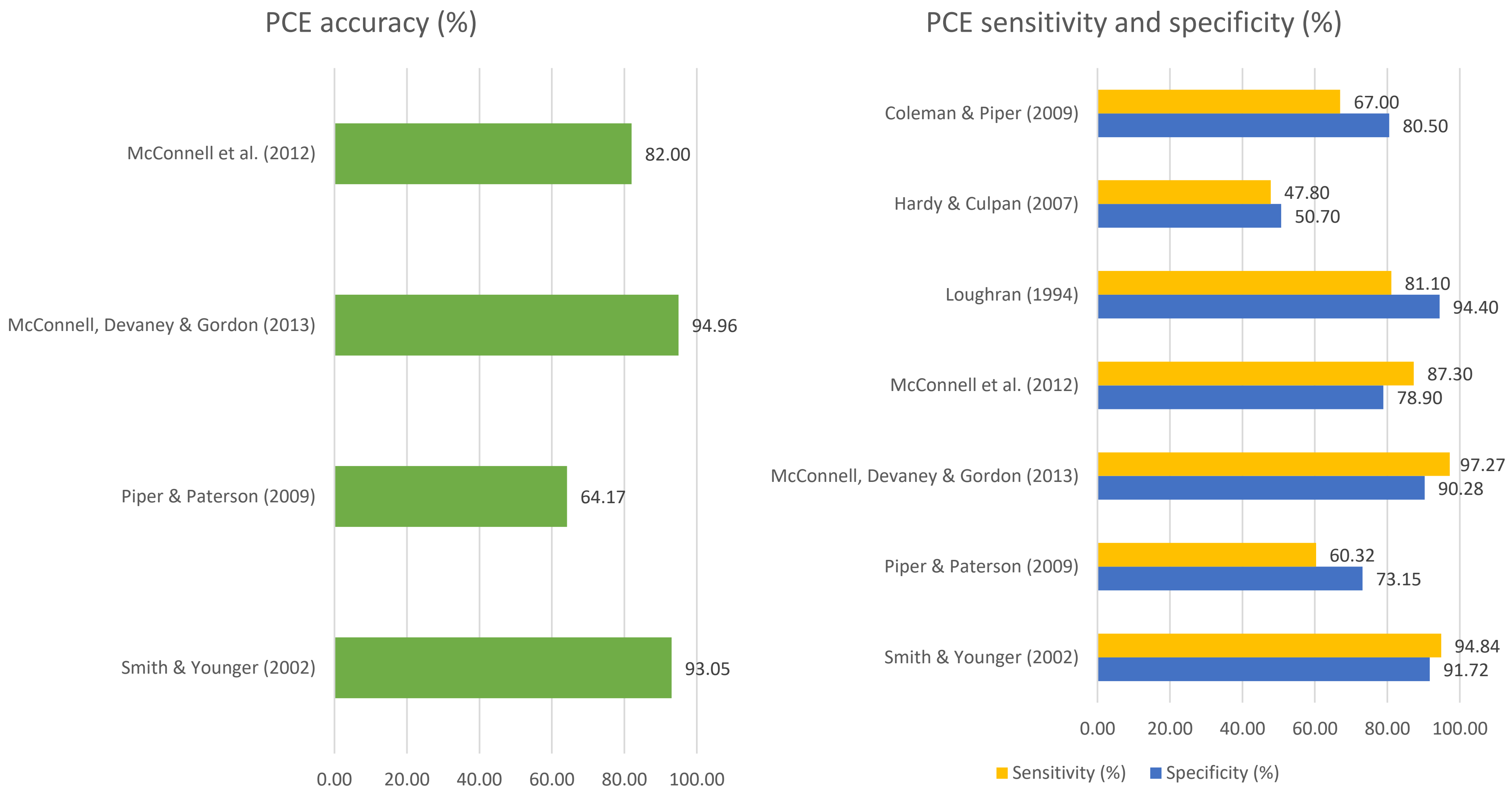
Results of the PCE studies.

PCE studies	Training	TP	FP	FN	TN	Accuracy (%)	Sensitivity (%)	Specificity (%)
Coleman & Piper (2009)	None	-	-	-	-	-	67.00	80.50
Hardy & Culpan (2007)[1]	Pre	-	-	-	-	-	47.80	50.70
	Post	-	-	-	-	-	74.40	51.40
Loughran (1994)	Pre	-	-	-	-	-	81.10	94.40
	Post	-	-	-	-	-	95.90	96.60
McConnell et al. (2012)	Pre	-	-	-	-	82.00	87.30	78.90
	Post	-	-	-	-	81.40	90.80	76.00
	8 - 10 weeks	-	-	-	-	86.80	93.50	82.90
McConnell, Devaney & Gordon (2013)	Post[2]	427	21	12	195	94.96	97.27	90.28
Piper & Paterson (2009)	Pre	152	29	100	79	64.17	60.32	73.15
	Post	173.5	18.5	78.5	89.5	73.06	68.85	82.87
Smith & Younger (2002)	None	331	39	18	432	93.05	94.84	91.72

[1] PCE sensitivity and specificity from Hardy & Culpan (2007)

[2] The participants had a training programme in McConnell et al. (2012).

Figure 2.3. Accuracy, sensitivity and specificity of the PCE studies.



Loughran (1994) was the first to examine radiographers' ability to detect abnormality and describe findings for skeletal radiographs from an A&E department. In this pre/post-education study, three radiographers took part in a six-month training programme (weekly X-ray tutorials) and their comments were audited over the same period of time. The study found that sensitivity and specificity of the radiographers improved during the training and audit period, from 81.1% to 95.9% and 94.4% to 96.6% respectively. Literature suggests that the notion for the 95% accuracy rule (Chapter 1) principally hinges on this study's findings. The results also indicated that the overall error rate (FNs and FPs) declined throughout the study period. The author suggested image evaluation of appendicular skeleton could be safely delegated to specially trained radiographers, although the author also noted that evaluation of more anatomically complex parts of the skeletal system (such as the skull and spine) may need to be restricted to radiologists. The educational intervention studies typically offer short courses (Hardy & Culpan, 2007: Mackay, 2006: McConnell & Webster, 2000: Piper & Paterson, 2009). The relatively longer education intervention (weekly X-ray tutorials over six months) was an exceptional component in the research design of Loughran (1994). This conceivably provided the participating radiographers with sufficient opportunities to reflect on their learning experience, thus resulting in their image evaluation performance being favourably comparable with radiologist. However, a drawback of the study design was that the radiogaphers comments were dichotomously classified (correct or incorrect), despite the common partially correct comments (Hazel et al., 2015). This is also a methodological obstacle in many PCE studies that the threshold to classify comments is set without scientific rationale. Pre-defined criteria used to decide the threshold could easily alter study results. The small sample of Loughran (1994) (n = 3) and its possible research consequences should be acknowledged.

In an Australian study, Smith and Younger (2002) argued that the Red-dot system lacked precision, especially for research purposes. They pointed out that, for example, the Red-dot system’s dichotomous classifier judges a decision as true positive when a radiographer incorrectly identifies a normal variant as abnormal and other subtle abnormalities on the same radiograph are missed. However, a truthful outcome of such a case is a combination of false positive (normal anatomy is incorrectly identified) and false negative (abnormality is missed). To surmount this ambiguity, they proposed the use of a radiographer opinion form in conjunction with the Red-dot system to detect errors caused by false positive decisions. This form consisted of tick boxes to indicate radiographers’ general and specific opinion with a comment section to clarify their reasoning for the findings. In their study, 26 self-selected radiographers completed the forms for 820 A&E radiographs (musculoskeletal, chest and abdomen) over a three-month period. The analysis of the forms demonstrated that the radiographers’ sensitivity and specificity were 94.8% and 91.7% respectively. The authors maintained that the use of a radiographer opinion form is a useful application to provide initial evaluation of radiographs and has the potential to go beyond the Red-dot system. Care must be taken to interpret the radiographers’ performance since the study included radiographs of appendicular skeleton as well as the axial skeleton, abdomen and chest. According to more detailed figures, sensitivity for upper and lower appendicular images were 99.1% and 98.2%, while specificity were 93.9% and 93.0% respectively. The radiographers’ evaluation skills for the appendicular skeleton appeared comparable with radiologists and reporting radiographers. However, as the authors noted, some of the radiographers might have self-selected examinations (e.g., chose to provide the forms for easier cases and opted out of challenging clinical cases). This might have overestimated the radiographers’ performance.

Also in Australia, McConnell et al. (2012) adopted a similar research method to Smith and Younger (2002). The authors in this pilot study first developed and delivered an educational programme to radiographers and then they determined the radiographers' image evaluation and descriptive skills by using a radiographer opinion worksheet. This worksheet was specifically developed for trauma radiographs and contained a comment section and tick boxes to provide Red-dot style decisions, types of abnormality and level of confidence. Ten radiographers provided their clinical decisions by using the form three times: before, immediately after and 8-10 weeks following the educational programme. The radiographers provided answers for an image bank consisting of 102 trauma appendicular skeleton. Each response for Red-dot was verified by the comments on the opinion sheets and categorised as either TP, FP, TN or FN. Overall, the radiographers improved their accuracy, sensitivity and specificity after the training programme, although specificity was the least improvement owing to a high false positive rate (Table 2.8). In the final assessment conducted 8-10 weeks after the educational programme, they achieved 86.80% accuracy, 93.50% sensitivity and 82.90% specificity. An analysis of the comments found that the radiographers' descriptive skills improved. The authors concluded that Australian radiographers with appropriate education and continuous audit of performance have the potential to assist in Emergency departments that are understaffed or depend on junior medical personnel. Following this pilot study, McConnell, Devaney and Gordon (2013) conducted an audit study using the same group of radiographers. The worksheets used in this study consisted of sections to provide Red-dot style decisions, types of abnormality and comments. The radiographers were audited for a total of 655 appendicular radiographs over a 22-day period, and their performance was compared with the emergency doctors' (n = 10) records of the patients. The radiographers' accuracy, sensitivity and specificity were 94.96%, 97.27% and 90.28% respectively. Statistical analysis showed no significant difference of performance between two groups. McConnell et al. (2012) and McConnell et al. (2013) emphasised that, while PCEs are being produced, the presence of patients with potential signs of injury positively influenced radiographers' abnormality detection and its description. The authors therefore suggested that involving radiographers in image evaluation and double-reading images with emergency department doctors at pre-radiologist stage could optimise abnormality detection rate. The authors in both studies noted a possible selection bias because of the self-selected radiographers with varying clinical experience and educational background. Adult appendicular radiographs were solely used to measure the radiographers' image evaluation performance in both studies. This biased selection of radiographs raises questions about the validity of the image bank and the study findings. The same limitations are also observed in the following two PCE studies.

Piper and Paterson (2009) examined the effect of a short training programme on nurses (n = 22) and radiographers (n = 18). The participants in two groups undertook a short course (a total of 12 hours) in image evaluation of the appendicular skeleton. A test bank comprised of 20 appendicular radiographs was assembled for the image evaluation test. The participants viewed the radiographs of the bank and expressed their decision using a five-point scale: definitely normal, probably normal, possibly abnormal, probably abnormal, definitely abnormal). They also provided comments to clarify the nature and location of the abnormality they identified. Unlike other PCE studies, they utilised a partial mark method to make judgement on partially correct comments provided for abnormal images. For example, this method recorded 1/2 TP when an abnormal image was correctly classified as abnormal but the comment did not satisfy all the pre-defined criteria in the expected answer. Sensitivity and specificity were calculated based on the sum of the whole and partial TPs, FPs, TNs and FNs. The authors also used a partial marking method for test bank scoring. Two marks were recorded when abnormal images were accurately identified or normal images were identified as normal. One mark was recorded when comments were partially correct, for example when one abnormality was correctly identified but another abnormality on the same radiograph was missed. A maximum of 40 marks was achievable in this scoring method. The results demonstrated that overall performance of the radiographers was better than the nurses. Although the findings were not statistically significant, the radiographers' sensitivity and specificity improved from 60% to 69% and 73% to 80% respectively. A comparison of the mean test scores of pre- and post-training indicated that the radiographers showed statistically significant improvement in their performance (pre-training: 25.7 and post-training: 29.1). The authors maintained that the positive educational impact on image evaluation skills should encourage radiographers in providing PCEs in emergency departments and minor injury units.

Coleman & Piper (2009) considered difference of image evaluation skills between radiographers, nurses and casualty officers when viewing a bank of 20 appendicular skeletal images. This study included 18 radiographers, 13 nurses and seven casualty officers. The method from Piper and Paterson (2009) was adopted to calculate sensitivity/specificity and test bank scores. Prior to the image evaluation test, the participants were asked about their levels of confidence in image evaluation, while in normal practice, on a scale of 1 to 10. The results revealed that the radiographers achieved higher mean test bank score (28.5/40;

71%) than the nurses (21.5/40; 54%) and the casualty officers (21.5/40; 54%). The radiographers demonstrated a higher mean value of sensitivity (67%) than the nurses (49%) and the casualty officers (51%). The mean specificity achieved by the radiographers (80.5%) was also greater than the nurses (54%) and the casualty officers (57%). Considering the three groups' perceived ability in image evaluation, the mean test bank scores of the casualty officer and nurse groups showed no correlation with their perceived capability to correctly evaluate radiographs. On the other hand, there was a moderate positive correlation between the radiographers' mean test bank score and their perceived capability, suggesting that the radiographers were more likely to accurately perceive their image evaluation skills than the casualty officers and nurses who overestimated their evaluation ability. However, the authors noted that the radiographers' mean sensitivity in the study was relatively lower than other mean sensitivity values in similar studies. They attributed this radiographers' low sensitivity to a lack of in-house training and therefore recommended that training programmes are essential to maintain sufficiently high sensitivity.

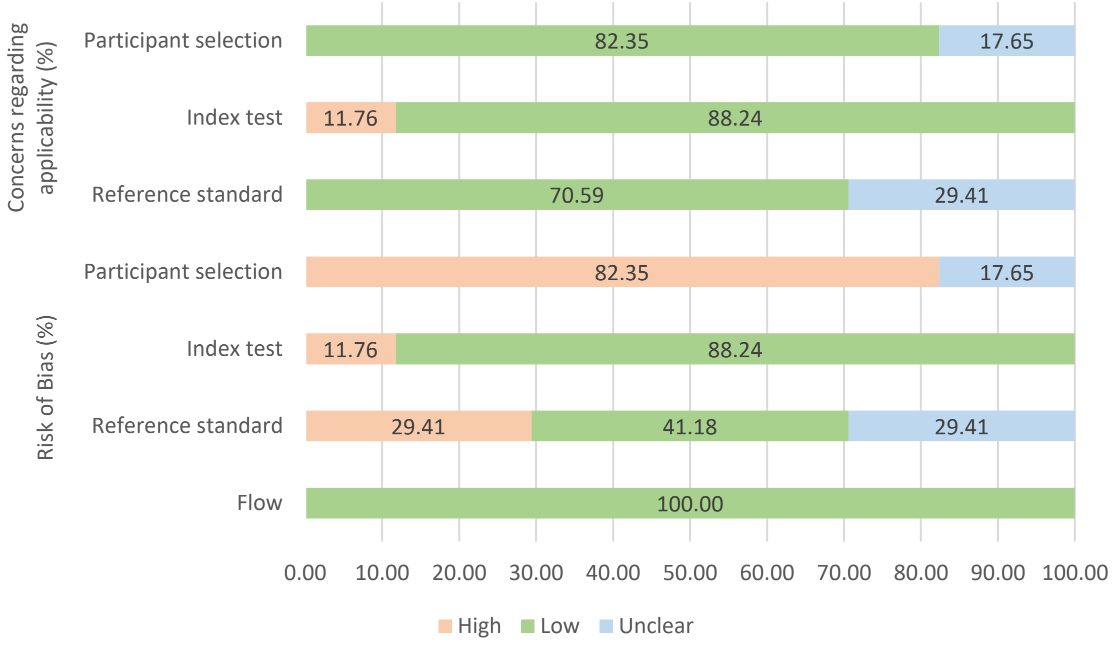
2.7. Discussion

This literature review assessed a total of 17 studies that evaluated diagnostic radiographers' performance in the two related tiers of plain film image evaluation: RADS (Red-dot) and PCE. The review has presented the evidence about radiographers' competencies in image evaluation. The primary finding suggests that most of the studies' results and conclusions support image evaluation practice by radiographers. Despite the variations in research methods (audit or image bank) and tiers of image evaluation practice, a majority of the authors advocated that radiographers with appropriate education and training have the potential to assist radiological reporting service for musculoskeletal radiographs. Furthermore, all the studies that evaluated the impacts of training programmes have concluded that radiographers' performance, especially sensitivity, in image evaluation improved after educational interventions. The following section discusses the findings of the literature review in more detail.

2.7.1. Evaluation of research quality – QUADAS-2 assessment

Prior to the literature review, QUADAS-2 was used to assess the quality of the studies concerning image evaluation by radiographers. Figure 2.5 summarises the result of the quality assessment. Domain 1 of QUADAS-2 assessed a risk of bias and concerns regarding applicability for participant selection. Differences in participating observers can affect the study outcomes. High risk of bias for participation selection was observed in 14 studies (82.35%). Although QUADAS-2 recommends a random sample of participants to prevent bias, most of the reviewed studies employed a convenient sampling method that relied on self-selected volunteer radiographers or radiographers who were attending an educational programme. The literature review indicated that this has been a generic limitation of the image evaluation studies that random or consecutive sampling of radiographers is difficult to achieve. It is conceivable that the self-selected participants had an avid interest in image evaluation and the estimate of their evaluation performance was greater than the performance of radiographers with other clinical interests. Three audit studies poorly documented the information of the participants (Brown & Leschke, 2012: du Plessis & Pitcher, 2015: Hlongwane & Pitcher, 2013). Risk of bias for participant selection therefore remained “Unclear”.

Figure 2.5. Results of research quality assessment by QUADAS-2.



On the other hand, there was a negligible concern for the applicability of participant selection since the target population of the review question (diagnostic radiographers) clearly matched with the participants of most of the reviewed studies (n = 14, 82.35%).

Domain 2 explored the index tests of the reviewed studies. In image evaluation studies, an index text measures performance of image observers (e.g., radiographers, radiologists, casualty officers and nurses). The measurement is either retrospective audit or image evaluation tests using X-ray image banks. The participants' accuracy, sensitivity and specificity are determined by comparing their evaluation results against the gold standard. Methodological variation in index tests may have different impacts on the study results. For example, research evidence suggests that prevalence of abnormality in image banks influences image reading accuracy (Pusic et al., 2012; Nocum, Brennan, Huang & Reed, 2013; Hardy, Flintham, Snaith & Lewis, 2016).

The reviewed Red-dot and PCE studies raised little concern for a risk of bias for the index test because of their relatively straightforward study methods. The Red-dot studies determined the participants' accuracy, sensitivity and specificity by dichotomously classifying their decisions: correct or incorrect. The binary classification of the Red-dot studies mechanically judges the image observers' decisions (correct or incorrect) without the classifiers' subjectivity. Therefore, inter-rater reliability (which is the extent to which the same results are obtained by different raters) (McHugh, 2012), is high. However, one limitation of this classification method is that it only examines the final decisions made by image observers. It lacks an adequate analytical power to inspect how observers arrive at their decisions. The participants' decisions of the reviewed Red-dot studies could have been falsely classified as correct even when the decision-making process involved partially correct or incorrect reasoning. The PCE studies were conducted with more methodological rigour. Two PCE studies used a partial marking system to address the issue related to the classification of decisions made by partially correct or incorrect reasoning (Coleman & Piper, 2009; Piper & Paterson, 2009). The marking system recorded fractional marks (e.g. 1/2 TP and 1/2 FN) when the participants correctly detected abnormality but failed to describe all key elements that were pre-defined in the gold standard. Three PCE studies used opinion forms that asked the observers to provide Red-dot style decisions and then comments to clarify the reasoning for their decision (McConnell et al., 2012; McConnell et al., 2013; Smith & Younger, 2002). Unlike the classification system of the Red-dot studies, the use of the partial marking system and opinion forms allowed the classifiers to verify whether the participants came to the decisions with correct findings. QUADAS-2 found a high risk of bias for the index tests in two PCE studies (Hardy & Culpan, 2007: Loughran, 1994). This is because they did not provide sufficient information regarding research methods to assess the reliability. For example, it is unknown whether the studies defined an acceptable level of agreement between the participants' comments and the gold standard to judge their concordance (i.e., how were partially correct reports dealt with?). It is possible to hypothesise that there were inconsistent levels of threshold to determine concordance between the participants' comments and the gold standard. The two studies could have under or overestimated the participants' performance. The bias in this domain is related to subjectivity of interpreting index tests (Whiting et al., 2011), which could compromise the reliability and validity of image evaluation studies. Neep, Steffens, Riley, Eastgate and McPhail (2017) pointed out that, there is little data provided to allow determination of reliability and validity in image evaluation studies. This literature review also found that the reviewed studies devoted a paucity of attention to the validity and reliability.

Domain 3 assessed the use of gold standard. In image evaluation studies, a gold standard (or reference standard) refers to a collection of radiological reports that is believed to have 100% sensitivity and specificity. An estimate of accuracy, sensitivity and specificity is made by comparing observers' decisions and the gold standard. An inappropriate application of the gold standard without acknowledging its limitations leads to inaccurate results (Brealey & Scally et al., 2005). For example, misclassification occurs when the gold standard contains erroneous reports. In Radiology, clinical reports with diagnostic consensus constitute the gold standard when assessing image evaluation accuracy (Onega et al., 2013). Risk of bias caused by the gold standard was low for seven (41.18%) studies because they used validated radiological reports (Brown & Leschke, 2012; Coleman & Piper, 2009; du Plessis & Pitcher, 2015; Mackay, 2006; McConnell & Baird, 2017; McConnell et al., 2012)

Risk of bias was considered high in five studies (29.41%) (Hargreaves & Mackay, 2003; Hazel et al., 2015; Hlongwane & Pitcher, 2013; Renwick et al., 1991; Smith & Younger, 2002). This was because the gold standards used in these studies were comprised of reports from a single radiologist or reporting radiographer. This method arbitrarily assumes that the gold standard has a zero-error rate but it can be a potential source of misclassification bias. Berlin (2007) estimated that radiologists' error rate in their daily practice is around 3.5% to 4%. Brady (2017) also maintained that radiologists' reports should not be assumed to be definitive or incontrovertible. Therefore, risk of bias was considered high for those studies which used the gold standard produced by one radiologist/radiographer. There was little concern about applicability since the target condition by the gold standard (presence or absence of abnormalities on plain radiographs) matched the review question for most of the studies (n = 12, 70.59%). Five studies (29.41%) did not describe the gold standard, therefore risk of bias and concerns regarding applicability remained unclear (Hardy & Culpan, 2007; Loughran, 1994; McConnell et al., 2013; McConnell & Webster, 2000; Piper & Paterson, 2009).

Domain 4 (Flow) assessed three types of biases: partial verification bias, differential verification bias and outcome reporting bias (Table 2.9). Overall, the reviewed studies showed little concern for these biases. Risk for partial and differential verification biases were low because methodology and results sections of each study suggested that all participants received the same reference standard. There were no studies that reported exclusion of certain participants for the analysis and the risk for outcome reporting bias was also low.

Table 2.9.

Three types of biases that are assessed in Domain 4 (Whiting et al., 2011).

Types	Definition
Partial verification bias	Only a portion of participants is evaluated against the gold standard.
Differential verification bias	Some participants receive different standards (gold and “brass”).
Outcome reporting bias	Some participants were excluded from the analysis.

2.7.2. X-ray image evaluation by diagnostic radiographers

This literature review elicited the current evidence about X-ray image evaluation by diagnostic radiographers. Regardless of the image evaluation types, the review found that many of the authors concurred that radiographers with appropriate training and education have the potential to accurately evaluate plain skeletal radiographs. The SCoR (2013) acknowledges that establishing the performance standard in a quantitative term is difficult (Chapter 1). However, clinical reporting appears to have its own performance standard. Robinson, Wilson, Coral, Murphy and Verow (1999) explained that an acceptable standard in image evaluation (or interpretation) implies a performance that is indistinguishable from that of a group of experienced consultant radiologists. They argued that, prior to defining such a performance standard, the variation of plain film clinical reporting among experienced radiologists must be established. They therefore investigated the variation of three experienced radiologists reporting for skeletal, chest and abdominal radiographs. The results found 9-10% of disagreement for the skeletal images with the average error rate between 3-6 % per radiologist. In response to this, Brealey (2001a) argued that the standard for image evaluation performance must reflect radiographers' clinical performance underpinned by research evidence and proposed that any professional group involved in clinical reporting of A&E skeletal radiographs should demonstrate 95% (ideal), 90% (optimal), and 80% (minimal) accuracy. The Special Interest Group in Radiographic Reporting (SIGRR) reviewed the study results of Loughran (1994) and Piper et al. (1999) and established that at least 95% of sensitivity and specificity for musculoskeletal radiographs can and should be maintained for clinical reporting (Paterson, Price, Thomas & Nuttall, 2004). The same standard for reporting of musculoskeletal plain films is proposed by Stephenson et al. (2012). It is also known that in the late 1990s, postgraduate programmes for image interpretation at six universities began to develop the 95% policy and expected their students to demonstrate 90% to 95% reporting accuracy at the end of the education (Prime et al., 1999). It is perhaps reasonable to expect that reporting radiographers, who hold a post-graduate qualification in image interpretation, should demonstrate minimum of 95% sensitivity and specificity. This expectation is undeniably supported by clinical reporting studies. Radiographers in clinical reporting studies consistently demonstrated sensitivity and specificity that are above or fairly close to 95% (Blakeley et al., 2008; Buskov et al., 2013; Carter & Manning, 1999; Piper et al., 1999; Piper et al., 2005; Robinson, 1996). A meta-analysis of clinical reporting studies by Brealey and Scally et al. (2005) also provided evidence that reporting performance of selectively trained radiographers was indistinguishable (92.6% sensitivity and 97.7% specificity) from radiologists of varying seniority. The clinical contribution to reporting service is in little doubt.

Contrary to clinical reporting, there is an absence of widely accepted performance standards for the Red-dot system and PCE. Paterson et al. (2004) explains this is due to the difficulty of establishing verifiable and absolute standards for image evaluation performance. Brealey (2001b) also maintained that establishment of the standard or acceptable level of error rate needs to take account of economic and social costs associated with diagnostic and therapeutic outcomes, and ultimately patient management at a societal level. The current literature indicates that there are still no performance standards that are underpinned by rigorous research in a socioeconomic context. In an empirical context, authors of six reviewed Red-dot/PCE studies acknowledged performance standards (ranging from 85% to 95%) (du Plessis & Pitcher, 2015; Hazel et al., 2015; Hlogwane & Pitcher, 2013; Mackey, 2006; Wright & Reeves, 2016; Smith & Younger, 2002). Authors of more than half of the Red-dot/PCE studies (n = 11) took a neutral stance and concluded their studies without setting a baseline of image evaluation performance.

If we accept that the 90% sensitivity and specificity (introduced in Chapter 1 and discussed in greater detail in Chapter 5) should be maintained for any type of image evaluation and applied the performance standard to the results of the 11 Red-dot studies, including Hardy and Culpan (2007) and Hazel et al. (2015), there is no study in which radiographers demonstrated over 90% of mean sensitivity and specificity at the same time. Overall, the radiographers in seven Red-dot studies demonstrated higher sensitivity than specificity (du Plessis & Pitcher, 2015; Hardy & Culpan, 2007; Hazel et al., 2015; Mackey, 2006; McConnell & Baird, 2017; McConnell & Webster, 2000; Wright & Reeves, 2016). This may illustrate radiographers' common tendency towards over-calling imaging examinations that leads to increased sensitivity caused by a high rate of false positives and subsequently reduced specificity. The radiographers in four studies achieved above 90% specificity (Brown & Leschke, 2012; Hargreaves & Mackay, 2003; Hlogwane & Pitcher, 2013; Renwick et al., 1991). However, these four studies were retrospective audit projects. The radiographers' possible selective participation in the Red-dot system and positively skewed specificity must be considered.

Since the introduction of the Red-dot system by Berman et al. (1985), Renwick et al. (1991) were the first to voice an unconvinced view on the feasibility of image evaluation by radiographers. They argued that the radiographers' false positive rate of 7% that they found in their study was too. Loughran (1994) critically responded that their findings were based on the performance of untrained radiographers with different levels of clinical experience and suggested that experienced radiographers with appropriate training could evaluate radiographs of the appendicular skeleton with a high degree of accuracy. In the study conducted by Hargreaves and Mackay (2003), the radiographers demonstrated a false positive rate of 3% for both pre- and post-training Red-dot assessments. They asserted that radiography had undergone substantial changes in qualification, educational level and programme contents, and a false positive rate of 3% was a noticeable improvement since 1991.

Brown and Leschke (2012) favoured a careful approach to image evaluation by radiographers owing to two reasons: 1) radiographers’ high false negative rate for subtle fractures, and 2) mitigation of radiologist shortage in Australia. They suggested that a cautious approach to clinical reporting by untrained radiographers was necessary. However, their approach has been subjected to criticism. The study did not determine the sample size and participation rate. Although they acknowledged the involuntary nature of the Red-dot system, they assumed that most radiographers participated in the Red-dot system in Australia. Conflicting research evidence was found by Neep, Steffens, Owen and McPhail (2014). They undertook a survey (n = 73) to investigate frequencies of Australian radiographers’ participation in the Red-dot system. The result showed that 41% (n = 30) of the radiographers participated in RADS in less than 20% of examinations. In audit studies, absence of red dots on abnormal radiographs are automatically judged as false negative decisions regardless of radiographers’ intention to participate in RADS. Neep et al. (2014) therefore argued that a large fraction of false negatives for the subtle fractures could be a result of the radiographers who chose not to participate. Smith (2013) called attention to the fact that the shortage of radiologists in Australia had not been alleviated. Smith explained that the number of radiologists between 2000 and 2010 grew only 35%, while the number of X-ray examinations increased by 54%. Furthermore, the number of imaging examinations is still increasing because of the aging population.

The arguments put forward by Renwick et al. (1991) and Brown and Leschke (2012) are similar, in that they stated that the practice of clinical reporting should be confined to radiologists because of the high error rate of the Red-dot system by untrained radiographers. The arguments are predicated on the understanding that the Red-dot system is a direct forerunner of clinical reporting. However, in the UK context, the SCoR (2013) now defines that clinical reporting is the practice of radiographers with a postgraduate qualification and appropriate training to produce diagnostic reports. The Red-dot system by untrained radiographers does not serve as an immediate substitute for clinical reporting, but an informal forerunner of the definitive reports. A direct comparison of the accuracy between the Red-dot system and clinical reporting is therefore misleading. They did not consider educational support for radiographers who have not reached the desired level of image evaluation performance. Instead of discouraging radiographers from giving their initial opinion, Smith (2013) emphasised the need for redesigning the initial image evaluation system and educating radiographers so that they could provide a short description of abnormal appearances.

Despite these conflicting views, the review of 11 Red-dot studies found there was no group of radiogaphers that achieved 90% sensitivity and specificity at the same time. Figure 2.2 indicated that many performed below 90% sensitivity (nine groups) and specificity (seven groups), perhaps indicating that more intense educational investment is necessary. Indeed, many research authors who advocate image evaluation by radiographers have pointed out the importance of providing appropriate educational opportunities to qualified radiographers (Hlongwane & Pitcher, 2013; Mackay, 2006; McConnell & Baird, 2017;

McConnell & Webster, 2000) and continuous audit to establish and maintain performance standards (Hardy & Culpan, 2007; Paterson et al., 2004; Wright & Reeves, 2016).

The literature review assessed seven PCE studies and found less consistent results than the Red-dot. Two groups of radiographers with a previous educational intervention demonstrated above mean sensitivity and specificity of 90% at the same time (Loughran, 1994; McConnell et al., 2013). Independent evaluation of sensitivity found dichotomised results. The radiographers in four studies achieved above 90% sensitivity (Loughran, 1994; McConnell et al., 2012; McConnell et al., 2013; Smith & Younger, 2002), although the rest of the radiographers in three studies demonstrated below 80% sensitivity (Coleman & Piper, 2009; Hardy & Culpan, 2007; Piper & Paterson, 2009). The findings for PCE specificity were inconsistent. Three groups of radiographers achieved above 90% specificity (Loughran, 1994; McConnell et al., 2013; Smith & Younger, 2002). However, specificity of the radiographers in other three studies was below 90% (Coleman & Piper, 2009; McConnell et al., 2012; Piper & Paterson, 2009) and 80% (Hardy & Culpan, 2007). These conflicting results from seven PCE studies may suggest that further research is necessary.

2.7.3. Educational impacts on X-ray image evaluation

Education appears a parallel research interest to image evaluation studies. The literature review found that nearly half (n = 8) of the reviewed studies (n = 17) investigated the impact of education or training programmes on radiographers’ image evaluation performance. This review will therefore provide an additional insight into the effect of educational intervention on radiogaphers' performance in image evaluation.

Brealey et al. (2006) pointed out an absence of evidence for the impacts of educational intervention. Since then several studies have been conducted to evaluate the effect of education on image evaluation performance. Overall, the study results suggest that a training programme positively influences radiographers' image evaluation performance, especially sensitivity (Table 2.10). The authors of seven studies attributed the improved radiographers' performance to their educational interventions.

Table 2.10.

Results of image evaluation studies with educational interventions.

Studies	Type	Training	Sensitivity (%)	Sensitivity (+/-)	Specificity (%)	Specificity (+/-)
Hardy & Culpan (2007)	Red-dot	Pre	72.10		50.10	
		Post	88.50	+16.40	53.40	+3.30
Hargreaves & Mackay (2003)	Red-dot	Pre	76.20		96.40	
		Post	81.30	+5.10	96.10	-0.30
Hazel, Motto & Chipeya (2015)	Red-dot	Pre	83.72		59.69	
		Post	87.31	+3.59	70.29	+10.60
Mackey (2006)	Red-dot	Pre	78.90		76.90	
		Post	88.20	+9.30	76.90	0.00
		6 months	76.50	-2.40	79.90	+3.00
McConnell & Webster (2000)	Red-dot	Pre	91.66		65.00	
		Post	100.00	+8.34	53.33	-11.67
		6 - 8 weeks	95.83	+4.17	75.00	+10.00
Hardy & Culpan (2007)	PCE	Pre	47.80		50.70	
		Post	74.40	+26.6	51.40	+0.70
Loughran (1994)	PCE	Pre	81.10		94.40	
		Post	95.90	+14.8	96.60	+2.22
McConnell et al. (2012)	PCE	Pre	87.30		78.90	

		Post	90.80	+3.50	76.00	-2.90
		8 - 10 weeks	93.50	+6.20	82.90	+4.00
Piper & Paterson (2009)	PCE	Pre	60.32		73.15	
		Post	68.85	+8.53	82.87	+9.72

Chapter 2.6.1 pointed out the absence of control groups in the educational intervention studies. When evaluating the impact of education on image evaluation performance, randomised control group pretest posttest design improves internal validity (Brealey, Scally & Thomas, 2002a) because the use of control and experimental groups avoids unfounded interpretation of research results (Marsden & Torgerson, 2012). The presence of a control group allows determination as to whether improved performance can be attributed to educational interventions rather than other known or unknown variables. Despite this methodological concern, randomisation of control and experimental groups may not be a practicable option (Harris et al., 2006) since image evaluation studies typically depend on small groups of self-selected volunteers. All the studies that evaluated the educational impacts on image evaluation performance used a single group pretest posttest design. None of the studies discussed or acknowledged that the use of single group design could potentially weaken the internal validity of their study results.

Five studies used the same image bank for pre-tests and post-tests. Two of the five studies acknowledged that recall bias (i.e. decision making in a post-test is influenced by the previous exposure to images in the pre-test) is a potential limitation of studies with pretest posttest design (McConnell & Webster, 2000; Piper & Paterson, 2009). A study which specifically investigated recall bias in mammogram evaluation found that recall bias was unlikely to affect studies especially when the same images were presented with other similar images (Hardesty et al., 2005). Notwithstanding this, the potential effect of recall bias is still poorly researched (Boone, Halligan, Mallett, Taylor & Altman, 2012), and care must be taken to interpret results of studies using the same radiographs for a pre- and post-test.

Two Red-dot studies included an additional third test to determine the effect of education for a longer time span (Mackay, 2006; McConnell & Webster, 2000). The radiographers' sensitivity in both studies showed similar outcomes. Their sensitivity increased for the second assessment but slightly decreased for the third assessment. On the other hand, the studies found inconsistent results for specificity. Mackay (2006)'s study showed very little fluctuation of specificity throughout the study period. McConnell and Webster (2000)'s study demonstrated that specificity decreased (-11.67%) for the second test but increased (+10.00%) for the third test compared to the specificity of the first test (65.00%). This 11.67% of reduction in specificity is particularly noticeable when compared with other studies. This decreased specificity may have been caused by a sudden impulse to look for abnormal appearances rather than normal (Mackay, 2006), thus resulting in 100% sensitivity with an increased rate of over-calling (false positives) in the second assessment. There were two studies that found a reduction in specificity for the post-tests, but the changes were negligible (Hargreaves & Mackay, 2003; McConnell et al., 2012). Interestingly, the results from McConnell and Webster (2000) and Mackay (2006) showed that the improved sensitivity deteriorated over time. Mackay (2006) hypothesised that the effect of a training programme may be short lived. Two longitudinal reporting studies (Carter & Manning, 1999; Kumar, 2007) found that the postgraduate students improved and retained their image interpretation skills while exposed to training programmes. These results may indicate that educational interventions positively influence image evaluation skills but the learnt skills could gradually deteriorate once the learning discontinues. Skill fade is a known phenomenon in the health sector (General Medical Council, 2014). McConnell and Webster (2000) and Mackay (2006) did not investigate if the participating radiographers regularly used their Red-dot skills and engaged in continuous learning between the second and third assessment. It is therefore possible to hypothesise that gradual skill fade occurs after a training programme if radiographers do not actively engage in image evaluation practice and learning.

An independent-samples t-test for sensitivity and specificity of reviewed pretest posttest studies found a statistically significant difference between the degrees of improvement for sensitivity (M = 8.68, SD = 7.59) and specificity (M = 2.39, SD = 6.21) conditions; t(22) = 2.22, p =.037. The mean difference of 6.29% indicates that education exerts greater improvement on radiographers' sensitivity than specificity. The reasons for this finding are not deducible from the limited research information of the reviewed studies. However, possible explanations for this might be that 1) the radiographers after an educational intervention were driven by a sudden instinct to look for abnormalities, which resulted in increased sensitivity and decreased specificity (Mackay, 2006), 2) Radiographers' specificity is generally lower than sensitivity regardless of educational interventions, and/or 3) insufficient emphasis was made on evaluating normal images in the training programmes.

Importantly, education should not only improve radiographers' image evaluation performance but also lead them to arrive at reliable clinical decisions. Mackay (2006) and

Hazel et al. (2015) acknowledged that their participating radiographers did not reach an ideal performance standard after the training programmes. Only one study (Loughran, 1994) showed that the radiographers' mean sensitivity and specificity exceeded 90% after the training programme. In other studies, the radiographers typically performed between 80 to 90% sensitivity and below 80% specificity (Table 2.12). Despite the differing degrees of improvement, the authors attributed the improved radiographers' performance to the educational intervention. However, Chapter 2.6.1 emphasised that the research has not scientifically highlighted the link between education and evaluation skills owing to the lack of control groups. Further studies are needed to clarify the link.

2.7.4. Research question of this study

In the UK, there appears to be a declining interest in Red-dot studies. In 2006, the SCoR conveyed its expectation that provision of written clinical comments on the examinations that radiographers conduct would become a core competence of the profession. The last Red-dot study that primarily aimed to measure radiographers' image evaluation performance in the UK was conducted in 2006. The SCoR's aspiration for image evaluation by radiographers was reiterated more explicitly in 2013 that the SCoR expected the Red-dot systems to phase out and be replaced by PCE. This series of announcements by the SCoR may explain the declining research interest in RADS in the UK.

Research should be conducted in accordance with the changes proposed by the SCoR so that research evidence could support undergraduate programmes and implementation of clinical commenting by unselected radiographers (Brealey et al., 2006). If necessary, the Red-dot system and PCE by radiographers should be regularly audited to maintain and improve the standards. However, these circumstances suggest that PCE studies have a high priority in the future research agenda. Implementation of PCE must be supported by evidence, training and audit (Hardy & Culpan, 2007). The SCoR (2013) aspires that PCE becomes a core competence of radiographers. The SCoR also considers that newly qualified radiographers are now equipped with necessary education and training to start participating in PCE. However, the literature review found that the feasibility of PCE by radiographers had not been vigorously explored in the UK, especially since the SCoR's announcement in 2013. Moreover, despite the SCoR's expectation, there was no evidence to describe newly qualified radiographers' competencies in PCE. Therefore, research was conducted with a research question: "What is the image evaluation performance of diagnostic radiography graduates relative to benchmarking standards?".

2.7.5. Limitations of the literature review

The target population of the review question was loosely defined as "diagnostic radiographers". The literature search did not specify characteristics of sample populations such as basic demographics, educational background, years of experience and areas of clinical interest that potentially influence radiographers' evaluation skills. The participants of many studies were self-selected volunteers with an interest in image evaluation. Results from some studies and subsequent analysis in the literature review potentially overestimated the radiographers' performance. The literature review was conducted with a primary interest in radiographers' evaluation skills for skeletal radiographs. In the study of

Renwick et al. (1991), the radiographers demonstrated lower sensitivity and specificity for the axial skeleton than the appendicular skeleton. Piper and Paterson (2009) also hypothesised that inclusion of images of the axial skeleton in the research design may result in lower specificity. Image evaluation studies that solely used images of the appendicular skeleton may have overestimated the radiographers' performance. However, the impacts of including or excluding images of the axial skeleton have not been extensively explored, and therefore the discussion of the reviewed studies did not consider the difference between the appendicular and axial skeletons. There were three early studies that included skeletal systems as well as chest and abdomen. This chapter independently assessed the research results for skeletal images in two studies although same assessment was not possible in one study.

The reviewed articles were published between 1991 to 2017. There have been substantial changes in radiography education and A&E departments since the introduction of the Red-dot system in 1985 (Mackay, 2006: Wright & Reeves, 2016). These changes may have possibly altered the radiographers' attitude, knowledge and skills for image evaluation but this was not considered in evidence synthesis. Similarly, the review did not consider potential regional differences. This chapter reviewed articles published in the UK, Australia and South Africa. Although possible differences of education and healthcare systems are expected among three countries, possible regional variations were not reflected.

2.8. Summary

17 studies were reviewed in this chapter. QUADAS-2 was used to assess the quality of each study prior to conducting the literature review. The assessment found several methodological concerns. A small sample size appeared a generic limitation of the reviewed studies, which posed a question about the generalisability of the results. There was another concern for the establishment of the gold standards. More than half of the reviewed studies did not use appropriate gold standard or provide information to determine the risk of bias.

Chapter 1 and 2 pointed out that there is now little doubt in the clinical contribution to reporting service by qualified reporting radiographers. Regardless of the basis of the image evaluation practice (The red-dot system or PCE), many of the authors in the reviewed studies agreed that radiographers with appropriate training and education have the potential to accurately evaluate plain skeletal radiographs. However, the review also found that, assuming that 90% sensitivity and specificity are ideal performance standards for the Red-dot system and PCE, many radiographers still require further training and education to achieve and maintain this hypothetical standard.

Several research authors extended their research interests beyond radiographers' image evaluation competencies and explored the impacts of educational interventions on their skills. Although the absence of control groups in these studies was a methodological limitation, the study results suggested that training programmes and educational interventions can positively influence radiographers' ability in image evaluation. However, research is still necessary to investigate the sustainability of the learnt skills. The next chapter introduces the research method of this study.

Chapter 3. Research method

3.1. Introduction

Professionals must contribute to the body of knowledge for themselves to progress (Malamateniou, 2008). Historically, radiographers' roles in research were non-existent until 1990s owing to the scant attention and obligation to research in the profession. Radiography was perceived as a research area exclusively for medical practitioners and medical physicists (Challen, Kaminski & Harris, 1996; Nixon, 2001). However, research activity by radiographers burgeoned in mid-1990s and this resulted in a significant improvement to the sense of obligation to research in the profession (Williams, 2002). Recent bibliometric evaluations also consistently found evidence for the continuous growth of the radiographic knowledge base by radiography itself (Ekpo, Hogg & McEntee, 2016; McKellar & Currie, 2015; Snaith, 2012; Snaith, 2013). X-ray image evaluation studies followed a similar research path. Effort has been devoted to determine the feasibility of image evaluation practice by radiographers since the introduction of the Red-dot system by Berman et al. in 1985.

Continuous growth of professional knowledge is essential and a deficit of research may restrain career progression of the profession and chance to improve healthcare delivery (Sim & Radloff, 2009; Snaith & Hardy, 2007). The literature review in the previous chapter found that, despite the SCoR's aspiration, the feasibility of PCE by general radiographers had not been vigorously explored in the UK; especially, there was a scarcity of research evidence to determine whether newly qualified radiographers possess enough knowledge and experience to take part in PCE without compromising the professional roles.

Therefore, this research was conducted to determine whether new radiography graduates at the point of qualification were capable of providing reliable PCE. The following sections in this chapter discuss the research method.

3.2. Philosophical underpinning to the research

Radiographers' ability to evaluate X-ray images has been quantitatively measured and expressed in forms of sensitivity, specificity and accuracy (method of calculation is discussed in the next section). The literature review of this study (Chapter 2) indicated that image evaluation studies have largely adopted this quantitative approach. "Positivism" refers to the philosophical disciplines that underpin quantitative research. The term was coined by a French 19 century sociologist, August Comte (1798–1857), who asserted that social research should aim to unveil decisive factors that govern human behaviour by collecting and analysing empirical data. Comte noted three stages that human knowledge goes through: theological, philosophical and scientific stages. Comte advocated that research must confine itself to empirical data (science) and repudiate metaphysical theories (theology and philosophy). The positivist method emphasises that absolute truth (positive knowledge) is attainable when laws and principles of phenomena are observed and measured. This notion of positivism underpins the bulk of the quantitative approach. Quantitative research focuses on a systematic and scientific investigation of phenomena and this approach often employs observations, measurements and numerical analysis (Curtis & Drenann, 2013; Maltby, 2010). Positivist researchers assert that the findings derived from their observation and measurement are the truth which mirrors reality. The researchers in image evaluation studies may appear to have predominantly applied the view of positivism in their quantitative research methods. However, they do not fully accept positivism. Despite the quantitative approach that stems from positivism, researchers can only explore performance of radiographers in particular groups and they do not aim to generalise the findings to a larger context (Chapter 2.8), thus the true performance of radiogaphers in the radiography population remain unknown. Positivism still remains as a core research philosophy to provide epidemiological information to the public health and healthcare service providers (WHO, 2016), although unconditional acceptance of positivism to image evaluation research is impractical.

Contrast to positivism, "interpretivism" is an epistemological belief that researchers' different values and beliefs determine the truth (Ryan, 2018). Qualitative studies commonly adopt interpretivism as it allows deeper understanding of phenomena in social context. For example, an interactive interview allows researchers to probe unobservable phenomena in quantitative research which promote further and prompt investigation of interviewees. One limitation of interpretivism is that it only attempts to understand phenomena in complex contexts rather than generalising the findings to wider social situations (Pham, 2018). Interpretivism adds little methodological advantage when researchers aim to quantitatively determine radiographers' competencies in image evaluation. However, these quantitative researchers could sway toward an interpretivist position when concluding their research. For example, in some PCE studies, the authors suggested the radiographers' potential to take part in image evaluation practice despite the noticeably low performance of the radiogaphers (Coleman & Piper, 2009; Hardy & Culpan, 2007; Piper & Paterson, 2009) (Table 2.8). There seems no rational link between their study findings and conclusions. In studies with educational interventions (Chapter 2.7.3), what defines “improvement” principally depends on researchers’ subjective point of view (e.g., can 1% increase in sensitivity after a training programme be considered “improvement”?). These exemplify another limitation of interpretivism that the subjectivity of researchers, such as personal belief and cultural preferences, may result in biased conclusions (Pham, 2018). Application of interpretivism to studies measuring radiographers’ performance in image evaluation is therefore inappropriate (or impossible), although it could provide additional depth to a study design (e.g., interview radiographers about their perception of image evaluation practice).

“Post-positivism” emerged as an alternative epistemological approach that can be argued to alleviate the limitations of positivism and interpretivism (Panhwar, Ansari and Shah, 2017). A post-positivist approach views that positivism, with strong reliance on empiricism and eradication of subjectivity, does not lead to the attainment of the truth. This does not indicate that post-positivism rejects the scientific and quantitative values of positivism. Post-positivists still strive to scientifically explore various phenomena. However, post-positivists acknowledge that researchers’ common humanity (such as belief, passion and politics in research) inevitably influences research results, and therefore the absolute truth is unattainable. Post-positivistic approach also encourages the triangulation of quantitative and qualitative methods to allow various investigations in many researchable fields and formulation of new knowledge (Ryan, 2006). In the context of image evaluation research (the Red-dot system and PCE), post-positivism commonly underpins the research philosophy. Radiographers’ performance must be measured quantitatively, but decisions as to whether the radiographers are clinically competent in the Red-dot system or PCE often depend on researchers' subjective beliefs, due to the lack of agreed performance standard (Chapter 2.7.2). Hazel et al. (2015) conducted qualitative analysis of PCE comment quality before and after an educational intervention along with quantitative measurement of image evaluation performance of radiographers (Chapter 2.6.1), although triangulation methods are rarely used in this research area. Nevertheless, post-positivism is an appropriate research philosophy for measuring radiographers' performance in image evaluation. Post-positivist researchers must devote an effort to maintain a good balance of authority and flexibility, and avoid dogmatic attitudes in the research design.

3.3. Quantitative approach to measure X-ray image evaluation ability

The previous section explained that a quantitative approach is traditionally adopted when attempting to understand how accurately image observers discriminate between presence and absence of abnormalities on radiographs. This is because observers' discriminative ability is measured, quantified and expressed in sensitivity and specificity as percentages. Accuracy is also frequently used in many image evaluation studies. Sensitivity and specificity are inherent properties of an image assessment and are not influenced by sample population (number of X-ray images) or prevalence of abnormalities. The first step in the calculation of sensitivity and specificity is to create a 2 x 2 contingency table with true conditions according to a gold standard in columns and image observers' binary decisions in rows (Table 3.1).

Table 3.1.

A 2 x 2 Table for the estimate of image observers' discriminative ability.

		Gold standard	
		Abnormal	**Normal**
Observer's decision	**Abnormal**	True positive	False positive
	Normal	False negative	True negative

As briefly discussed in Chapter 2.5, sensitivity of an image observer represents the ability to correctly identify abnormalities on radiographs. Sensitivity is the proportion of correct decisions (true positives) with a sum of decisions made for abnormal images (true positives + false negatives). Specificity of an image observer represents the ability to correctly identify normal appearances on radiographs. Specificity is the proportion of correct decisions (true negatives) with a sum of decisions made for normal images (true negatives + false positives). Accuracy is additionally computed in many image evaluation studies. Accuracy is a measure that incorporates sensitivity and specificity into a single index. Accuracy of an image observer represents the ability to correctly identify both normal and abnormal appearances on radiographs. Accuracy is therefore the proportion of correct decisions (true positives + true negatives) with a sum of decisions made for all images (true positives + true negatives + false positives + false negatives). Unlike sensitivity and specificity, accuracy is not an inherent property of an image evaluation assessment and is affected by prevalence of abnormalities (Ackobeng, 2006; Anvari, Halpern & Samir, 2015; Parikh, Mathai, Parikh, Sekhar & Thomas, 2017; Šimundić, 2008; Stojanović et al., 2014; Wong & Lim, 2011). Figure 3.1 summarises the formulae to compute sensitivity, specificity and accuracy.

Figure 3.1. Formulae to compute sensitivity, specificity and accuracy.

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives} = \frac{True\ positives}{Total\ number\ of\ radiographs\ with\ abnormal\ appearance}$$

$$= Probablity\ that\ an\ observer\ correctly\ identifies\ abnormal\ appearance$$

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} = \frac{True\ negatives}{Total\ number\ of\ radiographs\ with\ normal\ appearance}$$

$$= Probablity\ that\ an\ observer\ correctly\ identifies\ normal\ appearance$$

$$Accuracy = \frac{True\ positives + True\ negative}{True\ positives + True\ negatives + false\ positives + false\ negatives}$$

$$= \frac{True\ negatives + True\ negative}{Total\ number\ of\ radiographs\ with\ normal\ and\ abnormal\ appearance}$$

$$= \text{Probablity}\ that\ an\ observer\ correctly\ identifies\ normal\ and\ abnormal\ appearance$$

Radiographic findings are often subtle or complex. It is sometimes impractical to make clinical decisions by using dichotomous scales (normal/abnormal or negative/positive) while evaluating radiographs. In some image evaluation studies, ordinal rating scales are used to determine sensitivity and specificity. A typical example is seen when image observers are asked to evaluate radiographs and report their findings by using the Likert scale. For example: 1) definitely normal, 2) probably normal, 3) possibly abnormal, 4) probably abnormal, or 5) definitely abnormal (Lasko, Bhagwat, Zou & Ohno-Machado, 2005). A cut-off point or threshold is set to distinguish decisions for negative or positive findings, for instance, a threshold is set between “2) probably normal” and “3) possibly abnormal” to distinguish decisions for normal and abnormal findings (Figure 3.2).

Subsequently, sensitivity, specificity and accuracy are calculated by using the 2 x 2 table (Table 3.1) and the formulae (Figure 3.1).

Figure 3.2. A threshold to distinguish decisions for normal and abnormal findings.

$$\underbrace{\text{1) definitely normal, 2) probably normal,}}_{\text{Normal}}\ \underbrace{\text{3) possibly abnormal, 4) probably abnormal, 5) definitely abnormal}}_{\text{Abnormal}}$$

3.4. Exploration of PCE errors and classification

Radiology has devoted a continuous effort to understand types and etiology of diagnostic errors (Berlin & Hendrix, 1998; Bruno, Walker & Abujudeh, 2015). Several error classification systems with a wide spectrum of objectives have been established by many groups of researchers in radiology (Brook et al., 2010; Graber, Franklin & Gordon, 2005; Kim & Manfield, 2014; Renfrew, Franken, Berbaum, Weigelt & Abu-Yousef, 1992; Pinto & Brunese, 2010; Provenzale & Kranz, 2011; Taylor, Voss, Melvin & Graham, 2011). The goal of radiological error classification is to prevent errors. Prevention of errors improves the quality of patient care, healthcare service efficiency and professional satisfaction (Mankad, Hoey, Jones & Smith, 2009; Pinto et al., 2012).

Similar benefits are expected in PCE. Identifying and classifying PCE comment errors will set out practical strategies to portray radiographers' PCE commenting behaviour and boost their PCE performance. Identifying PCE comment errors and their sources aims to prevent radiographers from making the same mistakes again. Brook et al. (2010) argued that prevention of errors requires a gold standard, which can detect, classify and manage errors. Despite this, the current literature suggests there is no gold standard specifically developed to identify, classify and manage image evaluation errors in PCEs. The radiological error classification schemes are not adaptable for PCE error classification because they involve some error categories that are inapplicable to PCE. They encompass a wide range of radiologists' duties and associated latent conditions, some of which are irrelevant to, or beyond, the objectives PCE. Therefore, this study developed a PCE error classification that aims to provide a mechanism to detect, identify and manage errors. This classification system was made according to evaluation ratings for human error identification tools (Shorrock & Kirwan, 2002). Prior to classifying errors, all theoretically possible PCE outcomes were first considered and organised in a PCE taxonomy. Errors were then systematically classified by using this taxonomy.

Only four types of decision outcomes are possible in the Red-dot system (Table 3.1). PCE, on the other hand, needs to deal with more complex outcomes because of written articulation of findings. For example, some of the reviewed PCE studies have recognised partially correct comments, indicating that there are three types of comment outcomes: "correct" or "partially correct" or "incorrect" (Coleman & Piper, 2009; Piper, Piper & Paterson, 2009; Paterson & Godfrey, 2005). Wallis and McCoubrie (2011) maintains that radiological reports require accuracy without hiding behind ambiguous terms. The SCoR (2013) also requires that decisions made in PCE must be communicated in unambiguous written forms. Therefore, ambiguous comments must be considered as a potential type of outcome as this is a known issue in radiology (Berlin, 2000). The taxonomy additionally included a comment outcome, “complex”, in order to deal with outcomes that may be impractical to classify (e.g., a long and grammatically complex comment that produce multiple outcomes).

Most of the reviewed PCE studies in the previous chapter adopted a hybrid system in which participants provided their clinical decision (using the Red-dot system or Likert scale) followed by comments. As a result, four outcomes of the Red-dot decisions (Table 3.1) and nine outcomes of comments are expected (Table 3.2), resulting in a total of 33 combined outcomes of PCE: 10 patterns for normal images and 23 for abnormal images. A PCE taxonomy was established by assembling these possible PCE outcomes (Table 3.3). This PCE taxonomy incorporated 33 theoretically attainable PCE outcomes and served as the reference to systematically classify PCE comments obtained from the X-ray image evaluation test. Table 3.4 summarises the definitions of PCE error sources used in the taxonomy.

Table 3.2.

Two-by-five contingency table showing eight possible outcomes a PCE comment.

Comment	Normal images	Abnormal images
Correct	Absence of abnormality is clearly stated (TN).	Presence of identifiable abnormality is clearly stated (TP).
Partly correct	-	Presence of at least one identifiable abnormality is clearly stated but • Another identifiable abnormality is missed (FN). • Another identifiable abnormality is appreciated but dismissed as normal (FN). • Normal anatomical structure is described as abnormal (FP). • Combination of three criteria described above.
Incorrect	Normal anatomical structure is described as abnormal (FP).	Decision is incorrect because: • Identifiable abnormality is missed (FN) • Identifiable abnormality is appreciated but dismissed as normal (FN) • Abnormality is missed or dismissed as normal and Normal anatomical structure is described as abnormal (FP).
Ambiguous (or inconclusive)	Hedge words are used and presence or absence of abnormality is not clearly stated.	Hedge words are used or only indirect sign of injury (e.g. soft tissue swelling or raised fat pad) is commented and presence or absence of abnormality is not clearly stated.

Complex	Comment produces more than one outcome as described above (e.g., Incorrect + ambiguous).	Comment produces more than one outcome as described above (e.g., partly correct + ambiguous).

Table 3.3.

PCE taxonomy: theoretically attainable PCE outcomes.

Report	Decision	Comment	Outcome		Comment examples	Comment error sources
			Decision	**Comment**		
Normal	Normal	Present	**Correct**: Normal image is classified as normal (TN).	**Correct**: Absence of abnormality is clearly stated (TN).	“No abnormality seen.”	(No error)
			Correct: Normal image is classified as normal (TN).	**Incorrect**: Normal anatomical structure is described as abnormal (FP).	“A fracture on the radial styloid process.”*	Discrepancy
			Correct: Normal image is classified as normal (TN).	**Ambiguous**: Hedge words are used and presence or absence of abnormality is not clearly stated.	“A swelling on the wrist suggesting a fracture but may be normal.”	Ambiguity
			Correct: Normal image is classified as normal (TN).	**Unclassifiable**: Comment produces more than one possible outcome.		
		Absent	**Correct**: Normal image is classified as normal (TN).	**Correct** (assumed): There is no comment to clearly state the absence of abnormality (TN is assumed: absence of comment is understood as “normal”).	Comment is absent.	(No error)
	Abnormal	Present	**Incorrect**: Normal image is classified as abnormal (FP).	**Correct**: Absence of abnormality is clearly stated (TN).	“No abnormality seen.”*	Discrepancy

			Incorrect: Normal image is classified as abnormal (FP).	**Incorrect**: Normal anatomical structure is described as abnormal (FP).	"A fracture on the radial styloid process."	Over-calling
			Incorrect: Normal image is classified as abnormal (FP).	**Ambiguous:** Hedge words are used or only indirect sign of injury (e.g. soft tissue swelling or raised fat pad) is commented and presence or absence of abnormality is not clearly stated.	"Although there is no visually detectable abnormality, a slight swelling on the wrist may suggest a fracture."	Ambiguity
			Incorrect: Normal image is classified as abnormal (FP).	**Unclassifiable:** Comment produces more than one possible outcome.	"A fracture on the ulnar styloid process. May be another fracture on the distal radius" (Over-calling + ambiguous).	
		Absent	**Incorrect**: Normal image is classified as abnormal (FP).	**Incorrect** (assumed): There is no comment to indicate the presence of abnormality (FP is assumed).	Comment is absent.	No comment
Abnormal	Normal	Present	**Incorrect**: Abnormal image is classified as normal (FN).	**Correct**: Presence of identifiable abnormality is clearly stated (TP).	"An impacted transverse fracture on the distal radius and a fracture on the ulnar styloid process."	Discrepancy
			Incorrect: Abnormal image is classified as normal (FN).	**Partly correct**: Presence of at least one identifiable abnormality is clearly stated (TP) but		
				• Another identifiable abnormality is missed (FN).	"An impacted transverse fracture on the distal radius."	Perceptual error or SOS

	• Another identifiable abnormality is appreciated but dismissed as normal (FN).	“An impacted transverse fracture on the distal radius. There is also a subtle cortical irregularity on the ulnar styloid process but this is normal.”*	Under-calling
	• Normal anatomical structure is described as abnormal (FP)	“An impacted transverse fracture on the distal radius and scaphoid.”*	Over-calling
	• Combination of three criteria described above.	“An impacted transverse fracture on the distal radius and scaphoid. There is also a subtle cortical irregularity on the ulnar styloid process but this is normal.”*	
Incorrect: Abnormal image is classified as normal (FN).	**Incorrect**:		
	• Identifiable abnormality is missed (FN)	“No abnormality seen.”	Perceptual error or SOS
	• Identifiable abnormality is appreciated but dismissed as normal (FN)	“There are cortical irregularities on the distal radius and styloid process but these are normal anatomical variants.”*	Under-calling
	• Abnormality is missed or dismissed as normal and normal anatomical structure is described as abnormal (FP).	“A scaphoid fracture.”*	Perceptual error/SOS or under-calling and over-calling
Incorrect: Abnormal image is classified as normal (FN).	**Ambiguous**: Hedge words are used and presence or absence of abnormality is not clearly stated.	“A swelling on the wrist suggesting a fracture but may be normal.”	Ambiguity

		Incorrect: Abnormal image is classified as normal (FN).	**Unclassifiable:** Comment produces more than one possible outcome.	"A scaphoid fracture. There is also a lucent line through the scaphoid suggesting a fracture, but may be normal. " (over-calling + ambiguous)	
	Absent	**Incorrect**: Abnormal image is classified as normal (FN).	**Incorrect**: Comment is absent (FN is assumed).	Comment is absent.	No comment
Abnormal	Present	**Correct**: Abnormal image is classified as abnormal (TP).	**Correct**: Presence of identifiable abnormality is clearly stated (TP).	"An impacted transverse fracture on the distal radius and a fracture on the ulnar styloid process."	(No error)
		Correct: Abnormal image is classified as abnormal (TP).	**Partly correct**: Presence of at least one identifiable abnormality is clearly stated (TP) but		
			• Another identifiable abnormality is missed (FN).	"An impacted transverse fracture on the distal radius."	Perceptual error or SOS
			• Another identifiable abnormality is appreciated but dismissed as normal (FN).	"An impacted transverse fracture on the distal radius. There is also a subtle cortical irregularity on the ulnar styloid process but this is normal."*	Under-calling
			• Normal anatomical structure is described as abnormal (FP).	"An impacted transverse fracture on the distal radius and scaphoid."*	Over-calling
			• Combination of three criteria described above.	"An impacted transverse fracture on the distal radius and scaphoid. There is also a subtle cortical irregularity on the ulnar styloid process but this is normal."*	

	Correct: Abnormal image is classified as abnormal (TP).	**Incorrect**:		
		• Identifiable abnormality is missed (FN)	“No abnormality seen.”*	Discrepancy
		• Identifiable abnormality is appreciated but dismissed as normal (FN)	“There are cortical irregularities on the distal radius and styloid process but these are normal anatomical variants.”*	Under-calling
		• Abnormality is missed or dismissed as normal and Normal anatomical structure is described as abnormal (FP).	“A scaphoid fracture.”	Perceptual error or under-calling and over-calling
	Correct: Abnormal image is classified as abnormal (TP).	**Ambiguous**: Hedge words are used or only indirect sign of injury (e.g. soft tissue swelling or raised fat pad) is commented on and presence or absence of abnormality is not clearly stated.	“Although there is no visually detectable abnormality, a slight swelling on the wrist may suggest a fracture.”	Ambiguity
	Correct: Abnormal image is classified as abnormal (TP).	**Incorrect**: Comment is absent (FN is assumed).	Comment is absent.	No comment
	Correct: Abnormal image is classified as abnormal (TP).	**Unclassifiable:** Comment produces more than one possible outcome.	“May be an impacted transverse fracture on the distal radius?” (ambiguous + under-calling)	
Absent	**Correct**: Abnormal image is classified as abnormal (TP).	**Incorrect**: Comment is absent (FN is assumed).	Comment is absent.	No comment

Table 3.4.

Possible error sources in PCE

PCE error sources	Definition
Perceptual error (scanning error)	Failure to detect abnormality, caused by failing to note the absence of normal finding or failing to note the presence of abnormal finding.
Over-calling (evaluation error)	Normal anatomy is described as abnormal.
Under-calling (decision-making error/faulty reasoning)	Abnormality is missed or appreciated but dismissed as normal.
Satisfaction of search (SOS)	Premature termination of search after detecting at least one abnormality.
Ambiguity	Hedge words are used or only indirect sign of injury (e.g. soft tissue swelling or raised fat pad) is commented. Presence or absence of abnormality is not clearly stated.
Discrepancy error	Disagreement between diagnostic decision (abnormality is present or absent) and comment.
No comment error	Image is classified as abnormal but there is no comment to describe the abnormality

3.5. Data collection

3.5.1. Sampling frame

There were 24 HEIs that provided pre-registration undergraduate diagnostic radiography education in the UK in the 2014/2015 academic year. Due to the differences of educational structures and geographic factors, three HEIs from Scotland and one HEI from

Northern Ireland were first excluded from the sampling frame. 20 HEIs in England and Wales were considered to be the potential research sites. Course leaders of the institutions were first contacted regarding the number of the final year students in order to estimate the sampling frame. This estimated the output of radiography graduates in England and Wales in 2015 was 946.

3.5.2. Sample size estimation and sampling method

Sample size estimation determines how many participants are needed in a study. The reviewed studies in Chapter 2 seldom reported sample size estimation. These studies probably did not calculate necessary sample size at all since the studies depended on a convenience sampling method that resulted in involving small groups of self-selected radiographers (Chapter 2.7.1). However, a sufficiently large sample size with a proper statistical power is one of the key elements to obtain valid research results which have the potential to be extrapolated to the target population (e.g., diagnostic radiographers in the UK) (Nayak, 2010; Youssef, 2011). Therefore, this study performed a calculation of necessary sample size using the estimated number of potential participants (n = 946). Figure 3.3 presents the formulae for the calculation.

Figure 3.3. Cochran's sampling formula (1963) (for infinite population) and a modified formula (for small population (Israel, 1996).

$$SS = \frac{Z^2 \times p \times (1 - p)}{C^2}$$

$$New\ SS = \frac{SS}{1 + \frac{1 - SS}{Pop}}$$

Where,
SS = Sample size (for infinite population)
Z = Z Value = 1.96 and 2.576 for 95% and 99% confidence levels respectively
P = Percentage of population = 0.5
C = Confidence interval = 0.05
New SS = New sample size (modified for finite or small population)
Pop = population = 946 (estimated diagnostic radiography graduates in England and Wales in 2015)

The calculation yielded that the numbers of participants required were 274 and 391 for 95% and 99% confidence level, respectively. The course leaders of the 20 HEIs were asked for participation and nine agreed. Additionally, one HEI agreed to participate in a pilot study and was therefore excluded from the potential sample population. The participating HEIs held a total of 443 final year (third year) students. Each HEI was provided with all necessary documentation to facilitate the ethical approval process within the HEI. The documentation included the participant information sheet, informed consent form and the SHUREC1 letter of approval (Appendix C, D and E) as the evidence that the research proposal had been registered with Sheffield Hallam University (discussed in Chapter 3.8). One HEI required ethical approval and indemnity cover from its own university and these were obtained accordingly. The final year students of the agreeing HEIs were then recruited via the course leaders. The information sheet and the informed consent form in an electronic form were also distributed to the students by their course leaders.

3.5.3. Inclusion and exclusion criteria

This research included the final year diagnostic radiography students from the collaborating HEIs. Students who did not wish to take part in this research (or wished to withdraw from this research) were excluded. The students who wished to participate but were unable to attend the test were also excluded as no provision could be made for this.

3.5.4. Timing of data collection

Since this research aimed to benchmark the competencies of radiography students at the point of graduation/qualification, the data collection was conducted shortly before the end of the academic year (between April and June in 2015).

3.5.5. X-ray Image bank

The literature review presented that the use of an X-ray image bank is a standard method to measure evaluative competence of image observers. X-ray image evaluation studies with image banks typically use a mixture of normal and abnormal images encompassing a spectrum of pathologies and body areas (Brealey et al., 2002a). In these studies, improved reproducibility of the results is expected because measurement of evaluative performance is conducted under controlled conditions and performance of observers would be unlikely to differ greatly. Moreover, the chance of introducing observer review bias and reference standard review bias is absent. This is because the gold standards are established prior to image evaluation tests and observers evaluate images without the knowledge of the gold standards (Brealey, Scally & Thomas, 2002b).

One criticism of the use of image banks is that prevalence of abnormality influences observers' performance. The prevalence of musculoskeletal trauma is estimated around 20 to 30% (Hardy, Snaith & Scally, 2013; Hardy, Spencer & Snaith, 2008; McConnell et al, 2013; Renwick et al., 1991; Robinson, Culpan & Wiggins, 1999) in A&E settings. Evidence suggests that higher prevalence of abnormality in image banks results in increased sensitivity with concomitantly decreased specificity (Pusic et al., 2012). It has been questioned that image evaluation performance measured by high prevalence (around 70%) image test banks may not truly reflect image interpreters' ability in emergency settings (Hardy et al., 2016). Hardy et al. (2016) argued that an image bank that represents local clinical practice would measure evaluation performance more accurately. However, this study did not consider the development of an X-ray image bank that reflect typical daily image profile at A&E settings because of possible variations in image profile in different hospitals. For example, regional and seasonal characteristics may alter daily image profile across the UK. An image bank that reflects local clinical work load in a single hospital, as seen in Hardy et al. (2016) and McConnell and Baird (2017), may not reflect image profiles of other hospitals. Since the present study targeted multiple HEIs across England and Wales, development of an image bank that reasonably reflect daily image profile (e.g., abnormality prevalence, anatomical areas, ratios of adult/paediatric images) at all study sites was impractical.

It was expected that all students had experience of both adult and paediatric cases throughout their education and clinical training. Bony development and the identification of normal variants is a key element of learning. The test bank was therefore constructed to include some non-adult cases and present some normal variants. From a statistical perspective, in order provide equal prediction of sensitivity and specificity, disease prevalence for testing purposes should be 50% (Piper et al., 2004). This is consistent with the rapid reporting test used to assess radiologists (RCR, 2017b) which consists of 30 cases of good image quality and a definitive answer evidenced by blind double reporting. The X-ray image bank consisted of 30 X-ray images of the appendicular skeleton, which is similar to the structure used in the rapid reporting section (3b) of the Final Examination for the Fellowship in Clinical Radiology (Part B) developed by the Fellow and the Royal College of Radiologists (FRCR) (RCR, 2017b). The survey results of Snaith and Hardy (2009) suggested that all HEIs in the UK deliver image evaluation education for the appendicular skeleton at the undergraduate level. However, the results also found that not all HEIs provide education for the axial skeleton at both undergraduate and postgraduate levels. Therefore, this study focused on the X-ray images of the appendicular skeleton.

A rubric was developed to ensure a full range of appendicular anatomy was included in the image bank (Table 3.5). The images were selected from the RadBench (http://www.radbench.org/[3]) image bank. All images were blind double reported by radiologists. This was considered as the gold standard for the X-ray image evaluation test. In clinical practice, image quality can be variable and this might affect decision-making process. However, only good quality images were selected since test results with unknown effects of poor-quality images would not demonstrate true sensitivity and specificity of the participants. A test bank with reasonably good-quality images were expected to give a fairer reflection of the students' sensitivity and specificity before entering the preceptorship. Very

[3] At the time of writing (September 2018) the website was under construction.

obvious fractures were excluded because this was “a test” for the final year students at the point of graduation who should be able to identify a fracture that would be obvious to a total novice To be practical, the participants in the study needed to be able to take the test at their own institution, under supervision to ensure that they would not exchange opinions. Image evaluation at undergraduate level is routinely taught in computer labs, using standard resolution monitors and jpeg images in Microsoft PowerPoint. High resolution reporting monitors and X-ray images organised in Digital Imaging and Communications in Medicine (DICOM) and delivered through a Picture Archiving and Communication System (PACS) network might offer improved visual benefits. However, these devices were not available universally throughout the participating HEI’s. The tests were therefore constructed in MS PowerPoints.

All normal images presented no fracture or other identifiable pathologies. Abnormal images presented with no pathology other than bony trauma. Each abnormal image had at least one identifiable fracture. The abnormal images were similar to those encountered in EDs. Satisfaction of search (SOS) refers to a false negative error that occurs when at least one abnormality is missed in a multiple abnormality case (Berbaum et al., 2012). SOS is a well-known error type in skeletal radiology (Berbaum et al., 2001; Berbaum et al., 2007; Berbaum et al., 2012; Berbaum et al., 2013). The test bank therefore included three multiple abnormality cases to investigate SOS in PCE. The test bank did not contain abnormal images with subtle fractures that may pose great difficulties. Each case was presented with the typical projections expected for the body part, in most cases offering two views that were displayed together on one slide. Descriptions of each image including anatomical parts, status of normal/abnormal and radiologist's report are summarised in Table 3.6. The chosen 30 images were randomised by the List Randomizer (https://www.random.org/lists/) and arranged in a descending order in Microsoft PowerPoint slides. Appendix F shows a screenshot of a sample radiograph of the image bank.

The Image bank was verified by a qualified reporting radiographer who was also a lecturer of image evaluation at the host institution. The expectation was that the difficulty of the test would be fair and appropriate for final year students although being a single delivery test due to time constraints, averaging of performance across multiple tests, thus determining the difficulty of each image in the test bank was not possible.

Table 3.5.

A rubric for X-ray image selection.

Anatomical areas	No. of normal images	No. of abnormal images
Upper limb		
Thumb & Hand	2	2
Wrist	2	2
Radius & Ulna	1	1
Elbow	2	2
Humerus	0	0
Shoulder	1	1
Lower Limb		
Toe & Foot	3	3
Ankle	3	3
Tibia & Fibula	0	0
Knee	1	1
Femur	0	0
Hip	0	0
Subtotal	15	15

Total number of X-ray images included in the test bank = 30

Table 3.6.

X-ray images in the test bank.

Question No.	Anatomical area	Status	Views	Projections	Reports
1	Hand	Abnormal	2	PA and oblique	Fracture through the growth plate of the base of the first metacarpal, extending obliquely into the ulna aspect of the proximal metaphysis. Salter II fracture with slight dorsal distraction of the metacarpal.
2	Ankle	Normal	2	AP and lateral	No fracture seen.
3	Elbow	Normal	2	AP and lateral	No bony injury.
4	Wrist	Abnormal	2	PA and lateral	Minimally displaced intra-articular chip fracture to the central articulating surface of the dorsal aspect of the distal radius.
5	Wrist	Abnormal	2	PA and lateral	Undisplaced transverse fracture through the distal radius.
6	Foot	Abnormal	2	AP and oblique	Fractures to the first, second and third distal phalanges.
7	Ankle	Abnormal	2	AP and lateral	No ankle fracture seen. However, there is a displaced fracture through the base of the 5th metatarsal.
8	Foot	Normal	2	AP and oblique	Normal. Accessory ossicle (os peroneum) noted.

9	Foot	Abnormal	2	AP and oblique	Undisplaced oblique fracture through the proximal phalanx of the little toe.
10	Ankle	Normal	2	AP and lateral	No fracture seen.
11	Foot	Abnormal	2	AP and oblique	Undisplaced transverse fracture at the base of the 5th metatarsal.
12	Wrist	Normal	2	PA and lateral	No fracture seen.
13	Radius & Ulna	Normal	2	AP and lateral	No fracture seen.
14	Wrist	Normal	2	PA and lateral	No fracture seen.
15	Radius & Ulna	Abnormal	2	AP and lateral	Subtle fracture to the lateral aspect of the Radial neck.
16	Foot	Normal	2	AP and oblique	No fracture seen.
17	Elbow	Normal	2	AP and lateral	No acute bony injury or joint effusion seen.
18	Shoulder	Abnormal	1	AP	Fracture of the left clavicle. Multiple rib fractures. Mild left pneumothorax noted.
19	Ankle	Normal	2	AP and lateral	No fracture seen.

20	Knee	Normal	2	AP and lateral (horizontal beam)	No bony injury identified.
21	Elbow	Abnormal	2	AP and lateral	Minimally displaced intra-articular fracture of the coronoid process. Small irregularity to the lateral aspect of the radial head suggesting a bony injury.
22	Hand	Abnormal	2	PA and oblique	No fracture seen to the hand. However, there is a mildly impacted transverse fracture to the distal Radius with possible intra-articular involvement. Mild dorsal angulation. Probable undisplaced fracture of the ulnar styloid too.
23	Ankle	Abnormal	2	AP and lateral	Minimally displaced oblique fracture of the lateral malleolus with no talar shift.
24	Hand	Normal	2	PA and oblique	No fracture seen.
25	Elbow	Abnormal	2	AP and lateral	Radial head fracture with raised anterior and posterior fat pads.
26	Foot	Normal	2	AP and oblique	No fracture seen. Multipartite accessory navicular.
27	Shoulder	Normal	1	AP	No fracture seen.
28	Knee	Abnormal	2	AP and lateral (horizontal beam)	Minimally displaced intra-articular fracture of the medial tibial eminence with associated lipophaemarthrosis.

29	Ankle	Abnormal	2	AP and lateral	Intra-articular crush fracture of the right Calcaneum. No significant fragments.
30	Hand	Normal	2	PA and oblique	No fracture seen.

3.5.6. Pilot study

Prior to the data collection, a pilot study was conducted to assess the data collection instruments of this research. One English HEI agreed to take part in this pilot study. Five students registered with the RadBench website and took the X-ray image evaluation test. They viewed the 30 images and provided their clinical decision by using a five-point scale (“definitely normal”, “probably normal”, “possibly abnormal”, “probably abnormal” or “definitely abnormal”). Their responses were then entered into data entry sheets (discussed in Chapter 3.4.2.) to calculate accuracy, sensitivity and specificity. The data collection instruments worked as expected and therefore no modification was made.

3.5.7. X-ray image evaluation test

The course leaders of the participating HEIs were asked to arrange a test day and time that was most suitable for their academic staff and students. The Principal Investigator (PI) of this research visited each of the participating sites to supervise the test on an agreed test day. The students were provided with a printed version of the information sheet and consent form, and given enough time to think about their participation or ask questions about the research. A short PowerPoint presentation (five minutes) about the objectives of the test and test instruction was given to the participants. The participants were then asked to fill in the registration form (Appendix G). This form gathered their demographic data, including participant’s name, HEI’s name, age, previous education, expected degree and clinical placements.

Information regarding clinical history of X-ray images and the prevalence of the trauma cases of the image bank was not given to the participant. Each PowerPoint slide with test image(s) was displayed in a single screen with a zooming function. Each participant was given an answer booklet (Appendix H). The participants were then asked to evaluate each image and provide their clinical decision by using the five-point scale and also provide PCEs (or comments) in the comment section. The participants were instructed that they should only look for and record fractures. The participants were encouraged to view the X-ray images in the "Slide show" mode of PowerPoint (full screen view) although they had the option to switch to magnified view if they wished. The PI also requested and ensured that students did not exchange their opinions or share answers during the test. There is a negative correlation between speed of evaluation and accuracy (Sokolovskaya et al., 2015). 90 seconds per image was considered sufficient for the participants (70 seconds are allocated in the rapid reporting session by the FRCR). The test therefore run for approximately 45 minutes (90 seconds x 30 images). The students were allowed to revisit any image within this time frame. Research has investigated into several viewing conditions that may or may not affect observers' image evaluation performance (Awan, Safdar, Siddiqui, Moffitt & Siegel, 2011; Chen, James Turnbull & Gale, 2015; Ferranti et al., 2017; Laffranchi et al., 2018; Moshfeghi, Shahbazian, Sajadi, Sajadi & Ansari, 2015; Ohla et al., 2018; Tewes, Rodt, Marquardt, Evangelidou, Wacker & Flack, 2013). However, the present study could not control the physical properties of the viewing conditions (e.g., resolution of the monitors) since the image evaluation tests relied on computer monitors at each collaborating HEIs (Chapter 3.5.5). Surrounding illumination in the test rooms were also inconsistent.

When the test finished, immediate feedback on each image was given to the participants so that the test became a learning opportunity. Morton (2002) explained that incentives improve participation rate in research. A certificate (as a form of incentive) with summary of test performance was sent to each student by e-mail. A sample image of the certificate is shown in Appendix I.

3.6. Interview questionnaire with course leaders

Chapter 2 indicated that X-ray image evaluation studies typically accept a quantitative approach. Chapter 3.6 also explained why quantitative research methods have been traditionally considered appropriate. However, this approach does not allow researchers to investigate why some image observers more accurately evaluate X-ray images than others or how they acquire image evaluation skills. Many X-ray image evaluation studies fail to establish theoretical links between image evaluation skills and unknown parameters (e.g., educational background, experience and areas of specialities) that could influence evaluation performance. There is now evidence to suggest that many healthcare studies have adopted both quantitative and qualitative components (a mixed research method) to introduce more methodological rigour (Tariq & Woodman, 2010). This method provides more analytic depth (Albright, Gechter & Kempe, 2013) and bolster strengths or alleviate weaknesses of quantitative and qualitative approaches (Tariq & Woodman, 2010). Therefore, an interview questionnaire regarding image evaluation education was created to add a qualitative component to the research design (Appendix J). The interview with course leaders of the participating HEIs was conducted before the image evaluation test at each collaborating site.

This Interview questionnaire was an attempt to explore how different undergraduate course structures influence image evaluation skills and also update the work of Snaith and Hardy (2008). However, the interview questionnaire could not capture sufficient data to integrate with other quantitative components in this study. Reasons for the unfinished interview questionnaire from are reflected in Chapter 6.

3.7. Data analysis

3.7.1. Demographic data

Correlations between the demographic characteristics of the participants and their performance in image evaluation (explained in the following sections) were analysed by using inferential statistics. Chapter 4 describes the participants' demographic information. Descriptive statistics portray characteristics of a sample population and highlight a meaning of the data (Marshall & Jonker, 2010). However, descriptive statistics do not allow accurate extrapolation to a wider population (Breau, 2012). On the other hand, inferential statistics uses a variety of statistical tests to allow investigation of differences, examination of relationships and extrapolation to a wider population (Allua, YEAR; Marshall & Jonker, 2011).

3.7.2. Calculation of accuracy, sensitivity and specificity

Once the answer booklets were collected, the participants' decisions expressed by the five-point scale were entered into the Microsoft Excel data entry sheet. The data entry sheet was formatted so that it automatically calculated accuracy, sensitivity, specificity of each participant and entered the values into corresponding cells. The data were then transferred into the SPSS data entry sheet for further statistical analysis.

3.7.3. Calculation of accuracy for anatomical areas

Hargreaves and Mackay (2003) found that radiographers' accuracy varied with anatomical areas. For example, the hand and lower limb were the areas where the radiographers made more evaluation errors than other areas. They therefore encouraged that education should focus on anatomical areas where errors frequently occur. Chapter This research used the X-ray image bank which comprised of 30 X-ray images of two body parts (upper and lower body) with eight anatomical areas (hand, wrist, radius & ulna, elbow, shoulder, foot, ankle and knee) (Table 3.5). Accuracy for anatomical areas were independently calculated to investigate frequencies of evaluation errors with respect to anatomical areas.

3.7.4. Confidence in decision making

Confidence is partly associated with knowledge, amount of training, and expertise (Benvenuto-Andrade et al., 2006). A lack of confidence in decision making expressed with terms of uncertainty in clinical reports could potentially results in delayed diagnosis or clinical management and misdiagnosis if the reports are hard for readers to understand (Reiner, 2013). Overconfidence, on the other hand, is one of known factors to causes diagnostic errors (Mayer et al, 2013).

Little is also known about radiographers' confidence in decision making while evaluating radiographs and the clinical consequences. The literature review found three studies that used the same Likert scale items to allow the participants to express their opinions with different levels of confidence in decision making (1: definitely normal, 2: probably normal, 3: possibly abnormal, 4: probably abnormal, or 5: definitely abnormal) (Coleman & Piper, 2009; Piper & Paterson, 2008; Wright & Reeves, 2016). Understanding of confidence in decision making may be favourable when designing training (Wright & Reeves, 2016). Extreme response style (ERS) refers to the tendency to favour ends points or extreme categories of Likert-type scales more frequently than other available response items. Extreme responders represent between 25% to 30% of responders in survey studies (Naemi, Beal & Payne, 2009). ERS is generally considered a source of bias in surveys. Despite this, ERS may be an ideal behaviour for X-ray image evaluation tests if responses are supported by high sensitivity and specificity with correct reasoning. One study found that radiographers with ERS were a minority (Wright & Reeves, 2016). However, there is still insufficient research evidence to illustrate the relationship between radiographers' confidence and outcomes of their image evaluation practice. This research therefore analysed how the participants used the Likert items that expressed different levels of confidence to provide their decisions.

3.7.5. Analysis of radiographic comments: WHAT/WHERE/HOW scoring system

The literature review in the previous chapter found that radiographers' ability to accurately articulate radiographic findings in PCE had not been thoroughly evaluated. Lancaster and Hardy (2012) argued that the lack of evidence in relation to radiographers' skills in describing abnormality was a possible barrier to the implementation of PCE. The review also found that a binary classifier (judging comments either correct or incorrect) fails to deal with partially correct or incomplete comments. However, strategies to classify partially correct PCEs were poorly documented and discussed in the reviewed studies. Although researchers have established a classification system to address this dilemma (Coleman & Piper, 2009; Piper & Paterson, 2009; Piper et al., 2005), it does not evaluate accuracy and completeness of radiographic descriptions at the same time. Therefore, a new scoring system was developed to evaluate comment quality for this research.

Physicians prefer structured reporting (Bosmans et al., 2011; Plumb, Grieve & Khan, 2009; Schawartz, Panicek, Berk, Li & Hricak, 2011). “What, Where and How” is a conceptual framework that encourages image observers in structuring PCEs of musculoskeletal trauma (Harcus & Wright, 2013). This framework takes a simplistic and systematic approach to describe three essential components of comments: type of abnormality (What), location (Where) and displacement/angulation (How) (Appendix K). The ‘WWH Scoring system’ was developed on the “What, Where and How” concept (Akimoto, Wright, Reeves & Harcus, 2016) (Table 3.7) in order to add a granular scoring approach.

Table 3.7.

What/Where/How (WWH) scoring system (Akimoto, Wright, Reeves & Harcus, 2016).

Image type	Candidate response	Mark
Normal image	– Correctly classified and described.	+3
	– Correctly classified although comment indicates the presence of abnormality (discrepancy between the final diagnosis and comment).	+0
	– Incorrectly classified (appropriate false positive).	+0
	– No answer given.	+0
Abnormal image	– Correctly classified. Marks depend if the comment:	
	fully satisfies (+1),	
	partially satisfies (+0.25/+0.5/+0.75), or	
	fails to satisfy (+0) evaluation criteria of each category below:	
	Type of abnormality (WHAT)	+1 (max)
	Location of the abnormality (WHERE)	+1 (max)
	Displacement/angulation of the abnormality (HOW)	+1 (max)
	Correctly classified although comment indicates there is no abnormality (discrepancy between the final diagnosis and comment).	+0
	– Incorrectly classified although abnormalities are correctly identified.	+0
	– Incorrectly classified (false negative).	+0
	– No answer given.	+0
	Maximum score for normal images =	45
	Maximum score for abnormal images =	45
	Total score =	90

What/Where/How (WWH) scoring system aims to examine whether PCEs precisely describe type (What), location (Where) and a degree of angulation or dislocation (How) of abnormality. The WWH scoring system was used to examine the accuracy, completeness and precision of the PCEs that the participants provided for the test. Three marks were allocated to each image totalling 90 achievable marks for the test. For abnormal images, three marks were divided and allocated equally (one mark) to What, Where and How categories. Evaluation criteria (key elements of comments) were defined for each category and one mark was further distributed to the criteria (0.25, 0.5 or 1 mark, depending on the number of criteria). The number of the evaluation criteria and marks allocated varied depending on complexity of anatomy and abnormality. The evaluation criteria and score allocation for abnormal images are described in more details in Appendix L.

There are another three publicly available scoring systems for written comments/reports (Neep et al., 2017; Stevens & Thompson, 2018; The Royal College of Radiologists, n.d.). The Rapid Reporting session of the Final FRCR (Part B) examination is marked by its own scoring system (The Royal College of Radiologists, n.d.) (Table 3.8). For a normal image, one mark is recorded when correctly classified and 0.5 mark when incorrectly classified. one mark is recorded when abnormality is correctly classified and correctly identified for an abnormal image. The FRCR’s scoring system does not deal with partially correct comments for abnormal images. However, this was the only publicly available scoring system at the time of the data analysis of this research. Therefore, the participants’ comments were marked by using the FRCR’s scoring and WWH scoring systems for comparison.

Table 3.8.

FRCR scoring system.

Image type	Candidate response	Mark
Normal	– Correctly classified.	+1
	– Incorrectly classified (appropriate false positive).	+0.5
	– No answer given.	0
	–	
Abnormal	– Correctly classified and correctly identified.	+1
	– Correctly classified but incorrectly identified.	0
	– Incorrectly classified (false negative).	0
	– No answer given.	0
	Maximum score for normal images =	15
	Maximum score for abnormal images =	15
	Total score =	**30**

3.8. Ethical considerations

Researchers must safeguard human participants. Ethics in research refers to appropriate behaviour while designing and conducting research, particularly for participants. Unethical research adversely affects participants, target population, society and research realms. Researchers are therefore obliged to take an ethically responsible approach to protect participants and participating institutions (DePoy & Gitlin, 2016). Modern research ethics stem from the Nuremberg Code (1947). The Nuremberg Code was the first international document that upheld the importance of ethical principles in human subject research. Since then several ethics guidelines have been established to protect human participants in research (Belmont Report, 1979; Declaration of Helsinki, 1964, 2013;

European Union directive/regulation of the conduct of clinical trial, 2014; International Ethical Guidelines for Biomedical Research Involving Human Subjects, 2002). These guidelines advocate the protection of participants through voluntary consent, freedom from coercion, proper risk-benefit ratio, respect for autonomy, justice and fair selection (Doody and Noonan, 2016). Although these guidelines emphasise different ethical principles, four ethical standards are broadly recognised as essential in the ethics of human subject research and practice: (1) respect for autonomy; (2) beneficence; (3) non-maleficence; and (4) justice (Koyfman and Yom, 2017; Williams, 2015; Yuko & Fisher, 2015).

3.8.1. Full disclosure of research information and informed consent

Participants retain the right to self-determination and self-governance. Respect for autonomy connotes that participants in research can act freely without coercion from external force (Stephenson, Wagner & Bolton, 2012). Researchers must respect the autonomy of participants by seeking their consent. The ethical aim of consent is to ensure that potential research participants understand all the necessary information before agreeing to take part in a research project (Hain, 2016). Informed consent refers to an individual’s autonomous authorisation of participation in research. It comprises five elements: (1) competence; (2) disclosure; (3) understanding; (4) voluntariness; and (5) consent. Informed consent is obtained only when an individual is competent to act freely, receives full disclosure of information, has the capacity to understand the information, acts voluntarily, and consent to participate (Beauchamp & Childress, 2001). With enough time to read and comprehend, written information about research is invaluable to encourage potential participants to give informed consent (Hain, 2016).

In this research, a participant information sheet was developed prior to recruiting participants (Appendix C). This information sheet ensured full disclosure of research information, including the aims, procedure, length, confidentiality and expected benefits and risks. The information sheet also assured that participation was voluntary and withdrawal from the research would have no negative consequences. An informed consent form was also developed to ensure that the participants understood and were willing to participate in this research (Appendix M).

3.8.2. Risk management

The participants were asked to view 30 X-ray images in 45 minutes. There are potential risks caused by viewing X-ray images on the Display Screen Equipment (DSE) (e.g., a computer monitor). Viewing DSEs can be related with upper body muscle aches, fatigue and eyestrain. However, these symptoms are not unique to DSE and the risks to DSE viewers is low (Health and Safety Executive, 2003). Although the risks in this research were considered negligible, Research Ethics Checklist (SHUREC1) was used to ensure the ethical scrutiny of this research (Appendix M).

3.8.3. Confidentiality and anonymity

Confidentiality refers to management of private information. Researchers must establish a secure information management system that prevents identification of participants and unauthorised access to data (Williams, 2015). De-identification is an essential strategy to protect the confidentiality of participants (Chertoff, Pisano & Gert, 2009). Names of the participants and participating HEIs were anonymised by attaching unique identifiers to their information when the data was digitalised. Identifiable information of patients and radiographers on the X-ray images in the image bank was permanently removed or erased by using Adobe Photoshop.

3.8.4. Data management

Data management plans (DMP) for both digital and paper documents were developed by using DMPonline (https://dmponline.dcc.ac.uk/) (Appendix N).

3.8.5. Ethical and Research approval

Sheffield Hallam University requires that staff and doctoral research must be subjected to ethical scrutiny. Ethical approval was obtained from the Faculty Research Ethics Committee of Sheffield Hallam University on 4 November 2014 (Appendix O). The proposal for this research was then approved by the Research Degree Sub-Committee of Sheffield Hallam University on 13 May 2015 (Appendix E).

3.9. Summary

This research adopted a quantitative approach to determine the competencies of the final year undergraduate radiography students in PCE. This chapter suggested that a quantitative method, which is underpinned by the principles of post-positivism, was an appropriate method for benchmarking image observers’ evaluation performance. Identification and understanding of PCE comment errors are expected to prevent radiographers making the same mistakes. A PCE taxonomy was established to classify errors. This taxonomy incorporated 33 theoretically attainable PCE outcomes and served as the reference standard to classify PCE errors. The comments on radiographic appearances were further evaluated by a new scoring system developed specifically for this research. This scoring system was used to evaluate precision and completeness of PCE errors. Prior to carrying out the data collection, ethical standards were established to protect the participants and collaborating HEIs from ethical harm. The participants were given an information sheet and informed consent sheet to ensure that they understood the research aims and agreed to voluntarily participate. Anonymity was maintained by allocating pseudonyms to the participants and the HEIs.

Final year undergraduate students of diagnostic radiography programmes from England and Wales were invited to take an X-ray image evaluation test. The test was conducted by using an X-ray image bank that comprised 30 images of the appendicular skeleton. The results of the test were analysed to determine the competencies of the final year radiography students in PCE. The next chapter presents the results of the image evaluation test.

Chapter 4. Results

4.1. Introduction

This research aimed to benchmark the final year undergraduate diagnostic radiography students' competencies in PCE. The previous chapter discussed the methodologies of this research. An X-ray image evaluation test with an image bank comprised of 30 appendicular skeleton images was used to measure the evaluative performance of participating students. A quantitative approach was employed to determine accuracy, sensitivity and specificity of the participating students. Further analysis of PCE comments were performed by using the PCE error classification scheme and WHAT/WHERE/HOW scoring system.

This chapter presents the results of the image evaluation test. Most of the test results are presented with descriptive statistics as they are useful for assembling and summarising quantitative data (Marshall & Jonker, 2010). The data obtained from the answer booklets were entered in specially designed Excel spreadsheets with functions and conditional formatting to enable consistent data production and management. The quantitative data were subsequently transferred to SPSS for statistical analysis.

4.2. Overview of the results

Overall, a total of 87 students from nine universities agreed to participate in this study. The X-ray image evaluation test was conducted at each collaborating university, typically a computer lab or a lecture room with computers specifically booked for the test.

The mean accuracy, sensitivity and specificity were 73.37%, 79.62% and 67.13% respectively. As a result of the low response rate (24.58%), the study findings may not be statistically applicable to a wider sample. However, this study involved a larger sample size in comparison with most of other PCE studies (Coleman & Piper, 2009; Loughran, 1994; McConnell et al., 2013; Piper & Paterson, 2009; Smith & Younger, 2002).

4.3. Population and sample

The population of this study was defined as the final year undergraduate diagnostic radiography students at a point of graduation in England and Wales in 2015. In England and Wales, there were 20 HEIs with undergraduate diagnostic radiography programmes in 2014/2015 academic year. These programmes held a total of 946 final year students who were expected to graduate in 2015. Course leaders of 20 radiography programmes were contacted and nine (45.00%) agreed to take part. Agreed HEIs held a total of 443 final year students, which accounted for 46.83% of the target population.

A total of 87 students from nine HEIs participated in this study with the average response rates of 24.58% (*SD = 25.25)*. These students accounted for 9.20% of the whole diagnostic radiography students in the England and Wales and 19.64% of the students in participating nine HEIs in 2014/2015 academic year. Table 4.1 illustrates the number of participants, number of students in the course and response rates from each collaborating HEIs. No student withdrew from the test. Data collection started on 13 April 2015 and completed on 26 June 2015.

Table 4.1.

Comparison of the number of participating students, students in the course and response rates.

HEI IDs	No. of participants	No. of students in the course	Response rate (%)	% in sample (n = 87)
A	6	31	19.35	6.90
B	11	98	11.22	12.64
C	19	28	67.86	21.84
D	26	35	74.28	29.89
E	3	56	5.35	3.45
F	3	41	7.32	3.45
G	9	54	16.67	10.34
H	3	41	7.32	3.45
I	7	59	11.86	8.05
Total	87	443	-	100.00
Average	9.67	49.22	24.58	-

4.4 Demographics

4.4.1. Gender

Two thirds (72.41%, n = 63) of the participants were females and one quarter (24.59%, n = 24) were males (Figure 4.1).

4.4.2. Age

Participants' age was calculated based on the participants' dates of birth and the dates of the image evaluation test. The average age was 27.29 (*SD* = 6.77). Figure 4.2 shows the participants' age distribution. One student chose not to answer.

Figure 4.1. Gender distribution of the participants.

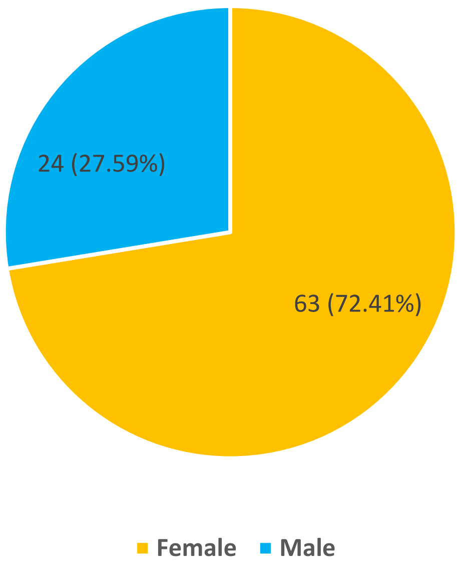
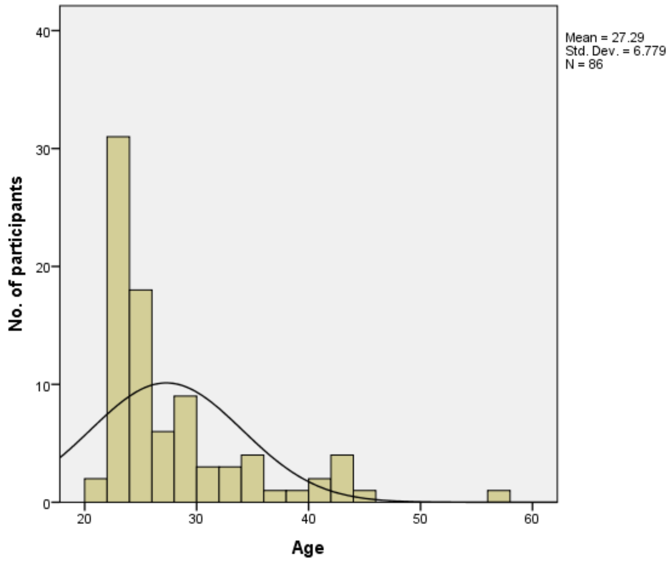


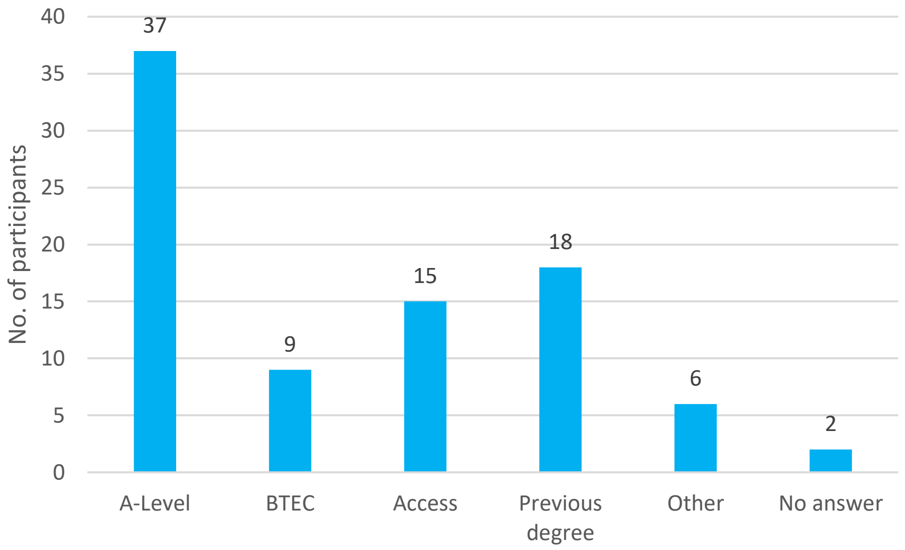
Figure 4.2. Age distribution of the participants.



4.4.3. Education prior to the programme

When asked about their education prior to the undergraduate programme, nearly half (42.53%, n = 37) responded they had A-level (Figure 4.3). Eighteen (20.69%) answered they had previous degrees. Participants with BTEC and Access accounted for 10.04% (n = 9) and 17.24% (n = 15) respectively. Only small numbers of the participants reported they had other academic qualifications (6.90%, n = 6) or chose not to answer (2.30%, n = 2).

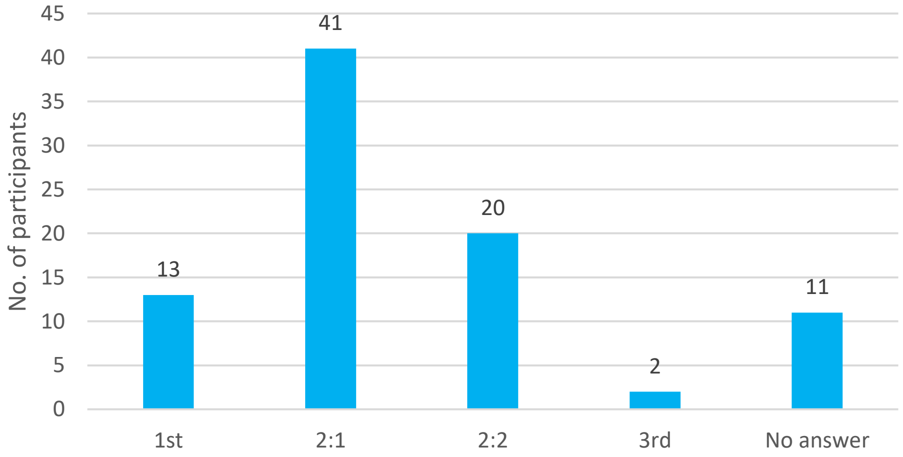
Figure 4.3. Education prior to the undergraduate programme.



4.4.4. Estimated degree

Nearly half of the participants (47.13%, n = 41) responded that they were expecting the upper second-class honours, followed by 20 participants (22.99%) with lower second-class honours at the time of the image evaluation test. These results were comparable with the figures published by Higher Education Statistics Agency (HESA) in 2016 (upper second-class honours: 49.52% and lower second-class honours: 22.98% in 2014/15). 14% (n = 13) of the participants responded that their estimated degree was first class honours. A small number of the participants (2.30%, n = 2) expected third class honours. Eleven participants (12.64%) chose not to answer (Figure 4.4).

Figure 4.4. Estimated degree.



4.4.6. X-ray images in the image bank

The participating students made a total of 2610 decision in the image evaluation test. Chapter 3.4.5 explained the development of the X-ray image bank used for the test. The image bank contained a total of 30 X-ray images with an equal ratio of normal and abnormal images. However, the image bank included three abnormal images with multiple fractures, and therefore the total number of abnormalities was 19. The demographic information of the image bank is summarised in Table 4.2.

Table 4.2.

Demographic information of the X-ray image bank.

	No. of images	%
Report		
Normal	15	50.00
Abnormal	15	50.00
Body parts		
Upper body	16	53.33
Lower body	14	46.67
Anatomical areas		
Hand	4	13.33
Wrist	4	13.33
Radius & Ulna	2	6.67
Elbow	4	13.33
Shoulder	2	6.67
Foot	6	20.00
Ankle	6	20.00
Knee	2	6.67

4.5. Results of the X-ray image evaluation test

4.5.1. True positives, true negatives, false positives and false negatives

The X-ray image evaluation test with the Likert scale (1) definitely normal, 2) probably normal, 3) possibly abnormal, 4) probably abnormal, or 5) definitely abnormal) produced a total of 2610 clinical decisions (87 participants x 30 X-ray images = 2610 decisions). The confusion matrix (Table 3.1) classified the participants' decisions into four categories: TPs, TNs, FPs and FNs. The most frequently occurred decisions were TPs, which accounted for 39.81% (n = 1039) of the total decisions. The second most frequent decisions were TNs (33.56%, n = 876), followed by FPs (16.44%, n = 429). FNs were the least common decisions (10.19%, n = 266) (Table 4.3 and Figure 4.5). The average number of TPs, TNs, FPs and FNs made by the participants for 30 X-ray images were 11.94 (*SD = 1.61*), 10.07 (*SD = 2.46*), 4.93 (*SD = 2.46*) and 3.06 (*SD = 1.61*) respectively (Table 4.4).

4.5.2. Accuracy, sensitivity and specificity

The mean accuracy was 73.37% with standard deviation of 8.01. The mean sensitivity and specificity were 79.62% (*SD = 10.78*) and 67.13% (*SD = 16.42*) respectively (Table 4.5). Accuracy ranged from 56.67% to 86.67%. Sensitivity and specificity ranged from 46.67% to 79.62% and 20.00% to 100.00% respectively (Figure 4.6). Table 4.6 and Figure 4.7 show the performance distribution of individual participants.

Table 4.3.

Total number, ratio (%), standard deviation (%) and 95% Confidence interval (%) of TP, FP, TN and FN.

Classifiers	Total No.	Ratio (%)	SD (%)	95% CI
TP	1039	39.81	5.39	[38.66, 40.96]
TN	876	33.56	8.21	[31.81, 35.31]
FP	429	16.44	8.21	[14.69, 18.19]
FN	266	10.19	5.39	[9.04, 11.34]

Figure 4.5. Distribution of TP, FP, TN and FN (%)

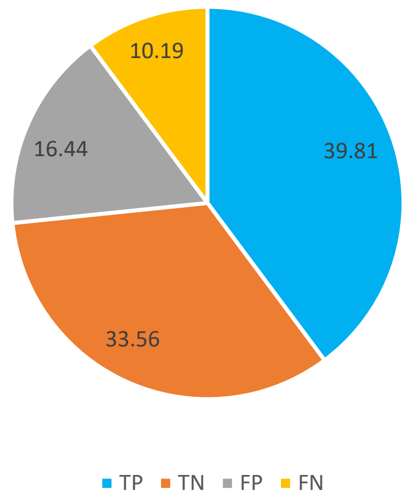


Table 4.4.

Average number of TPs, TNs, FPs and FNs made by the participants.

Classifiers	Min	Max	M (SD)	95% CI
TP	7	15	11.94 (1.61)	[11.60, 12.29]
TN	3	15	10.07 (2.46)	[9.54, 10.59]
FP	0	12	4.93 (2.46)	[4.41, 5.46]
FN	0	8	3.06 (1.61)	[2.71, 3.40]

Table 4.5.

Mean (M) values of accuracy, sensitivity and specificity.

	Min	Max	M(SD)	95% CI
Accuracy (%)	56.67	86.67	73.37 (8.01)	[71.66, 75.08]
Sensitivity (%)	46.67	100.00	79.62 (10.78)	[77.32, 81.91]
Specificity (%)	20.00	100.00	67.13 (16.42)	[63.63, 70.63]

Figure 4.6 Boxplots of mean accuracy, sensitivity and specificity with the performance standard at 90%.

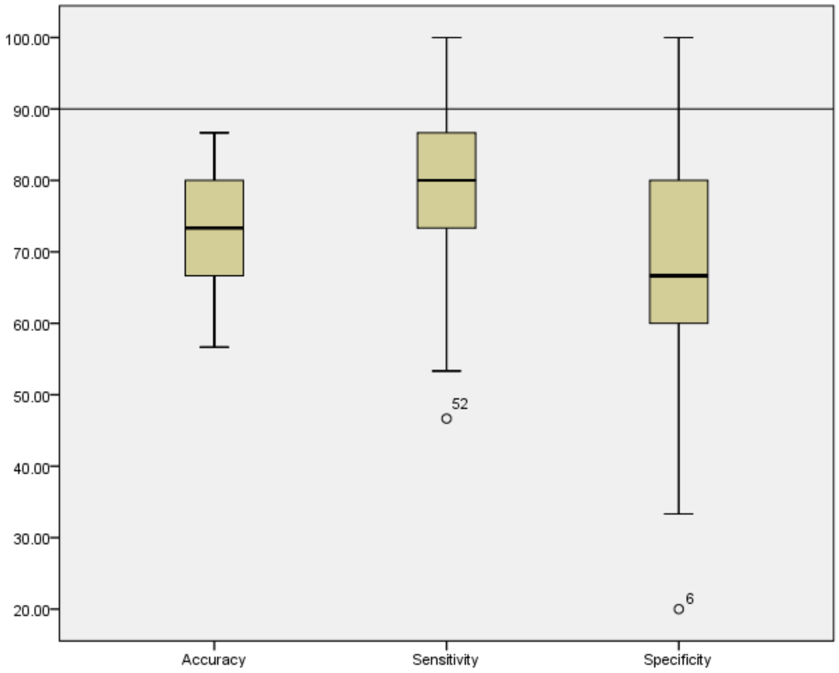
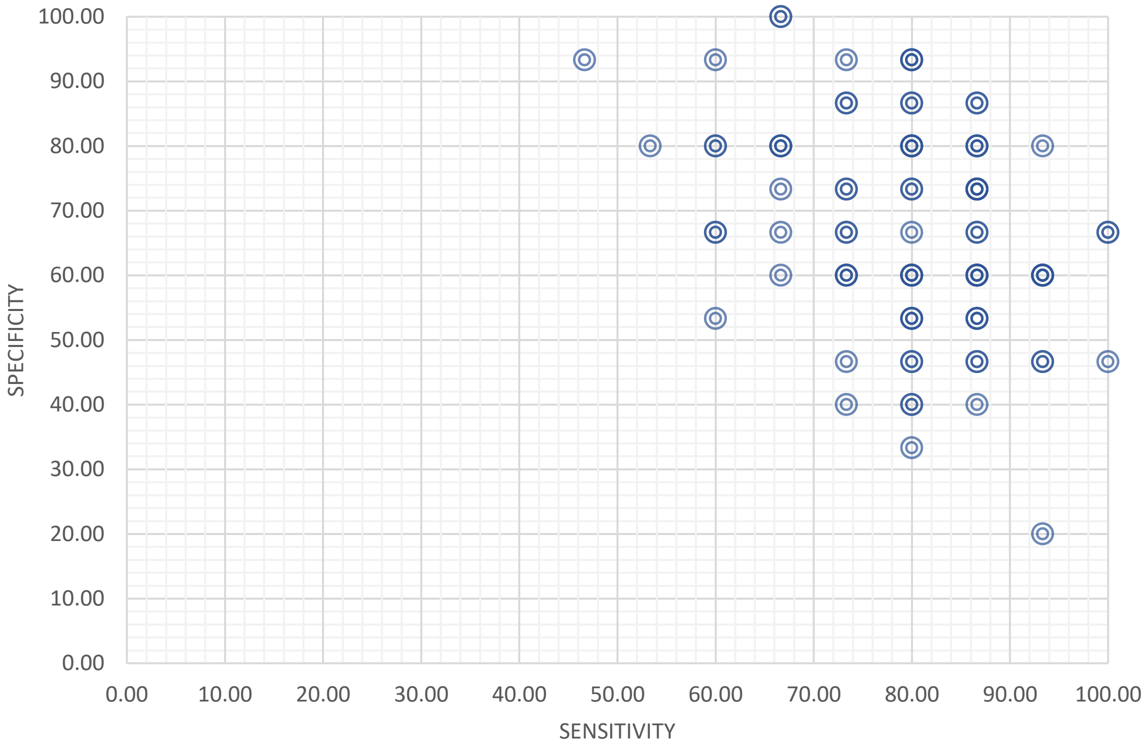


Table 4.6.

Performance distribution of individual participants.

	Accuracy (n/%)	Sensitivity (n/%)	Specificity (n/%)
Below 80%	60/68.97%	28/32.18%	59/67.82%
Between 80% and 90%	27/31.03%	47/54.02%	20/22.99%
Above 90%	0/0%	12/13.79%	8/9.20%

Figure 4.7 A scatter plot showing the distribution of sensitivity and specificity of individual participants.



4.5.3. Accuracy for body parts and anatomical areas

There was only 2.72% difference of accuracy between upper body images (74.64%) and lower body images (71.92%), although accuracy for different anatomical areas varied greatly from 47.13% (Radius & Ulna) to 90.23% (Hand) (Table 4.7).

4.6. Decision making confidence levels

The participants expressed their levels of decision confidence for image evaluation by using the Likert scales. Overall, “Definitely abnormal” was most commonly used (30.04%, n = 784) and closely followed by “Probably normal” (26.44%, n = 690) (Figure 4.8).

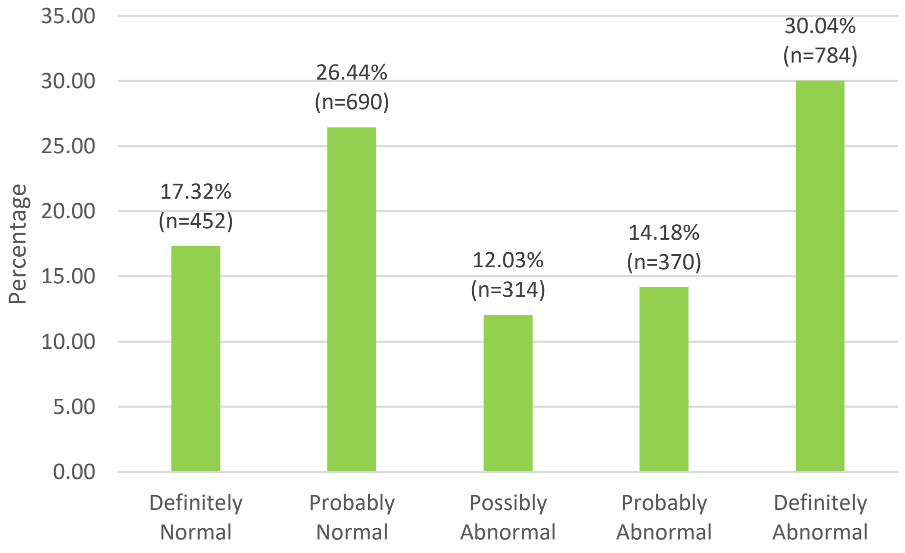
Table 4.7.

Mean accuracy (%) based on body parts and anatomical areas.

Body parts	M (SD)	95% CI
Upper body	74.64 (20.56)	[63.68, 85.60]
Lower body	71.92 (21.95)	[59.25, 84.59]

Anatomical areas	M (SD)	95% CI
Hand	90.23 (14.97)	[66.41, 100.00]
Wrist	83.91 (11.02)	[66.36, 100.00]
Radius & Ulna	47.13 (6.50)	[0.00, 100.00]
Elbow	64.37 (21.38)	[30.34, 98.39]
Shoulder	72.99 (25.19)	[0.00, 100.00]
Foot	74.52 (16.36)	[57.35, 91.69]
Ankle	72.80 (20.19)	[51.60, 93.99]
Knee	61.50 (51.20)	[0.00, 100.00]

Figure 4.8. Distribution of the Likert scale items used to express levels of decision confidence.



When the participants viewed the abnormal images, they were more likely to choose "Definitely abnormal" (26.97%, n = 704) than "Probably abnormal" (8.35%, n = 218) or "Possibly abnormal" (4.48%, n = 117). On the other hand, when they viewed normal images, they were more likely to choose "Probably normal" (19.50%, n = 509) than "Definitely normal" (14.06%, n = 367%) (Figure 4.9). The mean accuracy of each decision level was as follows: "Definitely normal" = 82.78%, "Probably normal" = 74.48%, "Possibly abnormal" = 38.69%, "Probably abnormal" = 60.34% and "Definitely abnormal" = 90.03% (Figure 4.10). Relationship between the participants' image evaluation performance and their decision confidence are summarised for sensitivity and specificity separately in Table 4.8 and 4.9.

Figure 4.9. Distribution of the Likert scale items used to express levels of decision confidence for normal and abnormal images.

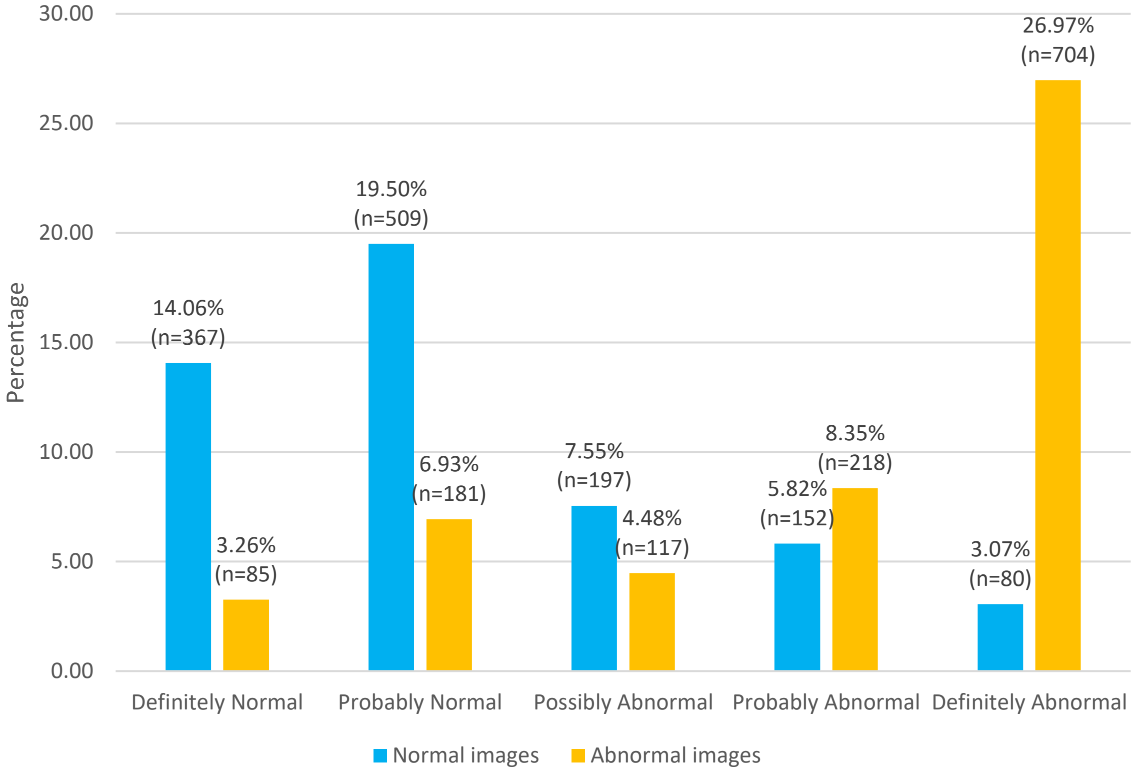


Figure 4.10. Mean accuracy (%) of decision confidence levels.

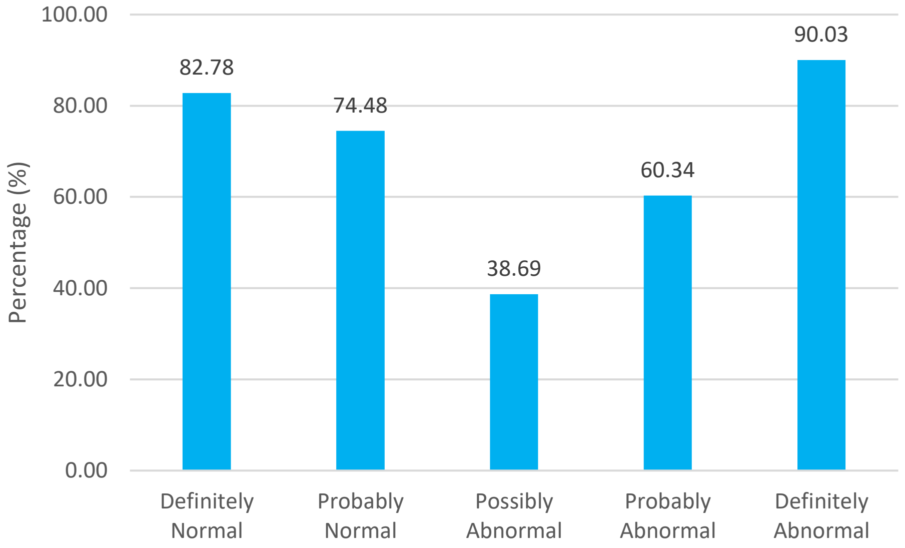


Table 4.8.

Sensitivity (%) and the participants' answers for abnormal X-ray images (X-ray images are sorted in a descending order of sensitivity).

Image No.	Sensitivity	Definitely Normal: 1	Probably Normal: 2	Possibly Abnormal: 3	Probably Abnormal: 4	Definitely Abnormal: 5
23	100.00	0	0	1	11	75
1	98.85	0	1	2	2	82
22	97.70	1	1	1	2	82
28	97.70	0	2	5	22	58
4	94.25	1	4	10	34	38
6	93.10	2	4	0	4	77
11	90.80	1	7	5	20	54
18	90.80	1	7	16	16	47
25	88.51	1	9	10	20	47
7	85.06	3	10	3	12	59
5	78.16	6	13	13	22	33
9	58.62	10	26	15	17	19
15	42.53	14	36	6	19	12
21	39.08	21	32	14	11	9
29	39.08	24	29	16	6	12
Mean		5.67	12.07	7.80	14.53	46.93
Total		85	181	117	218	704

Table 4.9.

Specificity (%) and the participants' answers for normal X-ray images (X-ray images are sorted in a descending order of specificity).

Image No.	Specificity	Definitely Normal: 1	Probably Normal: 2	Possibly Abnormal: 3	Probably Abnormal: 4	Definitely Abnormal: 5
24	96.55	45	39	2	1	0
12	91.95	37	43	7	0	0
16	77.01	30	37	11	5	4
8	74.71	31	34	12	4	6
17	73.56	17	47	12	7	4
10	71.26	21	41	7	14	4
14	71.26	26	36	14	11	0
19	71.26	29	33	11	11	3
2	70.11	25	36	11	7	8
30	67.82	28	31	15	10	3
3	56.32	23	26	16	15	7
27	55.17	16	32	20	12	7
26	52.87	14	32	20	13	8
13	51.72	16	29	21	17	4
20	25.29	9	13	18	25	22
Mean		24.47	33.93	13.13	10.13	5.33
Total		367	509	197	152	80

4.7. PCE error classification

Chapter 3.4 discussed the intention and development of the PCE taxonomy. PCE error classification scheme with the taxonomy aimed to determine types and frequencies of PCE errors by systematically categorising theoretically attainable outcomes of PCE and extracting erroneous decisions. The ultimate goal of this classification scheme is to reduce PCE errors by highlighting frequencies and causes of the errors. This study classified a total of 2610 PCEs from the image evaluation test by using the PCE taxonomy and decision tree classifier. Classified PCEs are organised based on their frequencies in Table 4.10. Following sections presents the results of each classified PCE category.

Table 4.10.

The results of PCE classification, organised by the frequencies of PCE types.

	n	%
Correctly classified and described	1451	55.59
Normal anatomy is described as abnormal (FP)	299	11.46
Abnormality is missed (FN)	255	9.77
Partially correct	209	8.01
Ambiguous	200	7.66
Correctly classified but description is incorrect (FN or FP or FN + FP)	130	4.98
No comment	49	1.88
Unclassified due to complexity	9	0.34
Discrepancy	8	0.31
Total	2610	100.00

4.7.1. PCEs with correct classification and description

Correctly classified and correctly described PCEs most frequently occurred (55.79%, n = 1451). Correctly classified and described PCEs for normal and abnormal images accounted for 33.10% (n = 861) and 22.68% (n = 590) of the total PCEs respectively. The mean of correctly classified and described PCEs produced by the participants was 16.68 (*SD* = 3.12) (Table 4.11).

Table 4.11.

Summary of PCEs with correct classification and description.

	n	%	Min	Max	M (SD)	95% CI
PCEs with correct classification and description	1451	55.79	10	22	16.68 (3.12)	[16.01, 17.34]
Normal images	861	33.10	3	15	9.90 (6.51)	[9.35, 10.44]
Abnormal images	590	22.68	2	12	6.78 (2.06)	[6.34, 7.22]

4.7.2. PCEs with incorrect description: False positive and false negative errors

Incorrectly described PCEs (including correctly classified PCEs) constituted quarter (26.30%, n = 684) of the total PCEs. False positives were the most frequent type of PCE errors (11.5%, n = 229), followed by false negatives (9.80%, n = 255). PCEs with correct classification but incorrect description (decision is correct but for incorrect evaluation reasoning) occurred less frequently (5%, n = 130) (Table 4.12).

Table 4.12.

Summary of PCEs with incorrect description (including correctly classified PCEs).

	n	%	Min	Max	M (SD)	95% CI
Normal anatomy is described as abnormal (FP)	299	11.50	0	11	3.44 (2.16)	[2.60, 3.26]
Abnormality is missed (FN)	255	9.80	0	8	2.93 (1.55)	[2.60, 3.26]
Correctly classified but description is incorrect (FN or FP or FN + FP)	130	5.00	0	4	1.49 (1.11)	[1.26, 1.73]

4.7.3. PCEs with partially correct description

Analysis of PCEs with the taxonomy found different types of PCEs with correct image classification with partially correct description of the abnormality. Such PCEs occurred when at least one identifiable abnormality was correctly described but the PCEs also expressed an erroneous or ambiguous judgement about presence or absence of abnormality.

Such PCEs occurred when the following criteria are met:

1. Image is correctly classified.
2. At least one identifiable abnormality is correctly described.
3. Normal anatomical structure is described as abnormal (false positive), or another abnormality is missed (false negative), or a combination of false positive and false negative, or the rest of the PCE remains ambiguous or inconclusive.

Overall, partially correct PCEs accounted for 8.04% (n = 209) of the total PCEs. Most of these partially correct PCEs occurred when at least one identifiable abnormality was correctly described, but the PCE also contained either false positive (4.15%, n = 108) or false negative errors (2.85%, n = 74). Although very rarely, partially correct PCEs also occurred when the PCEs had a combination of false positive and false negative (0.58%, n = 15) or ambiguous description of radiographic appearances (0.46%, n = 12) (Table 4.13).

Table 4.13.

Summary of PCEs with partially correct PCEs.

	n	%	Min	Max	M (SD)	95% CI
At least one abnormality is correctly described (TP), but	**209**	**8.04**	**0**	**5**	**2.40 (1.05)**	**[2.18, 2.63]**
Another abnormality is missed (FN)	108	4.15	0	4	1.24 (0.88)	[1.05, 1.43]
Normal anatomy is described as abnormal (FP)	74	2.85	0	4	0.85 (0.96)	[0.65, 1.05]
Another abnormality is missed (FN) + normal anatomy is described as abnormal (FP)	15	0.58	0	2	0.17 (0.41)	[0.09, 0.26]
Rest of the PCE is ambiguous and it is incorrect	8	0.31	0	1	0.09 (0.29)	[0.03, 0.15]
Rest of the PCE is ambiguous but it is actually correct	4	0.15	0	1	0.05 (0.21)	[0.00, 0.09]

4.7.4 PCEs with ambiguous description

Radiographic descriptions must be written without ambiguity. The SCoR's (2013) also expect that radiographers articulate radiographic findings in unambiguous written forms. In this study, a PCE was considered "ambiguous" if it expresses a clinical judgement with one or both of the following criteria:

1. Hedge words used to avoid clear clinical answer, or/and

2. Indirect sign of trauma (e.g., lipohaemarthrosis or raised fat pads) was noted but precise description of abnormality was absent.

PCEs with the following hedge words were considered ambiguous in the analysis: "possible", "possibly", "probable", "probably", "potential", "potentially", "may", "maybe", "might", "might be" and a question mark "?".

PCE analysis with the taxonomy identified that 7.69% (n = 200) of the PCEs provided ambiguous or evasive clinical judgements. A closer look at the results also showed that only a small fraction of ambiguous descriptions of abnormality were correct (0.04%, n = 1 for normal images and 1.04%, n = 27 for abnormal images) (Table 4.14).

Table 4.14.

Ambiguous PCEs for normal and abnormal images. Table also shows a small fraction of PCEs that were ambiguous but actually correct.

	n	%	Min	Max	M (SD)	95% CI
Ambiguous PCEs: total	**200**	**7.70**	**0**	**10**	**2.30 (2.22)**	**[1.83, 2.77]**
Ambiguous PCEs for normal images	109	4.20	0	8	1.25 (1.59)	[0.91, 1.59]
Ambiguous but correct: normal	1	0.04	0	1	0.01 (0.10)	[0.00, 0.03]
Ambiguous PCEs for abnormal images	91	3.50	0	5	1.05 (1.12)	[0.81, 1.28]
Ambiguous but correct: abnormal	27	1.04	0	2	0.31 (0.57)	[0.76, 1.24]

4.7.5. Other PCE errors

Other three error types appeared fairly infrequently. No comment error (PCEs with no description of radiographic appearances) occurred in 1.88% (n = 49) of the total PCEs.

“Unclassifiable” PCEs accounted for 0.34% (n = 9) of the PCEs. These arose when grammatical structures of PCEs were highly complex and long (e.g., a combination of false positive and false negative expressed in a vague language). The least frequent PCE errors were discrepancy errors (0.31%, n = 8). This type of error occurred when clinical decision and description of radiographic appearances contradicted each other (e.g., image is classified as abnormal but described as normal).

4.8. PCE comment analysis: FRCR and WWH scoring systems

Chapter 3.5.5 discussed the background and justification of the development of WWH scoring system. For comparative purposes, FRCR and WWH scoring systems were used to assess the quality of PCEs. Two scoring systems have different maximum scores (FRCR = 30 points and WWH = 90 points) and therefore comment scores are expressed in percentage terms to allow comparative data presentation.

Overall, the FRCR scores for total, normal images and abnormal images were 65.54% (*SD* = 7.53), 83.56% *(SD* = 8.21) and 47.51% *(SD* = 13.48) respectively. The WWH scores for total, normal images and abnormal images were 49.02% (*SD* = 9.75), 67.13% (*SD* = 16.42) and 30.91% (*SD* = 9.87) respectively (Table 4.15).

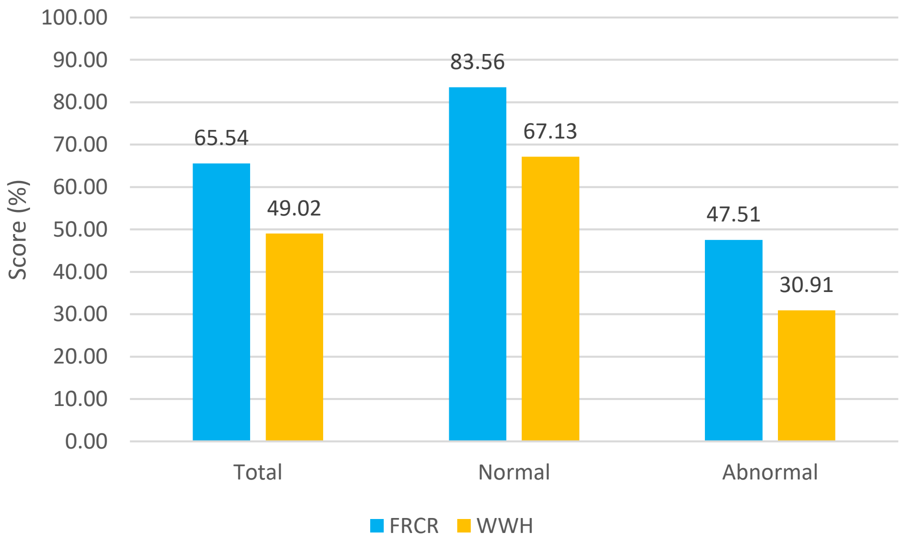
Table 4.15.

FRCR and WWH scores (%) for total, normal images and abnormal images.

	Min	Max	M (SD)	95% CI
Total				
FRCR	50	85	65.54 (7.53)	[63.93, 67.14]
WWH	23.33	72.5	49.02 (9.75)	46.94, 51.09]
Normal Images				
FRCR	60	100	83.56 (8.21)	[81.81, 85.31]
WWH	20	100	67.13 (16.42)	[63.63, 70.63]
Abnormal images				
FRCR	20	86.67	47.51 (13.48)	[44.64, 50.38]
WWH	10	63.89	30.91 (9.87)	[28.80, 33.01]

A graphical comparison of the scores (Figure 4.11) highlights that FRCR scoring system yielded approximately 15% more points for total, normal images and abnormal images than WWH scoring system.

Figure 4.11. Comparison of FRCR and WWH scores (%).



4.8.1 FRCR and WWH scores for normal images

FRCR and WWH scoring systems take a slightly different approach to quantify evaluation performance for normal images. The scoring system of FRCR rapid reading session suggests that it acknowledges classified decisions with faulty reasoning, and therefore records zero mark. On the other hand, WWH scoring system first judges whether decisions are correct, then continues to investigate completeness and precision of descriptions for radiographic appearances. Despite the difference of scoring systems for normal images, specificity (M = 67.13%) and FRCR/WWH scores demonstrated an exact positive linear relationship (FRCR: $r = 1$, $n = 87$, $p < 0.00$ and WWH: $r = 1$, $n = 87$, $p< 0.00$) (Figure 4.12 and 4.13).

However, FRCR's mean normal image score (83.6%) was 16.43% higher than WWH's mean normal score (67.13%). This is because FRCR records a half mark (+0.5) for incorrect classification (FP). Figure 4.14 found that participants with specificity as low as 20% achieved 60% of FRCR normal score, while WWH's normal scoring system perfectly mirrored specificity.

Figure 4.12. Correlation between specificity and FRCR normal score (%) (r = 1, n = 87, p < 0.00).

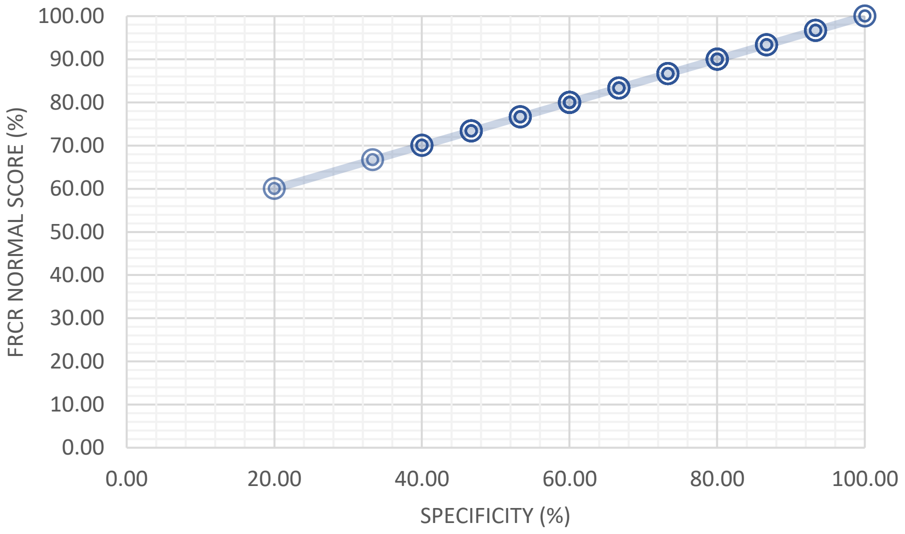
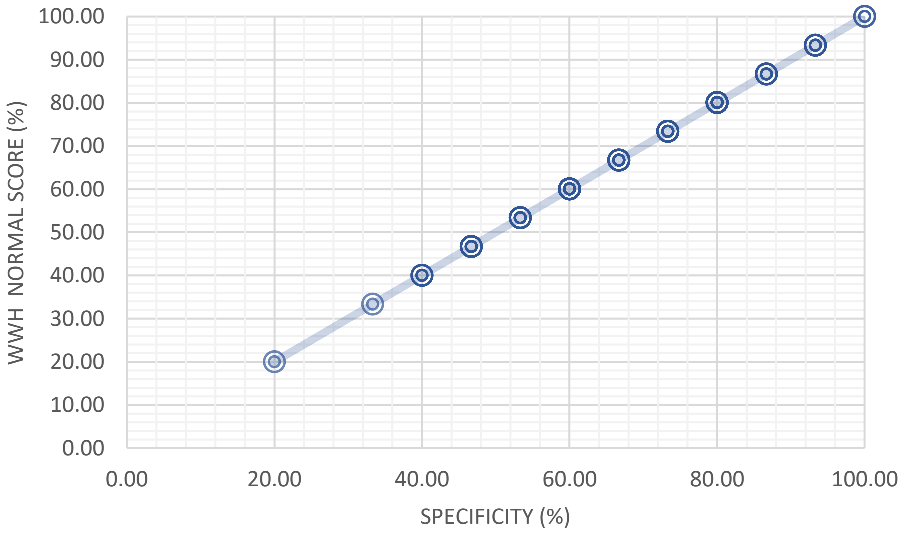


Figure 4.13. Correlation between specificity and WWH normal score (%) (r = 1, n = 87, p < 0.00).



4.8.2 FRCR and WWH scores for abnormal images

There was also a positive relationship between sensitivity and FRCR/WWH abnormal scores (FRCR: r = 0.603, n = 87, p < 0.00 and WWH: r = 0.543, n = 87, p < 0.00). However, this was evidently weaker than the relationship between specificity and FRCR/WWH normal image scores (Figure 4.14 and Figure 4.15), indicating the binary logic classified some decisions as correct although the reasoning was incorrect.

4.8.3 WHAT, WHERE and HOW scores

WWH scoring system allocates 1 point to each WHAT, WHERE and HOW categories. This point allocation assigns three points to every abnormal image and 45 points for the abnormal images (n = 15) in the image bank. The mean score for WHAT (abnormality type) and WHERE (abnormality location) were 5.89 (*SD* = 1.76) and 6.52 (*SD* = 1.79) respectively. Very few referred to the abnormality's angulation or dislocation (M = 1.50, 10.00%) (Table 4.15).

Figure 4.14. Correlation between sensitivity and FRCR abnormal scores (%) (r = 0.603, n = 87, p < 0.00).

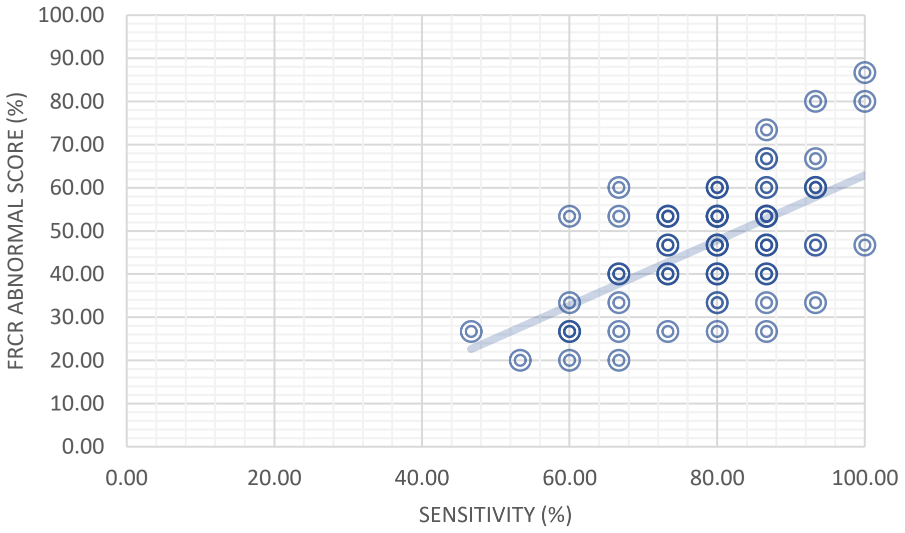


Figure 4.15. Correlation between sensitivity and WWH abnormal scores (%) (r = 0.543, n = 87, p < 0.00).

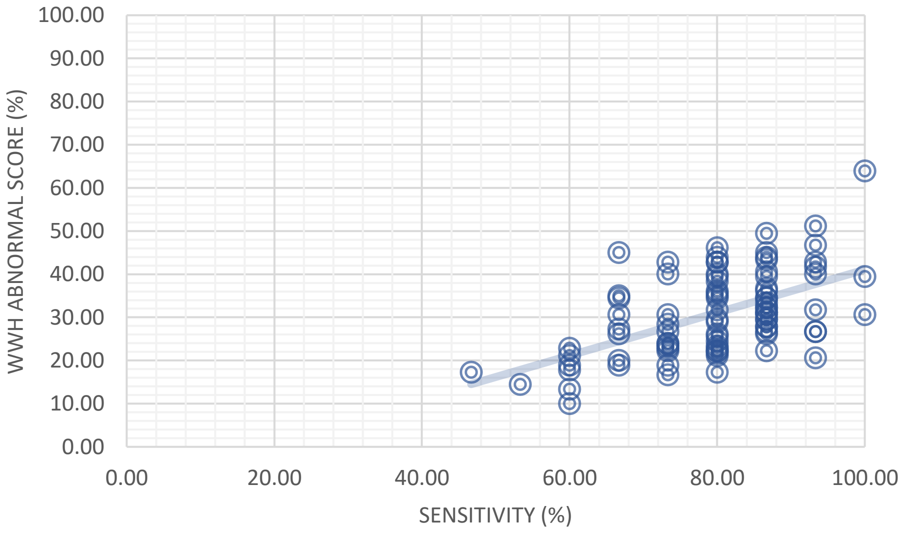


Table 4.16.

Minimum, maximum and mean scores for WHAT, WHERE and HOW categories.

	Min	Max	M (SD)	(%)	95% CI
WHAT	1.75	10.75	5.89 (1.76)	39.27	[5.51, 6.26]
WHERE	2.75	10.75	6.52 (1.79)	43.47	[6.14, 6.90]
HOW	0	7.25	1.50 (1.85)	10.00	[1.11, 1.90]

4.9. Summary

This chapter demonstrated the image evaluation competencies of the final year radiography students at the point of graduation. The image evaluation test demonstrated that their mean accuracy, sensitivity and specificity were 73.37%, 79.62% and 67.13% respectively. The participants commonly expressed their decision using “Definitely” for abnormal images. On the other hand, “Probably” was more frequently used for normal image evaluation. Scrutiny of comments using the PCE taxonomy showed that more than half (55.79%, n = 1451) were correctly classified and described, although incorrect PCEs constituted approximately quarter (26.30%, n = 684). The scoring of PCE comments indicated that FRCR tended to yield more points (approximately 15%) than WWH. The WWH scoring found that some correct Red-dot style decisions were made with incorrect reasoning. The results also indicated that angulation or dislocation of abnormalities were very rarely articulated in the comments. Next chapter provides an in-depth discussion of the research findings.

Chapter 5. Discussion

5.1. Introduction

This chapter discusses potential competencies of newly qualified radiographers in X-ray image evaluation and PCE. The literature review (Chapter 2) is also reflected upon with the findings of this study to explore the current state of newly qualified radiographers with respect to their image evaluation skills.

5.2. Sample size

The literature review found that the convenient sampling method, which does not allow statistical inference to larger samples, was a generic limitation of image evaluation studies. A small sample size is also a limitation of many image evaluation studies since it does not allow generalisation of the findings. Post-positivist researchers generally do not show a tangible interest in generalisability of their study findings. However, a representative sample generalisability must be at least considered in the study design.

Two studies managed to recruit relatively large numbers of radiographers (Hardy & Culpan, 2007: n = 155 and Mackey, 2006: n = 133), the sample sizes of other reviewed studies (Table 5.1) remained small, ranging from three (Loughran, 1994) to 34 (Wright & Reeves, 2016). Non-audit studies largely depended on a non-probability sampling method, either using volunteering radiographers (Hazel et al., 2015; McConnell et al., 2012; McConnell & Baird, 2017; McConnell et al., 2013; Smith & Younger, 2002; Wright & Reeves, 2016) or self-selected radiographers in training programmes (Hargreaves & Mackay, 2003;

Loughran, 1994; McConnell & Webster, 2000; Piper & Paterson, 2009; Piper et al., 2005). The small sample sizes and their representativeness should therefore be carefully considered.

Table 5.1.

Number of participants of other image evaluation studies.

Authors	No. of participants
Red-dot studies	
Brown & Leschke (2012)	n/a
du Plessis & Pitcher (2015)	9
Hardy & Culpan (2007)	115
Hargreaves & Mackay (2003)	7
Hazel, Motto & Chipeya (2015)	9
Hlogwane & Pitcher (2013)	n/a
Mackey (2006)	133
McConnell & Baird (2017)	16
McConnell & Webster (2000)	22
Renwick, Butt & Steel (1991)	n/a
Wright & Reeves (2016)	34
PCE studies	
Coleman & Piper (2009)	18
Loughran (1994)	3
McConnell et al. (2012)	10
McConnell, Devaney & Gordon (2013)	10
Piper & Paterson (2009)	18
Smith & Younger (2002)	26

n/a: Audit studies without sample information.

The audit studies poorly documented their sample population and evaluation of sample representativeness was impossible (Brown & Leschke, 2012; Hlogwane & Pitcher, 2013; Piper et al., 1999; Renwick et al., 1991).

There are several conditions that result in small sample sizes. Hazell et al., (2015) noted that radiographers' work commitment and staffing requirements hindered radiographer recruitment in their study. McConnell and Baird (2017) also observed that radiographers' decisions not to participate in the image evaluation test were because they felt their absence from the workplace would have a negative impact on the service. Another cause may be that the reviewed studies targeted a small size of radiographer population during the participant recruitment stage. Many of the reviewed studies were conducted at one or two hospitals, where only a small number of radiographers may have been available for recruitment (Brown & Leschke, 2012; Coleman & Piper, 2009; du Plessis & Pitcher, 2015; McConnell et al., 2012; McConnell et al., 2013; Hargreaves & Mackay, 2003; Hazel et al., 2015; Hlogwane & Pitcher, 2013; Loughran, 1994; Renwick et al., 1991; Smith & Younger, 2002). The potential size of participants was relatively large (n = 75) in the study conducted by Hazel et al. (2015). However, the resulting sample size was small (n = 9) due to a low response rate (12.00%). This indicates that image evaluation studies must ensure not only a higher response rate, but also a large number of radiographers who can potentially be recruited.

Coordinating across multiple study sites (e.g., undergraduate courses or hospitals) is one possible solution to achieve a larger sample size. The sample size of the present study (n = 87) was less than anticipated (274 students were required for 95% confidence level). However, the size was one of the largest in the reviewed image evaluation studies. There are two conceivable reasons for this. Firstly, this study initially targeted multiple institutes (n = 20) in England and Wales and the response rate was high. Nearly half (n = 9) of the institutes agreed to collaborate, with a potential sample size of 443 (46.83%) of the final year diagnostic radiography students in 2014/2015 academic year. Although the student response rates ranged from 5.35% to 74.28%, sampling from multiple institutes ensured a larger sample size than most other image evaluation studies. Secondly, incentives were given. Incentives improve a response rate; said to be especially effective for population referred as “Generation Y” (born from the 1980s and onwards) (Morton, 2002). The participant information sheet of this study (Appendix C, Part 1.7) explained that each student would receive a learning opportunity as well as a certificate after completing the image evaluation test. Although there is no evidence to suggest incentives work for image evaluation tests, it is reasonable to suggest that the incentives may have exerted a positive influence on the students’ response rate in this study.

The main objective of this study was to benchmark the final year students’ newly qualified radiographers’ competencies in image evaluation. Therefore, final year radiography students at a point of graduation (between April and June) were recruited. However, it appeared that, this timing of data collection (image evaluation test) was the cause of low students’ response rates. A few months before the graduation period, students at many study sites were in clinical placements and therefore absent from their home institutes. The response rates could have been higher if the study had been conducted when students still regularly attended their institutes. However, this poses a dilemma that the earlier timing of data collection would have placed recruited students outside the boundary of the target population (newly qualified radiographers/graduates at the point of qualification).

5.3. Binary decision accuracy

Mean accuracy, sensitivity and specificity of the participants based on their binary decisions (normal or abnormal) were 73.37%, 79.62% and 67.13% respectively. The mean specificity was 12.49% lower than sensitivity, which generally indicates that evaluating normal images was more challenging than abnormal images for the participating students. Hazel et al. (2015) reported a similar trend that radiographers may lack an ability to confidently identify normality. The test results of this present study demonstrated that the frequencies of false positive errors were greater (16.44%, n = 429) than false negative errors (10.19%, n = 266). Some image evaluation studies observed a tendency for the radiographers to err on the side of caution, which resulted in a greater chance of making false positive decisions and reduced specificity (McConnell et al., 2013; Buksov et al., 2013). This may also be the case in the present study.

There is currently absence of widely accepted performance standards for the Red-dot and PCE. For clinical reporting, three independent groups agreed that 95% sensitivity and specificity constituted the performance standard for reporting (Brealey, 2001a; Paterson et al., 2004; Stephenson et al., 2012). Despite its arbitrariness, this performance standard is a reasonable expectation for radiologists and qualified reporting radiographers since their qualifications require formal specialised postgraduate education and many years of clinical experience. Many radiographers in Red-dot and PCE studies that accepted the "95% rule" did not reach the desired performance level (du Plessis & Pitcher, 2015; Hlogwane & Pitcher, 2013; Mackey, 2006; Piper et al., 2005; Piper et al., 1999; Smith & Younger, 2002; Wright & Reeves, 2016;). Likewise, if we accepted that 95% was the absolute standard, the performance of the participants of this study fell short of the standard. However, there has been insufficient discussion about the applicability of the 95% rule to research, regardless of study types (audit or image bank) and image evaluation types (Red-dot system or PCE). It is probably unreasonable to expect that general radiographers (especially newly qualified radiographers before entering their preceptorship) should achieve the same performance standard of radiologists and reporting radiographers. Instead, 90% sensitivity and specificity may be more realistic and commendable goals for general and newly qualified radiographers who participate in the Red-dot system and PCE of the musculoskeletal examinations. Wright and Reeves (2017) maintained that radiographers are now reasonably expected to demonstrate 90% accuracy in any form of RADS (Chapter 1). Brealey (2001a) also argued that 90% accuracy is optimal for radiographers (and 95% is ideal). The literature review found that 14 groups of radiographers (82.35%) in the Red-dot and PCE studies (n = 17) achieved above mean sensitivity of 80%. Table 4.5 and Figure 4.7 indicate that 67.82% (n = 59) participants of the present study also demonstrated more than 80% sensitivity at the point of qualification with HCPC. Further improvement is conceivable if appropriate training is delivered during and after the preceptorship. 90% sensitivity should therefore be the minimum performance standard that radiographers should strive for, considering that:

1) PCE is a non-definitive forerunner of clinical reporting.

2) it is irrational to expect that radiographers have the same performance standard of radiologists and reporting radiographers.

3) the reviewed literature and the results of this present study indicate many radiographers demonstrate more than 80% sensitivity.

4) appropriate training before and after qualification could boost the performance of radiographers who have not achieved 90% sensitivity.

Whether radiographers should demonstrate 90% specificity in the current state is open to debate. Table 4.5 indicated that only 32.18% (n = 28) of the participants in the image evaluation test achieved above 80% specificity. Figure 4.7 also showed that there were no participants who achieved 90% sensitivity and specificity at the same time in the test. Furthermore, the literature review (Chapter 2) found that low specificity is a long-standing occupational characteristic of diagnostic radiographers. However, since normal radiographs constitute a larger proportion of radiographs in A&E settings, minimum of 90% specificity should be ideal as the performance standard when PCE becomes a more widely accepted role of radiographers. Overall accuracy of 90% is also in line with the pass mark of Final FRCR Part B Examination (54/60 marks).

Little is understood about the impact of clinical experience on image evaluation performance. Some studies attempted to investigate the relationship between clinical experience and image evaluation competencies, but the results were inconsistent (du Plessis & Pitcher, 2015; Mackay, 2006; Wright & Reeves, 2016). However, Hargreaves and Mackay (2003) pointed out the likelihood that radiographers continue to improve their image evaluation skills through informal learning after qualification. This is plausible since radiographers regularly view radiographs they take and consider if extra images are necessary for better visualisation of abnormalities when they are present. Hardy et al. (2016) also hypothesised that radiographers may acquire a mental library of normal radiographic appearances after viewing normal radiographs. They argued that radiographers with such a mental library of common normal radiographs could be less prone to over-call normal radiographs. The prevalence of musculoskeletal trauma in A&E settings is estimated to be around 20 to 30% (Hardy et al., 2012; Hardy et al., 2008; McConnell et al, 2013; Renwick et al., 1991; Robinson, Culpan & Wiggins, 1999), clearly indicating that radiographers in A&E settings view a large volume of normal images during their practice. If we accept the theory proposed by Hardy et al. (2016), radiographers should possess a mental library that allows them to evaluate radiographs with high specificity.

The evidence shows otherwise. The literature review of the Red-dot studies found that specificity of many radiographers remained below the ideal performance standard (90%) (Figure 2.2). The X-ray image evaluation test of this study similarly demonstrated that specificity of the students at the point of graduation fell short of the desired standard. In the current state, it could be argued that radiography students complete the undergraduate education without necessarily acquiring a mental library of normal radiographs that is large enough to correctly make decisions for normality.

Further analysis of the test results suggested a possible influence of education on how undergraduate students evaluate normal images. A Kruskal-Wallis test demonstrated very strong evidence of a difference of FP decisions (p = .006) between the mean ranks of at least one pair of universities. Dunn’s pairwise tests were carries out for the two pairs of universities: University A (Uni. A) and University D (Uni. D). There was very strong evidence (p = .034, adjusted using the Bonferroni correction) of a difference between the two groups of students in Uni. A and D. The median frequencies of FP decisions for Uni. A was 7 compared to 3 for Uni D. Another set of Kruskal-Wallis test and Dunn’s pairwise test also indicated very strong evidence of a difference of TN decisions (p = .034) between Uni. A (Mdn = 8) and D (Mdn = 12). These statistically significant differences indicated that the students in Uni. A were less likely to correctly evaluate normal images than the students in Uni. D. The differences were also reflected in statistically significant differences of the median specificity (Uni. A: 53.30% vs. Uni. D: 80.00%) (p = .034), median FRCR normal scores (Uni. A: 11.5 points vs. Uni. D: 13.5 points) (p = .034) and median WWH normal scores (Uni. A: 24 points vs. Uni. D: 36 points) (p = .034). The present study could not determine the possible reasons that created the differences between two universities owing to a lack of information for further analysis. However, different modes of education delivery and clinical placements could be probable explanations for the performance gap. Descriptive statistics also showed that the mean specificity of the students from Uni. A (48.89%) was noticeably lower than the rest of the student groups (68.47%), perhaps suggesting a lack of educational emphasis on normal image evaluation.

The mean specificity at each university was invariably lower than sensitivity except for Uni. F (n = 3). If high specificity is attributed to education with routine exposure to normal radiographs as seen for reporting radiographers, should a greater educational focus be brought into evaluating normality to boost specificity at the undergraduate level? Implications of the findings suggest that undergraduate education providers should collaborate in partnership with clinical placement sites to devote sufficient focus on evaluation of normal radiographs. Furthermore, preceptorship holds promising potential. Literature suggests that 70% of radiographs that newly qualified radiographers view during preceptorship in A&E settings would be normal. Despite the lack of documentation in diagnostic radiography (Chapter 1), preceptorship provides ample opportunity for newly qualified radiographers to reinforce their cognitive libraries by viewing and evaluating a large volume of normal radiographs under supervision. The possible values of the preceptorship are considered in Chapter 6.2.7.

5.4. Decision making confidence levels

Chapter 3.7.4 pointed out that there is still limited research evidence to illustrate the relationship between radiographers’ confidence for image evaluation practice and accuracy. The X-ray image evaluation test of this research therefore asked the participants to express their levels of confidence for their decisions by using the five-point Likert scales. The test results demonstrated that “Definitely abnormal” (30.04%, n = 784) and “Probably normal” (26.44%, n = 690) constituted nearly half of the responses.

For abnormal images, the participants were far more likely to use “Definitely” than “Probably” and “Possibly” for abnormal images (Figure 4.8). This was probably because when the participants recognised (or they thought they recognised) obvious signs of fractures (e.g., a clear cortical disruption), the most sensible choice was “Definitely”, rather than “Probably” or “Possibly”. Previous sections demonstrated that evaluation of normal images posed more challenges to the participants than abnormal images. For normal images, Figure 4.8 showed that the participants were more likely to use “Probably” than “Definitely”. This may be a fair indication of uncertainty and an insufficient size of a cognitive library that the participants had for normal images.

There seems to be a positive relationship between image evaluation accuracy and decision-making confidence (Figure 4.9). The figure suggests that the participants’ accuracy exceeded 80% for normal and 90% for abnormal when they expressed their decisions using “Definitely”. A positive correlation between the participants’ performance for individual images and confidence was also demonstrated by further breakdown of the test results (Table 4.8 and Table 4.9). Table 4.8 shows concomitantly decreased or increased sensitivity and levels of decision-making confidence. “Definitely abnormal” was used for more than half of the responses (53.94%, n = 704) for the abnormal images. There was a very strong and positive correlation between sensitivity and frequencies of “Definitely abnormal” for abnormal images, which was statistically significant (r = .888, n = 15, p < .00). On the other hand, there was moderate negative correlation between sensitivity and frequencies of “Possibly abnormal”, which was also statistically significant (r = -.591, n = 15, p = .02). These indicate that decisions expressed with stronger confidence for abnormal images are likely to result in higher sensitivity.

The participants responded differently for the normal images. Table 4.8 shows that the participants tended to use “Probably” (39.00%, n = 509) more frequently than “Definitely” (28.12%, n = 367) for normal images, perhaps suggesting that they evaluated the normal images with added caution and uncertainty. The table also indicates that responses were almost equally distributed to “Definitely” and “Probably” even when specificity was high, such as in image 24 (96.55%) and 12 (91.95%) in the test. This indicates that, unlike the behaviour observed for abnormal images, many achieved high specificity with weaker confidence or increased degree of difficulty. The analysis of levels of decision-making confidence and sensitivity showed that their relationship was proportional to each other. If levels of decision-making confidence and resulting accuracy of radiographs are proportional to each other, as seen for abnormal images, the choice of “Definitely normal” should have a stronger positive correlation with specificity than “Probably normal”. However, this was not observed for the normal images in this study. There was only a negligible difference of positive correlations between specificity and the use of “Definitely” (r = .887, n = 15, p < .00) and “Probably” (r = .837, n = 15, p < .00), indicating that there is no linear relationship between levels of decision-making confidence and specificity. The literature review of this study pointed out that radiographers generally demonstrate lower specificity than sensitivity. A similar result was found in the image evaluation test of this study (Table 4.5). The previous section of this chapter also discussed the need for a greater educational focus on normality to improve specificity. The participants' lack of confidence in evaluating normal images in this study may be important supplemental evidence that suggests insufficient emphasis on normality in the current undergraduate education. If confidence is partly associated with knowledge, amount of training, and expertise (Benvenuto-Andrade et al., 2006), education providers must acknowledge that radiographers and radiography students will benefit from additional training on evaluating normal radiographs at pre- and post-qualification, such as preceptorship, to confidently evaluate normal radiographs, and thus boost specificity.

Chapter 3.7.4 pointed out that ERS (Extreme Response Style) is an ideal behaviour for image evaluation tests when decisions are based on correct observation. For example, radiographers with high sensitivity and specificity might dichotomously classify images as "Definitely normal" or "Definitely abnormal", which results in ERS. However, the participants of this study did not show ERS in the image evaluation test. The test results for the abnormal images implied a possibility of ERS. However, more frequent use of "Probably" than "Definitely" for the normal images suggests ERS was not the case (Figure 4.8). The participants appeared to have avoided extreme responses and preferred a more careful evaluation approach for the normal images, which is consistent with Wright and Reeves (2016) that radiographers rarely show ERS while evaluating radiographs. Mayer et al. (2013) evaluated how physicians' diagnostic accuracy and confidence changed with increasing difficulty of clinical cases. They found that physicians' levels of confidence were unaffected by diagnostic accuracy and difficulties of clinical cases, implying that physicians may wrongly establish diagnoses without a feeling of uncertainty. Research evidence suggests the opposite for radiographers. The image evaluation test conducted by Coleman and Piper (2007) compared perceived and actual image evaluation accuracy of the participants. Perceived accuracy of the radiographers, nurse practitioners and casualty officers were 67.8%, 63.9% and 64.2%, while the actual accuracy figures were 71.5%, 52.1% and 53.8%, respectively. Although the radiographers' perceptions about image evaluation skills were more realistic than other two groups, their perceived accuracy was low. Lancaster and Hardy (2012) conducted a survey study which explored radiographers' (n = 53) attitudes to commenting. They found that radiographers generally showed a positive attitude towards PCE. However, nearly half (47.2%, n = 25) of the respondents felt they would require additional training programmes. The image evaluation test of the present study also indicated that the difficulties of clinical cases influenced the participants' confidence in image evaluation (Table 4.8 and Table 4.9). These findings may suggest that radiographers generally hold a candid view about their own image evaluation skills and they tend to take a non- ERS with a cautious approach when evaluating normal radiographs.

5.5. PCE error classification

This study was the first to explore types and frequencies of errors that may occur in PCE. PCE errors were extracted by using the PCE taxonomy (Table 3.3). The taxonomy comprised of four main classes: The gold standard (clinical reports), Decision (observers' binary decisions: normal or abnormal), Comment (present or absent) and Outcome (accuracy of the binary decisions and comments). Each class was assessed by using a decision-tree classifier. The comments were analysed to determine whether they were correct, partially correct, incorrect, ambiguous or complex. The errors were then extracted from partially correct and incorrect comments to construct the PCE error classification. The extracted errors were then organised by descending order of error frequency (Table 5.2).

Table 5.2.

PCE error classification.

Error types	Descriptions of errors	Possible causes
Over-calling	Wrongly described that abnormality is present	evaluation error
Under-calling	Wrongly described that abnormality is absent	Perceptual error/Scanning error
Incomplete	Correctly classified and identified at least one abnormality but there is one of (or combination of other errors)	One of or combination of other possible causes
Ambiguous	Description is inconclusive due to ambiguous language	Uncertainty/lack of knowledge
Faulty reasoning	Correctly classified but reasoning is wrong	Lack of knowledge/decision-making error
Miscellaneous	Minor errors	No comment, unclassifiable or discrepancy between decision and comment

Radiologists rarely over-call normal radiographs. Kim and Mansfield (2014) retrospectively reviewed a total of 1269 radiologists' errors and found that over-calling normal radiographs accounted for only 2% of the errors. The most frequent radiologists' error is under-calling abnormal radiographs (missed abnormalities). It accounts for between 60% to 80% of the errors made by radiologists (Berlin & Hendrix, 1998; Bruno et al., 2015). The present study observed differing trends. Over-calling normal radiographs was the most frequent type of error in the image evaluation test (11.50%, n = 299). This finding is perhaps predictable from the low mean specificity (67.13%) with higher false positive rate (16.44%) in the test. Nevertheless, a combination of these findings further supports the argument that the evaluation of normal images was more challenging than abnormal images for the participants.

Subtle or inconspicuous fractures may not be the cause of false negative decisions for the image evaluation test of this study since subtle fractures were intentionally excluded from the image bank (Chapter 3.5.5). Rapid reading of radiographs was a more plausible cause of the perceptual errors for the test. There is a positive correlation between faster speed of evaluation and increased errors (Skolovkaya et al., 2015). The participants of this study spent 45 minutes on 30 cases (an average of 90 seconds per case). In FRCR's rapid reporting session, radiologists spend 35 minutes on 30 cases (an average of 70 seconds per case). Despite the 10 minutes advantage over radiologists, it could be argued that the participants of this study needed more than an average of 90 seconds per case to thoroughly evaluate and provide comments for an imaging examination (typically two different views of the same area of interest).

Fixation is a retained focus on a single location. Visual information is gained during fixations (Bertram et al., 2016). A scanning error occurs when an abnormality is outside the area of interest and image observers fail to fixate their attention to the location of the abnormality (Kim & Mansfield, 2014; Pinto & Brunese, 2010). One limitation of the PCE taxonomy and error classification scheme is that comment analysis does not allow eye movement tracking, and it is therefore impossible to determine whether missed fractures are caused by scanning errors. If fixation and a scanning error are particularly concerned in PCE, an analysis of eyeball movements is a more scientifically robust method.

Incomplete, or partially correct comments, accounted for 8.04% (n = 209). Partially correct decisions have been recognised by some researchers (Coleman & Piper, 2009; Hazel et al., 2015; Piper & Paterson, 2009; Piper et al., 2005). Analysis of comments in this study also confirmed that the final decision may be expressed by a combination of correct and incorrect reasons. Overall, one implication of this finding is that a cautious approach is recommended when assessing the validity of Red-dot studies, because their binary classifier judges partially correct comments as “correct” and therefore sensitivity is positively skewed. More than half (n = 108/209, 51.67 %) of the partially correct comments were made when at least one abnormality was correctly described but another abnormality was missed. The previous section pointed out that perceptual and scanning errors are possible reasons for the missed fracture. However, satisfaction of search (SOS) should be considered as additional aetiology of missed fractures in multiple injury cases. SOS is a common error in radiology (Kim & Mansfield, 2014). Ashman et al. (2001) found that radiologists’ detection rates for the second and third abnormalities were considerably lower than the first in musculoskeletal examinations. The same result was observed for multiple injury cases of the image evaluation test. For example, one of the multiple fracture radiographs in the test bank contained anteroposterior (AP) and oblique radiographs of the left foot with fractures to the first, second and third distal phalanges. Sensitivity for this case was high (95.40%). However, scrutiny of the comments showed that 59.77% (n = 52) of the participants only identified the fracture on the first phalanx. 21.84% (n = 19) found the fractures on the first and second phalanges. 13.79% (n = 12) found all. The radiographs clearly demonstrated separation of the distal ends of the phalanges and these fractures were adjacent to each other. It is therefore plausible to suggest that the missed fractures were caused by SOS based on premature closure of visual inspection, resulting in reduced true positive rates with an accompanying reduction of false positive rates (or increased false negative rates) (Berbaum et al, 2013). Although there are several other postulated explanations for SOS (such as fatigue, severity of abnormality and faulty pattern recognition), investigation into other possible reasons of SOS in PCE may be beyond the scope of this study. However, one obvious implication of this finding is that radiographers always need to consider multiple injury cases and avoid premature closure of visual inspection.

4.98% (n = 130) of binary decisions (normal or abnormal) were correct but reasoning for the decisions was incorrect. Although this type of error was the least frequent of the errors that emerged, the finding agrees with Hardy and Culpan (2005) that correct classifications may be associated with incorrect reasons. Despite the rare occurrence, this may pose a challenge to the validity of Red-dot studies since participants’ reasons for their decisions are not recorded.

Ambiguous PCEs constituted 7.69% (n = 200) of the total PCEs. Although this study arbitrarily pre-defined ambiguous terms (Chapter 4.7.4) (and ambiguity in comments may not be regarded as errors), they occurred more frequently than faulty reasoning. Education providers should encourage their undergraduate students to steer clear of ambiguous terms for PCEs. Once qualified and placed in preceptorship or appropriate supervision, newly qualified radiographers will gain more knowledge, experience and confidence for PCE practice. With suitable education and guidelines for commenting, ambiguity error could be reduced relatively easily. Alternatively, a commenting taxonomy with controlled vocabularies and ontologies on a structured web base interface will expunge ambiguity from PCEs (Cosson & Dash, 2015). Other minor errors, including “No comment” (1.88%, n = 49), “Unclassifiable” (0.34%, n = 9) and “Discrepancy” (0.31%, n = 8) errors, were grouped as “Miscellaneous” since they only accounted for 2.53% of the total PCEs. Most of these errors were minor mistakes and eradicating them is impractical.

Chapter 3.4 discussed the absence of a gold standard specifically developed to identify and classify errors in PCE. Researchers have attempted to create different error classification systems with a wide spectrum of objectives in Radiology. However, adopting radiological error classification systems for PCE appeared illogical because they encompass broader error categories that are inapplicable to PCE. A review of literature also found some limitations of the error classification scheme in radiology. Selection bias in error classification schemes occurs when classified errors are unrepresentative of possible errors that may be encountered. It is a common limitation in some radiology error classification systems (Brook et al., 2010; Pinto & Brunese, 2010; Provenzale & Kranz, 2011; Renfrew et al., 1992). These systems used unknown criteria to select error types that entailed possible risks of introducing a selection bias or arbitrary selection and omission of errors. For example, Renfrew et al. (1992) reviewed 182 cases presented at a problem case conference and classified several error types: complacency, faulty reasoning, lack of knowledge, under-reading, poor communication, miscellaneous and complications. Provenzale and Kranz (2011) pointed out that this classification omits two possible error mechanisms: under-calling and lack of knowledge of study limitations. Brook et al. (2010) also argued that the error classification did not fully take account of some latent conditions (systemic failures) such as work volume and understaffing, which The Royal College of Radiologists (2014) acknowledged as major and ongoing difficulties in NHS Radiology departments. The PCE classification system of this study depended on the PCE taxonomy (Chapter 3.3 and Table 3.4) that incorporates all the theoretically attainable PCE outcomes. The taxonomy served as the reference to systematically identify and classify PCE errors which minimised the risks of a selection bias or arbitrary selection and omission of errors.

Latent conditions (failures in department, management and equipment) must be acknowledged as sources of errors (Brook et al., 2010). However, including latent errors into classification systems seems to result in a complex classification design. For example, the classification systems developed by Graber et al. (2006) and Taylor et al. (2011) included 38 and 18 error categories respectively. Comprehensiveness is vital for discriminating and classifying a wide range of possible errors. Nonetheless, an exceedingly comprehensive and complex classification design will pose a risk of reduced reliability because the more choices are available to error classifiers, the more diverse results of the classification may become.

This will result in lower inter and intra-reliability. Low usability (greater difficulty of use) is also a known problem for human error identification tools (Shorrock & Kirwan, 2002). Overly complex error classification that encompasses both human and latent errors may challenge the usability and other criteria that need to be satisfied. Individual practitioners have little control over latent errors and these errors must be addressed at a departmental level. A holistic approach to quality management that focuses on both human and latent errors in PCE is essential for maintaining service quality. Despite this, dealing with the latent errors is beyond the clinical responsibilities of individual radiographers.

Several benefits were found for the use of the PCE taxonomy (Table 3.4) and the resulting PCE error classification system:

- Comprehensiveness: The taxonomy, which incorporates all the theoretically attainable PCE outcomes, enabled a comprehensive classification of PCE errors.
- Structure: The taxonomy specifically targets evaluation errors in PCE. Although the size of the taxonomy is large, the resultant classification structure is simple (six error categories).
- Consistency: Good consistency is expected because classification of comments relies on classifiers' objective judgement with a decision-tree concept.
- Predictive accuracy: All the potential PCE errors are predicted by the taxonomy, therefore the predictability of errors is high.
- Training requirement: There is little training requirement for anyone who has a basic understanding of image evaluation practice and associated terms. Radiographers

should be able to use the taxonomy to classify errors in audit studies with minimum training.

- Auditability: Considering the benefits listed above, the taxonomy and classification technique are suitable for auditable documentation.

Error classifiers must assess each class of the taxonomy and glean errors by scrutinising each comment. The amount of time needed to extract errors is one limitation of the error classification technique. Another limitation is that classification systems cannot directly contribute to a reduction of PCE errors. Although the system delineates error types and frequencies, and suggests possible causes of them, it depends on educational interventions for actual error reduction.

5.6. PCE comment analysis: FRCR and WWH scoring systems

PCE is twofold. In a Red-Dot system, the only measurable performance standard is radiographers' ability to identify or rule out abnormalities. In PCE, the task goes beyond the simple red-dot (RADS) style decision making. PCE further requires radiographers to coherently articulate radiographic appearances after viewing radiographs that they have taken. However, the results of several studies (Hardy & Culpan,2007; Neep et al., 2014; Smith et al., 2009) and the present study demonstrated that radiographers maintain less confidence and accuracy for providing comments than making Red-dot style decisions.

Descriptive skills are of central importance. If physicians' clinical decisions partly depend on PCEs until final clinical reports are available, descriptive quality of PCEs must follow similar quality standards to those which radiologists and reporting radiographers strive for. Several studies have delved into radiologists' reporting styles and physicians' preference for them. In spite of radiologists' propensity for free-form reporting, physicians prefer structured reporting (Bosmans et al., 2011; Grieve & Khan, 2009; Schawartz et al. 2011). However, there is still a dearth of information to determine whether radiographers and radiography students follow certain templates or style guidelines for PCE.

"WHAT, WHERE and HOW" framework (Harcus & Wright, 2014) (Appendix K) breaks down three important components of radiographic appearances: type of abnormality (WHAT), location (WHERE), and presence or absence of separation/angulation (HOW). The format of this framework conceptually resembles a tabular report that allows image observers to focus on the image appearances rather than descriptive writing styles. Based on this concept, the WWH scoring system (Akimoto et al., 2016) was developed as a tool to quantitatively determine quality of PCEs, especially for abnormal images (Table 3.7). Literature suggests that quantification of PCE (comments) is a newly emerged research interest (Neep et al., 2017; Stevens & Thomas, 2018).

Chapter 4.8 presented a comparison of FRCR and WWH scores for the image evaluation test. The final examination for the Fellowship in Clinical Radiology consists of three parts: rapid reporting, reporting and orals. The scoring system of FRCR's rapid reading session does not assess the quality of comments. The reason for this is not explicitly clarified by the RCR. The RCR explains that the rapid reporting session reflects the radiologists' tasks in an A&E setting and tests their ability to rapidly decide whether an image is normal or abnormal (RCR, n.d.). It appears that a more detailed focus is given in the reporting session where responses were categorised into "No answer offered", "Fail", "Borderline", "Pass", "Good Pass" or "Excellent" and marks allocated accordingly. Unlike Red-dot style assessments, the scoring system of FRCR rapid reading session acknowledges correctly classified decisions with faulty reasoning, and therefore records zero mark. However, the scoring system lacks the analytical power to detect partially correct comments. It also neglects descriptive quality. WWH scoring system takes a markedly different approach. In WWH, once abnormalities are correctly identified, it continues to examine whether image observers articulate the abnormality type, location and presence of dislocation or angulation. WWH does not only assess the observers' ability to identify abnormalities, but it also evaluates other essential elements of PCEs.

The results of WWH scoring demonstrated that the mean scores for normal and abnormal images were 67.13% and 30.91% respectively. The mean score for abnormal images was considerably lower than the mean score for normal images. This is because three marks are automatically recorded when normal images are correctly classified and described likewise. On the other hand, three marks for an abnormal image are immediately lost when abnormality is missed. It appeared that earning full marks for an abnormal image was a challenging task because the participants needed to satisfy many pre-defined evaluation criteria for each image (Appendix L). For example, only one mark is given to a simple comment “fracture (+0.5) on the first metacarpal (+0.5)”, while full marks are given to a more detailed comment “Oblique (+0.5) fracture (+0.5) on the base (+0.5) of the first metacarpal (+0.5) with minimum (+0.5) dorsal (+0.5) angulation”.

The mean scores of the three categories (WHAT/WHERE/HOW) remained below 50% of the maximum score (45 points) (Table 4.16), suggesting that precision and completeness of written descriptions were less than anticipated. The score for HOW was particularly low (M = 1.50, 10.00%). Further break down of the scores demonstrated that, even when the participants correctly identified abnormalities, they only achieved an average of 0.16 points per image (maximum of one point is achievable) for HOW. More than one third of the participants (34.48%, n = 30) lost a full 15 points because they never described the severity of angulation or dislocation for the test. This clearly displays that the participants gave very little attention to the angulation or dislocation of the abnormalities. The finding may suggest a radiographers’ lack of understanding about the significance of angulation and dislocation in fracture management. This is probably because radiographers traditionally took a proactive role in image acquisition rather than care and management of fractures. It could be argued that the current undergraduate education providers do not instruct their students to assess and describe alignment of bones sufficiently. Much higher average points for WHAT and WHERE per image were observed: 0.64 and 0.70 points respectively. The figures generally indicate that, when the participants identified fractures, some tended to give detailed descriptions (e.g., “an oblique fracture to the proximal epiphysis of the first metacarpal bone”, rather than “a fracture to the first metacarpal bone”). PCE could influence physicians’ decisions until official clinical reports are available. Radiographers are now, although indirectly, involved in care and treatment of trauma patients though PCE. Therefore, radiographers should articulate their clinical decisions in a precise and complete form. WHAT, WHERE and HOW concept provides an ideal commenting template that ensures that PCE encompasses necessary clinical information in a proper structure.

One limitation of the scoring system is its limited flexibility. Each abnormal image requires abnormality specific evaluation criteria for each WHAT, WHERE and HOW category. This method allows subjective quantification of written descriptions of radiographic appearances. However, the types and numbers of criteria (and how points are allocated to them) may be arbitrarily decided and the same criteria cannot be used for different image banks.

5.7. Summary

Since the late 1990s, the SCoR's position has been consistent that image evaluation should be a core clinical practice of diagnostic radiographers. In 2013, the SCoR elaborated on their expectation further that newly qualified radiographers have the necessary education and training to take part in PCE. However, there was no evidence to support the SCoR's assumption. Transition from the Red-dot system to PCE requires, not only the ability to provide reliable clinical decisions, but also the skills to precisely articulate radiographic findings in a written form. Therefore, this study benchmarked image evaluation performance of the final year radiography students at the point of graduation.

The test results demonstrated that the participants' image evaluation performance fell short of the ideal level (90% sensitivity and specificity). Specificity was particularly low, which is consistent with the results of other studies reviewed in Chapter 2. This suggests that normal images posed more cognitive challenges to the participants than abnormal images. Analysis of decision-making confidence also supports this finding that the participants evaluated normal images with less confidence and added caution. The error classification scheme found that the most frequent PCE error was false positives. This further assists the theory that participants tended to err on the side of caution while evaluating normal images, resulting in a high rate of false positive decisions with low specificity. The WWH scoring system found substantially lower HOW scores than WHAT and WHERE scores, indicating that many of the participants inadequately addressed the presence or absence of angulation/dislocation of fractures. This study recommends the use of commenting guidelines or templates, such as the WHAT, WHERE and HOW framework, to enable the systematic provision of structured comments that encompass necessary clinical information.

Is PCE by newly qualified radiographers feasible? The test results indicated that immediate participation at the point of qualification is questionable in the current state Radical education reform specifically for PCE is impractical. However, a long-term implication of the study finding is that undergraduate HEIs must continue to devote sufficient academic effort so that their students, at the point of graduation, are equipped with 90% sensitivity and specificity in local A&E settings. Hardy and Culpan (2007) predicted that most imaging departments, that considered introduction of PCE into practice, would depend on experienced A&E radiographers to set the practice in motion. It is perhaps more reasonable to suggest that departments considering PCE implementation provide learning opportunities to less experienced radiographers until they achieve 90% sensitivity and specificity (or locally established performance standard), while more experienced radiographers take the lead in PCE from the outset. Preceptorship (or alternative forms of clinical supervision) is therefore a valuable transitional phase for new graduates to consolidate their knowledge for the forthcoming professional roles in PCE (Stevens & Thompson, 2018). This study therefore recommends intense PCE training and audit of image evaluation performance during the preceptorship (discussed in Chapter 6.2.7). The next chapter presents reflections on the research and recommendations for future studies.

Chapter 6. Reflections and recommendations for future studies

6.1. Introduction

This chapter aims to reflect upon the research and put forward recommendations for future image evaluation studies. The purpose of reflection is to identify what is already known and formulate new knowledge (Moon, 2004). Although there is a myriad of definitions, most interpret reflection as a mental process of analytic strategies to solve problems and create meaning (Roessger, 2014). Reflection is a vital stage of learning from experience and developing one's competencies (Paterson & Chapman, 2013). Leijen, Valtna, Leijen and Pedaste (2012) distinguished four different hierarchical levels of reflection: description (descriptive information), justification (logic or rationale), critique (explanation and evaluation) and discussion (incorporating all lower levels of reflection). Leijen et al. (2012) explained that reflection could be summarised with two metaphorical terms: deepening and broadening. Deepening ensures thorough reflection through consecutive stages: describing, justifying, evaluating and discussing, while broadening encompasses the transition of reflection to a wider social context.

Research into PCE must continue. Research evidence is necessary to promote successful implementation of PCE. The review of literature (Chapter 2) found a paucity of research evidence since 2013 to determine the feasibility of PCE by radiographers. This chapter reflects upon this research and makes recommendations to promote better research models for future image evaluation studies.

6.2. Reflections on the research and recommendations for future studies

6.2.1. Participant recruitment

Participant recruitment appears one of the methodological challenges for image evaluation studies. The literature review consistently found that a small sample size was a common limitation in most of the studies. This is because most studies were conducted at one or two research sites (hospitals) and the number of radiographers that could be recruited was small. This study targeted multiple institutes and achieved a larger sample size than most of other image evaluation studies. Nearly half of the undergraduate course leaders in England and Wales agreed to participate in this study. The high participation rate may have indicated their positive attitude toward image evaluation education. However, geographical barriers were potential limitations of this multicentre sampling model, since the PI needed to visit multiple study sites to supervise the tests. Long distance travel and coordination of tests at multiple study sites posed methodological challenges.

Development of an online platform or software to allow image evaluation tests may become a future research agenda. Such an online testing platform would alleviate the geographical barriers for both investigators and participants, although it would not control for environmental factors. It would also allow concurrent testing of multiple cohorts within tighter timescales than was possible in this study. Participants could also take tests at any time if appropriate devices (e.g., a computer and a monitor with sufficient size and resolution) are available. A digital library of radiographs could be incorporated into the platform to allow more consistent research methods, which will further promote meaningful meta-analysis of image evaluation studies.

6.2.2. Timing of sampling data

The potential participants of the present study were narrowly defined as the final undergraduate diagnostic radiography students at the point of graduation/qualification. The time frame for the image evaluation tests (sampling) was therefore scheduled between April and June in 2015 at the collaborating institutes. However, it appeared that students were in clinical placements and many were absent from the study sites, which might have resulted in the low response rates. However, earlier timing of the tests was inappropriate as the students would not have been at the point of qualification. Timing of sampling could be extended (e.g., six months before the graduation) if drastic improvement in image evaluation performance is not expected at the end of the final year.

6.2.3. Development of X-ray image bank

The prevalence of abnormality in the image bank was 50% as recommended by Piper et al. (2004) (Chapter 3.4.5). There are mixed views about the appropriate prevalence of abnormality in image banks. Hardy et al. (2014) argued that constructing a clinical workload image bank (prevalence of 20-30% abnormality) that reflected local image profile was a better approach. However, the validity of such low prevalence test banks should be questioned. For example, considering that typical image banks contain around 30 radiographs, only six abnormal images would need to be used to measure sensitivity if the prevalence of abnormality is 20%. Validity of image evaluation tests (ability to truly measure sensitivity) with such low prevalence of abnormality and a severely limited range of anatomical parts is questionable (e.g., six abnormal images cannot examine the full range of anatomical parts in the skeletal system). A statistically reasonable number of abnormal images could be used, but this results in a larger image bank size and a longer test duration. For example, McConnell and Baird (2017) aimed to develop an image bank that reflected local image profile with 95% confidence and 0.05 precision in their Red-dot study. The image bank contained 209 images: 148 normal (70.8%) and 61 abnormal (29.2%) images. An adequate number of abnormal images were selected to represent the average clinical load (anatomical areas, prevalence of normal and abnormal cases, and ratio of adult and paediatric patients). However, a limitation of this Red-dot study was a possibly prolonged test duration. Considering that the participants spent an average of 60 seconds on each image (without the time required for describing findings), the test required approximately 3.5 hours to complete. The radiographer participation rate of McConnell and Baird (2017) was 4% (n = 16). They noted that the reason for this was due to the time required for the test.

In this present study, the selection of X-ray images in the test bank was design to encompass the entire appendicular skeleton. Consistent with other benchmarking tests, the test image bank was not designed to reflect clinical workload, where the prevalence of abnormality is often not 50%. This was necessary for the equal statistical assessment of sensitivity and specificity (Chapter 3.5.5).

Brown and Leschke (2012) found that radiographers' ability to identify subtle fractures (displacement or distraction < 1mm) sharply drops (Chapter 2.6.1). The test bank of this present study did not contain abnormal images with subtle fractures that may pose great difficulties (Chapter 3.5.5). Inclusion of some subtle fracture images could have provided additional values to the study findings.

6.2.4. Measurement and analysis of X-ray image evaluation competencies

Traditionally, image evaluation studies adopted a quantitative approach and this study was no exception. The measurement of image observers' performance is relatively simple and should pose few methodological difficulties. Image observers' decisions are classified using the binary classifier and subsequently calculate accuracy, sensitivity and specificity. The use of computers is an obvious advantage in image evaluation studies since results are presented in a quantitative term. Chapter 3.7.2 explained that this study used two computer software packages, Microsoft Excel and SPSS. These programmes were extensively used for summarising literature (Chapter 2), developing the data processing tools (Chapter 3), presenting the test results (Chapter 4) and discussing statistical significance of the findings (Chapter 5) without a major complication. It appeared that the use of spreadsheet and statistical analysis software would generally suffice to manage and process data from the test. Drawbacks of using computers and software packages may be the cost and time to learn how to use them, although these posed little methodological challenge in this study.

Chapter 3.7.1 explained that inferential statistics were used to examine the correlation between demographic characteristics and image evaluation skills of the participants. The most noteworthy finding was the statistically significant difference of specificity between Uni. A and D, and thus pointed out a possible lack of educational emphasis on evaluating normal images at Uni. A (Chapter 5.3). The results of inferential statistics also indicated, for example, that the female participants took a more cautious approach when evaluating abnormal images than the male participants (p = .016) or the participants with estimated 1st class degree demonstrated better WWH scores than the rest of the group (p = .003). Gender-based differences in image perception and expression of decision-making confidence may be an interesting research topic. However, discussions of these statistically significant differences were omitted because the findings were deemed to be irrelevant to the research aim and question of this study. Many of the reviewed image evaluation studies (Chapter 2) and this present study focused on statistically significant findings. This does not suggest that statistically non-significant results carry no meaning or impact. Perhaps, the flexibility in a post-positivism approach (Chapter 3.2) could allow additional insight into findings without statistical significance to explore future research agenda. However, this was not considered in this study in order to concentrate on research topics that are relevant to the aim and objectives (Chapter 1).

One methodological challenge was the qualitative component (comments) of the test results. The comments for abnormal images were scrutinised and quantified by using the WWH scoring system and PCE taxonomy. This process appeared to be simple but also prone to researcher bias owing to its monotonous and laborious nature. The quantification of the comments had to be performed twice for the scoring and error classification to ensure that the results were accurate and consistent. Several other authors have published different scoring systems for PCE since the start of this project (Neep et al., 2017; Stevens & Thompson, 2018), perhaps suggesting that quantification of comments is a newly expanding research field.

6.2.5. Validity and reliability of the study

Neep et al. (2017) argued that image evaluation studies provide little research information to determine reliability and validity of research results. The literature review of this present study also found that the reviewed studies hardly discussed the reliability and validity of their own study results. However, image evaluation research must determine (or at least consider) reliability and validity of findings in order to support radiographers' clinical practice with persuasive research evidence.

Reliability in image evaluation studies means the ability to produce consistent results more than once (stable performance of sensitivity, specificity and accuracy). Chapter 3.5.5 explained that improved reproducibility of test results is expected when using image banks under controlled conditions. Test-retest reliability of this present study is expected to be high if the same sample is tested. On the other hand, inter-rater reliability would vary because participants' knowledge and experience in image evaluation could greatly alter the outcomes. Different levels of image evaluation performance are foreseeable when two markedly different groups of participants take the same test (e.g., first year undergraduate diagnostic radiography students vs. qualified reporting radiographers). However, inter-rater validity of this study could be high because the study specifically targeted the final year radiography students at the point of graduation and their education and clinical experience may not diverge vastly every year.

Validity in image evaluation studies signifies the ability to measure true performance of image observers. Studies with test banks that do not reflect clinical cases in A&E settings may not provide a true reflection of radiographers' performance in daily practice. Chapter 6.2.3 discussed that this was a possible threat to the internal validity of this present study. An audit study with real clinical cases may have the potential to provide better internal validity especially when high participation rate of radiographers' in image evaluation practice. The use of a red-dot style binary classifier in this study could also have affected the validity of the findings (discussed further in Chapter 6.3). Chapter 2.7.1 and Chapter 5.2 pointed out that small sample size is a typical methodological limitation in image evaluation studies, and explained that generalisability (external validity) of the study results must be interpreted with caution. In this study, the participants accounted for 9.20% (n = 87) of the whole final year diagnostic radiography students in the England and Wales in 2014/2015 academic year (Chapter 4.3). The sample size appears too small to consider that the participants represented the target population.

6.2.6. Interview questionnaire with course leaders

Chapter 3.5 explained that an interview questionnaire was developed to add more analytic depth. This study could not incorporate the results of the questionnaire with other quantitative components due to a lack of necessary information to complete the analysis. There are several reasons for this discarded element of the study work. Some questions in the questionnaire asked detailed information (e.g., precise number of credits and hours allocated to X-ray image evaluation modules) that some course leaders could not immediately answer. At many study sites, course leaders were not responsible for (or directly involved in) the education of X-ray image evaluation. These course leaders suggested that lecturers who delivered image evaluation education were more likely to be able to provide the necessary information. The interview was also conducted before the X-ray evaluation test within a fairly short time frame at each HEI. It was therefore unreasonable to expect that the course leaders could answer all questions at the time of the interview without referring to course syllabi or curricula. Email of the questionnaire prior to the image evaluation test might have been a more sensible approach as the questionnaire could have been handed to the most appropriate person which would have allowed enough time to elicit information. A follow-up questionnaire with modified questions was conducted by e-mail. However, this did not resolve the missing information because of a low response rate (seven out of nine HEIs did not respond). An analysis of incomplete data was deemed to be misleading, and therefore a discussion of the interview questionnaire was omitted. A complete set of data from the questionnaire with other quantitative components could have provided more meaningful research findings.

Despite its incompleteness, the elicited partial information from the collaborating HEIs could update the current knowledge about image evaluation education. According to the survey by Hardy and Snaith (2009), image evaluation of appendicular skeleton was taught at all the responding HEIs, although provision of education for the axial skeleton, chest and abdomen was inconsistent across the UK. However, the questionnaire of this present study found that all the responding HEIs (n = 9) delivered education for both appendicular and axial skeletal systems. Image evaluation of chest and abdomen was also taught at all the HEIs except for one. Moreover, evaluation of CT head and contrast examinations (e.g., barium enema) were also taught at one or two institutes. This is a fair indication that education of image evaluation has been expanding since 2009. A larger and more robust survey/questionnaire is recommended to officially update the work by Hardy and Snaith (2009).

6.2.7. Potential values of preceptorship

This study has questioned the feasibility of immediate participation in PCE by newly qualified radiographers (Chapter 5.7). However, it does not entirely preclude them from PCE practice. The educational value of preceptorship must be acknowledged. Preceptorship is a period for newly qualified radiographers to "consolidate knowledge (educative), to be induced into the policies and procedures of the workplace (normative) and to reflect on their practice, especially on challenging experience (restorative)" (SCoR, 2003).

SCoR (2013) acknowledges its value for newly qualified radiographers before taking part in PCE. Stevens and Thompson (2017) evaluated the impacts of focused training on image evaluation skills of radiographers (n = 4) who were in a preceptorship period. The radiographers demonstrated a statistically significant improvement of abnormality detection rate from 42% (pre-training) to 56% (post-training). The results also showed 50% reduction in false positive errors for normal images. Despite the low abnormality detection rate, the findings indicate that radiographers shortly after qualification still improve their image evaluation skills when a preceptorship programme incorporates appropriate training. Radiographers’ participation in PCE after the preceptorship may be a more pragmatic approach if the current undergraduate education does not satisfy the vision of the SCoR. However, the general benefits of the preceptorship in diagnostic radiography have not been fully understood, owing to a lack of research and documentation (Chapter 1). Nisbet (2008) documented and published the development of a preceptorship programme for therapeutic radiographers and suggested various pedagogical strategies. Unfortunately, the evaluation of the programme was not conducted prior to the publication. Literature suggests there has been no other research or published record to indicate the general benefits of the preceptorship in diagnostic radiography. On the other hand, research has identified many benefits of the preceptorship in medicine, nursing and other allied healthcare professions (Billay & Myrick, 2008). The primary advantage is the opportunity for new graduates to practice skills under close supervision by clinical experts (Tan et al., 2011). There are many other known benefits including: application of theoretical knowledge to clinical situations, development of communication, clinical and problem-solving skills, alleviation of mental distress as well as retention and utilisation of future workforce (Billay & Myrick, 2008;

Marks-Maran et al., 2013; Nielsen et al., 2017; Quek & Shorey, 2018; Tan et al., 2011). Similar benefits are conceivable for newly qualified radiographers or anyone who is considered as a preceptee, although more research is needed to confirm this.

In-depth discussion about the development of the preceptorship in diagnostic radiography is beyond the focus of this study. However, the principal objective of the preceptorship in the context of PCE is the provision of reliable decisions (e.g., above 90% sensitivity and specificity). Research has consistently found that radiographers improve their image evaluation skills after educational interventions (Chapter 2.7.3). Adequately constructed preceptorship will provide ample educational opportunities for preceptees to enhance their image evaluation skills.

A team of preceptors must include reporting radiogaphers. They could share their professional knowledge and work experience in image evaluation with new practitioners. Reporting radiogaphers should play a proactive role in teaching image evaluation skills and allowing preceptees to identify their own learning needs. Chapter 5.3 pointed out the possible lack of emphasis on evaluating normal images at the undergraduate education. Preceptors, particularly reporting radiographers, are in an advantageous position to utilise the clinical image profile (estimated 70-80% prevalence of normality) to reinforce preceptees' skills to evaluate normal images, and thus enhance their specificity with higher decision-making confidence. In the context of image evaluation and PCE, preceptors must establish goals and ensure that preceptees attain the goals within a specified time frame (typical preceptorship length is between six to twelve months).

Monitoring of performance may be useful to assess whether preceptees have achieved the goal. Two clinical reporting studies by Carter and Manning (1999) and Kumar (2007) demonstrated that monitoring of performance allows close observation of the changes in reporting accuracy for individual radiographers. Carter and Manning (1999) also pointed out that monitoring radiographer performance would allow developers of training courses to identify the effect of training and learning activities that enhanced competencies. They made four recommendations that are expected to improve radiographers' performance in training programmes:

1) attendance at radiologists reporting sessions to understand appropriate wording and reporting structures,

2) modification of the report writing to a more concise style,

3) discussion of images and search strategies and revision of assessment, and

4) the use of learning materials to introduce normal variants.

The recommendations were made by the monitoring of a postgraduate radiographer in a training programme for clinical reporting. However, the recommendations could be applicable to preceptees who are preparing to take part in PCE. Chapter 5.6 pointed out that radiogaphers and radiography students do not seem to follow certain writing styles to articulate radiographic findings. Preceptors could introduce appropriate terminologies and a precise writing style (recommendation 1 and 2). Preceptees are expected to view X-ray images under close supervision (by reporting radiographers) which can promote more interactive and reflective learning than the conventional classroom learning (recommendation 3). The literature review (Chapter 2) and the results of this present study (Chapter 5.3) found that radiographers' specificity is generally lower than sensitivity. Preceptees will benefit from viewing a large volume of images, which consolidates their cognitive libraries (Hardy et al., 2016), if preceptors develop image banks of clinically challenging cases that may not be sufficiently addressed in daily clinical image load (e.g., rare appearances of normal variants) (recommendation 4). Quantitative monitoring of performance (sensitivity and specificity) is ideal, although there are some possible complications. Monitoring could be lengthy. Carter and Manning (1999) and Kumar (2007) spent nine weeks and 10 months respectively to complete their studies. Preceptorship developers may question about allocating a long time-frame solely for monitoring of image evaluation or PCE performance, while preceptees must accustom themselves to other areas of practice, such as CT, MRI, mobile and theatre radiography. Quantitative monitoring could cause mental stress to preceptees if their performance must be regularly evaluated (e.g., weekly audit or test to quantitatively monitor the progress). The monitoring process could also impose tighter constraints on preceptors' work time. Daily or weekly reflection may be a more reasonable option to evaluate the progress of preceptees. Preceptors and preceptees could discuss and reflect on the learned experiences and skills, and subsequently identify next learning objectives. Nisbet (2008) recommended that each preceptee develops a personal portfolio to record and reflect on clinical experience. Stevens and Thompson

(2017) highlighted that radiographers immediately after their qualification still improve their image evaluation skills. Foreseeable implications of their findings may be that radiographers who completed well developed preceptorship could confidently provide reliable decisions for PCE which ultimately result in improved patient management.

Preceptors must fulfil their responsibilities along with their usual clinical duties. Their commitment with respect to time and effort to develop preceptorship programme and supervising preceptees might become burdensome. However, preceptors will be rewarded in return for their dedication to assist new practitioners. Studies in nursing have found that preceptors increase their knowledge base, find personal satisfaction, improve leadership/organisational skills, expand awareness toward professionalism and gain a perception of their contribution to the profession and being recognised as role models (Cloete & Jeggels, 2014; Usher et al., 1999). Preceptorship is not only a complex interplay between preceptors and preceptees. Departmental support is therefore essential for both preceptors and preceptees to optimise teaching/learning conditions.

The SCoR (2013) holds a view that image evaluation or PCE will be a core competence of diagnostic radiographers. Nisbet (2008) maintained that departments could submit their preceptorship programmes to the SCoR for their accreditation and validation so that the programmes would maintain acceptable standard. The SCoR could provide a specific guideline for the preceptorship in order to encourage more active research and documentation; or, alternatively, they could grant autonomy to individual departments, which will encourage the development of unique preceptorship programmes.

6.3. Limitations of the study

The sample size was a limitation of this study, even though a relatively larger sample than most of other image evaluation studies was recruited. The participants were restricted to the final year radiography students from England and Wales. Nearly half of the HEIs offering undergraduate radiography programmes in England and Wales agreed to participate in this study. Response rate of 50% may generally appear to be high in health and social care research. However, the rest of the half of the contacted HEIs opted not to take part and this restricted the geographical coverage of this study. Therefore, it is unrealistic to consider that the participants of this study represented the final year radiography students in the UK.

The radiographs in the image bank were presented in Microsoft PowerPoint. This viewing condition may not replicate typical digital imaging workstations (e.g., high resolution monitors with tools to manipulate window level or depth settings) at which students normally view radiographs in departmental settings. Chapter 3.5.7 explained that the present study could not consistently control viewing conditions (e.g., darkness in the test rooms and resolution of the displays) which could potentially affect the participants' image evaluation performance. Literature suggests that there are still conflicting views and research results of how differing viewing conditions affect the performance of image observers. However, it is possible that images presented in Microsoft PowerPoint in dissimilar test environments could have affected the results of the image evaluation tests. Chapter 6.2.3 also discussed that the X-ray image bank that did not reflect typical clinical workload is a limitation of this study.

A red-dot style binary classifier was used to calculate the image evaluation performance of the participants. One limitation is that calculation of sensitivity ignores whether decisions were made based on correctly identified locations of abnormalities. Hardy and Culpan (2007) (and this study) pointed out that radiographers may arrive at correct red-dot style decisions with wrong reasons (e.g., a combination of false negative and false positive decisions for abnormal images). It is therefore likely that the binary classifier overestimated the sensitivity of the participants.

6.4. Summary

This chapter reflected on the research process. The research process required continuous identification of methodological threats and their modification or exclusion to achieve better validity and reliability of the results. It also seemed that anyone exploring radiographers' performance in image evaluation needs to possess, not only the ability to conduct evaluation tests, but also the capacity to search databases for extensive literature acquisition and summarise previous studies, consider ethical problems, develop statistical strategies, establish evidence and deduce new theories. The reflections in this chapter established three pillars towards which scrupulous attention must be directed:

1) Method of participant recruitment
2) Development of image banks
3) Measurement and analysis of PCE performance

Timing of sampling is also important if research is to specifically target a certain group of radiographers, such as newly qualified radiographers.

Limitations of this research were also acknowledged. A small sample was one limitation of this study, whilst still being larger than that of comparable studies. This study (and other reviewed studies) indicated that a small sample size may be an inherent limitation of any image evaluation studies. Another limitation was the varying viewing and environmental conditions during the tests at nine collaborating HEIs that may have differently affected the participants' evaluation performance. Finally, methodological concern for using a red-dot style decision classifier (which may overestimate sensitivity) in PCE studies was raised and therefore adoption or development of a more location-sensitive study model was recommended. The next chapter summarises the key findings of this study, contributions to knowledge and implications for practice. Potential areas of future research are also explored.

Chapter 7. Conclusion

7.1. Introduction

This chapter summarises the overarching themes that emerged in this study. This chapter also accentuates this study's contributions to the current knowledge, implications for practice and recommendation for future studies.

7.2. Conclusion of this study

This study aimed to benchmark new graduate radiographers' competencies in evaluation of plain appendicular X-ray images. The SCoR (2013) holds an expectation that new graduates of diagnostic radiography could begin PCE at the point of qualification. However, the literature review of this study found no empirical evidence to support the SCoR's vision since 2013 (Chapter 2.7.4). This study therefore evaluated PCE performance of the final year diagnostic radiography students at the point of graduation in order to determine the feasibility of implementation of the SCoR strategy.

The research question of this study was "What is the image evaluation performance of diagnostic radiography graduates relative to benchmarking standards?" (Chapter 2.7.4). This study was the first to benchmark PCE performance of the final year diagnostic radiography students at the point of graduation and qualification with HCPC (first objective of this study, Chapter 1 on p. 29). The results of this study provide the initial insight into newly qualified radiographers' competencies in PCE. The results of the X-ray image evaluation test showed that the participants' mean sensitivity and specificity were 79.62% and 67.13% respectively (Chapter 4.5.2, Table 4.5). This study established that 90% sensitivity and specificity are ideal performance standards for radiographers who are taking part in PCE (Chapter 5.3). The test results indicated that the participants of this study did not achieve the ideal PCE performance standards at the point of graduation and qualification with HCPC. The test also found that the mean specificity was considerably lower than sensitivity. This finding is consistent with other reviewed image evaluation studies that radiogaphers' specificity is generally lower than sensitivity (Chapter 2.6, Table 2.6 and Table 2.8). The difficulty in the generalisability of the findings was acknowledged (Chapter 6.3). Nevertheless, this study concludes that the SCoR's prospect of PCE by newly qualified radiographers may be implausible in the current state.

Literature suggests a recent increasing research interest in the quantification of descriptive performance in image evaluation (Neep et al., 2017; Stevens & Thompson, 2018). Neep et al. (2017) pointed out that an image evaluation test must evaluate not only the ability to detect but also the ability to describe the presence or absence of abnormalities. This study established the WWH scoring system to evaluate written description of PCE (Chapter 3.7.5). This was the first attempt to quantify radiographic comment quality (second objective of this study, p. 29) without depending on a binary classifier and subsequent calculation of accuracy, sensitivity and specificity, thus introducing a new evaluation model into PCE benchmarking. The results of the scoring system could exemplify the current descriptive skills of newly qualified radiographers. The most prominent finding from the analysis was that the participants very rarely paid attention to the extent of dislocation and angulation when they detected abnormality (Chapter 5.6). This was a common trend observed in all the collaborating HEIs.

This study added an empirically assembled PCE error classification to the current sparse evidence relating to radiographers' PCE errors (third objective of this study, p. 29). The most frequently made PCE error was false positive decisions. Although this finding was predictable from quantitatively obtained low specificity, qualitative scrutiny of the comments also suggested that the participants tended to raise red flags when in doubt (Chapter 5.5). The error classification system also delved into "grey zone comments". Researchers in image evaluation studies have acknowledged that correct image classifications could be made by partially correct decisions or wrong reasoning. These outcomes of radiograph evaluation fall in the grey zone. Dichotomous classification of clinical decisions is therefore imprecise and image evaluation studies should not dismiss this methodological concern. This study found that partially correct PCEs accounted for 8.04% (n = 209) of the total PCEs in the image evaluation test (Chapter 4.7.3, Table 4.13), indicating that quantitatively calculated sensitivity may not be the true reflection of the ability to identify abnormalities.

More than 10 years have elapsed since researchers expended the last effort to synthesise the evidence on radiographers' competencies in the Red-dot system (Brealey et al., 2005) and clinical reporting (Brealey et al., 2006). The literature review of the present study filled this knowledge gap (fourth objective of this study, p. 29). The SCoR's 2013 document explicitly introduced the definition of PCE and drew clear distinctions between three tiers of decision-making practice by radiographers: Red-dot system (RADS), PCE and clinical reporting. With the new addition of PCE, this study was the first to delineate and review separately the literature of image evaluation studies (the Red-dot system and PCE). The review properly categorised image evaluation studies that had not been formally acknowledged as PCE studies prior to 2013 (Chapter 2.6.2). The SCoR (2013) expects that PCE will gradually replace the role of the Red-dot system. The literature review found possible declining research interests in RADS since 2006 in the UK (Chapter 2.7.4). On the other hand, there is now tangible research evidence to indicate that reporting radiographers provide clinical reports with a high degree of accuracy that is favourably comparable with radiologists (Chapter 1). This inevitably means that PCE will become the central theme of future image evaluation studies.

7.3. Contributions to knowledge

This section briefly summarises the contributions to the current knowledge.

1. This study benchmarked PCE performance of the final year undergraduate diagnostic radiography students at the point of graduation and qualification with HCPC (Chapter 3.5.7). The results of the benchmarking could suggest the current PCE performance standard of the final year students (Chapter 4.5), although the participants of this study may not be the epitome of other students in England and Wales.

2. This study developed an evidence-based error classification scheme (Chapter 3.4) to highlight PCE error types and frequencies (Chapter 4.7).
3. This study developed a unique scoring model (Chapter 3.7.5) to allow quantification of qualitative components (PCE comments) as part of a benchmarking process (Chapter 4.8).
4. This study updated the previous literature reviews of the Red-dot system (Brealey et al., 2005) and added a new PCE literature review (Chapter 2.6). This first PCE literature review could be perceived as a baseline for the future PCE studies.

7.3. Implications for practice

1. Benchmarking competence is key to establishing image evaluation practice within the HCPC standards of proficiency for diagnostic radiographers and fulfilling the SCoR 2013 professional vision. The scoring model and error classification system developed as part of this research are ideally suited to assessment at any learning stages.
2. Radiography graduates are unlikely to meet the benchmark standards required for accurate and reliable participation in PCE schemes. This finding has two key implications:

   a) Universities may need to review their undergraduate education and work in partnership with clinical placement sites to ensure that students develop image evaluation skills throughout the course and graduate with higher level performance to meet benchmarking standards.

b) PCE needs to become a key component of preceptorship. The potential values of preceptorship in the context of PCE were considered in Chapter 6.2.7. A small-sample study by Stevens and Thompson (2017) indicated that newly qualified radiographers continue to improve their image evaluation skills after their qualification. More research and documentation are desirable to illustrate local preceptorship schemes and their impacts.

3. Improving undergraduate education and preceptorship will result in a stronger pool of radiography workforce, ready and able to progress to post graduate reporting.

7.4. Future research areas

Calculation of accuracy, sensitivity and specificity based on the Red-dot style binary decision classifier poses a methodological threat. This method only examines the final decisions made by image observers and ignores locations of abnormalities, therefore it disregards the reasons for the decisions. Literature suggests that FROC based assessment (such as JAFROC) is a more location sensitive approach; perhaps this could be incorporated in an online platform or software with pre-defined statistical tools and image banks. Development and a wide-spread use of such a location sensitive benchmarking platform could alleviate the methodological limitations discussed in Chapter 6.3.

The radiography workforce in the UK does not seem to have openly voiced concern about the SCoR's vision that PCE will become a core competence of radiographers. Moreover, research has not explored radiographers' attitude toward PCE. Radiographers with formal education or training programmes of image evaluation may hold favourable perceptions. However, radiographers could show resistance against a commenting scheme if they felt a lack of experience in articulating radiographic findings. Introduction of mandatory PCE is imprudent if a large proportion of the radiographer population hold differing opinions. Further research investigating the feasibility of mandatory PCE is desirable.

Research should not overlook the service users of PCE (referring physicians and ultimately patients). A complete portrayal of PCE needs the understanding of, not only the competencies of the service providers (radiographers), but also its clinical value to the service users. The SCoR expects that PCE will allow referring physicians to expedite patient admission and clinical treatment. However, this assumption has not been challenged. Research is therefore recommended to explore and confirm the value of PCE to physicians and patients.

7.5. Dissemination

One primary purpose of a research project is to publicise research findings and new knowledge. Two published conference posters were involved in this study. A brief summary of the WWH scoring model (Akimoto et al., 2016) (DOI: 10.13140/RG.2.2.20043.18729) and the results of the image evaluation test (Akimoto, Wright & Reeves, 2017) (DOI: 10.13140/RG.2.2.36820.40328) were presented in the UK Radiological Congress (UKRC). The present study, including the literature reviews of the image evaluation practice by diagnostic radiographers, the results of the image evaluation test, WWH scoring system and PCE error classification scheme have the potential for peer-reviewed journal papers.

References

Aberdour, K. R. (1976). Must radiologists do all the reporting? *British Journal of Radiology, 49*(582), 573. doi:doi.org/10.1259/0007-1285-49-582-573

Akimoto, T., Wright, C., Reeves, P., & Harcus, J. (2015, July). Preliminary clinical evaluation: The What/Where/How (WWH) approach to scoring. Paper presented at the UK Radiological and Radiation Oncology Congress (UKRC), Liverpool.

Akobeng, A. K. (2007). Understanding diagnostic tests 1: Sensitivity, specificity and predictive values. *Acta Paediatrica (Oslo, Norway: 1992), 96*(3), 338. doi:10.1111/j.1651-2227.2006.00180.x

Albright, K., Gechter, K., & Kempe, A. (2013). Importance of Mixed Methods in Pragmatic Trials and Dissemination and Implementation Research. *Academic Pediatrics*, *13*(5), 400–407. doi.org/10.1016/j.acap.2013.06.010

Anvari, A., Halpern, E., & Samir, A. E. (2015). Statistics 101 for radiologists. *Radiographics, 35*(6), 1789-1801. doi:10.1148/rg.2015150112

Awan, O., Safdar, N., Siddiqui, K., Moffitt, R., & Siegel, E. (2011). Detection of Cervical Spine Fracture on Computed Radiography Images: A Monitor Resolution Study. *Academic Radiology*, *18*(3), 353–358. doi.org/10.1016/j.acra.2010.11.011

Baratloo, A., Hosseini, M., Negida, A., & El Ashal, G. (2015). Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emergency (Tehran, Iran), 3*(2), 48.

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (seventh ed.). New York: Oxford University Press.

Benvenuto-Andrade, C., Dusza, S., Agero, A., Kopf, A., Hay, J., & Marghoob, A. A. (2005). Level of confidence in diagnosis: Clinical examination vs. dermoscopy examination. *Journal of Investigative Dermatology, 124*(4), A142-A142. DOI: 10.1111/j.1524-4725.2006.32149.x

Berbaum, K. S., Brandser, E. A., Franken, E. A., Dorfman, D. D., Caldwell, R. T., & Krupinski, E. A. (2001). Gaze dwell times on acute trauma injuries missed because of satisfaction of search. *Academic Radiology, 8*(4), 304-314. doi:10.1016/S1076-6332(03)80499-3

Berbaum, K. S., El-Khoury, G., Ohashi, K., Schartz, K. M., Caldwell, R. T., Madsen, M., & Franken, E. A. (2007). Satisfaction of search in multitrauma patients: Severity of detected fractures. *Academic Radiology, 14*(6), 711-722. doi:10.1016/j.acra.2007.02.016

Berbaum, K. S., Schartz, K. M., Caldwell, R. T., El-Khoury, G., Ohashi, K., Madsen, M., & Franken, E. A. (2012). Satisfaction of search for subtle skeletal fractures may not be induced by more serious skeletal injury. *Journal of the American College of Radiology, 9*(5), 344-351. doi:10.1016/j.jacr.2011.12.040

Berbaum, K. S., Schartz, K. M., Caldwell, R. T., Madsen, M. T., Thompson, B. H., Mullan, B. F., . . . Franken, E. A. (2013). Satisfaction of search from detection of pulmonary nodules in computed tomography of the chest. *Academic Radiology, 20*(2) doi:10.1016/j.acra.2012.08.017

Berlin, L. (2000). Pitfalls of the vague radiology report. *American Journal of Roentgenology, 174*(6), 1511. doi:10.2214/ajr.174.6.1741511

Berlin, L. (2007). Accuracy of diagnostic procedures: Has it improved over the past five decades? *American Journal of Roentgenology*, 188(5), 1173-1178. doi:10.2214/AJR.06.1270

Berlin, L., & Hendrix, R. W. (1998). Perceptual errors and negligence. *American Journal of Roentgenology, 170*(4), 863-867. doi:10.2214/ajr.170.4.9530024

Berman, L., Lacey, G. D., Twomey, E., Twomey, B., Welch, T., & Eban, R. (1985). Reducing errors in the accident department: A simple method using radiographers. *British Medical Journal (Clinical Research Ed.), 290*(6466), 421-422. doi:10.1136/bmj.290.6466.421

Billay, D., & Myrick, F. (2008). Preceptorship: An integrative review of the literature. *Nurse Education in Practice*, *8*(4), 258–266. doi.org/10.1016/j.nepr.2007.09.005

Blakeley, C., Hogg, P., & Heywood, J. (2008). Effectiveness of UK radiographer image reading. *Radiologic Technology, 79*(3), 221-226.

Boone, D., Halligan, S., Mallett, S., Taylor, S. A., & Altman, D. G. (2012). Systematic review: Bias in imaging studies - the effect of manipulating clinical context, recall bias and reporting intensity. *European Radiology, 22*(3), 495-505. doi:10.1007/s00330-011-2294-0

Bosmans, J., Peremans, L., Menni, M., Schepper, A., Duyck, P., & Parizel, P. (2012). Structured reporting: If, why, when, how—and at what expense? results of a focus group meeting of radiology professionals from eight countries. *Insights into Imaging, 3*(3), 295-302. doi:10.1007/s13244-012-0148-1

Brady, A. (2017). Error and discrepancy in radiology: Inevitable or avoidable? *Insights into Imaging, 8*(1), 171-182. doi:10.1007/s13244-016-0534-1

Brealey, S. (2001a). Measuring the effects of image interpretation: An evaluative framework. *Clinical Radiology, 56*(5), 341-347. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1053/crad.2001.0678

Brealey, S. (2001b). Quality assurance in radiographic reporting: A proposed framework. *Radiography, 7*(4), 263-270. doi:dx.doi.org/10.1053/radi.2001.0342

Brealey, S., King, D. G., Hahn, S., Crowe, M., Williams, P., Rutter, P., & Crane, S. (2005). Radiographers and radiologists reporting plain radiograph requests from accident and emergency and general practice. *Clinical Radiology, 60*(6), 710-717. doi:10.1016/j.crad.2004.11.013

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. doi:dx.doi.org/10.1016/j.crad.2004.07.012

Brealey, S., Scally, A., Hahn, S., Thomas, N., Godfrey, C., & Crane, S. (2006). Accuracy of radiographers red dot or triage of accident and emergency radiographs in clinical practice: A systematic review. *Clinical Radiology, 61*(7), 604-615. doi:10.1016/j.crad.2006.01.015

Brealey, S., Scally, A. J., & Thomas, N. B. (2002a). Review article: Methodological standards in radiographer plain film reading performance studies. *The British Journal of Radiology BJR., 75*(890), 107-113. doi:10.1259/bjr.75.890.750107

Brealey, S., Scally, A. J., & Thomas, N. B. (2002b). Presence of bias in radiographer plain film reading performance studies. *Radiography, 8*(4), 203-210. doi:10.1053/radi.2002.0386

Brindle, M. J. (1976). Should we report every film? *British Journal of Radiology, 49*(579), 298-299. doi:doi.org/10.1259/0007-1285-49-579-298-c

Brook, O. R., O'Connell, A. M., Thornton, E., Eisenberg, R. L., Mendiratta-Lala, M., & Kruskal, J. B. (2010). Anatomy and pathophysiology of errors occurring in clinical radiology practice. *Radiographics, 30*(5), 1401-1410. doi:10.1148/rg.305105013

Brown, N., & Leschke, P. (2012). Evaluating the true clinical utility of the red dot system in radiograph interpretation. *Journal of Medical Imaging and Radiation Oncology, 56*(5), 510-513. doi:10.1111/j.1754-9485.2012.02398.x

Bruno, M. A., Walker, E. A., & Abujudeh, H. H. (2015). Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *Radiographics, 35*(6), 1668. doi:10.1148/rg.2015150023

Brusco, J. M. (2010). Effectively conducting an advanced literature search. *AORN Journal, 92*(3), 264-271. doi:10.1016/j.aorn.2010.06.008

Buskov, L., Abild, A., Christensen, A., Holm, O., Hansen, C., & Christensen, H. (2013). Radiographers and trainee radiologists reporting accident radiographs: A comparative plain film-reading performance study. *Clinical Radiology, 68*(1), 55-58. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1016/j.crad.2012.06.104

Carter, S., & Manning, D. (1999). Performance monitoring during postgraduate radiography training in reporting -- a case study. *Radiography, 5*(2), 71-78. doi:doi.org/10.1016/S1078-8174(99)90034-2

Challen, V., Kaminski, S., & Harris, P. (1996). Research-mindedness in the radiography profession. *Radiography, 2*(2), 139-151. doi:10.1016/S1078-8174(96)90005-X

Chen, Y., James, J., Turnbull, A., & Gale, A. (2015). The use of lower resolution viewing devices for mammographic interpretation: implications for education and training. *European Radiology*, *25*(10), 3003–3008. doi.org/10.1007/s00330-015-3718-z

Chertoff, J., Pisano, E., & Gert, B. (2009). Core curriculum: Research ethics for radiology residents. *Academic Radiology, 16*(1), 108-116. doi:10.1016/j.acra.2008.06.011

Coleman, L., & Piper, K. J. (2009). Radiographic interpretation of the appendicular skeleton: A comparison between casualty officers, nurse practitioners and radiographers. *Radiography, 15*(3), 196-202. doi:doi.org/10.1016/j.radi.2007.12.001

Cooper, C. P. (1976). Must radiologists do all the reporting? *British Journal of Radiology, 49*(584), 740. doi:doi.org/10.1259/0007-1285-49-584-740

Cosson, P., & Dash, R. (2015). A taxonomy of anatomical and pathological entities to support commenting on radiographs (preliminary clinical evaluation). *Radiography, 21*(1), 47. doi:10.1016/j.radi.2014.06.013

Council for International Organizations of Medical Sciences (CIOMS). (2002). *International ethical guidelines for biomedical research involving human subjects.* Geneva, Switzerland: Council for International Organizations of Medical Sciences.

Curtis, E. A., & Drennan, J. (2013). *Quantitative health research issues and methods*. Maidenhead, Berkshire, England; Maidenhead: Open University Press.

Department of Health. (2004). *The NHS knowledge and skills framework (NHS KSF) and the development review process (October 2004)*. London: Department of Health Publication.

DePoy, E., & Gitlin, L. N. (2016). *Introduction to research* (Fifth ed.). St. Louis: Mosby. doi:doi.org/10.1016/B978-0-323-26171-5.00003-3

Doody, O., & Noonan, M. (2016). Nursing research ethics, guidance and application in practice. *British Journal of Nursing, 25*(14), 803. doi:10.12968/bjon.2016.25.14.803

du Plessis, J., & Pitcher, R. (2015). Towards task shifting? A comparison of the accuracy of acute trauma-radiograph reporting by medical officers and senior radiographers in an African hospital. *Pan African Medical Journal, 21.* doi:10.11604/pamj.2015.21.308.6937

Ekpo, E. U., Hogg, P., & Mcentee, M. F. (2016). A review of individual and institutional publication productivity in medical radiation science. *Journal of Medical Imaging and Radiation Sciences, 47*(1), 13-20. doi:10.1016/j.jmir.2015.11.002

European Parliament and The Council of the European Union. (2014). Regulation no 536/2014 of the European parliament and of the council on clinical trials on medicinal products for human use, and repealing directive 2001/20/EC. *Official Journal of the European Union*, L158:1-76.

Ferranti, C., Primolevo, A., Cartia, F., Cavatorta, C., Ciniselli, C., Lualdi, M., … Scaperrotta, G. (2017). How Does the Display Luminance Level Affect Detectability of Breast Microcalcifications and Spiculated Lesions in Digital Breast Tomosynthesis (DBT) Images? *Academic Radiology*, *24*(7), 795–801. doi.org/10.1016/j.acra.2017.01.014

Freckleton, I. (2012). Advanced practice in radiography and radiation therapy: Report from the inter-professional advisory team. Retrieved from https://www.sor.org/system/files/news_story/201205/IPAT%20Final%20Report%2012%2034%204%2028%20(2).pdf

Friedenberg, R. M. (2000). The role of the supertechnologist. *Radiology, 215*(3), 630. doi:doi.org/10.1148/radiology.215.3.r00jn49630

General Medical Council. (2014). Skills fade: A review of the evidence that clinical and professional skills fade during time out of practice, and of how skills fade may be measured or remediated. Retrieved from https://www.gmc-uk.org/-

/media/about/skills-fade-literature-review-full-report.pdf?la=en&hash=8B32071AF03167EE588EE574F6DCC4C85B1FEF0B

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine, 165*(13), 1493-1499. doi:10.1001/archinte.165.13.1493

Hain, R. (2016). Consent. *Medicine, 44*(10), 593-595. doi:10.1016/j.mpmed.2016.07.009

Harcus, J., & Wright, C. (2013, June). WHAT, WHERE and HOW, A proposal for structuring preliminary clinical evaluations. Paper presented at the UK Radiological and Radiation Oncology Congress (UKRC), Liverpool.

Hardesty, L. A., Ganott, M. A., Hakim, C. M., Cohen, C. S., Clearfield, R. J., & Gur, D. (2005). “Memory effect” in observer performance studies of mammograms. *Academic Radiology, 12*(3), 286-290. doi:doi-org.lcproxy.shu.ac.uk/10.1016/j.acra.2004.11.026

Hardy, M., & Culpan, G. (2007). Accident and emergency radiography: A comparison of radiographer commenting and 'red dotting'. *Radiography, 13*(1), 65-71. doi:10.1016/j.radi.2005.09.009

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. doi:10.1016/j.radi.2015.10.010

Hardy, M., & Snaith, B. (2009). Radiographer interpretation of trauma radiographs: Issues for radiography education providers. *Radiography, 15*(2), 101-105. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1016/j.radi.2007.10.004

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. doi:10.1259/bjr.20120112

Hardy, M., Spencer, N., & Snaith, B. (2008). Radiographer emergency department hot reporting: An assessment of service quality and feasibility. *Radiography, 14*(4), 301-305. doi:10.1016/j.radi.2007.10.003

Hargreaves, J., & Mackay, S. (2003). The accuracy of the red dot system: Can it improve with training? *Radiography, 9*(4), 283-289. doi:dx.doi.org/10.1016/j.radi.2003.09.002

Harris, A. D., McGregor, J. C., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., & Finkelstein, J. (2006). The use and interpretation of quasi- experimental studies in medical informatics. *Journal of the American Medical Informatics Association, 13*(1), 16-23. doi:10.1197/jamia.M1749

Harris, R., & Paterson, A. (2016). Exploring the research domain of consultant practice: Experiences of consultant radiographers. *Radiography, 22*(1), e25-e33. doi:10.1016/j.radi.2015.07.003

Harvey-Lloyd, J., Morris, J., & Stew, G. (2019). Being a newly qualified diagnostic radiographer: Learning to fly in the face of reality. *Radiography*, *25*(3), e63–e67. doi.org/10.1016/j.radi.2019.01.007

Hazell, L., Motto, J., & Chipeya, L. (2015). The influence of image interpretation training on the accuracy of abnormality detection and written comments on musculoskeletal radiographs by South African radiographers. *Journal of Medical Imaging and Radiation Sciences, 46*(3), 302-308. doi:10.1016/j.jmir.2015.03.002

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

Health and Safety Executive. (2003). *Work with display screen equipment: Health and safety (display screen equipment) regulations 1992 as amended by the health and safety (miscellaneous amendments) regulations 2002*. Norwich: The Stationery Office publications.

Higgins, J. P. T., & Green, S. (2011). Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated march 2011]. Retrieved from http://handbook-5-1.cochrane.org/

Higher Education Statistics Agency. (2016). Higher education statistics for the UK 2015/16. Retrieved from https://www.hesa.ac.uk/data-and-analysis/publications/higher-education-2015-16

Hinde, S., & Spackman, E. (2015). Bidirectional citation searching to completion: An exploration of literature searching methods. *Pharmacoeconomics, 33*(1), 5-11. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1007/s40273-014-0205-3

Hlongwane, S. T., & 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), 638-640. doi:10.7196/SAMJ.6267

Hyde, E. (2015). A critical evaluation of student radiographers' experience of the transition from the classroom to their first clinical placement. *Radiography*, *21*(3), 242–247. doi.org/10.1016/j.radi.2014.12.005

Institute of Medicine. (2009). In Nass S. J., Levit A. L. and Gostin L. O. (Eds.), *Beyond the HIPAA privacy rule: Enhancing privacy, improving health through research.* (doi.org/10.17226/12458 ed.). Washington, DC: The National Academies Press. doi:doi.org/10.17226/12458.

Jain, V., & Raut, D. (2011). Medical literature search dot com. *Indian Journal of Dermatology, Venereology, and Leprology, 77*(2), 135-140. doi:10.4103/0378-6323.77451

Kim, Y. W., & Mansfield, L. T. (2014). Fool me twice: Delayed diagnoses in radiology with emphasis on perpetuated errors. *American Journal of Roentgenology, 202*(3), 465. doi:10.2214/AJR.13.11493

Knopf, J. W. (2006). Doing a literature review. *PS: Political Science Politics; APSC, 39*(1), 127-132. doi:10.1017/S1049096506060264

Koyfman, S. A., & Yom, S. S. (2017). Clinical research ethics: Considerations for the radiation oncologist. *International Journal of Radiation Oncology, Biology, Physics; International*

Journal of Radiation Oncology, Biology, Physics, 99(2), 259-264. doi:10.1016/j.ijrobp.2017.06.001

Kumar, R. D. (2007). *Evaluating medical radiation technologists' image interpretation accuracy and clinical practice relative to their postgraduate educational experience in New Zealand.*

Laffranchi, A., Cicero, C., Lualdi, M., Ciniselli, C., Calareso, G., Canestrini, S., … Marchianò, A. (2018). Different pixel pitch and maximum luminance of medical grade displays may result in different evaluations of digital radiography images. *La Radiologia Medica*, *123*(8), 586–592. doi.org/10.1007/s11547-018-0891-6

Lancaster, A., & 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*(2), 105. doi:doi.org/10.1016/j.radi.2011.08.003

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics, 38*(5), 404-415. doi:10.1016/j.jbi.2005.02.008

Leijen, Ä., Valtna, K., Leijen, D. A. J., & Pedaste, M. (2012). How to determine the quality of students' reflections? *Studies in Higher Education, 37*(2), 203-217. doi:10.1080/03075079.2010.504814

Loughran, C. F. (1994). Reporting of fracture radiographs by radiographers: The impact of a training programme. *British Journal of Radiology, 67*(802), 945-950. doi:doi.org/10.1259/0007-1285-67-802-945

Mackay, S. J. (2006). The impact of a short course of study on the performance of radiographers when highlighting fractures on trauma radiographs: "The Red Dot System". *The British Journal of Radiology, 79*(942), 468. doi:10.1259/bjr/53513558

Moshfeghi, M., Shahbazian M., Sajadi, S., Sajadi, S., & Ansari, H. (2016). Effects of Different Viewing Conditions on Radiographic Interpretation. *Journal of Dentistry of Tehran University of Medical Sciences*, *12*(11), 853–858.

Malamateniou, C. (2009). Radiography and research: A United Kingdom perspective. *European Journal of Radiography, 1*(1), 2-6. doi:doi.org/10.1016/j.ejradi.2008.12.003

Maltby, J. (2010). *Research methods for nursing and healthcare*. Harlow: Pearson Education.

Mankad, K., Hoey, E. T. D., Jones, J. B., Tirukonda, P., & Smith, J. T. (2009). Radiology errors: Are we learning from our mistakes? *Clinical Radiology, 64*(10), 988-993. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1016/j.crad.2009.06.002

Manning, D., & Hogg, P. (2006). Writing for publication. *Radiography, 12*(2), 77-78. doi:10.1016/j.radi.2006.01.008

Marks-Maran, D., Ooms, A., Tapping, J., Muir, J., Phillips, S., & Burke, L. (2013). A preceptorship programme for newly qualified nurses: A study of preceptees’ perceptions. *Nurse Education Today*, *33*(11), 1428–1434. doi.org/10.1016/j.nedt.2012.11.013

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post- test research designs: Some methodological concerns. *Oxford Review of Education, 38*(5), 583-616. doi:10.1080/03054985.2012.731208

Marshall, G., & Jonker, L. (2010). An introduction to descriptive statistics: A review and practical guide. *Radiography, 16*(4), e1-e7. doi:10.1016/j.radi.2010.01.001

Marshall, G., & Sykes, A. E. (2011). Systematic reviews: A guide for radiographers and other health care professionals. *Radiography, 17*(2), 158-164. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1016/j.radi.2010.08.007

McConnell, J. R., & Baird, M. A. (2017). Could musculo-skeletal radiograph interpretation by radiographers be a source of support to Australian medical interns: A quantitative evaluation. *Radiography; Radiography, 23*(4), 321-329. doi:10.1016/j.radi.2017.07.001

McConnell, J., Devaney, C., & Gordon, M. (2013). Queensland radiographer clinical descriptions of adult appendicular musculo-skeletal trauma following a condensed education programme. *Radiography, 19*(1), 48-55. doi:10.1016/j.radi.2012.09.002

McConnell, J., Devaney, C., Gordon, M., Goodwin, M., Strahan, R., & Baird, M. (2012). The impact of a pilot education programme on Queensland radiographer abnormality description of adult appendicular musculo-skeletal trauma. *Radiography, 18*(3), 184-190. doi:10.1016/j.radi.2012.04.005

McConnell, J. R., & Webster, A. J. (2000). Improving radiographer highlighting of trauma films in the accident and emergency department with a short course of study--an evaluation. *The British Journal of Radiology, 73*(870), 608-612. doi:10.1259/bjr.73.870.10911784

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Mckellar, C., & Currie, G. (2015). Publication productivity in the medical radiation sciences. *Journal of Medical Imaging and Radiation Sciences, 46*(3), S52-S60. doi:10.1016/j.jmir.2015.06.013

Mckenna, P. G., O, Neill, C., & Mcintyre, I. (1995). Research funding for radiography—The rules of the game. *Radiography, 1*(2), 145-149. doi:10.1016/S1078-8174(95)80024-7

Medical Radiation Practice Board of Australia. (2013). Professional capabilities for medical radiation practice. Retrieved from https://www.medicalradiationpracticeboard.gov.au/documents/default.aspx?record=WD13%2F12534&dbid=AP&chksum=OluB81d6eQCqo%2BewP9PHOA%3D%3D

Meline, T. (2006). Selecting studies for systematic review: Inclusion and exclusion criteria. *Contemporary Issues in Communication and Disorders, 33*, 21-27.

Meyer, A. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *The Journal of the*

American Association Internal Medicine, 173(21), 1952-1958. doi:10.1001/jamainternmed.2013.10081

Moon, J. (2004). Using reflective learning to improve the impact of short courses and workshops. *Journal of Continuing Education in the Health Professions, 24*(1), 4-11. doi:10.1002/chp.1340240103

Morton, L. P. (2002). Targeting generation Y. (segmenting publics). *Public Relations Quarterly, 47*(2), 46.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality, 77*(1), 261-286. doi:10.1111/j.1467-6494.2008.00545.x

Naylor, S., Ferris, C., & Burton, M. (2016). Exploring the transition from student to practitioner in diagnostic radiography. *Radiography*, *22*(2), 131–136. doi.org/10.1016/j.radi.2015.09.006

Neep, M. J., Steffens, T., Owen, R., & 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*(2), 69-77. doi:10.1002/jmrs.48

Neep, M. J., Steffens, T., Riley, V., Eastgate, P., & McPhail, S. M. (2017). Development of a valid and reliable test to assess trauma radiograph interpretation performance. *Radiography, 23*(2), 153-158. doi:10.1016/j.radi.2017.01.004

Nisbet, H. (2008). A model for preceptorship – the rationale for a formal, structured programme developed for newly qualified radiotherapy radiographers. *Radiography*, *14*(1), 52-56. doi:10.1016/j.radi.2006.07.004

Nightingale, J. (2016). Establishing a radiography research culture – are we making progress? *Radiography, 22*(4), 265-266. doi:10.1016/j.radi.2016.09.002

Nisbet, H. (2008). A model for preceptorship – the rationale for a formal, structured programme developed for newly qualified radiotherapy radiographers. *Radiography, 14*(1), 52-56. doi:10.1016/j.radi.2006.07.004

Nielsen, K., Finderup, J., Brahe, L., Elgaard, R., Elsborg, A., Engell-Soerensen, V., … Sommer, I. (2017). The art of preceptorship. A qualitative study. *Nurse Education in Practice*, *26*, 39–45. https://doi.org/10.1016/j.nepr.2017.06.009

Nixon, S. (2001). Professionalism in radiography. *Radiography, 7*(1), 31-35. doi:10.1053/radi.2000.0292

Nocum, D. J., Brennan, P. C., Huang, R. T., & Reed, W. M. (2013). The effect of abnormality-prevalence expectation on naïve observer performance and visual search. *Radiography, 19*(3), 196-199. doi:10.1016/j.radi.2013.04.004

Nuremberg Code. (1947). *Trials of war criminals begore the Nuremberg military tribunals under control law no. 10, vol 2 (p. 181-182).* Washington, DC: US Government Printing Office.

Ohla, H., Dagassan-Berndt, D., Payer, M., Filippi, A., Schulze, R., & Kühl, S. (2018). Role of ambient light in the detection of contrast elements in digital dental radiography. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, *126*(5), 439–443. doi.org/10.1016/j.oooo.2018.08.003

Onega, T., Anderson, M. L., Miglioretti, D. L., Buist, D. S. M., Geller, B., Bogart, A., . . . Yankaskas, B. C. (2013). Establishing a gold standard for test sets: Variation in interpretive agreement of expert mammographers. *Academic Radiology, 20*(6), 731. doi:10.1016/j.acra.2013.01.012

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology, 56*(1), 45-50. doi:10.4103/0301-4738.37595

Paterson, C., & Chapman, J. (2013). Enhancing skills of critical reflection to evidence learning in professional practice. *Physical Therapy in Sport, 14*(3), 133-138. doi:10.1016/j.ptsp.2013.03.004

Paterson, A. M., Price, R. C., Thomas, A., & Nuttall, L. (2004). Reporting by radiographers: A policy and practice guide. *Radiography, 10*(3), 205-212. doi:dx.doi.org/10.1016/j.radi.2004.03.004

Pines, J. M., Hilton, J. A., Weber, E. J., Alkemade, A. J., Al Shabanah, H., Anderson, P. D., . . . Schull, M. J. (2011). International perspectives on emergency department crowding. *Academic Emergency Medicine, 18*(12), 1358-1370. doi:10.1111/j.1553-2712.2011.01235.x

Pinto, A., & Brunese, L. (2010). Spectrum of diagnostic errors in radiology. *World Journal of Radiology, 2*(10), 377-383. doi:10.4329/wjr.v2.i10.377

Pinto, A., Brunese, L., Pinto, F., Reali, R., Daniele, S., & Romano, L. (2012). The concept of error and malpractice in radiology. *Seminars in Ultrasound, CT and MRI, 33*(4), 275-279. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1053/j.sult.2012.01.009

Piper, K. J., & Paterson, A. (2009). Initial image interpretation of appendicular skeletal radiographs: A comparison between nurses and radiographers. *Radiography, 15*(1), 40-48. doi:10.1016/j.radi.2007.10.006

Piper, K. J., Paterson, A., & 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. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1016/j.radi.2004.05.004

Piper, K. J., Paterson, A., & Ryan, C. (1999). *The implementation of a radiographic reporting service, for trauma examinations of the skeletal system, in 4 national health service trust.* Canterbury Christ Church University, Canterbury:

Plumb, A. A. O., Grieve, F. M., & Khan, S. H. (2009). Survey of hospital clinicians' preferences regarding the format of radiology reports. *Clinical Radiology, 64*(4), 386-394. doi:10.1016/j.crad.2008.11.009

Price, R. C. (2001). Radiographer reporting: Origins, demise and revival of plain film reporting. *Radiography, 7*(2), 105-117. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1053/radi.2001.0281

Prime, N. J., Paterson, A. M., & Henderson, P. I. (1999). The development of a curriculum—a case study of six centres providing courses in radiographic reporting. *Radiography, 5*(2), 63-70. doi:10.1016/S1078-8174(99)90033-0

Provenzale, J., & Kranz, P. (2011). Understanding errors in diagnostic radiology: Proposal of a classification scheme and application to emergency radiology. *Emergency Radiology; A Journal of Practical Imaging Official Journal of the American Society of Emergency Radiology, 18*(5), 403-408. doi:10.1007/s10140-011-0974-3

Pusic, M. V., Andrews, J. S., Kessler, D. O., Teng, D. C., Pecaric, M. R., Ruzal-Shapiro, C., & Boutis, K. (2012). Prevalence of abnormal cases in an image bank affects the learning of radiograph interpretation. *Medical Education, 46*(3), 289-298. doi:10.1111/j.1365-2923.2011.04165.x

Quek, G., & Shorey, S. (2018). Perceptions, Experiences, and Needs of Nursing Preceptors and Their Preceptees on Preceptorship: An Integrative Review. *Journal of Professional Nursing*, *34*(5), 417–428. doi.org/10.1016/j.profnurs.2018.05.003

Reiner, B. I. (2013). Strategies for radiology reporting and communication. part 1: Challenges and heightened expectations. *Journal of Digital Imaging, 26*(4), 610-613. doi:10.1007/s10278-013-9615-6

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), 145-150. doi:10.1148/radiology.183.1.1549661

Renwick, I. G. H., Butt, W. P., & Steele, B. (1991). How well can radiographers triage x ray films in accident and emergency departments? *British Medical Journal, 302*(6776), 568-569. 10.1136/bmj.302.6776.568

Robinson, P. J. (1996). Short communication: Plain film reporting by radiographers--a feasibility study. *The British Journal of Radiology, 69*(828), 1171-1174. doi:doi.org/10.1259/0007-1285-69-828-1171

Robinson, P. J. A., Culpan, G., & Wiggins, M. (1999). Interpretation of selected accident and emergency radiographic examinations by radiographers: A review of 11,000 cases. *British Journal of Radiology, 72*(JUN.), 546-551. doi:10.1259/bjr.72.858.10560335

Robinson, P. J., Wilson, D., Coral, A., Murphy, A., & Verow, P. (1999). Variation between experienced observers in the interpretation of accident and emergency radiographs. *British Journal of Radiology*, 72(856), 323-330. doi:10.1259/bjr.72.856.10474490

Roessger, K. M. (2014). The effect of reflective activities on instrumental learning in adult work-related education: A critical review of the empirical research. *Educational Research Review, 13*, 17-34. doi:10.1016/j.edurev.2014.06.002

Sampson, M., & McGowan, J. (2006). Errors in search strategies were identified by type and frequency. *Journal of Clinical Epidemiology, 59*(10), 1057-1063. doi:10.1016/j.jclinepi.2006.01.007

Saxton, H. M. (1992). Should radiologists report on every film? *Clinical Radiology, 45*(1), 1-3. doi:doi.org/10.1016/S0009-9260(05)81457-6

Schwartz, L. H., Panicek, D. M., Berk, A. R., Li, Y., & Hricak, H. (2011). Improving communication of diagnostic radiology findings through structured reporting. *Radiology, 260*(1), 174-181. doi:10.1148/radiol.11101913

Shorrock, S. T., & Kirwan, B. (2002). Development and application of a human error identification tool for air traffic control. *Applied Ergonomics, 33*(4), 319-336. doi:10.1016/S0003-6870(02)00010-8

Sim, J., & Radloff, A. (2009). Profession and professionalisation in medical radiation science as an emergent profession. *Radiography, 15*(3), 203-208. doi:10.1016/j.radi.2008.05.001

Šimundić, A. (2009). Measures of diagnostic accuracy: Basic definitions. *Electronic Journal of International Federation of Clinical Chemistry and Laboratory Medicine, 19*(4), 203.

Smith, T. (2013). To dot or not?: The need to redesign frontline image interpretation. *Journal of Medical Imaging and Radiation Oncology, 57*(2), 205-205. doi:10.1111/1754-9485.12052

Smith, T. N., & Baird, M. (2007). Radiographers' role in radiological reporting: A model to support future demand. *Medical Journal of Australia, 186*(12), 629-31.

Smith, V., Devane, D., Begley, C. M., & Clarke, M. (2011). Methodology in conducting a systematic review of healthcare interventions. *BMC Medical Research Methodology, 11.* doi:10.1186/1471-2288-11-15

Smith, T., & Younger, C. (2002). Accident and emergency radiological interpretation using the radiographer opinion form. *The Radiographer, 47*, 27-31.

Snaith, B. (2012). Collaboration in radiography: A bibliometric analysis. *Radiography, 18*(4), 270-274. doi:10.1016/j.radi.2012.07.003

Snaith, B. (2013). *Development of the radiographer evidence base: An examination of advancing practice* (PhD). Retrieved from https://bradscholars.brad.ac.uk/handle/10454/6314

Snaith, B., & Hardy, M. (2007). How to achieve advanced practitioner status: A discussion paper. *Radiography, 13*(2), 142-146. doi:10.1016/j.radi.2006.01.001

Snaith, B., & Hardy, M. (2008). Radiographer abnormality detection schemes in the trauma environment -- an assessment of current practice. *Radiography, 14*(4), 277-281. doi:doi.org/10.1016/j.radi.2007.09.001

Snaith, B., Hardy, M., & Lewis, E. F. (2014). Reducing image interpretation errors – do communication strategies undermine this? *Radiography, 20*(3), 230-234. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1016/j.radi.2014.03.006

Sokolovskaya, E., Shinde, T., Ruchman, R. B., Kwak, A. J., Lu, S., Shariff, Y. K., . . . Talangbayan, L. (2015). The effect of faster reporting speed for imaging studies on the number of misses and interpretation errors: A pilot study. *Journal of the American College of Radiology, 12*(7), 683-688. doi:10.1016/j.jacr.2015.03.040

Stephenson, L. A., Wagner, S. J., & Bolton, D. (2012). Understanding autonomy within philosophical tradition and modern medical ethics. *British Journal of Hospital Medicine, 73*, C190-C192. doi:10.12968/hmed.2012.73.Sup12.C190

Stephenson, P., Hannah, A., Jones, H., Edwards, R., Harrington, K., Baker, S. -., . . . Belfield, J. (2012). An evidence based protocol for peer review of radiographer musculoskeletal plain film reporting. *Radiography, 18*(3), 172-178. doi:doi.org/10.1016/j.radi.2012.03.004

Stevens, B. J., & Thompson, J. D. (2018). The impact of focused training on abnormality detection and provision of accurate preliminary clinical evaluation in newly qualified radiographers. *Radiography; Radiography, 24*(1), 47-51. doi:10.1016/j.radi.2017.08.007

Stevinson, C., & Lawlor, D. A. (2004). Searching multiple databases for systematic reviews: Added value or diminishing returns? *Complementary Therapies in Medicine, 12*(4), 228-232. doi:10.1016/j.ctim.2004.09.003

Stojanovic, M., Apostolovic, M., Stojanovic, D., Milosevic, Z., Toplaovic, A., Lakusic, V., & Golubovic, M. (2014). Understanding sensitivity, specificity and predictive values. *Vojnosanitetski Pregled; Vojnosanit.Pregl., 71*(11), 1062-1065. doi:10.2298/VSP1411062S

Swinburne, K. (1971). Pattern recognition for radiographers. *The Lancet, 297*(7699), 589-590. doi:dx.doi.org.lcproxy.shu.ac.uk/10.1016/S0140-6736(71)91180-9

Tan, K., Feuz, C., Bolderston, A., & Palmer, C. (2011). A literature review of preceptorship: A model for the medical radiation sciences? *Journal of Medical Imaging and Radiation Sciences, 42*(1), 15-20. doi:10.1016/j.jmir.2010.08.004

Tariq, S., & Woodman, J. (2013). Using mixed methods in health research. *JRSM Short Reports*, *4*(6), 2042533313479197. doi.org/10.1177/2042533313479197

Taylor, G., Voss, S., Melvin, P., & Graham, D. (2011). Diagnostic errors in pediatric radiology. *Pediatric Radiology, 41*(3), 327-334. doi:10.1007/s00247-010-1812-6

The PRISMA group. (2009). PRISMA 2009 flow diagram. Retrieved from http://prisma-statement.org/documents/PRISMA%202009%20flow%20diagram.pdf

The Royal College of Radiologists. (1995). *Statement on reporting in departments of clinical radiology*. London: The Royal College of Radiologists.

The Royal College of Radiologists. (2015). Unreported X-rays, computed tomography (CT) and magnetic resonance imaging (MRI) examinations: Results of the September 2015 snapshot survey of English NHS acute trusts. Retrieved from https://www.rcr.ac.uk/sites/default/files/unreported_studies_feb2015.pdf

The Royal College of Radiologists. (2017a). Clinical radiology UK workforce census 2016 report. Retrieved from https://www.rcr.ac.uk/system/files/publication/field_publication_files/cr_workforce_census_2016_report_0.pdf

The Royal College of Radiologists. (2017b). Final examination for the fellowship in clinical radiology (part B) guidance note for candidates. Retrieved from https://www.rcr.ac.uk/sites/default/files/docs/radiology/pdf/CR2B_Candidate_Guidance_Notes.pdf

The Royal College of Radiologists. (n.d.). Final examination for the fellowship in clinical radiology (part B) scoring system. Retrieved from https://www.rcr.ac.uk/sites/default/files/cr2b_scoring_system.pdf

The Royal College of Radiologists and The College of Radiographers. (1998). *Inter-professional roles and responsibilities in a radiology service*. London: The Royal College of Radiologists and The Society and College of Radiographers.

The Society and College of Radiographers. (1997). *Reporting by radiographers: A vision paper*. London: The Society and College of Radiographers.

The Society and College of Radiographers. (2003). Clinical supervision framework. Retrieved from https://www.sor.org/system/files/article/201202/sor_clinical_supervision_framework.pdf

The Society and College of Radiographers. (2006). Medical image interpretation and clinical reporting by non-radiologists: The role of the radiographer. Retrieved from https://www.sor.org/system/files/document-library/public/sor_Definitive_Guidance_May_2010.pdf

The Society and College of Radiographers. (2013). Preliminary clinical evaluation and clinical reporting by radiographers: Policy and practice guidance. Retrieved from https://www.sor.org/learning/document-library/preliminary-clinical-evaluation-and-clinical-reporting-radiographers-policy-and-practice-guidance

The Society and College of Radiographers. (2015). 2016-2021 Society and College of Radiographers research strategy. Retrieved from https://www.sor.org/sites/default/files/document-versions/research_strategy_final_4.pdf

Tewes, S., Rodt, T., Marquardt, S., Evangelidou, E., Wacker, F., & von Falck, C. (2013). Evaluation of the use of a tablet computer with a high-resolution display for interpreting emergency CT scans. *Rofo*, *185*(11), 1063–1069. doi.org/10.1055/s-0033-1350155

United States Department of Health, Education and Welfare. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, DC: US Department of Health, Education and Welfare Publication.

Wallis, A., & Mccoubrie, P. (2011). The radiology report — are we getting the message across? *Clinical Radiology, 66*(11), 1015-1022. doi:10.1016/j.crad.2011.05.013

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

Williams, J. R. (2015). *Ethics for biomedical research involving humans: International guidelines.* Ottawa: Elsevier Ltd. doi:10.1016/B978-0-08-097086-8.11015-3

Williams, P. (2002). Research, radiography and the RAE: Lessons from the 2001 research assessment exercise. *Radiography, 8*(4), 195-200. doi:10.1053/radi.2002.0388

Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: Sensitivity, specificity, PPV and NPV. *Proceedings of Singapore Healthcare, 20*(4), 316-318. doi:10.1177/201010581102000411

World Medical Association. (1964). *WMA Declaration of Helsinki - ethical principles for medical research involving human subjects*. Helsinki, Finland: 18th WMA General Assembly.

World Medical Association. (2013). *WMA Declaration of Helsinki - ethical principles for medical research involving human subjects*. Fortaleza, Brazil: WMA General Assembly.

Wright, C., & Reeves, P. (2016). RadBench: Benchmarking image interpretation skills. *Radiography, 22*(2), e131-e136. doi:10.1016/j.radi.2015.12.010

Wright, C., & 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

Yuko, E., & Fisher, C. B. (2015). Research ethics: Research. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences (second edition)* (pp. 514-522). Oxford: Elsevier. doi:doi-org.lcproxy.shu.ac.uk/10.1016/B978-0-08-097086-8.11022-0

Appendices

Appendix A – Results of literature searches

Database: PubMed

Last updated: 21/03/2017

Search result: 1,723

Literature retrieved: 15

#	Keywords	Results	Rationale
1	Radiographer*[Title/Abstract] **OR** Radiography[Title/Abstract]	61,650	Primary theme of the study.
2	Accuracy[Title/Abstract] **OR** Competenc*[Title/Abstract] **OR** Education*[Title/Abstract] **OR** Program*[Title/Abstract] **OR** Sensitivity[Title/Abstract] **OR** Specificity[Title/Abstract] **OR** Training[Title/Abstract]	2,298,243	Free-text keywords related to diagnostic radiographers' skills in X-ray image evaluation.
3	Comment*[Title/Abstract] **OR** Interpret*[Title/Abstract] **OR** PCE[Title/Abstract] **OR** "Preliminary Clinical Evaluation"[Title/Abstract] **OR** "Red dot"[Title/Abstract] **OR** "Red-dot"[Title/Abstract] **OR** Report*[Title/Abstract]	3,498,360	Free-text keywords related to radiographers' clinical roles in X-ray image evaluation.
4	1 **AND** 2 **AND** 3	1,300	A Boolean operator (AND) is used to combine the free-text keyword search concepts.
5	"Allied Health Personnel"[MAJR] **OR** "Diagnostic Services/standards"[MeSH] **OR** "Emergencies"[MeSH] **OR** "Emergency Medicine/standards"[MAJR] **OR** "Fractures, Bone/diagnostic imaging"[MAJR] **OR** "Medical Staff, Hospital/standards"[MAJR] **OR** "Radiography"[MAJR] **OR** "Radiology Department, Hospital/standards"[MeSH] **OR** "Technology, Radiologic"[MAJR]	355,872	MeSH terms related to radiographers and X-ray image evaluation.
6	"Clinical Competence/standards"[MAJR] **OR** "Diagnostic Errors"[MeSH] **OR** "Education, Continuing"[MAJR] **OR**	264,772	MeSH terms related to radiographers' clinical

	"Educational Measurement"[MeSH] **OR** "Medical Audit"[MeSH]		competencies and professionalism.
7	"False Negative Reactions"[MeSH] **OR** "False Positive Reactions"[MeSH] **OR** "ROC Curve"[MeSH] **OR** "Sensitivity and Specificity"[MeSH]	510,770	MeSH terms related to analysis of radiographers’ clinical competencies.
8	5 **AND** 6 **AND** 7	6,542	A Boolean operator (AND) is used to combine the MeSH term search concepts.
9	4 **OR** 8	7,755	A Boolean operator (OR) is used to combine the free-text keyword and MeSH term search concepts.
10	"Humans"[MeSH]	16,203,337	A keyword that specifies human studies.
11	9 **AND** 10	7,390	A Boolean operator (AND) is used to restrict the search to human studies.
12	Abdom*[Title/Abstract] **OR** Angiograph*[Title/Abstract] **OR** Barium[Title/Abstract] **OR** "Computed Tomography"[Title/Abstract] **OR** "Computer-assist*"[Title/Abstract] **OR** CT[Title/Abstract] **OR** Dental[Title/Abstract] **OR** "Magnetic Resonance Imaging"[Title/Abstract] **OR** Mammograph*[Title/Abstract] **OR** MRI[Title/Abstract] **OR** Screen*[Title/Abstract] **OR** Ultrasonograph*[Title/Abstract] **OR** Ultrasound*[Title/Abstract] **OR** US[Title/Abstract]	2,194,739	Common free-text keywords used in diagnostic radiography but irrelevant to X-ray image evaluation.
13	"Angiography"[Mesh] **OR** "Barium Compounds"[Mesh] **OR** "Barium Enema"[Mesh] **OR** "Dental Caries"[Mesh] **OR** "Diagnosis, Computer-Assisted"[Mesh] **OR** "Image Interpretation, Computer-Assisted"[Mesh] **OR** "Image	1,488,206	Common MeSH terms used in diagnostic radiography but

	Processing, Computer-Assisted"[Mesh] **OR** "Magnetic Resonance Imaging"[Mesh] **OR** "Mammography"[Mesh] OR "Mass Screening"[Mesh] **OR** "Radiation Protection"[Mesh] **OR** "Radiographic Image Enhancement"[Mesh] **OR** "Radiography, Abdominal"[Mesh] **OR** "Radiography, Thoracic"[MeSH] **OR** "Tomography, X-Ray Computed"[Mesh] **OR** "Ultrasonography"[Mesh]		irrelevant to X-ray image evaluation.
14	12 **OR** 13	2,948,315	A Boolean operator (OR) is used to combine the irrelevant free-text keyword and MeSH term search concepts.
15	11 **NOT** 14	1,705	A Boolean operator (NOT) is used to exclude search items with irrelevant search concepts.

Database: CINAHL

Last updated: 18/01/2017

Search result: 743

Literature retrieved: 18

#	Keywords	Results	Comment
1	Radiographer* **OR** Radiography	120,113	Primary theme of the study.
2	Accuracy **OR** Competenc* **OR** Education* **OR** Program* **OR** Sensitivity **OR** Specificity **OR** Training	991,406	Free-text keywords related to diagnostic radiographers' skills in X-ray image evaluation.
3	Comment* **OR** Interpret* **OR** PCE **OR** "Preliminary Clinical Evaluation" **OR** "Red dot" **OR** "Red-dot" **OR** Report*	575,422	Free-text keywords related to radiographers' clinical roles in X-ray image evaluation.
4	1 **AND** 2 **AND** 3	3,937	A Boolean operator (AND) is used to combine the free-text keyword search concepts.
5	MH "Radiologic Technologists" **OR** MM "Emergency Care" **OR** MM "Emergency Service" **OR** MM "Fractures/RA" **OR** MM "Radiography" **OR** MM "Trauma/RA"	40,627	CINAHL headings related to radiographers and X-ray image evaluation.
6	MH "Audit" **OR** MH "Competency Assessment" **OR** MH "Diagnostic Errors" **OR** MH "Professional Role" **OR** MM "Clinical Competence" **OR** MM "Education, Continuing" **OR** MM "Staff Development"	77,578	CINAHL headings related to radiographers' clinical competencies and professionalism.
7	MH "Analysis of Variance" **OR** MH "Confidence Intervals" **OR** MH "Descriptive Statistics" **OR** MH "Paired T-Tests" **OR** MH	661,169	CINAHL headings related to analysis of

	"Pearson's Correlation Coefficient" **OR** MH "P-Value" **OR** MH "ROC Curve" **OR** MH "Sensitivity and Specificity"		radiographers' clinical competencies.
8	5 **AND** 6 **AND** 7	408	A Boolean operator (AND) is used to combine the CINAHL heading search concepts.
9	4 **OR** 8	4,273	A Boolean operator (OR) is used to combine the free-text keyword and CINAHL heading search concepts.
10	MH "Human"	1,503,328	A keyword that specifies human studies.
11	9 **AND** 10	3,144	A Boolean operator (AND) is used to restrict the search to human studies.
12	Abdom* **OR** Angiograph* **OR** Barium **OR** "Computed Tomography" **OR** "Computer-assist" **OR** "CT" **OR** Dental OR "Magnetic Resonance Imaging" **OR** Mammograph* **OR** "MRI" **OR** Screen* **OR** Ultrasonograph* **OR** Ultrasound* **OR** "US"	519,072	Common free-text keywords used in diagnostic radiography but irrelevant to X-ray image evaluation.
13	MH "Angiography+" **OR** MH "Barium Compounds+" **OR** MH "Barium" **OR** MH "Dental Caries" **OR** MH "Diagnosis, Computer Assisted+" **OR** MH "Health Screening+" **OR** MH "Image Processing, Computer Assisted+" **OR** MH "Magnetic Resonance Imaging+" **OR** MH "Mammography" **OR** MH "Radiographic Image Enhancement+" **OR** MH "Radiography, Abdominal+" **OR** MH "Radiography, Dental+" **OR** MH "Radiography, Thoracic+" **OR** MH "Ultrasonography+"	302,855	Common CINAHL headings used in diagnostic radiography but irrelevant to X-ray image evaluation.
14	12 **OR** 13	582,288	A Boolean operator (OR) is used to combine the irrelevant free-text

			keyword and CINAHL heading search concepts.
15	11 **NOT** 14	743	A Boolean operator (NOT) is used to exclude search items with irrelevant search concepts.

Database: ScienceDirect

Last updated: 19/01/2017

Search result: 215

Literature retrieved: 15

#	Keywords	Results	Comment
1	TITLE-ABSTR-KEY(Radiographer* **OR** Radiography)	14,026	Primary theme of the study.
2	TITLE-ABSTR-KEY(Accuracy **OR** Competenc* **OR** Education* **OR** Program* **OR** Sensitivity **OR** Specificity **OR** Training)	1,148,899	Free-text keywords related to diagnostic radiographers' skills in X-ray image evaluation.
3	TITLE-ABSTR-KEY(Comment* **OR** Interpret* **OR** PCE **OR** {Preliminary Clinical Evaluation} **OR** {Red dot} **OR** {Red-dot} OR Report*)	1,418,409	Free-text keywords related to radiographers' clinical roles in X-ray image evaluation.
4	1 **AND** 2 **AND** 3	591	A Boolean operator (AND) is used to combine the free-text keyword search concepts.
5	TITLE-ABSTR-KEY(Abdom* **OR** Angiograph* **OR** Barium **OR** {Computed Tomography} **OR** {Computer-assisted} **OR** CT OR Dental **OR** {Magnetic Resonance Imaging} **OR** Mammogra* OR MRI OR Screen* **OR** Ultrasonograph* **OR** Ultrasound* **OR** US)	3,577,893	Common free-text keywords used in diagnostic radiography but irrelevant to X-ray image evaluation.
6	4 **AND NOT** 5	215	A Boolean operator (AND NOT) is used to exclude search items with irrelevant search concepts.

Database: Web of Science

Last updated: 19/01/2017

Search result: 654

Literature retrieved: 15

#	Keywords	Results	Comment
1	TOPIC: (Radiographer* **OR** Radiography)	4,6043	Primary theme of the study.
2	TOPIC: (Accuracy **OR** Competenc* **OR** Education* **OR** Program* **OR** Sensitivity **OR** Specificity **OR** Training)	4,402,548	Free-text keywords related to diagnostic radiographers' skills in X-ray image evaluation.
3	TOPIC: (Comment* **OR** Interpret* **OR** PCE **OR** "Preliminary Clinical Evaluation" **OR** "Red dot" **OR** "Red-dot" **OR** Report*)	4,502,652	Free-text keywords related to radiographers' clinical roles in X-ray image evaluation.
4	1 **AND** 2 **AND** 3	2,765,588	A Boolean operator (AND) is used to combine the free-text keyword search concepts.
5	TOPIC: (Abdom* **OR** Angiograph* **OR** Barium **OR** "Computed Tomography" **OR** "Computer-assisted" **OR** CT **OR** Dental **OR** "Magnetic Resonance Imaging" **OR** Mammogra* **OR** MRI **OR** Screen* **OR** Ultrasonograph* **OR** Ultrasound* **OR** US)	1,860	Common free-text keywords used in diagnostic radiography but irrelevant to X-ray image evaluation.
6	4 **NOT** 5	654	A Boolean operator (NOT) is used to exclude search items with irrelevant search concepts.

Database: ProQuest

Last updated: 29/03/2017

Search result: 1,259

Literature retrieved: 16

#	Keywords	Results	Comment
1	ti(Radiographer* **OR** Radiography) **OR** ab(Radiographer* **OR** Radiography)	87,385	Primary theme of the study.
2	ti(Accuracy **OR** Competenc* **OR** Education* **OR** Program* **OR** Sensitivity **OR** Specificity **OR** Training) **OR** ab(Accuracy **OR** Competenc* **OR** Education* **OR** Program* **OR** Sensitivity **OR** Specificity **OR** Training)	35,579,418	Free-text keywords related to diagnostic radiographers' skills in X-ray image evaluation.
3	ti(Comment* **OR** Interpret* **OR** PCE **OR** "Preliminary Clinical Evaluation" **OR** "Red dot" **OR** "Red-dot" **OR** Report*) **OR** ab(Comment* **OR** Interpret* **OR** PCE **OR** "Preliminary Clinical Evaluation" **OR** "Red dot" **OR** "Red-dot" **OR** Report*)	47,914,193	Free-text keywords related to radiographers' clinical roles in X-ray image evaluation.
4	1 **AND** 2 **AND** 3	2,823	A Boolean operator (AND) is used to combine the free-text keyword search concepts.
5	ti(Abdom* **OR** Angiograph* **OR** Barium **OR** "Computed Tomography" **OR** "Computer-assisted" **OR** CT **OR** Dental **OR** "Magnetic Resonance Imaging" **OR** Mammogra* **OR** MRI **OR** Screen* **OR** Ultrasonograph* **OR** Ultrasound* **OR** US) **OR** ab(Abdom* **OR** Angiograph* **OR** Barium **OR** "Computed Tomography" **OR** "Computer-assisted" **OR** CT OR Dental **OR** "Magnetic Resonance Imaging" **OR** Mammogra* **OR** MRI **OR** Screen* **OR** Ultrasonograph* **OR** Ultrasound* **OR** US)	47,914,193	Common free-text keywords used in diagnostic radiography but irrelevant to X-ray image evaluation.
6	4 **NOT** 5	1,259	A Boolean operator (NOT) is used to exclude search items with irrelevant search concepts.

Appendix B – QUADAS-2

Review ID: ________________

Quality Assessment Tool

Review question
Study title
Author
Year
Publisher
Imaging modality
Index test(s)
Participants
Reference standard
Study type
Date of assessment

Risk of bias and applicability judgements

Domain 1: Participant selection

A. Risk of Bias

Describe methods of participant selection:

	Yes	No	Unclear
❖ Was a consecutive or random sample of participants enrolled?	□	□	□
❖ Were inclusion criteria of participants sufficiently described?	□	□	□
❖ Did the study avoid inappropriate exclusions?	□	□	□

Risk:	Low	High	Unclear
Could the selection of participants have introduced bias?	□	□	□

B. Concerns regarding applicability

Describe included participants:

Concern:	Low	High	Unclear
Is there concern that the included participants do not match the review question?	□	□	□

Domain 2: Index Test(s)

If more than one index test was used, please complete for each test.

A. Risk of Bias

Describe the index test and how it was conducted and interpreted:

	Yes	No	Unclear
❖ Was the execution of the index test described in sufficient detail to permit replication of the test?	□	□	□
❖ If a threshold was used, was it pre-specified?	□	□	□

Risk:	**Low**	**High**	**Unclear**
Could the conduct or interpretation of the index test have introduced bias?	□	□	□

B. Concerns regarding applicability

Concern:	**Low**	**High**	**Unclear**
Is there concern that index test, its conduct, or interpretation differ from the review question?	□	□	□

Domain 3: Reference Standard (Radiological reports)

A. Risk of Bias

Describe the reference standard and how it was conducted and interpreted:

	Yes	No	Unclear
❖ Is the reference standard likely to correctly classify the target condition?	□	□	□
❖ Were the reference standard produced without knowledge of the results of the index test?	□	□	□

Risk:	Low	High	Unclear
Could the reference standard, its conduct, or its interpretation have introduced bias?	□	□	□

B. Concerns regarding applicability

Concern:	Low	High	Unclear
Is there concern that the target condition as defined by the reference standard does not match the review question?	□	□	□

Domain 4: Flow

A. Risk of Bias

Describe any participants who did not receive the index test(s) and/or reference standard:

Risk:	Yes	No	Unclear
❖ Did all participants receive a reference standard?	□	□	□
❖ Did participants receive the same reference standard?	□	□	□
❖ Were all participants included in the analysis?	□	□	□

Concern:	Low	High	Unclear
Could the participant flow have introduced bias?	□	□	□

Appendix C – Information sheet

Participant information sheet

We would like to invite you to take part in our research study. Before you decide we would like you to understand why the research is being done and what it would involve for you. Talk to others about the study if you wish. **Part 1** tells you the purpose of this study and what will happen to you if you take part. **Part 2** gives you more detailed information about the conduct of the study. Ask us if there is anything that is not clear.

Study title: Image Interpretation Performance of Diagnostic Radiographers: Benchmarking new graduates

Principal investigator: Tatsuhito Akimoto

Contact: tatsuhito.akimoto@student.shu.ac.uk, +44 7476 908040.

Please read the following carefully before you decide to take part in this research.

Part 1:

1. **What is the purpose of this study?**

   The purpose of this study is to investigate final year diagnostic radiography students' competencies in plain musculoskeletal image interpretation before your graduation/qualification.

2. **Why have I been invited?**

   This is because you are in the final year of your diagnostic radiography programme.

3. **Do I have to take part?**

   Your decision to take part in this study is entirely voluntary. You may refuse to participate or you can withdraw from the study at any time. Your refusal to participate or wish to withdraw would not influence in any way your current or potential future and your progress on your education course.

4. **Expenses and payments**

   You will not be paid for taking part in this study.

5. **What will I have to do?**

You will be asked to view 30 musculoskeletal radiographic images, state if you see any abnormalities and provide a short preliminary clinical evaluation. The test will take approximately 45 minutes.

6. What are the possible disadvantages and risks of taking part?

There is a very low risk of eye strain relating to viewing images on a PC monitor. Therefore, it is recommended that you do the test under an optimum viewing condition.

7. What are the possible benefits of taking part?

Your participation will provide better understanding of the current diagnostic radiography students' competencies in musculoskeletal image interpretation.

When you complete the test, you will be able to compare the correct answers with your own, reflect on your performance and address any development needs. You will receive a certificate to demonstrate your image interpretation competence which can be used for continuous professional development. It can also be useful for job interviews, particularly if you meet the recommendations of the SCoR 2013 in providing reliable preliminary clinical evaluation.

8. What if there is a problem or I want to complain?

Any complaint about this research will be addressed. The detailed information on this is given in Part 2.

9. Will my taking part in the study be kept confidential?

Yes. We will follow ethical and legal practice and all information about you will be handled in confidence. The details are included in Part 2.

This completes Part 1. If the information in Part 1 has interested you and you are considering participation, please read the additional information in Part 2 before making any decision.

Part 2:

1. What will happen if I don't want to carry on with the study?

If you withdraw from the study, we will destroy all your data.

2. What if there is a problem?

If you have a concern about any aspect of this study, please contact me: Tatsuhito Akimoto (the principal investigator): tatsuhito.akimoto@student.shu.ac.uk, +44 7476 908040, Centre for Health and Social Care Research, Sheffield Hallam University.

Alternatively, you can contact my supervisor Dr Chris Wright: chris.wright@shu.ac.uk, 0114 225 5488.

If you would rather contact an independent person, you can contact Peter Allmark (Chair Faculty Research Ethics Committee) p.allmark@shu.ac.uk; 0114 225 5727.

3. Will my taking part in this study be kept confidential?

We have a duty of confidentiality to you as a research participant and we will do our best to meet this duty. Your personal information and test result will be treated with complete confidentiality and only looked at by authorised persons from the research supervision team. Your data may be looked at by authorised people to check that the study is being carried out correctly. Your unique ID code will not be revealed, even after the research is completed.

4. What will happen to the data collected in the research study?

Your data will be kept in a password-protected computer file and it will be kept for a maximum of seven years. If you would like your data to be excluded from this research study, please contact us. The results of this research will be presented as part of a PhD dissertation and subsequent publications.

5. Who is sponsoring and funding the research?

This is a self-funded PhD project supervised by Sheffield Hallam University.

6. Who has reviewed the study?

All research based at Sheffield Hallam University is looked at by a group of people called a Research Ethics Committee. This Committee is run by Sheffield Hallam University but its members are not connected to the research they examine. The Research Ethics Committee has reviewed this study and given a favourable opinion.

Appendix D – Informed consent form

Sheffield
Hallam
University

Participant Consent form

Study title: Image Interpretation Performance of Diagnostic Radiographers: Benchmarking new graduates

Principal investigator: Tatsuhito Akimoto

Please read the following statements and tick the box to show that you have read and understood them and that you agree with them.

Please tick each box

1. I confirm that I have read and understood the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. □

2. I understand that my involvement in this study is voluntary and that I am free to withdraw at any time, without giving any reason and without my legal rights being affected. □

3. I understand that relevant sections of my data collected during the study may be looked at by individuals from Sheffield Hallam University and the Research Ethics Committee where it is relevant to this research. I give permission for these individuals to have access to my records. □

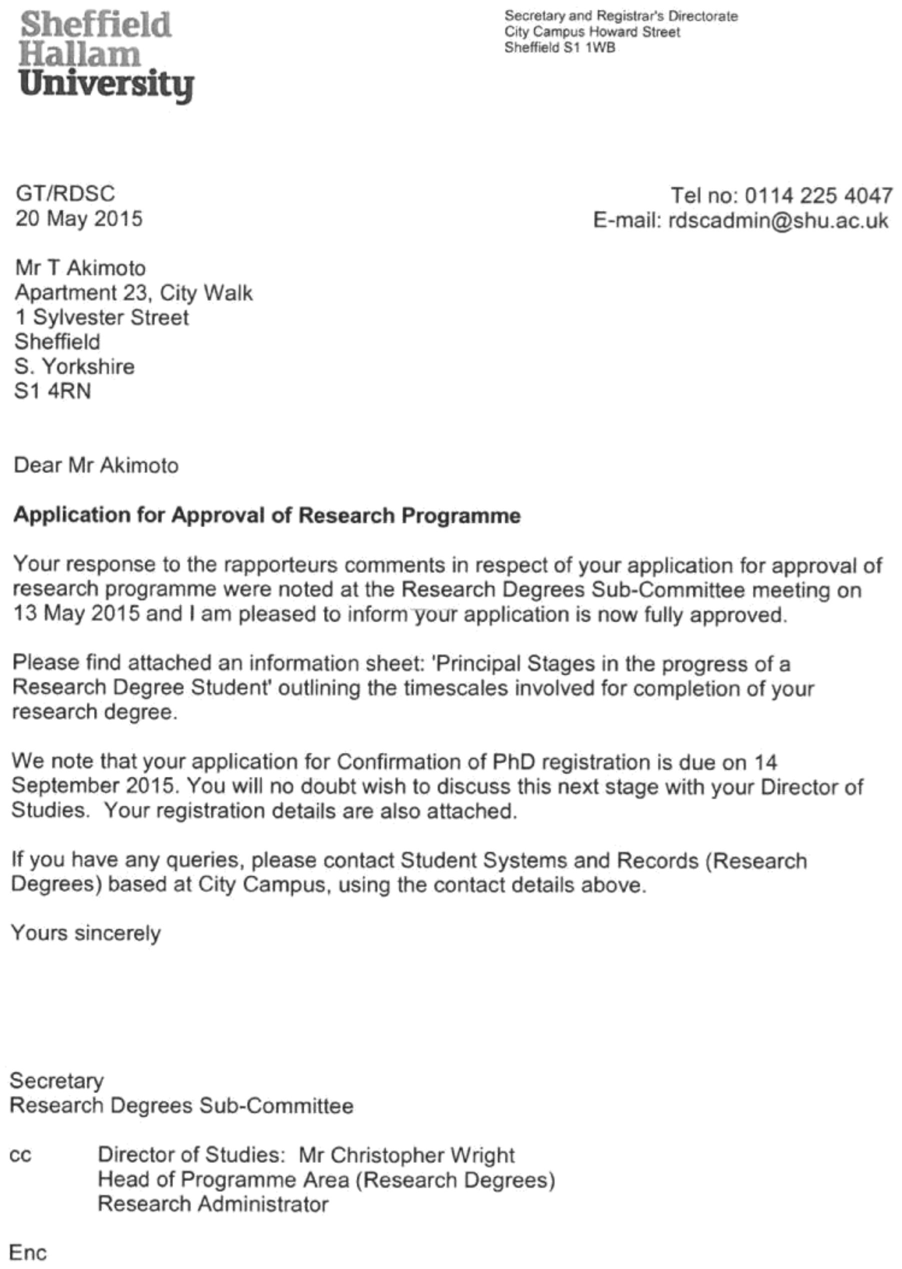
4. I agree to take part in the above study. □

To be filled by the participant

I agree to take part in the above study.

Signed: **Date**:

Appendix E – Research approval



Sheffield
Hallam
University

Secretary and Registrar's Directorate
City Campus Howard Street
Sheffield S1 1WB

GT/RDSC
20 May 2015

Tel no: 0114 225 4047
E-mail: rdscadmin@shu.ac.uk

Mr T Akimoto
Apartment 23, City Walk
1 Sylvester Street
Sheffield
S. Yorkshire
S1 4RN

Dear Mr Akimoto

Application for Approval of Research Programme

Your response to the rapporteurs comments in respect of your application for approval of research programme were noted at the Research Degrees Sub-Committee meeting on 13 May 2015 and I am pleased to inform your application is now fully approved.

Please find attached an information sheet: 'Principal Stages in the progress of a Research Degree Student' outlining the timescales involved for completion of your research degree.

We note that your application for Confirmation of PhD registration is due on 14 September 2015. You will no doubt wish to discuss this next stage with your Director of Studies. Your registration details are also attached.

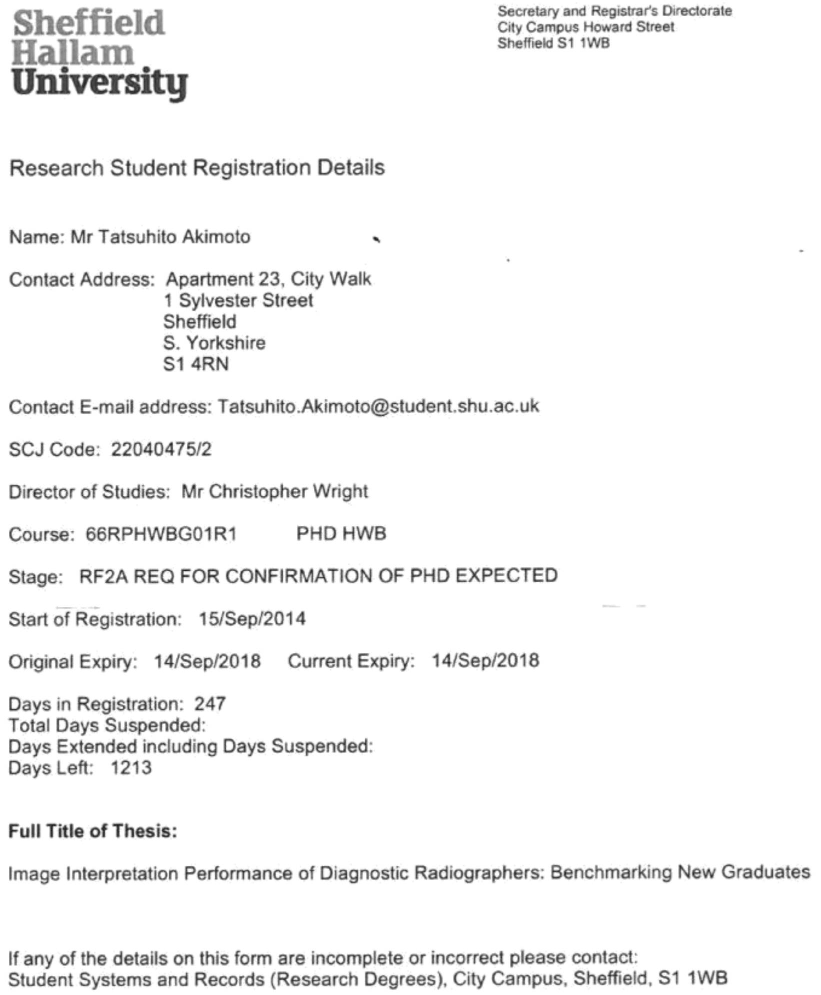
If you have any queries, please contact Student Systems and Records (Research Degrees) based at City Campus, using the contact details above.

Yours sincerely

Secretary
Research Degrees Sub-Committee

cc Director of Studies: Mr Christopher Wright
Head of Programme Area (Research Degrees)
Research Administrator

Enc



Sheffield
Hallam
University

Secretary and Registrar's Directorate
City Campus Howard Street
Sheffield S1 1WB

Research Student Registration Details

Name: Mr Tatsuhito Akimoto

Contact Address: Apartment 23, City Walk
1 Sylvester Street
Sheffield
S. Yorkshire
S1 4RN

Contact E-mail address: Tatsuhito.Akimoto@student.shu.ac.uk

SCJ Code: 22040475/2

Director of Studies: Mr Christopher Wright

Course: 66RPHWBG01R1 PHD HWB

Stage: RF2A REQ FOR CONFIRMATION OF PHD EXPECTED

Start of Registration: 15/Sep/2014

Original Expiry: 14/Sep/2018 Current Expiry: 14/Sep/2018

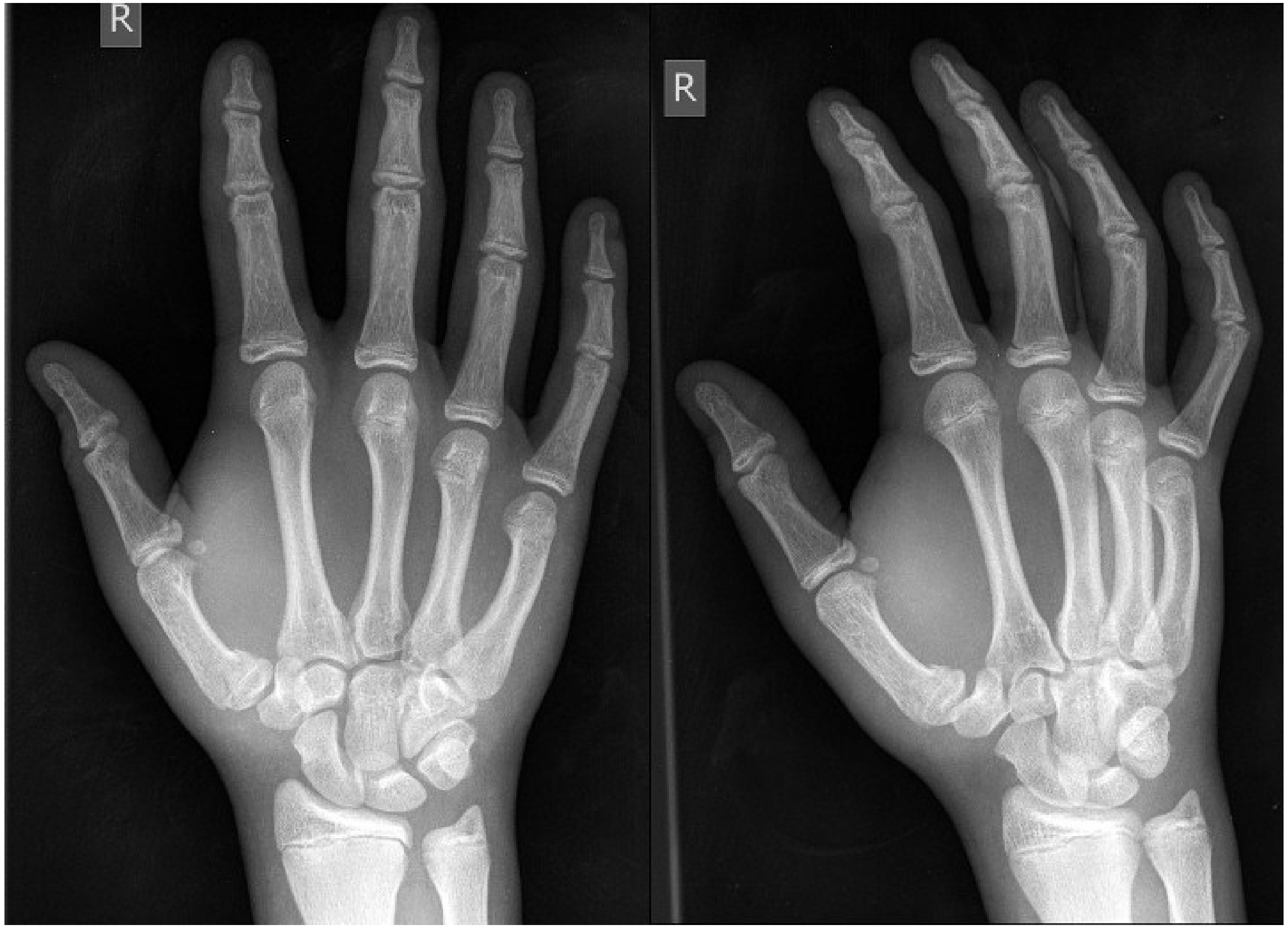
Days in Registration: 247
Total Days Suspended:
Days Extended including Days Suspended:
Days Left: 1213

Full Title of Thesis:

Image Interpretation Performance of Diagnostic Radiographers: Benchmarking New Graduates

If any of the details on this form are incomplete or incorrect please contact:
Student Systems and Records (Research Degrees), City Campus, Sheffield, S1 1WB

Appendix F – Sample radiographs in the image bank



Appendix G – Registration form

Registration Form

ID#	

for admin use only

Recipient data is held securely. All research data is anonymised.

First name		**Last name**	
e-mail address		(your certificate will be sent here)	
Date of birth		**Gender**	Male / Female
University		**Estimated degree classification**	1st / 2:1 / 2:2 / 3rd

Main education prior to university	A-Level / BTEC / Access / Previous Degree / other

Clinical placement(s)	

Appendix H – Answer booklet

Radiographic Image Interpretation Test

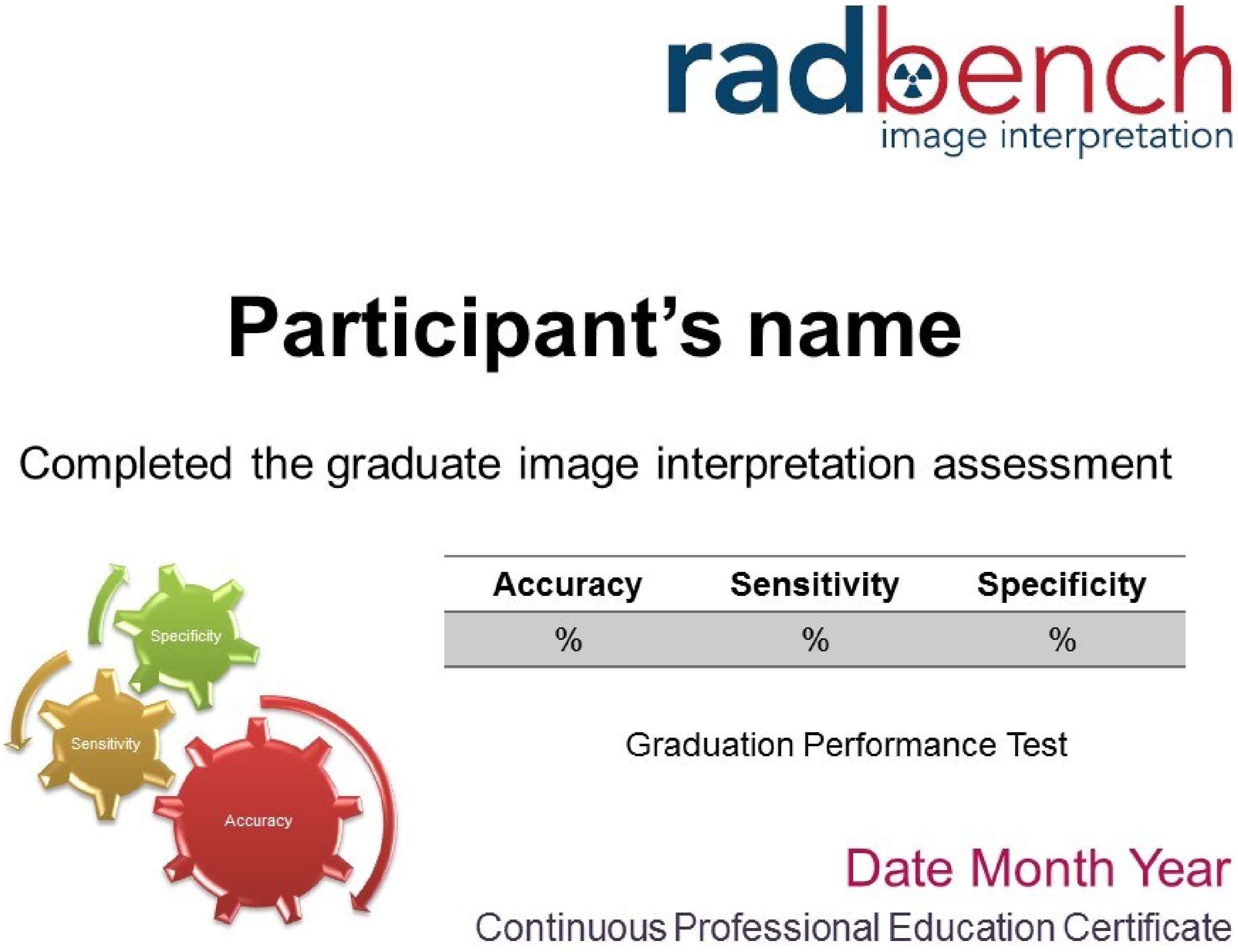
All questions must be answered. Select ONE ranking choice only and comment on your interpretation of the image

	Anatomical Region	Ranking					Preliminary Clinical Evaluation (PCE)
		1	2	3	4	5	
1	**Hand**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
2	**Ankle**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
3	**Elbow**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
4	**Wrist**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
5	**Wrist**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
6	**Foot**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
7	**Ankle**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
8	**Foot**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
9	**Foot**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	

10	**Ankle**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
	Anatomical Region	**Ranking**					**Preliminary Clinical Evaluation (PCE)**
		1	**2**	**3**	**4**	**5**	
11	**Foot**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
12	**Wrist**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
13	**Radius & Ulna**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
14	**Wrist**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
15	**Radius & Ulna**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
16	**Foot**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
17	**Elbow**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
18	**Shoulder**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
19	**Ankle**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
20	**Knee**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	

	Anatomical Region	**Ranking**					**Preliminary Clinical Evaluation (PCE)**
		1	**2**	**3**	**4**	**5**	
21	**Elbow**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
22	**Hand**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
23	**Ankle**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
24	**Hand**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
25	**Elbow**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
26	**Foot**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
27	**Shoulder**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
28	**Knee**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
29	**Ankle**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	
30	**Hand**	Definitely Normal	Probably Normal	Possibly Abnormal	Probably Abnormal	Definitely Abnormal	

Appendix I – Certificate for completing the X-ray image evaluation test.



radbench
image interpretation

Participant’s name

Completed the graduate image interpretation assessment

Specificity
Sensitivity
Accuracy

Accuracy	Sensitivity	Specificity
%	%	%

Graduation Performance Test

Date Month Year

Continuous Professional Education Certificate

Appendix J – Interview questionnaire

Survey on education of Preliminary Clinical Evaluation at diagnostic radiography courses in the UK

Information of the university

1. Name of the university:

2. Name of the course:

3. Number of the final year students:

4. Number of lecturers in the course:

X-ray image evaluation education for PCE

5. X-ray image evaluation is taught at this university.

□ Yes.
□ No.

If **No**, what are the reasons for excluding X-ray image evaluation from the curricula of the course? (skip to Q.15)

6. Is plain X-ray image evaluation is taught as discrete modules?

□ Yes.
□ No.

If **Yes**, how many modules, credits and teaching hours (for individual students) are allocated each year?

	Modules	Credits	Hours
1st year			
2nd year			
3rd year			

7. Is plain X-ray image evaluation education incorporated in other modules?

□ Yes.
□ No.

If **Yes**, how many modules, credits and teaching hours (for individual students) allocated to the modules each year?

	Modules	Credits	Hours

1st year
2nd year
3rd year

8. Is plain X-ray image evaluation taught in clinical placements?

□ Yes.
□ No.

If **Yes**, how many credits and teaching hours (for individual students) allocated to the clinical placements each year?

	Credits	Hours
1st year		
2nd year		
3rd year		

9. How is X-ray image evaluation education delivered?

□ Academic lectures/tutorials
□ Informal academic lectures/tutorials
□ Small group activity
□ Clinical lectures/tutorials
□ Informal clinical lectures/tutorials

Other:

10. Which of the following anatomical areas are included in the X-ray image evaluation education?

□ Appendicular skeleton
□ Chest
□ Axial skeleton
□ Abdomen

Other:

11. Is an X-ray image search strategy adopted in the X-ray image evaluation education? (example: ABBCS)

□ Yes.
□ No.

If **Yes**, what is the search strategy used in this course? :

12. How are students' competencies in X-ray image evaluation assessed?

□ Written examinations
□ Assignment
□ Clinical examinations with oral reports
□ Oral examinations/viva
□ Clinical examinations with written reports
□ Computer-based assessment

Other:

Is the quality of student's comments assessed? (example: if the students' comments include fracture types, fracture locations and presence of displacement/angulation)

□ Yes.
□ No.

13. How many lecturers/clinical supervisors are involved in X-ray image evaluation education? :

14. In terms of X-ray image evaluation education, what are the strengths of the course?

15. Does the course have ideas or future plans X-ray image evaluation education?

□ Yes.
□ No.

If **Yes**, what are the ideas/plans?

Clinical placements

16. Are radiographers involved in formal reporting of X-ray images in hospitals where student's clinical placements are taken place?

□ Yes.
□ No.

17. If **Yes**, do those reporting radiographers supervise the students in clinical placements?

□ Yes.
□ No.

18. Are reporting radiographers invited to teach students in formal lectures at the university?

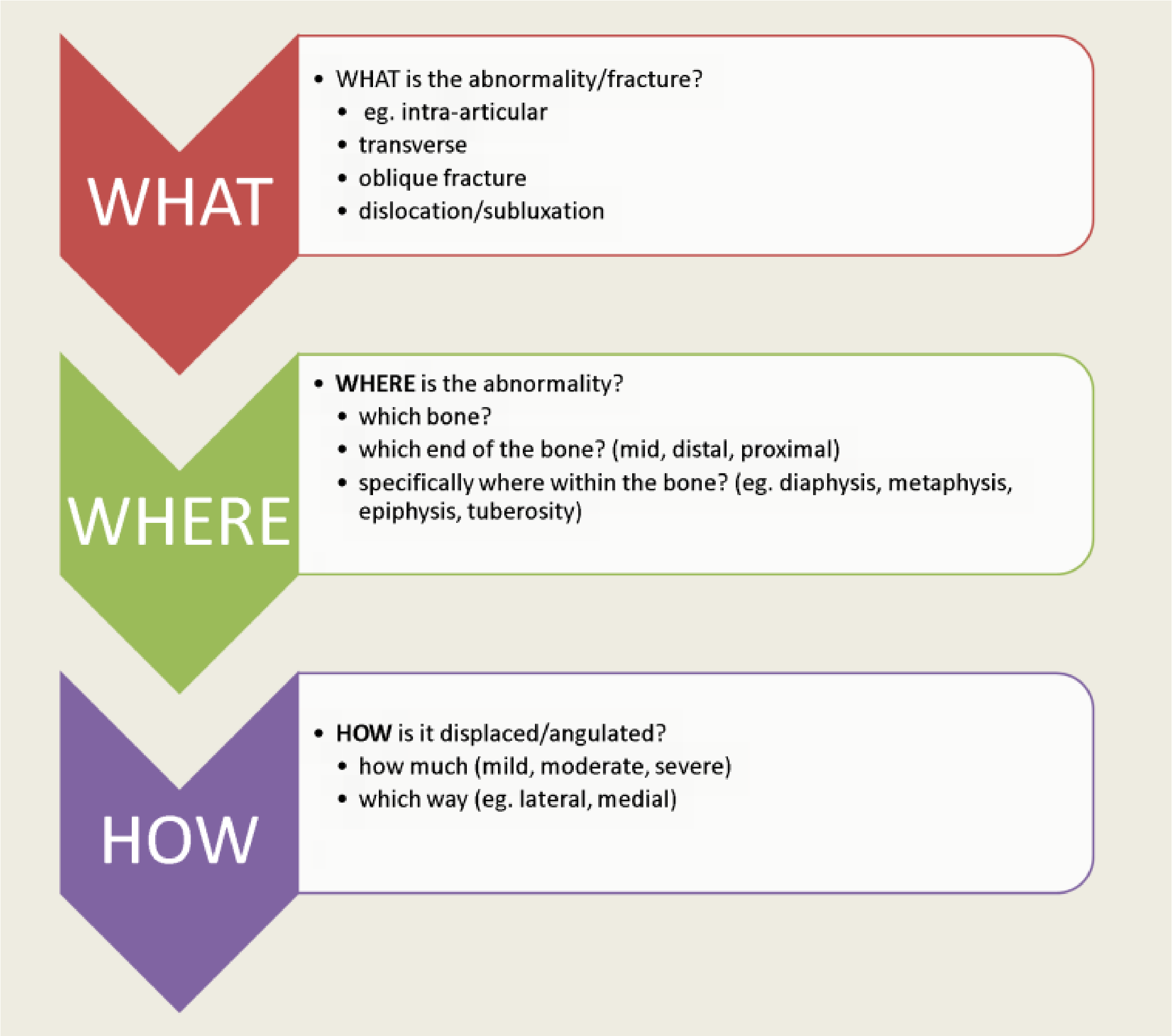
□ Yes.
□ No.

Additional comments (optional)

19. Do you have any particular interests in the data of this research?

20. Other comments for this study.

Appendix K – WHAT/WHERE/HOW conceptual framework



WHAT

- WHAT is the abnormality/fracture?
  - eg. intra-articular
  - transverse
  - oblique fracture
  - dislocation/subluxation

WHERE

- **WHERE** is the abnormality?
  - which bone?
  - which end of the bone? (mid, distal, proximal)
  - specifically where within the bone? (eg. diaphysis, metaphysis, epiphysis, tuberosity)

HOW

- **HOW** is it displaced/angulated?
  - how much (mild, moderate, severe)
  - which way (eg. lateral, medial)

Appendix L – Evaluation criteria for WWH scoring system

Image #	WHAT	Score	WHERE	Score	HOW	Score
1	Fracture	0.5	First metacarpal	0.5	Minimum or no displacement	0.5
	Oblique or Salter Harris (SA) 4	0.5	Base or proximal epiphysis	0.5	Dorsal	0.5
4	Fracture	0.5	Radius	0.5	No displacement or angulation	1
	Transverse	0.25	Distal epiphysis	0.5		
	Intra-articular	0.25				
5	Fracture	0.5	Radius	0.5	No displacement or angulation	1
	Transverse	0.25	Distal epiphysis or styloid	0.5		
	Intra-articular	0.25				
6	Fracture	0.5	(Distal end of) First distal phalanx	0.5	No displacement or angulation	0.5
	Fracture	0.25	(Distal end of) Second distal phalanx	0.25	No displacement or angulation	0.25
	Fracture	0.25	(Distal end of) Third distal phalanx	0.25	No displacement or angulation	0.25
7	Fracture	0.5	Fifth metatarsal	0.5	No displacement or angulation	1
	Transverse	0.5	Base or proximal end	0.5		
9	Fracture	0.5	Fifth proximal phalanx	0.5	No displacement or angulation	1
	Oblique	0.5	Mid-shaft	0.5		
11	Fracture	0.5	Fifth metatarsal	0.5	No displacement or angulation	1
	Transverse	0.5	Base or proximal end	0.5		
15	Fracture	0.5	(Lateral) radial neck	1	No displacement or angulation	1
	Oblique or SH 2	0.5				
18	Fracture	0.5	Clavicle	0.25	Minimum displacement	0.25
			Distal third	0.25	Inferior	0.25
	Multiple fractures	0.5	(Posterior) ribs (3-7)	0.5	No / minimum displacement or angulation	0.5
21	Fracture	0.5	Coronoid process	1	Minimum displacement	1
	Intra-articular	0.5				
22	Fracture	0.25	Radius	0.25	Slight angulation	0.25
	Transverse or impacted	0.25	Distal epiphysis or styloid	0.25	Dorsal	0.25

	Intra-articular	0.25				
	Fracture	0.25	Ulna	0.25	Minimum displacement	0.5
			styloid process	0.25		
23	Fracture	0.5	Lateral malleolus	1	No displacement or angulation	0.5
	Oblique or spiral	0.5			Not affecting the syndesmosis	0.5
25	Fracture	0.5	Radius	0.5	No displacement or angulation	1
	Vertical	0.25	Head	0.5		
	Intra-articular	0.25				
28	Fracture	0.5	Tibial eminence	0.5	Minimum displacement	1
	(Avulsion) Intra-articular	0.5	Medial	0.5		
29	Fracture	0.5	Calcaneus	1	No fragments or fragments intact	1
	Comminuted	0.25				
	Intra-articular	0.25				

Appendix M – Research Ethics Checklist (SHUREC1): Section 2 and 3

Section 2: Research with human participants

Question		Yes/No
1.	Does the research involve human participants? This includes surveys, questionnaires, observing behaviour etc.	YES
Note	*If YES, then please answer questions 2 to 10*	
If NO, please go to Section 3		
2.	Will any of the participants be vulnerable?	NO
Note	*'Vulnerable' people include children and young people, people with learning disabilities, people who may be limited by age or sickness or disability, etc. See definition*	
3	Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind?	NO
4	Will tissue samples (including blood) be obtained from participants?	NO
5	Is pain or more than mild discomfort likely to result from the study?	NO
6	Will the study involve prolonged or repetitive testing?	NO
7	Is there any reasonable and foreseeable risk of physical or emotional harm to any of the participants?	
Note	*Harm may be caused by distressing or intrusive interview questions, uncomfortable procedures involving the participant, invasion of privacy, topics relating to highly personal information, topics relating to illegal activity, etc.*	
8	Will anyone be taking part without giving their informed consent?	NO
9	Is it covert research?	NO
Note	*'Covert research' refers to research that is conducted without the knowledge of participants.*	
10	Will the research output allow identification of any individual who has not given their express consent to be identified?	NO

Section 3: Research in organisations

Question		Yes/No
1	Will the research involve working with/within an organisation (e.g. school, business, charity, museum, government department, international agency, etc.)?	YES

2	If you answered YES to question 1, do you have granted access to conduct the research? *If YES, students please show evidence to your supervisor. PI should retain safely.*	YES
3 *Note*	If you answered NO to question 2, is it because: A. you have not yet asked B. you have asked and not yet received an answer C. you have asked and been refused access. *You will only be able to start the research when you have been granted access.*	

Appendix N – Data Management Plan (DMP)

Data Collection

What data will you collect or create?

Data type: Quantitative and qualitative data.

Data format: All digital data formats used in this research are accepted by The UK Data Archive for long-term data preservation. The formats include: .sav, .doc/.docx, .txt, .xls/.xlsx and .jpeg/.jpg. Paper documents (answer booklets and registration forms) are digitalised into .docx and .xlsx formats.

Data volumes: 65.2MG (last updated on 10 September 2018).

How will the data be collected or created?

Research methodology

∉ Primarily a quantitative research.

Folder/file naming and versioning

∉ Unique, indicative and brief names are used for folders and files.

∉ Each data file has a version number.

∉ Older version files are kept in another folder.

Data quality assessment

∉ Same standardised tests (X-ray image interpretation test with an X-ray image bank) are used for consistent and reliable data acquisition.

Documentation and metadata

What documentation and metadata will accompany the data?

Documentation

∉ All contextual information and data description are summarised in data files themselves. Readme text files may be created and placed in the same location as the data files.

Metadata

∉ Disciplinary metadata standards for Social Science and Humanities, developed by The Digital Curation Centre (DCC), are used if necessary.

Ethics and Legal Compliance

How will you manage any ethical issues?

This research complies with Data Protection Act (DPA). Ethical approval was obtained from the Faculty Research Ethics Committee of Sheffield Hallam University on 4 November 2014. The proposal of this research was then fully approved by the Research Degree Sub-Committee of Sheffield Hallam University on 13 May 2015. Information sheets and consent forms will be used to ensure that informed consent is gained that allows for the preservation and sharing of the anonymised data. The names of participating students, course leaders and their universities will be anonymised accordingly to *Anonymisation managing data protection risk code of practice* (Information Commissioner's Office, 2012). Patient information and radiographer IDs are removed from the X-ray images used in this research.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

SHU will own the primary data that it collects, but the secondary data will be owned by the principal investigator of this research. The analysed data is owned by SHU, but will not be published without the agreement and support of our project partners. When the results of this research are published, the copyright on the article will be held by the publisher.

Storage and Backup

How will the data be stored and backed up during the research?

Digital data is regularly backed up in two USB flash devices. This research does not use any Cloud Storage services, such as Google Drive and Dropbox, as the university does not guarantee the quality of access controls. The devices are kept offline. The paper documents are stored in a plastic box file.

How will you manage access and security?

Security of the master data

∉ Digital data: Regularly backed up and kept in two offline USB flash devices

∉ Paper documents: The plastic box file with paper documents is kept in a locked cabinet.

Access to the master data

∉ The master data are only looked at by the principal investigator and authorised persons from the research supervision team.

Selection and Preservation

What data are of long-term value and should be retained, shared, and / or preserved?

All data (raw and analysed) will be deposited in the University's Repository for Data (SHURDA) before the end of the research project. The data will be retained in the archive for a period of 10 years. When depositing the data, no further changes to data formatting will be required as all necessary actions will have been conducted as the research progresses.

What is the long-term preservation plan for the dataset?

All 'raw' data (with appropriate documentation), and the analysed data will be made available to legitimate researchers or practitioners - in particular for the benefit of (ex) service personnel and/or those in recovery - after the embargo period has expired. This approach to open access will ensure the legacy of the project by enabling follow-up and/or longitudinal studies to be compared with these initial raw data sets.

Data Sharing

How will you share the data?

A data sharing agreement with re-users of the data will not be required, as the raw anonymized data and the data collection methodologies will be made available on a Creative Commons with Attribution (CC-BY) or equivalent license. While a robust approach to ensuring consent is received from all respondents in the study to allow raw data to be shared, should some respondents refuse permission, these data will be removed before depositing the data in the SHU Repository for Data (SHURDA). The project manager will keep the

Project Director informed during data collection of those respondents refusing permission for data sharing. The responsibility for ensuring extraction of data from those declining will ultimately be the Project Director.

Are any restrictions on data sharing required?

We will deposit and share our data at the end of the project without any delay. Any research outputs that are published will contain a statement that refers to the underlying datasets and how these datasets can be accessed; any restrictions to access will be outlined and justified in this statement. A data sharing agreement with re-users of the data will not be required, as the raw anonymized data and the data collection methodologies will be made available on a Creative Commons with Attribution (CC-BY) or equivalent license. While a robust approach to ensuring consent is received from all respondents in the study to allow raw data to be shared, should some respondents refuse permission, these data will be removed before depositing the data in the SHU Repository for Data (SHURDA). The project manager will keep the Project Director informed during data collection of those respondents refusing permission for data sharing. The responsibility for ensuring extraction of data from those declining will ultimately be the Project Director.

Responsibility and Resources

Who will be responsible for data management?

The responsibility for research data management lies with the Director of Studies (DoS). The research supervision team has the responsibility for implementation and supervision of each data management activity conducted by the principal investigator of this research.
What resources will you require to deliver your plan?

The research will use Research Data Management Advisory Service (rdm@shu.ac.uk) if necessary.

Appendix O – Ethical approval

4 November 2014

Tatsuhito Akimoto email
b2040475@my.shu.ac.uk
Collegiate Crescent Campus
Sheffield

Research proposal number: 2014-5/HWB/HSC/STAFF/7

Dear Tatisuhito

This letter relates to your research proposal: **Image Interpretation Performance of Diagnostic Radiographers: Benchmarking New Graduates**

This proposal was submitted to the Faculty Research Ethics Committee with a standard SHREC1 form. This indicates that your project does not require formal ethics and scientific review. As such, it has been added to the register of projects and given a reference number. You do not need any further review from the Ethics Committee. You will need to ensure you have all other necessary permission in place before proceeding, for example, from the Research Governance office of any sites outside the University where your research will take place. This letter can be used as evidence that the proposal has been registered within Sheffield Hallam University.

The documents reviewed were:

SHUREC1

Good luck with your project.

Yours sincerely

Peter Allmark
Chair Faculty Research Ethics Committee
Faculty of Health and Wellbeing
Sheffield Hallam University
32 Collegiate Crescent
Sheffield
S10 2BP 0114
224 5727
p.allmark@shu.ac.uk